# Gender Differences in Undergraduate Engineering Applicants: A Text Mining Approach

Shivangi Chopra, Hannah Gautreau, Abeer Khan, Melicaalsadat Mirsafian and Lukasz Golab
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
{s9chopra,hvgautre,a383khan,mmirsafi,lgolab}@uwaterloo.ca

## ABSTRACT

It is well known that post-secondary science and engineering programs attract fewer female students. In this paper, we analyze gender differences through text mining of over 30,000 applications to the engineering faculty of a large North American university. We use syntactic and semantic analysis methods to highlight differences in motivation, interests and background. Our analysis leads to three main findings. First, female applicants demonstrate a wider breadth of experience, whereas male applicants put a greater emphasis on technical depth. Second, more female applicants demonstrate a greater desire to serve society. Third, female applicants are more likely to mention personal influences for studying engineering.

## Keywords

Gender differences, engineering, admissions, text mining, clustering.

## 1. INTRODUCTION

The failure of science and engineering programs to attract equal numbers of women and men is well-documented; only 23% of women with high scores in mathematics pursue Science, Technology, Engineering and Mathematics (STEM) degrees as compared to 45% of men with the same scores [9]. As a result, there has been a great deal of research on understanding why this is the case; see, e.g., [1, 3, 4, 13, 18, 19, 20]. The major findings of prior work are that women are less likely to pursue STEM degrees because they do not see how this leads to societal improvement, and that women are more often led to study engineering because of influences from family and friends. Prior work has also found that the gender gap in STEM fields is *not* due to a difference in technical ability.

One weakness of existing work is that it is based on small datasets collected through surveys and longitudinal studies. In this paper, we present a large-scale text mining study of this topic. Our analysis is enabled by a unique dataset of over 30,000 undergraduate applications to the engineering faculty of a large North American university. Applicants are required to describe why they are interested in studying engineering, and provide other relevant information such as reading interests, extracurricular activities and programming experience. Our goal is to *determine whether female applicants identify different reasons for applying to an engineering program, and whether female applicants have different technical and extracurricular backgrounds.*

To answer these questions, we use text mining to extract the *reasons* why students apply to engineering programs. As in other text mining applications, challenges arise due to the ambiguity of natural language. To overcome these challenges, we rely on word embeddings and clustering to partition the text into semantically meaningful groups. We also analyze gender differences in programming languages and extracurricular activities through classification models and word frequency analyses. To the best of our knowledge, there is no prior work on large-scale text mining to obtain insights about students' motivation and interests.

The main findings of this paper are that women differentiate themselves through breadth of experience and men differentiate themselves through technical depth; women more often display a desire to serve society; and that women are more likely to mention interpersonal relationships when discussing their engineering goals.

The remainder of this paper is organized as follows. Section 2 summarizes related work; Section 3 discusses our dataset and methodology; Section 4 presents our results; Section 5 discusses the implications of our findings; and Section 6 concludes the paper with directions for future work.

## 2. RELATED WORK

There are three areas of work on gender differences in STEM. First are qualitative studies on small populations of students through interviews and surveys. Second are statistical studies that use census data or other summary data. Third, there are data mining studies on student performance. These works span students who are in high school, already enrolled in STEM programs, and who are working in a STEM profession.

First, we discuss qualitative survey-based studies.

Diekman et al. [3] studied 360 students from STEM and non-STEM fields consisting of 57.5% women. Each participant was asked about their mathematics and science experience and their perception of the degree to which different careers fulfill their personal goals. Participants' answers reflected that STEM careers impede communal-goal endorsement, which refers to how much a field enables achieving the goal of helping people and society. It was found that gender can predict communal-goal endorsement, and that communal-goal endorsement can negatively predict interest in STEM and positively predict interest in female-dominated programs with higher accuracy than other metrics such as gender or self-efficacy. Eccles [4] found similar results on a larger, more comprehensive dataset. They presented a longitudinal study of 1500 participants from south eastern Michigan from 6th grade to adulthood. They found that the main source of gender differences in entry to STEM careers is not gender differences in mathematical ability, but differences in inclinations towards society-oriented jobs. Women who aspire to math-related or engineering careers place a lower value on society-oriented job characteristics than their female colleagues who did not aspire to STEM careers.

Matusovich et al. [13] examined gender differences in values, but only within engineering. The study was conducted on 6 women and 5 men who majored in engineering. Each student was interviewed once a year throughout their undergraduate degree, and asked how his or her values affect their decision to earn an engineering degree. Values were classified under 4 groups: Attainment (ability to see oneself as an engineer), Cost (time and effort involved in their studies), Interest (enjoyment of understanding how math and science can be applied to every day life), and Utility (potential for future earnings). It was found that women were less likely to see themselves as engineers but continued to pursue an engineering degree due to the other values.

More reasons to pursue engineering were observed by Smith [19]. Smith interviewed 17 women who were studying engineering at four different colleges in the United States. Smith observed that participants were influenced to study engineering by family or friends. These influences played a pivotal role in helping the women build self confidence in their mathematical and science ability. They found an expression of "love" towards mathematics in many cases, despite the fact that these courses were also considered difficult. An interest in physics was found to be instrumental in their decision to study engineering. Women chose engineering because it allowed them to utilize the concepts covered in physics without having to major in physics. However, gender differences were not considered.

In terms of quantitative studies based on summary statistics, Hango [9] found that while mathematical ability plays a role, it does not explain gender differences in STEM career choices. Women with high mathematical ability are less likely to enter STEM fields than even men with a lower mathematical ability. He also supported the findings of Eccles suggesting that the gender gap in STEM programs is due to other factors.

There is prior work on gender differences in STEM using data mining techniques [16, 5, 10, 12]. However, these findings focus on student performance, whereas our work focuses on students' motivations for studying STEM, and their non-academic experiences and backgrounds.

Finally, there exists work on gender differences in computing, but it focuses on attitudes toward computing and proficiency with basic tasks [20, 1, 18]. Instead, we focus on reported programming language knowledge.

To the best of our knowledge, our work is the first one that conducts a data driven analysis of the reasons why students want to pursue engineering, and calculates the gender differences in these reasons. We also study past employment experiences, and programming knowledge in an effort to capture a more holistic view of the personalities of women and men who apply to engineering. In our conclusions, we verify some of the results of previous studies, and add to others.

