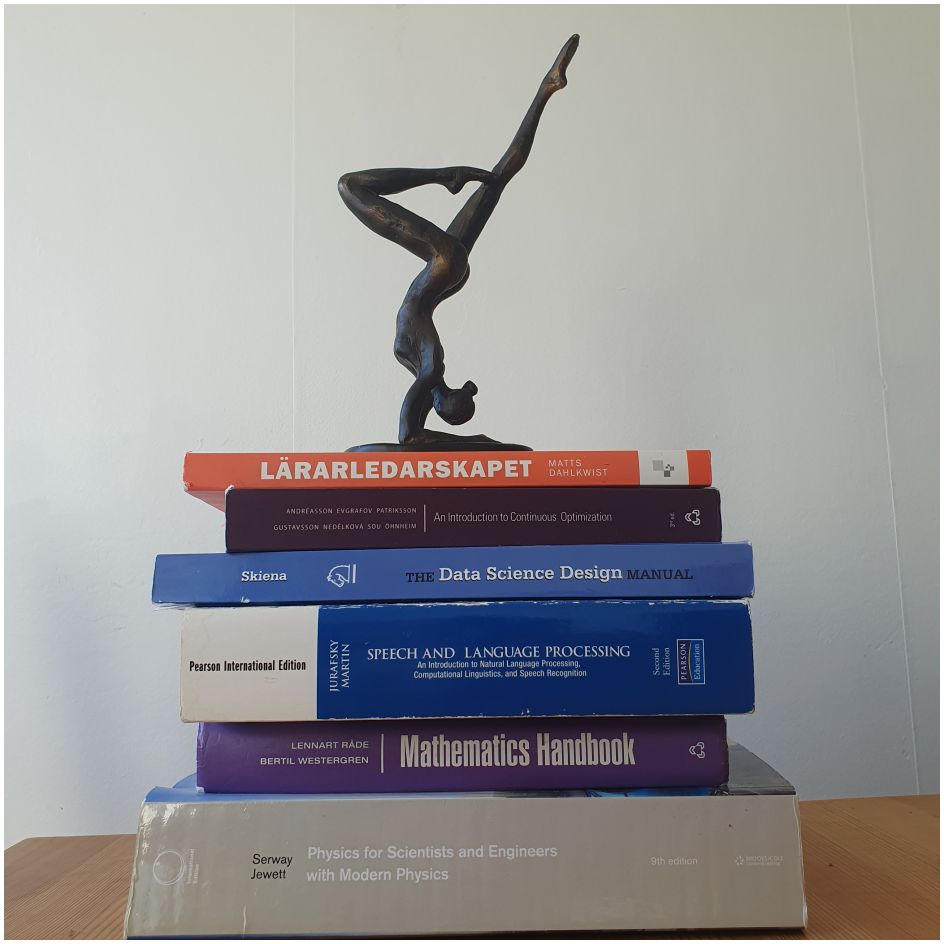

Machine Learning for Detecting Gender Bias at Chalmers

Using Course Evaluations to Fight the Patriarchy

Master's thesis in Computer science and engineering

LINNEA NILSSON

SARAH LINDAU

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's Thesis 2023

Machine Learning for Detecting Gender Bias at Chalmers

Using Course Evaluations to Fight the Patriarchy

Linnea Nilsson Sarah Lindau

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Machine Learning for Detecting Gender Bias in Course Evaluations
Using Course Evaluations to Fight the Patriarchy
Linnea Nilsson Sarah Lindau

Supervisor: Peter Ljunglöf, Department of Computer Science and Engineering
Examiner: Moa Johansson, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A statue of a woman in an elbow stand on top of various literature from courses at Chalmers. The picture is taken by Sarah Lindau.