## 3. DATA AND METHODOLOGY

### 3.1 Data

Our dataset comes from the engineering faculty of a large North American university. It contains all applications – both accepted and rejected – to the 14 available engineering programs from 2013 to 2016 inclusive. Table 1 shows the number of applications and the gender distribution of the applicants to each program, sorted by percentage of female students. The dataset includes gender, first choice program, and short free text responses to the following fields:

1. Engineering interests and goals: explain why you are interested in engineering and the specific program to which you applied.
2. Reading interests: discuss a book or an article you enjoyed or that has had an impact on you (preferably something that was not part of a course at school).
3. List any extracurricular activities or areas of significant interest.
4. List any jobs you held throughout high school.
5. Only mandatory for applicants to Software Engineering: list any programming experience you have.
6. Additional information: tell us anything else about yourself that you would like us to know when we review your application.

We report results for three groups of applicants: Biomedical and Environmental Engineering (BEE), Software Engineering (SE) and all other programs (OTHER). We initially analyzed applications to each program separately but observed applicants to programs within OTHER to be similar in the trends they display. Notably, the gender split in BEE is equitable, unlike other programs which are male-dominated. Furthermore, SE has unique application requirements (programming knowledge) and requires additional analysis.

Table 1: Gender breakdown by program

| Program | Applicants | % Women | % Men |
|---|---|---|---|
| Environmental | 1021 | 53% | 47% |
| Biomedical | 2015 | 52% | 48% |
| Chemical | 3612 | 38% | 62% |
| System Design | 957 | 38% | 62% |
| Management | 1040 | 36% | 64% |
| Civil | 3375 | 28% | 72% |
| Geological | 361 | 25% | 75% |
| Nanotechnology | 1670 | 24% | 76% |
| Electrical | 3782 | 17% | 83% |
| Computer | 3931 | 16% | 84% |
| Software | 3635 | 14% | 86% |
| Mechanical | 5473 | 12% | 88% |
| Mechatronics | 2886 | 12% | 88% |
| **Total** | 33758 | 23% | 77% |

## 3.2 Methodology

We use *syntactic* and *semantic* methods to analyze the free text responses. Syntactic methods identify words mentioned by more men or women, or words that can predict gender. Additionally, we apply semantic methods to "Engineering Interests and Goals" to capture context and extract the reasons why men and women want to study engineering.

### *3.2.1 Syntactic Analysis*

For each of the six free text fields, we first perform standard pre-processing: we remove stop words, tokenize the text, and stem the tokens using the NLTK snowball stemmer[1]. We then perform two syntactic analyses on each field:

**Document Frequencies:** we identify words used at least once by a larger fraction of men or women (where each response is considered a document). We only report statistically significant differences with a P-value of 0.05 using a proportion test [6].

**Gender Prediction:** we build classifiers to predict gender based on the words or contiguous sequences of words (bigrams and trigrams) appearing in a free text response. Following previous work on text classification, we use logistic regression [8] where the dependent variable is gender, and the explanatory variables correspond to the possible words (or word bigrams/trigrams), and their values correspond to their TF-IDF scores [15, 21]. To calculate a TF-IDF score for a given word and a given response, we divide the number of times the word appears in the response by the Inverse Document Frequency - the fraction of responses in the entire dataset containing this word. TF-IDF is a useful measure because it balances the uniqueness of a term in the corpus and the importance of the term to the specific document. For each free text field except programming experience, we report the F-measure, which is the weighted harmonic mean of precision and recall [2], and accuracy, both calculated using 10-fold cross validation. We use oversampling for SE and OTHER to control for gender imbalance; otherwise, a classifier that always predicts gender as "male" would have a high accuracy on any male-dominated dataset.

Table 2: Families of programming languages

| Family | Constituent Programming Languages |
|---|---|
| Java | java, bluej, jython, android |
| C++ | c++, beta |
| Python | python |
| HTML/CSS | html, html5, css, css3 |
| C | C, objective-C, robotc |
| JavaScript | javascript, jscript, jquery, angularjs |
| Turing | turing, touring |
| C# | c#, visual c# |
| Php | php |
| SQL | sql, pl/sql |
| Other | .net, ada, alice, applescript, bash, etc |

### *3.2.2 Analysis of Programming Experience*

In the "Programming Experience" field, SE applicants are asked to list their programming experience. The structure of this question elicits not only specific programming languages, but also encourages applicants to share details about their programming experience. Thus, in addition to the document frequency analysis mentioned earlier, we perform the following detailed analyses:

- Programming Language analysis: we calculate the number of responses that mention a given programming language. We start with a list of known languages from Wikipedia[2]. We then add common misspellings of these languages, and we group them into families in consultation with a domain expert. Table 2 shows the language families whose frequencies we will report.
- Programming Concept analysis: we compile a list of computing concepts, a sample of which is shown in Table 3, group them into categories, and calculate the number of responses that mention a given concept.
- Learning Method analysis: we compile a list of online programming courses, and common variations of "high school", "self taught", "higher education", and "employment". We then categorize these terms according to how an applicant learned programming: "online", "high school", "self taught", "higher education", "work", and "other". Finally, we calculate the number of responses that mentioned each learning method.
- Experience analysis: we extract the amount of experience reported by an applicant by searching for the words "hour", "day", "month", "year", as well as common abbreviations and misspellings of these words. We use the token immediately preceding these words to determine the length of time. We convert all of the times into months.

### *3.2.3 Semantic Analysis of Engineering Interests*

Using the responses to "Engineering Interests and Goals", we want to identify the reasons why students apply to engineering programs. However, reasons cannot be inferred

[1] http://www.nltk.org/_modules/nltk/stem/snowball.html

[2] https://en.wikipedia.org/wiki/List_of_programming_languages

Table 3: Sample of programming concepts

| Concept Category | Constituent Concepts |
|---|---|
| Basic | array, list, loop, if-statement |
| Data Structures | stack, queue, linked list |
| Sorting | merge sort, bubble sort, quick sort |
| Searching | linear, binary, breadth first searches |
| OOP | object, class, abstraction, encapsulation |
| Data Science | machine learning, NLP |
| Other | storage, memory management |

Table 4: Nine questions used with the QA API

| Question No. | Question Variant |
|---|---|
| 1 | Why are you interested in Engineering? |
| 2 | What inspired you to study Engineering? |
| 3 | What do you find inspiring about Engineering? |
| 4 | What are the reasons you like Engineering? |
| 5 | Why do you feel the need to pursue Engineering? |
| 6 | Why are you passionate about Engineering? |
| 7 | Why does Engineering interest you? |
| 8 | Why do you want to study Engineering? |
| 9 | Why do you like Engineering? |

simply by counting occurrences of certain keywords; for example, family influence may be expressed by using words such as "father", "mother", "uncle", or simply, "family". Furthermore, an applicant may mention things other than the exact reason as to why they are interested in engineering in their response. Our semantic approach deals with these issues through the use of *Question Answering* to isolate topics being mentioned that could be considered indicative of reasons, followed by *Clustering using Word Embeddings* to analyze these. Figure 1 shows the steps in our semantic analysis, and they are explained in detail below.