Typeset in LaTeX
Gothenburg, Sweden 2023

Machine Learning for Detecting Gender Bias in Course Evaluations
Using Course Evaluations to Fight the Patriarchy
Linnea Nilsson
Sarah Lindau
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis studies gender bias in course evaluations through the lens of machine learning and NLP. Different methods are used to examine and explore the data and find differences in what students write about courses depending on the gender of the examiner. The data is also examined using more traditional statistical methods to get an understanding of how the students' impressions of the courses are related to the gender of the examiner. Other aspects related to gender and gender bias are also examined, such as how the proportion of female students relates to the gender of the examiner and whether male or female examiners give different grades to their students. Student grades and teaching language are also factors that are being examined to see whether there is any bias against female examiners or students that is easily detectable in the data. The main findings are that courses with female examiners seem to get lower overall impression scores than those with male examiners. Courses taught in Swedish also receive lower scores, compared to the English courses. No clear patterns as to what words are used when writing comments about a course with a male or female examiner were found. When trying to predict the author gender the patterns were clearer, finding that men write more words directly related to the course and women write more words related to communication.

Sammanfattning

En masteruppsats som studerar könsbias i kursvärderingar med hjälp av maskininlärningsmetoder och NLP. Sammantaget undersöks datan på olika sätt för att se om några tecken på bias mot kvinnliga examinatorer eller studenter är lätta att hitta. En stor del av arbetet går ut på att träna olika maskininlärningsmodeller på textkommentarer för att ta reda på vilka ord som är viktiga för att förutsäga hur stor del kvinnor det är bland eleverna i en kurs och vilket kön examinatorn har. Med mer traditionella metoder undersöks huruvida kvinnliga och manliga examinatorer ger olika betyg till sina elever beroende på elevens kön och om språket som kursen hålls på har någon betydelse för resultaten. De primära resultaten är att kurser med kvinnlig examinator får lägre betyg i helhetsbedömningen än de kurser med manlig examinator. Kurer som lärs ut på svenska får också lägre resultat i bedömningen, jämfört med engelska kurser. Inga tydliga mönster för hur studenter skriver om kurser med kvinnliga respektive manliga examinatorer hittades. Vid förutsägandet av könet på kommentarsförfattaren hittades tydligare mönster om att kvinnor skriver ord mer relaterade till kommunikation, medan män skriver mer ord direkt relaterade till kursen och dess innehåll.

Acknowledgements

We want to thank our supervisor Peter Ljunglöf for all his support and patience during this project. Linnea Nilsson, Sarah Lindau, Gothenburg, May 2023

Contents

List of Figures

List of Tables

1
Introduction

1.1 Gender Issues in Academia

In [1], we can read that women in academia report struggling in the workplace because of their gender. They find themselves denied opportunities that are given to their male colleagues. Among the American female faculty members that were surveyed, 29% found that they were passed over for promotions or other opportunities by their male colleagues. The same number for the male faculty members was 11%. Similar patterns are found in Norway and Italy in [2], where they find that the proportion of women among staff is in the shape of a pyramid, with fewer women in the top positions. This pattern is stronger in Italy than in Norway, but still clear for their Norwegian data. When the universities in Sweden were compared, they found similar numbers as those for Norway [3].
Female scientists publish less than their male colleagues [2]. One reason for this seems to be that women are more often in positions where they need to devote more of their time to teaching. Some of the explanations for why women tend to publish less are that women get a higher workload since students often perceive them as easier to approach, but also since women are often given extra activities from their superiors where they can show that they have female faculty [4]. These extra activities take time from the time that the women could have used for research, and can thus harm their ability to publish at the same rate as their male peers.
These issues seem to be especially apparent in departments working with any of the STEM, i.e. Science, Technology, Engineering, and Mathematics subjects [4]. Women are both very underrepresented in terms of numbers, but many also report that they feel unwelcome or like their working spaces aren't designed for them. These findings are also consistent with an equality ranking of Swedish universities where all of the Technology directed universities were at the bottom [3].

1.1.1 Gender Issues and Equality Work at Chalmers

In [3], they rank the different universities in Sweden on the distribution of gender in different academic positions. All the universities with a focus on technology were near the bottom, and out of all the universities, Chalmers was ranked the lowest. However, the report also lifted that the university is working on mitigating the equality issues and have put aside a large amount of money to combat it over the coming ten years (from 2019). To combat this, Chalmers has taken some different directed initiatives [5]. Some of these are the goal that 40% of the professors should be women in 2028 and *Gender Initiative for Excellence*, GENIE.

The *Plan for Equality and Equal opportunities* (Swedish title: "*Jämställdhets-och lika villkorsplan*") outlines the different ways in which Chalmers works to mitigate these equality issues [6]. The plan identifies the different problems related to equality that the university faces, such as discrimination, and sexual harassment, and the fact that the more senior a professional position in academia is, the less likely it is to be held by a woman. In conjunction with the presentation of the issues, they also present goals for resolving the issues and strategies for reaching the goals. For example, one of the strategies for mitigating the issue of inequality in the recruitment and promotion of staff is to always have at least one man and one woman present at an interview. Some of the goals and issues in this plan are related to GENIE, or *Gender Initiative for Excellence*, which is a project directed at improving equality at Chalmers [7]. The project has two main goals, creating and presenting ways of changing the culture in the departments and directly increasing the number of female faculty. There are many different projects that have been executed under the umbrella of this initiative[1] and this thesis is part of one of these projects. Other than projects, GENIE also produces yearly reports on the current equality situation of the different departments at Chalmers. They contain information about salaries, sick leave, and what proportion of different gendered staff there are for different categories of staff. There is also a program for hiring talented female guest researchers to get a higher proportion of female staff in the departments and to bring more female role models to other staff and students.

1.2 Gender Bias against teachers in Course Evaluations

Previous work indicates that female teachers are rated lower than male teachers in teaching evaluations [8], [9]. This has real-world consequences for female teachers that are consistently rated lower than their peers. These evaluations are often used as a metric in hiring or promoting decisions, which means that a low result can have serious effects on the careers of academics [8], [10]. Lower teaching ratings can hinder the careers of academics in several ways: by getting perceived by their superiors as performing worse than they do, by making them doubt their own ability, and by making them put more time and effort into teaching as opposed to research when they are not truly struggling in their teaching abilities.
The gender bias that was found in [8] was directed mainly at younger teachers that are still early on in their careers, meaning that these effects can have a significant impact on their future in academia. When it came to lecturers they in practice found no bias of statistical significance. Furthermore, they found that it was mainly male students that rated female instructors lower, while female students - while still giving the female teachers lower grades did so to a smaller extent.
However, gender bias is not only present in how the teachers are rated but also in what students base their evaluations on. In [10] they found, by analyzing the comments related to the evaluations, that women tend to get evaluated based on

[1] `https://www.chalmers.se/en/about-chalmers/organisation-and-governance/equality/genie-gender-initiative-for-excellence/current-projects-and-reports/`

their looks or personality to a higher extent than men, while men were more often evaluated based on their skill and competence. Women are also generally viewed as having less seniority and are more commonly referred to as the more casual “teacher” as opposed to “professor”.

1.3 Problem Description

In this thesis we broadly examine the issue of gender bias at Chalmers and whether or not any bias can be detected in the course evaluation data, using statistical analysis and machine learning methods. There are a few different aspects of this issue that we look into, mainly, we investigate if there is a difference in the students’ opinions of their teachers based on the teacher’s gender. Part of this pursuit is also to explore whether the students’ genders affect their opinion of the examiners. We explore whether the students’ opinions and other characteristics can be detected in the language they use when answering the course evaluations. Part of this is also to investigate what other factors that may be related to these issues and see if there are any differences in language or the students’ impressions of the courses and examiners, such as the teaching language of the course or the grades that the students receive in the course.

2
Theory

2.1 Sex and Gender

When working with gender it's important to note that sex and gender are not the same, while they are related [11]. Gender is a much more complex concept than a simple "woman" or "man" label, and there are many more possible gender identities. Gender has a few different definitions or parts, such as being a social construct or part of an identity someone presents [12]. It is important to note that a person's gender may or may not align with their biological sex. There are also many possible gender identities, including the lack thereof. A person's sex is typically either determined as male or female and is determined by biological factors such as chromosomes. Gender, on the other hand, is a wider term and is rather concerned with cultural and behavioral aspects related to sex [11]. There are many possible gender identities, but the terms related to male and female would be man and woman [12]. While we are well aware that sex and gender are two related, but distinct features of a person, we use sex as a proxy for gender in this thesis. When working with the students' responses, self-reported gender would have been ideal, but we only have access to the number of students with "man" or "woman" respectively as legal genders in the class. When working with the examiners, however, it is the students' perception of the examiner's gender that is interesting. As that data isn't readily available without making a large survey, we have to use a proxy. We deem using sex as a proxy for gender is a good enough solution as only approximately 0.4% of the Swedish population define themselves as transgender according to a report from *Folkhälsomyndigheten* [13]. Transgender in this case refers to people whose gender in any way differs from the sex they were assigned at birth. While 0.4% isn't a great number, we still understand that this means that some people and their identities are potentially not properly represented in the work we have performed in this thesis. Further, as this is a thesis on a technological subject, and some of the main tasks are binary classification, using sex as a proxy for gender becomes very practical from a technological standpoint as it would be difficult to implement technical solutions for a more inclusive gender representation.