**1. Question Answering (QA):** Here, we extract sentences that are most likely to contain the topics indicative of the applicants' underlying reasons for applying to engineering. We use a state of the art QA network [17] which is available as an open source API[3]. Given a question and a text document, this QA API extracts a sentence that may answer the question. However, we discovered that while asking the question that directly appeared on the entrance application - why are you interested in engineering - yielded *some* relevant sentences, there were additional relevant sentences that were not identified. To address this problem, we consulted with domain experts at the institution and formulated additional variants of this question. Depending on the applicant, not every variant identified a unique sentence. Overall, we observed that the number of unique sentences extracted per applicant plateaued at nine question variants. Table 4 lists the nine variants we use and Table 5 gives an example of the sentences extracted from a particular response using each question.

**2. Stop Word Removal:** Next, we remove stop words from the sentences extracted in the previous step because these do not contain any meaningful information about the underlying reasons. Similarly, we remove words excessively used by both genders such as "engineering" and the name of the university. This step happens after QA because QA requires the complete text, stop words included, as input.

Table 5: Sentences extracted from a particular response using all 9 question variants

| Question No. | Answer produced by the QA API |
|---|---|
| 1 | future entrepreneurship ventures |
| 2 | designing & building complicated solutions |
| 3 | future entrepreneurship ventures |
| 4 | intellectual curiosity and satisfaction is core to my personality |
| 5 | i think i fit in well in the tight culture of the engineering class |
| 6 | intellectual curiosity and satisfaction is core to my personality |
| 8 | intellectual curiosity and satisfaction is core to my personality |
| 7 | know people much closer |
| 9 | it's the best program available |

**3. Sentence Vector Computation:** At this point, each response has produced up to nine relevant sentences. We use *word embeddings* to capture semantic proximity between sentences. Specifically, we use the word2vec model [14], trained on the Google news corpus[4], to convert each word into a 300-dimensional vector that encodes the underlying semantics. We then use the average of all word vectors in a sentence as its *sentence vector.* If two sentence vectors are close, the sentences are also semantically similar [7, 11].

**4. Clustering of Sentence Vectors:** Next, we cluster the sentence vectors received from the previous step using $K$-Means clustering with Euclidean distance as the similarity metric and $K = 200$, where $K$ is the number of clusters (the rationale behind this choice of $K$ will be discussed shortly). The clusters converge around similar topics. For example, sentences containing words related to family such as "brother", "father", or "sister" have similar word vectors and are more likely to be assigned to the same cluster. Note that this would not be the case had we clustered the sentences themselves according to their *syntactic* similarity.

**5. Cluster Representative Extraction:** After computing clusters of sentence vectors, we extract representative words from each cluster to identify the topic of that cluster. First, we map sentence vectors back to the original sentences, which creates 200 sets of sentences, one set for each cluster. We then tokenize and stem the text in each set, as described in Section 3.2.1. The word2vec model consumes unstemmed words, compelling us to postpone tokenization and stemming until this step. The trigrams in each set are ranked using their TF-IDF scores calculated considering all 200 sets as the corpus. Finally, we represent each cluster with a list of 10 top ranking trigrams, an example of which is shown in Table 6.

[3] https://github.com/allenai/bi-att-flow

[4] https://code.google.com/archive/p/word2vec/

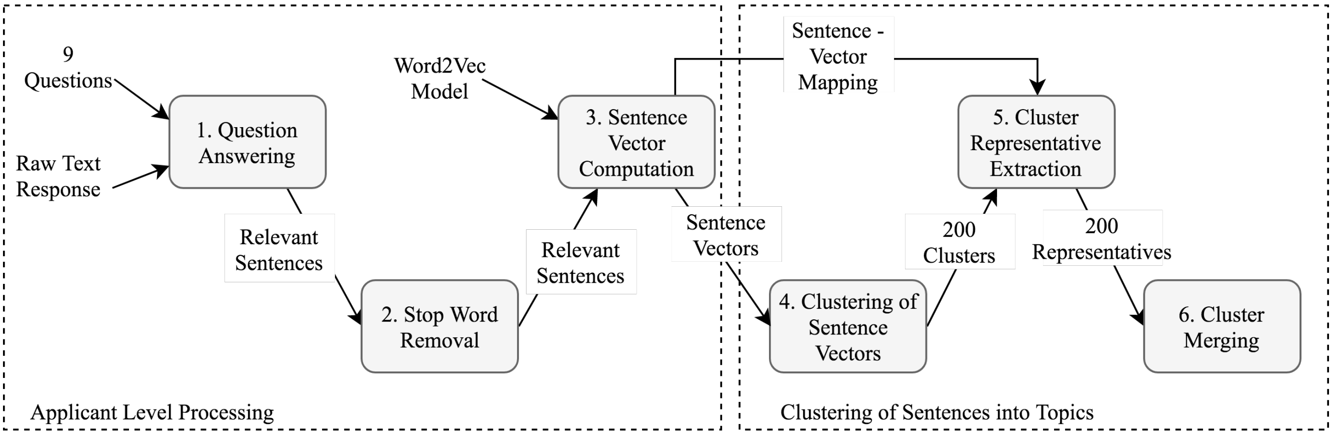
9 Questions
Raw Text Response
1. Question Answering
Relevant Sentences
2. Stop Word Removal
Relevant Sentences
Word2Vec Model
3. Sentence Vector Computation
Sentence - Vector Mapping
Sentence Vectors
4. Clustering of Sentence Vectors
200 Clusters
5. Cluster Representative Extraction
200 Representatives
6. Cluster Merging
Applicant Level Processing
Clustering of Sentences into Topics

Figure 1: Semantic analysis methodology

Table 6: An example of ten trigrams representing a cluster

| Rank | Trigram |
|---|---|
| 1 | solv problem solv |
| 2 | problem solv problem |
| 3 | enjoy problem solv |
| 4 | problem solv enjoy |
| 5 | enjoy solv problem |
| 6 | love solv problem |
| 7 | problem solv love |
| 8 | love problem solv |
| 9 | problem problem solv |
| 10 | solv problem problem |

Table 7: Examples of a mixed cluster using $K = 50$ and its pure equivalent using $K = 200$

| Mixed Cluster (50 Clusters) | Pure Cluster (200 Clusters) |
|---|---|
| kid work day | kid work day |
| apart tri togeth | young age father |
| love thing apart | visit construct site |
| countless hour spent | watch father work |
| use everi day | older brother mechan |
| pay close attent | dad electr took |
| decid high school | work day dad |
| work day saw | like help father |
| day day basi | expos young age |
| year high school | uncl civil engin |

**Choice of $K$:** In **Step 4**, we experimented with values of $K$ ranging from 50 to 200. When choosing a small $K$ and proceeding to **Step 5** with fewer clusters, many clusters were represented by trigrams that were not semantically similar enough to warrant being in the same cluster, and some uncommon trigrams were overpowered by extremely common ones. Thus, some nuanced topics were lost as they could not form a cluster of their own. Larger values of $K$ resulted in the splitting out of semantically similar topics. These resulted in pure but redundant clusters, i.e., several clusters representing the same topic. For instance, Table 7 shows a cluster of mixed topics on the left when K is 50, and a rather pure cluster on the right when $K$ is 200. A bigger $K$ made it possible for topics like “kid work day” to be grouped with similar semantic contexts like “watch father work”. The topics on the right consistently speak of the influence of a family member, indicative of family influence as a reason for engineering, whereas no single reason can be deduced from the cluster on the left. The first $K$ value that produced adequately pure clusters was 200. Thus, the decision was made to stop testing larger values and creating further unnecessary redundancy. To eliminate the unnecessary redundancies at $K = 200$, the clusters were merged in **Step 6**.

**6. Cluster Merging:** At this point we have 200 clusters of sentences, where each cluster is represented by the 10 highest ranking trigrams. To make the clusters interpretable and to group them under more general topics, we manually merge similar clusters based on their 10-trigram representations to produce ten final clusters. This process of merging follows the Card-sorting approach. Card-sorting has been widely used to systematically derive taxonomies from data, to reach a higher level of abstraction, and identify common themes [22]. For instance, it can be used to sort responses to an open-ended question into bins to deduce themes over the responses. We perform card-sorting on the representative trigrams, then we brand each of the ten final themes with human interpretable labels and consider these our final **topics**. In this process, a number of small clusters whose representatives were vague were disregarded, but even then, 99.5% of applicants were labelled with at least one topic. Table 8 shows two examples of representatives of vague clusters. Since the QA in **Step 1** used questions probing the reasons why the applicant was applying to engineering, our topics can be considered indicative of the same.

Table 9 shows the final set of topics along with sample trigram representations of clusters that were classified under each topic. *Technical Interests* refers to characteristics inherent to engineering along with topics related to specific engineering disciplines. For instance, the trigram “water treatment plant” in Table 9 is part of *Technical Interests* while being specifically related to Environmental Engineering.

All the sentences classified under a specific topic in Table 9 are tracked back to the applicants who mentioned them.

Table 8: Examples of discarded vague clusters

| Example #1 | Example #2 |
|---|---|
| appli program appli | pursu decid pursu |
| appli chemic program | experi inspir pursu |
| program appli appli | pursu motiv pursu |
| program program appli | pursu passion believ |
| program appli program | pursu wish pursu |
| program appli electr | encourag pursu pursu |
| appli mechan program | hope continu pursu |
| mechan program appli | encourag pursu believ |
| appli electr program | passion inspir pursu |
| program appli chemic | desir pursu educ |

The statistics presented in the next section are based on the number of applicants who mention a given topic, and hence indicate the same underlying reason for their interest in engineering

# 4. RESULTS

We now describe our results, treating applicants to BEE, SE and OTHER separately, as mentioned in Section 3.1. Section 4.1 presents syntactic (word frequencies and logistic regression) and semantic (question answering & clustering) results for "Engineering Interests and Goals". Section 4.5 describes the detailed analyses of programming experience (only for applicants to SE). The remaining sections discuss the results of frequency analysis and logistic regression for the remaining fields: job titles, reading interests, extracurricular activities, and additional information.

## 4.1 Engineering Interests and Goals

### 4.1.1 Syntactic Analysis

**Document Frequencies:** Overall, there are more terms that are used predominantly by women, indicating that women use a wider variety of terms. We see more women using non technical terms to express themselves, and men using more technical terms.

In BEE, more men mention "mechanical" (11.5% of men vs. 8.2% of women), and "compute" (8.5% of men vs. 5.3% of women. More women mention "health" (16.4% of women vs. 10.6% of men), "improve" (23.6% of women vs. 18% of men), "love" (24.8% of women vs. 20.5% of men), and "research" (20.6% of women vs. 16.5% of men).

In SE, more men mention "system" (14.2% of men vs. 9.6% of women), "problem" (25.5% of men vs. 20.9% of women), "game" (19.1% of men vs. 14.9% of women), and "goal" (25.7% of men vs. 21.5% of women). More women mention "science" (49.9% of women vs. 43.0% of men), "research" (11.0% of women vs. 6.9% of men), "challenge" (18.6% of women vs. 14.7% of men), and "people" (20% of women vs. 16.3% of men).

In the OTHER group of engineering programs, more men mention "mechanical" (28.9% of men vs. 7.5% of women), "compute" (25.8% of men vs. 17.2% of women), "robot" (16.2% of men vs. 9.9% of women),"car" (9.9% of men vs. 4.1% of women), and "goal" (24.3% of men vs. 19.4% of women). More women mention "chemical" (21.9% of women vs. 10.7% of men), "science" (41% of women vs. 32.6% of men), "creative" (16.1% of women vs. 10.2% of men), "study" (30.7% of women vs. 25.3% of men), and "love" (24.2% of women vs. 19.4% of men).