2.1.1 Name as an indicator of gender

To truly capture a person's gender one would need to have the self-reported gender [14]. This is, however, not always a feasible solution. Using first names to estimate is typically a relatively simple solution and the necessary data (i.e. the name of the person) is easy to obtain. This method also excludes the nuances of gender by

making it a simple binary variable. To be able to work with the examiners' genders in this thesis we have made an estimation using their names, according to the process described in section 3.2.5. While still leaving a lot to wish for regarding inclusivity and exactness, this method at least might capture the gender of transgender people that have changed names to something related to their gender. Still, we do see that this method still excludes a lot of the possible gender identities that people may have. We map the examiner names to legal genders and then use these legal genders as a proxy for how the students perceive the examiners' gender identities. Thus, there are two places where errors may be induced. However, as it is not the absolute truth about the examiner's gender that is interesting in this case, but rather the students' perception of the examiner's gender, this might better correlate to the examiner's name and how they present themselves in public. This work is only done for the examiners' genders and the information we have about students is their legal gender (or rather the number of registered students in a course that respectively have male and female as their legal gender).

2.2 Gender Bias

Bias is to unfairly favor or disfavor something or someone based on personal opinions or prejudices. There are many forms of bias towards other people, for example, based on race or nationality, age, or, as is focused in this thesis, gender.

A new report from the UN concludes that almost 90% of the population of the world holds a bias against females [15]. It is based on the Gender Social Norms Index (GSNI), analyzing gender bias in 75 countries in areas like education, work, health, and politics. The report also includes GSNI trends for countries representing 59% of the total population, which show that while some countries have improved their views on women, people in other countries have attitudes that have worsened in recent years. The bias consists of everything from a third of the global population being accepting of violence against women to thinking that men are better suited for leadership roles in politics, a belief held by roughly half the global population. This shows that bias and prejudice toward women permeate all layers of society and contribute to and fortify the gender power gaps in society, consisting of gender pay gaps and the fact that while women make up for half the global population and vote to the same extent as men they only hold 24% of the parliament seats globally.

Beyond the purely humanitarian and fairness factors of striving for gender equality in our world, our society would grow and thrive if all people were given the same chance of success in their field not limited by gender bias [16]. Groups that are more diverse tend to work better and enjoy work more than more homogenous groups. In the *Women Matter* report from 2007 they showed the importance of women in higher-up and leadership positions [17]. People tend to want to work in positions where they can see people similar to themselves represented. This can mean people of the same gender presentation, skin color, disability, or background.

2.2.1 Gender Bias in an Academic Context

In [2], they compare the performance of male and female scientists in Norway and Italy and explore some factors that may play a part in differences in performance. In their Norwegian data 41.5% of the assistant professors were women, 46.5% associate professors, and 26.1% full professors. The same numbers for the Italian data were 47.2%, 35.2%, and 18.1%. While there is a higher proportion of women in higher positions in Norway, there is still a clear gender gap for the top positions. They found that there is something that stops women (or helps men) to perform better, by their metrics, and that this should be investigated further. Their study takes into account a lot of different factors, such as the number of publications, positions in author lists in publications, and the publication patterns in the field, to be able to compare the different scientists' performance. In the report [3], they found that the proportion of women in different academic ranks in Swedish universities was compared. Around 60% of the students are women, 47% of the Ph.D. students, 46% of the lecturers, but only 28% of the professors.
Male scientists tend to publish more than their female colleagues, something that has been witnessed consistently for many years and across different fields according to [2]. This difference does, however, seem to be getting smaller. There also seems to be some relation to the career patterns of men and women (women typically work for fewer years) and that women are more often in positions where a proportion of their work time needs to be dedicated to teaching. [2] found that while most scientists perform similarly across men and women, the difference in the number of publications becomes apparent for the scientists that publish most and least, i.e. on the outer ends of the spectrum. Among those who publish the least, there are a larger proportion of women and there is a larger proportion of men among the scientists who publish the most papers. In the findings of the study comparing Norway and Italy, it is important to note that these differences were noticeably smaller for the Norwegian scientists, which could be related to the family responsibilities being more equal for the genders. However, there is still some difference between the top-performing scientists.