Table 9: The final set of ten topics, with representative word trigrams of the clusters classified under each topic

| Reason | Trigrams (stemmed) |
|---|---|
| Family | follow footstep father<br>older brother mechan |
| Contribution to Society | improv peopl live<br>make world better<br>make contribut societi |
| Outreach | attend open hous<br>talk student professor |
| Technical Interests | creat new technolog<br>water treatment plant<br>use dismantl toy<br>develop medic technolog |
| Love of Science | math physic chemistri<br>love math scienc |
| Extracurriculars | book watch video<br>robot competit team<br>particip extracurricular activ |
| Prior Accomplishments | leadership communic skill<br>profici skill mathemat |
| High School | talk physic teacher<br>high school student |
| Professional Development | pursu graduat studi<br>job opportun engin<br>futur career goal |
| Childhood Dream | began young age<br>dream childhood dream |

Table 10: F-Measure/Accuracies for predicting gender using Engineering Interests & Goals (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| BEE | 60/60.7 | 60/59.1 | 57/58 |
| OTHER | 72/78.8 | 76/80.4 | 80/77.3 |
| SE | 88/86 | 98/97.2 | 94/94 |

**Logistic Regression:** Table 10 shows the results for predicting gender using words from responses to "Engineering Interests and Goals". The predictive power of logistic regression decreases with increasing gender balance within a group, even after oversampling to compensate for the initial gender imbalance. In other words, in programs with an even gender split, it is more difficult to guess the gender.

### 4.1.2 Semantic Analysis

We classified the sentences extracted from students' responses under one of ten topics shown in Table 9. Table 11 shows the percentage of applicants to BEE who mentioned each topic. The most common topics are Technical Interests and Love of Science. More women mention **Love of Science**, which is statistically significant with a P-value of 0.03. No other topic had a statistically significant gender difference. On average, female students in this group

Table 11: BEE applicants' topics

| Topic | %All | %Women | %Men | P-value |
|---:|---|---|---|---|
| Family | 10.9% | 11.3% | 10.5% | 0.47 |
| Contribution to Society | 20.7% | 20.5% | 20.9% | 0.77 |
| Outreach | 8.5% | 9.3% | 7.7% | 0.12 |
| Technical Interests | 86.8% | 86.8% | 86.8% | 0.97 |
| **Love of Science** | **32.3%** | **34.1%** | **30.4%** | **0.03** |
| Extracurriculars | 5.7% | 5.6% | 5.9% | 0.69 |
| Prior Accomplishments | 6.1% | 5.9% | 6.3% | 0.61 |
| High School | 8.8% | 8.4% | 9.2% | 0.46 |
| Professional Development | 25.3% | 25.9% | 24.7% | 0.47 |
| Childhood Dream | 2.5% | 2.7% | 2.2% | 0.32 |

Table 12: SE applicants' topics

| Topic | %All | %Women | %Men | P-value |
|---:|---|---|---|---|
| **Family** | **7.6%** | **11.0%** | **7.0%** | **0.00** |
| Contribution to Society | 12.1% | 12.5% | 12.0% | 0.77 |
| Outreach | 8.7% | 9.1% | 8.6% | 0.703 |
| Technical Interests | 92.6% | 92.3% | 92.6% | 0.77 |
| Love of Science | 13.9% | 16.6% | 13.4% | 0.05 |
| Extracurriculars | 9.2% | 10.9% | 9.0% | 0.17 |
| Prior Accomplishments | 6.1% | 7.1% | 5.9% | 0.30 |
| High School | 11.0% | 12.9% | 10.7% | 0.15 |
| Professional Development | 25.0% | 27.7% | 24.6% | 0.13 |
| Childhood Dream | 2.7% | 2.8% | 2.6% | 0.87 |

Table 13: OTHER applicants' topics

| Topic | %All | %Women | %Men | P-value |
|---:|---|---|---|---|
| Family | 12.3% | 13.0% | 12.1% | 0.064 |
| **Contribution to Society** | **14.7%** | **16.1%** | **14.3%** | **0.00** |
| **Outreach** | **8.1%** | **9.9%** | **7.6%** | **0.00** |
| Technical Interests | 88.4% | 89.0% | 88.3% | 0.149 |
| **Love of Science** | **22.7%** | **26.6%** | **21.7%** | **0.00** |
| **Extracurriculars** | **9.0%** | **7.8%** | **9.3%** | **0.00** |
| Prior Accomplishments | 6.6% | 7.0% | 6.5% | 0.18 |
| High School | 10.3% | 9.8% | 10.5% | 0.13 |
| Professional Development | 26.6% | 27.5% | 26.3% | 0.07 |
| **Childhood Dream** | **3.7%** | **3.0%** | **3.9%** | **0.00** |

Table 14: Female students' topics across all groups

| Topic | % SE | % BEE | % OTHER |
|---:|---|---|---|
| Family | 11.1% | 11.3% | 13.0% |
| Contribution to Society | **12.5%** | **20.5%** | **16.1%** |
| Outreach | 9.1% | 9.3% | 9.9% |
| Technical Interests | **92.3%** | **86.8%** | **89.0%** |
| Love of Science | **16.6%** | **34.1%** | **26.6%** |
| Extracurriculars | **10.9%** | **5.6%** | **7.8%** |
| Prior Accomplishments | 7.1% | 5.9% | 7.0% |
| High School | **12.9%** | 8.4% | 9.8% |
| Professional Development | 27.7% | 25.9% | 27.5% |
| Childhood Dream | 2.8% | 2.7% | 3.0% |

mention 2.12 topics whereas male students mention 2.05, a statistically insignificant difference with a P-value of 0.06.

Table 12 shows the percentage of applicants to SE who mentioned each reason. The most common reasons are Technical Interests and Professional Development. Women mention **Family** more frequently than men, which is statistically significant with a P-value of 0.00. No other reason had a statistically significant gender difference. On average, female students in this program mention 2.04 reasons whereas male students mention 1.87, a statistically significant difference with a P-value of 0.00.

Table 13 shows the percentage of applicants to OTHER engineering programs who mentioned each topic. The most common topics are Technical Interests and Professional Development. Female students mention **Contribution to Society**, **Outreach**, and **Love of Science** more than male students, which is statistically significant with a P-value of 0.00. Male students mention **Extracurriculars** and **Childhood Dream** more than female students, which is statistically significant with a P-value of 0.00. No other topic had a statistically significant gender difference. On average, female students in this group mention 2.1 topics whereas male students mention 2.0 reasons, a statistically significant difference with a P-value of 0.00.