2.2.2 Gender Bias in Teaching Evaluations

Previous work has shown that there is indeed gender bias in how students evaluate their teachers in teaching evaluations [8], [9], [18]. In [8], they found that despite students getting the same results and working the same amount to get there, their female teachers received lower ratings than the male teachers. The bias against female teachers seems to be strongest with male students. The female teachers were perceived as 21% of a standard deviation worse than their male colleagues by the male students, while female students only thought of the female teachers as 8% worse than their male counterparts. This effect was most prominent for the more junior female teachers, such as Ph.D. students, where the females got a 28% of a standard deviation lower rating than the males. Similar results were found by Boring in [9]. She studies what overall satisfaction scores students give their professors in the teaching evaluations. In general, the students rate the male professors higher than they do the female professors, the male students by 0.354 points in the scoring system

and the female students by 0.192. Boring concludes that there is a double standard to what students take into account when assessing male and female professors and hold them to different standards. This leads to female teachers having to spend more of their time on improving their teaching, despite having teaching that is of the same quality as their male colleagues that receive higher scores in the evaluations. This can have serious consequences for women in academia who also need to spend their time doing research. As has been noted earlier, women in science publish less than men, and having to spend more time on teaching is one of the explanations for why that is [4].

[18] argues that there may be some ambiguity as to whether or not female faculty receive lower evaluation scores than their male colleagues, and that it may not hold true in all cases. However, it is clear that there is evidence pointing to female teachers in male-dominated fields receiving lower scores from male students, compared to male teachers. One reason may be that male students hold female teachers to a higher standard compared to male teachers.

2.2.3 Gender Bias in Teaching

The gender bias that teachers have can really affect the students they are teaching. This was examined on Greek high school students in [19]. The findings indicated that not only did the gender bias of the teachers have an effect on the students' grades, but also on what line of study (if any) they chose to pursue later on. Another interesting finding was that biased teachers also tended to be worse at teaching compare to their less biased counterparts. In their study, they found that regardless of subject, all teachers were biased against boys, in favor of girls. With the largest bias in computer science and the smallest in economics. Both male and female teachers had similar amounts of gender bias in their study, although male teachers seemed to be a bit more biased in economics. Female teachers were, however, more biased in mathematics and physics. Something very interesting is that although the bias in general was for girls, it also affected girls more negatively than it did boys, by for example making girls attend fewer classes – an effect that was less prominent in male students. A similar result has been found in [20], in regard to education in mathematics in middle school students in Italy. Female students with biased teachers don't only get worse results, but they also lose their confidence in their abilities and choose less demanding education. This means that teacher bias earlier in their education can have a significant impact on future results. Teacher bias in a student's high school years can have a statistically significant impact on their future choice of education and whether or not they decide to pursue higher education at all [19]. This means that teacher gender bias has an effect on their students' financial situation throughout their lifetime.

[20] found that male teachers were more biased in mathematics, while female teachers were more biased in literature. The bias was for men in mathematics and for women in literature, which might be related to teachers favoring the positive bias for their own sex. [19] found indications of the bias by female teachers having a larger impact on male students than the bias of male teachers did, while male teachers affected female students to a larger degree [20]. Having teachers biased in favor of male

students had a significant negative impact on female students. Interestingly, teacher bias did not in fact seem to have a lot of impact on their recommendation for the students' continued education, but rather on the students' belief in their own ability, which in turn affected their results and choices for the future.
Another interesting finding in [19] was that the teachers' bias seemed to have some effect on the absence of a student. Namely, if a teacher was biased toward male students, boys would attend more classes and girls would attend fewer. This effect was especially prominent in unexcused absences. That can, in turn, affect a student's results since they miss classes and will not have the same possibilities to learn the material.