Table 14 highlights the differences between women who applied to BEE vs. women who applied to SE vs. women who applied to OTHER programs. The bold values show percentage differences from the other two groups that are statistically significant with a P-value of less than 0.05. Female applicants to SE, BEE, and OTHER programs differ from each other in their mentions of **Contribution to Society, Technical Interests, Love of Science, and Extracurriculars** with a P-value of less than 0.05. Mentions of **High School** are only different in SE applicants compared to other groups, which is statistically significant with a P-value of less than 0.05. No other topic had a statistically significant difference.

Table 15 highlights the differences between men who applied to BEE vs. men who applied to SE vs. men who applied to OTHER. The bold values show percentage differences from the other two groups that are statistically significant with a P-value of less than 0.05. Male applicants to SE, BEE, and OTHER programs differ from each other in their mentions of **Contribution to Society** and **Love of Science** with a P-value of less than 0.05. Mentions of **Family** and **Technical Interests** are only different for SE applicants compared to applicants to other programs, which is statistically significant with a P-value of less than 0.05. Mentions of **Extracurriculars** are different for BEE applicants compared to applicants to other program groups, which is statistically significant with a P-value of less than 0.05. No other topic had a statistically significant difference.

## 4.2 Reading Interests

**Document Frequencies:** Overall, men tend to report reading technical content such as research papers and women report reading novels and writing that has a societal focus. Words that are predominantly used by men include "article" (17.6% of men vs. 13.4% of women), "enjoy" (29.5% of men vs. 25.6% of women), "compute" (5.6% of men vs. 2.2% of women), and "science" (12.3% of men vs. 10.3% of women). Words that are predominantly used by women include "love" (20.3% of women vs. 12.6% of men), "novel"

Table 15: Male students' topics across all groups

| Topic | % SE | % BEE | % OTHER |
|---|---|---|---|
| Family | **7.0%** | 10.5% | 12.1% |
| Contribution to Society | **12.0%** | **20.9%** | **14.3%** |
| Outreach | 8.6% | 7.7% | 7.6% |
| Technical Interests | **92.6%** | 86.8% | 88.3% |
| Love of Science | **13.4%** | **30.4%** | **21.7%** |
| Extracurriculars | 9.0% | **5.9%** | 9.3% |
| Prior Accomplishments | 5.9% | 6.3% | 6.5% |
| High School | 10.7% | 9.2% | 10.5% |
| Professional Development | 24.6% | 24.7% | 26.3% |
| Childhood Dream | 2.7% | 2.2% | **3.9%** |

Table 16: F-Measures/Accuracies for predicting gender using words from Reading Interests (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| BEE | 64/63 | 62/60 | 60/54.2 |
| OTHER | 79/77.8 | 93/89.8 | 95/91.8 |
| SE | 92/88.9 | 96/95.6 | 93/91.8 |

(31.2% of women vs. 24.6% of men), "character" (20.3% of women vs. 15.2% of men), "women" (6.1% of women vs. 1.1% of men), "people" (29.1% of women vs. 24.9% of men), and "family" (10.7% of women vs. 6.8% of men).

**Logistic Regression:** The results for predicting gender based on Reading Interests are shown in Table 16. As before, the predictive power of logistic regression decreases with increasing gender balance within the group.

## 4.3 Extracurricular Activities

**Document Frequencies:** Overall, male applicants' extracurricular activities have a technical focus, and female applicants have a wide breadth of experiences ranging from leadership to artistic pursuits.

In BEE, more men mention "robot" (7% of men vs. 3.6% of women) and "coach" (7.1% of men vs. 4.8% of women). More women mention "dance" (8.7% of women vs. 1.7% of men), "art" (11.3% of women vs. 6.9% of men), "council" (21.5% of women vs. 15.6% of men), and "lead" (21.1% of women vs. 16.8% of men).

In SE, more men mention "compute" (20.9% of men vs. 13.7% of women). More women mention "art" (14.5% of women vs. 4.8% of men), "council" (20.5% of women vs. 11.9% of men), "dance" (8.3% of women vs. 2.2% of men), and "lead" (18.7% of women vs. 14.3% of men).

In the OTHER group of engineering programs, more men mention "robot" (11.1% of men vs. 6.3% of women), "compute" (5.8% of men vs. 2.4% of women). More women mention "dance" (10.7% of women vs. 2.1% of men), "council" (20% of women vs. 12.1% of men), "art" (11.9% of women vs. 4.8% of men), "volunteer" (22.9% of women vs. 16.3% of men), and "lead" (19% of women vs. 13.1% of men).

**Logistic Regression:** The results for predicting gender based on Extracurricular Activities are shown in Table 17.

Table 17: F-Measures/Accuracies for predicting gender using words from Extracurricular Activities (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| BEE | 72/72.9 | 69/66.6 | 62/59.5 |
| OTHER | 81/81.1 | 80/77.8 | 78/71.4 |
| SE | 85/83.3 | 85/82 | 94/93.4 |

Table 18: F-Measures/Accuracies for predicting gender using words from Job Titles (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| BEE | 59/57.9 | 58/52.6 | 63/51 |
| OTHER | 65/61.5 | 64/59.1 | 67/61.9 |
| SE | 67/63.7 | 66/58.7 | 68/51.7 |

The predictive power of logistic regression decreases with increasing gender balance within the group.

## 4.4 Job Titles

**Document Frequencies:** Across all programs, men are more likely to mention terms that imply technical work or manual labour, whereas women are more likely to mention terms that imply customer service and caring professions. Example words in job titles from male applicants include "referee" (4.1% of men vs. 2% of women), "labor" (2.6% of men vs. 0.5% of women), and "technician" (3.1% of men vs. 1.2% of women). Example words in job titles from female applicants include "cashier" (12.8% of women vs. 6.8% of men), "teacher" (6.2% of women vs. 2.7% of men), and "assist" (17.6% of women vs. 14.3% of men).

**Logistic Regression:** As shown by the logistic regression scores in Table 18, Job Titles do not provide as much predictive power as other fields.

## 4.5 Programming Experience

### *4.5.1 Syntactic Analysis*

**Document Frequencies:** In general, women use more non technical terms, and men use more technical terms. Examples of terms that are more commonly used by male applicants include "game" (30.8% of men vs. 22.3% of women) and "develop" (21.5% of men vs. 14.4% of women), and terms more commonly used by female applicants include "mark" (39.9% of women vs. 30.6% of men) and "attend" (4.2% of women vs. 1.4% of men). Through manual inspection, we discovered that "mark" was used in the context of earning a certain mark in a course. "attend" was used to indicate attendance in a programming workshop or event.

**Logistic Regression:** As shown in Table 19, the words used to describe programming experience can be used to predict the gender of the applicant.

### *4.5.2 Programming Languages*

Table 20 shows a comparison of specific language knowledge between male and female applicants. All languages except for SQL are slightly skewed toward male applicants; however, only **Java, C++, C, Turing, C#** have statistically

Table 19: F-Measures/Accuracies for predicting SE applicants' gender using Programming Experience (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| SE | 91/88.8 | 98/98 | 97/95.7 |

Table 20: Comparison of reported programming language knowledge

| Language | % Women | % Men | Difference | P-value |
|---|---|---|---|---|
| **Java** | **58.9%** | **65.6%** | **-6.7%** | **0.00** |
| **C++** | **23.3%** | **28.5%** | **-5.2%** | **0.01** |
| Python | 25.1% | 28.1% | -3.0% | 0.18 |
| HTML/CSS | 19.0% | 19.5% | -0.5% | 0.75 |
| Basic | 16.1% | 18.6% | -2.5% | 0.114 |
| **C** | **12.5%** | **17.0%** | **-4.5%** | **0.01** |
| JavaScript | 12.7% | 15.0% | -2.3% | 0.17 |
| **Turing** | **10.8%** | **14.3%** | **-3.5%** | **0.03** |
| **C#** | **6.1%** | **9.4%** | **-3.3%** | **0.01** |
| **Php** | **3.9%** | **8.3%** | **-4.4%** | **0.00** |
| SQL | 3.9% | 3.2% | -0.7% | 0.50 |
| Other | 31.1% | 16.4% | -3.3% | 0.07 |
| None | 4.7% | 3.2% | +1.4% | 0.07 |

significant differences with a P-value of less than 0.05. In these cases, we only see differences ranging from 4% to 6%.

Men on average report experience with 2.43 programming languages, whereas women report experience with 2.05 languages, a significant result with a P-value of less than 0.05.

### 4.5.3 Programming Concepts

Among applicants who mentioned specific programming concepts, women reported **Basic Language Knowledge**, which includes loops, if-statements, and variables, 14% more than male applicants did. This result is significant with a P-value of less than 0.05.

There are small differences in mentions of data science, object oriented programming, sorting, searching, and data structures. However, these results were not statistically significant, so we cannot conclude that there is a gender difference in any mention of programming concepts.

### 4.5.4 Learning Method

We found that men were slightly more likely to learn how to program through employment or self-learning, and women were more likely to learn how to program in high school, through higher education, and through online courses. This result is not statistically significant with a P-value of greater than 0.05, so we cannot conclude that there is a gender difference in how men and women learn how to program.

### 4.5.5 Experience

On average, women report 6 months of programming experience, and men report 8 months of programming experience. This result is not significant with a P-value of greater than 0.05, so we cannot conclude that there is a gender difference in the amount of experience within applicants to SE.

Table 21: F-Measures/Accuracies for predicting gender using Additional Information (in %)

| Group | Unigram | Bigram | Trigram |
|---|---|---|---|
| BEE | 60/58.4 | 52/51.3 | 53/50 |
| OTHER | 78/77.3 | 81/77 | 93/89.2 |
| SE | 86/83.7 | 93/86.3 | 93/95.2 |

## 4.6 Additional Information

**Document Frequencies:** We see a difference in word choice between men and women when answering a question with no restrictions on the content of their answer.

In BEE, more men mention "sport" (10.9% of men vs. 7.1% of women) and "compute" 4.7% of men vs. 2.3% of women). More women mention "educate" (17.2% of women vs. 12.2% of men), "science" (17.9% of women vs. 13.4% of men), "develop" (15.1% of women vs. 10.7% of men), "community" (14.8% of women vs. 10.8% of men), and "create" (8.5% of women vs. 5.0% of men).

In SE, more men mention "compute" (27.8% of men vs. 20.8% of women) and "game" (9.2% of men vs. 3.8% of women). More women mention "attend" (16.7% of women vs. 10.4% of men), "English" (12.8% of women vs. 7.2% of men), "study" (21.5% of women vs. 16.5% of men), "parent" (8.7% of women vs. 3.7% of men), "love" (14.2% of women vs. 10.1% of men), and "creative" (8.7% of women vs. 4.6% of men).

In the OTHER programs, more men mention "sport" (10.2% of men vs. 5.7% of women) and "team" (16.3% of men vs. 12.4% of women). More women mention "art" (7.3% of women vs. 3.3% of men), "volunteer" (9.9% of women vs. 6.4% of men), and "passion" (13.6% of women vs. 10.4% of men).

**Logistic Regression:** The results for predicting gender based on Additional Information are shown in Table 21. As before, the predictive power of logistic regression decreases with increasing gender balance within the group.

# 5. DISCUSSION

## 5.1 Similarities

Regardless of gender, the most commonly mentioned topic in responses to "Why are you interested in engineering?" is Technical Interests. Female and male applicants seem to share the same interest in Engineering in all program groups. SE applicants show more technical interest in engineering than other programs.

In general, female and male applicants to SE mention the same motivation for studying engineering. Family is more popular among female applicants, not because female applicants to SE mention it more compared to other programs, but because male applicants talk about it less than men in other programs, as can be seen in Tables 15 and 14.

In SE, we do not see a large gender gap in self reported programming experience, as shown in Table 20. This suggests that students who are exposed to computer science do not differentiate themselves through the number of languages they learn, nor in the amount of programming experience.

In BEE, the differences between female and male applicants are minimal. We see evidence for this in the semantic analysis presented in Section 4.1.2 where there is only one topic that shows a gender difference, and we observe this in our inability to reliably predict gender based on any question as shown in Tables 11, 18, 17, and 21.

Based on Tables 14 and 15, Contribution to Society and Engineering Interests are inversely proportional, regardless of gender.

## 5.2 Differences

### *5.2.1 Depth vs. Breadth*

The overarching gender difference throughout the analysis is that men differentiate themselves through depth of experience, and women through breadth of experience. To study engineering, all applicants must demonstrate knowledge in mathematics and sciences through their academic work. However, we see male applicants differentiating themselves by highlighting their initiative to acquire more technical skills through their work experience, extracurricular activities, reading interests, and the topics they mention when asked why they are studying engineering. Female applicants differentiate themselves through demonstrating a wide range of experiences and capabilities. This is suggested by the fact that women mention a wider variety of topics when asked why they are studying engineering, their extracurricular activities place an emphasis on leadership and artistic pursuits, they often take service jobs, and they choose to discuss more non technical reading material.

In SE, men are more likely to report technical extracurriculars, as seen in Section 4.3, even though there is only a small gender difference in the reported amount of programming experience. This provides further justification that women differentiate themselves through breadth of experience even when they are extremely technically focused.

The gender difference in depth versus breadth is much smaller in BEE. The difference in the number of topics mentioned between men and women is the smallest across these two programs. We also only see a statistically significant difference between men and women in one topic, love of science, which is extremely common across all applicants. The small difference is consistent with our inability to predict gender in BEE.

We also see this in the syntactic analysis of reasons, where women mention “improve” and “health” more in the BEE group, and “people” more in the SE group. It is an interesting difference because BEE includes programs that focus on helping others, and SE is often the farthest removed from directly working with people.

### *5.2.2 Desire to Serve Society*

Women show a stronger desire to contribute to society and improve the world around them. We see this in their motivation to study engineering in “Engineering Interests and Goals” in the OTHER group of programs where they are more likely to mention “Contribution to society”. We also see this in the syntactic analysis of this field where they mention “improve” and “health” in the BEE group, and “people” in the SE group. This is also evident in their work experience where women mention “assist” and “teacher” more often than men. Finally, we see this in extracurricular activities, where women mention “volunteer” more frequently than men. Our findings in this section agree with [3, 4].

### *5.2.3 Influence*

Women are more likely to mention personal influences in their decision to study engineering. This is prevalent in SE, where women mention “Family” reasons more than men. This expands on the findings in [19].

# 6. CONCLUDING REMARKS

The main findings of this paper are that men differentiate themselves through having technical depth in their experiences, and women differentiate themselves through having a breadth of experiences. We see similar behavior in Software Engineering, even though women and men show similar levels of technical know-how. We see smaller gender differences in applicants to Biomedical and Environmental Engineering where there is gender equity. Finally, women mention more of a desire to serve society, and they mention more interpersonal reasons for studying engineering than men.

We infer that in order to attract more women to study engineering, it must be presented as a profession that can help others and allow for a broad range of careers and learning opportunities. A key part in fostering this new image of engineering lies in encouragement from family and role models who practice engineering.

## 6.1 Future Work

In future work, we intend to conduct data driven analysis of gender differences at various stages in STEM students’ academic careers; e.g., investigating the effects of university-sponsored outreach and mentorship programs on applicants, and correlating depth and breadth of expression at the time of admission to academic and career success. We also plan to investigate and compare gender differences in graduate school applications to those in undergraduate applications. We also want to expand the scope of our studies to include non STEM programs in our analysis, and conduct comparisons of differences in STEM vs. non-STEM programs.

# 7. REFERENCES

[1] T. Busch. Gender differences in self-efficacy and attitudes toward computers. *Journal of Educational Computing Research*, 12(2):147–158, 1995.

[2] N. Chinchor. Muc-4 evaluation metrics. In *Proc. of the 4th Conf. on Message Understanding*, 1992.

[3] A. B. Diekman, E. R. Brown, A. M. Johnston, and E. K. Clark. Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. *Psychological Science*, 21(8):1051–1057, 2010.

[4] J. Eccles. Where are all the women? gender differences in participation in physical science and engineering. In *Why aren’t more women in science?:*

*Top researchers debate the evidence*, pages 199–210. American Psychological Association, 2007.

[5] M. Feng, J. Roschelle, C. Mason, and R. Bhanot. Investigating gender differences on homework in middle school mathematics. In *Proc. of the Int. Conf. on Educational Data Mining (EDM)*, pages 364–369, 2016.

[6] J. L. Fleiss, B. Levin, and M. C. Paik. Determining sample sizes needed to detect a difference between two proportions. In *Statistical Methods for Rates and Proportions*, pages 64–85. John Wiley & Sons, Inc., 2004.

[7] P. W. Foltz, W. Kintsch, and T. K. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307, 1998.

[8] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[9] D. Hango. Gender differences in science, technology, engineering, mathematics, and computer science (STEM) programs at university. *Insights on Canadian Society*, 12 2013.

[10] S. Hussain, J. Hazarika, and P. Buragohain. Educational data mining on performance of under graduate students of dibrugarh university using r. *International Journal of Computer Applications*, 114(11):10–16, 2015.

[11] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 04 1997.

[12] Z. J. Kovacic. Early prediction of student success: Mining students enrollment data. In *Proc. of Informing Science & IT Education Conference (InSITE)*, 2010.

[13] H. M. Matusovich, R. A. Streveler, and R. L. Miller. Why do students choose engineering? a qualitative, longitudinal investigation of students' motivational values. *Journal of Engineering Education*, 99(4):289–303, 2010.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[15] J. Ramos. Using TF-IDF to determine word relevance in document queries. In *Proc. of the Instructional Conf. on Machine Learning*, volume 242, pages 133–142, 2003.

[16] M. Saarela and T. Karkkainen. Discovering gender-specific knowledge from finnish basic education using PISA scale indices. In *Proc. of the Int. Conf. on Educational Data Mining (EDM)*, pages 60–67, 2014.

[17] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603, 2016.

[18] L. Shashaani. Gender differences in computer attitudes and use among college students. *Journal of Educational Computing Research*, 16(1):37–51, 1997.

[19] A. Y. Smith. *They chose to major in engineering: A study of why women enter and persist in undergraduate engineering programs*. PhD thesis, 2012.

[20] A. Sullivan and M. U. Bers. Girls, boys, and bots: Gender differences in young children's performance on robotics and programming tasks. *Journal of Information Technology Education: Innovations in Practice*, 15:145–165, 2016.

[21] B. Trstenjak, S. Mikac, and D. Donko. KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, 69:1356–1364, 2014.

[22] T. Zimmermann. Card-sorting: From text to themes. In *Perspectives on Data Science for Software Engineering*, pages 137–141. Elsevier Science, 2016.