2.2.4 Examining gender bias using NLP

Most work studying gender bias in text data so far is based on English data, such as for example [21]. One of the most prominent aspects of this work is regarding finding lists of words that can be either assigned as feminine or masculine, for some different contexts.
There are many interesting techniques that can be used for the purpose of finding different types of gender bias in text, such as word embeddings and working with different types of language modeling as in [22]. Further, more traditional statistical methods are also used for this purpose.

2.3 Courses and Grading at Chalmers

The Chalmers University of Technology has educational programs on three levels: Master, Bachelor, and preparatory years [23]. Courses in the programs on the bachelor's degree level are mainly held in Swedish even if there are some exceptions to this. The master's courses on the other hand are held in English with some exceptions. The preparatory year is held completely in Swedish.
Chalmers has two grading scales. The simplest scale is either pass or fail. The other scale has passing numeric grades ranging from 3 to 5, and a failing grade of "U" [24]. When working with student grades in this thesis, we have excluded all courses with the pass/fail grading system, and instead only worked with the courses with number grades. We then assigned the failing grade, "U", a value of 0 in all calculations.
There are some different ways in which students' results in courses can be evaluated at Chalmers, where a written exam is the most commonly used [25]. Other forms of examination include oral assessment, take-home exams, shorter tests, and projects that can be conducted in a group or individual format. The written examinations can both be in-person at the campus or online. The study year at Chalmers is divided into four study periods that all end with an examination period or "exam week" as it's commonly known. Students have the possibility to write re-examinations, which can either be done during the ordinary exam week or at one of the re-exam periods that are spread out throughout the year. As long as a course is still running, a student may take the exam as many times as they please, so they can improve their results even if they already have received a passing grade.

Chalmers has a close connection with the University of Gothenburg and they work together in many ways, both in terms of research and education [26]. Among other things, they have two departments that are integrated between the two universities, *Computer Science and Engineering*, as well as *Mathematical Sciences*. This means that the universities are closely knit and there are several courses that are taken by students from both universities.

2.4 The Course Evaluation Process at Chalmers

Central to this thesis is understanding what the course evaluation at the Chalmers University of Technology is and what consequences it has for the examiners of the courses that are reviewed. The information used in this section mainly comes from the process description provided by Chalmers, which is only available in Swedish. An interested non-Swedish speaker can read some of the information on the Chalmers student website.[1]

The main purpose of the course evaluation process is to continuously improve the quality of the education and courses provided at Chalmers [27]. The process of the evaluation is described in the *Processbeskrivning* [28]. Each course is evaluated by a group consisting of students, the examiner, the program responsible for the program the course belongs to, and another representative from that program. This group is called the "Kursnämnd". At least two of the students in the Kursnämnd are selected student representatives from the course, given that the course has more than ten students registered. The student representatives are most often selected "randomly" among the registered students, but can also be selected in other ways. When selected "randomly", the choice is not completely random, but the computer program that makes the selection takes some parameters into account, such as the program of the student, their biological sex, and whether or not they have been selected for a Kursnämnd earlier. However, the examiner can always ask if there are any volunteers and then these students can be used as representatives for the students in the course instead. For some of the study programs, these student representatives are instead selected by a committee of students from the program called a studienämnd.

There should be in total three meetings dedicated to the course evaluation process, a start-up, a middle, and a final evaluating meeting [28]. The first meeting is focused on getting the student representatives on board with what the examiner wants the course to teach and they should also explain the course evaluation process to the student representatives and what is expected of them. The middle meeting is focused on what is going on in the course, how it's going, and if there is anything specific that should be asked in the course evaluation forms later on. This also gives the examiner a chance to make changes that can make the course better before the course has ended. The final meeting is held by the entire Kursnämnd after the results from the course evaluation forms sent to the students have been collected and presented in a pdf report that can be discussed at the meeting. During the meeting, a protocol

[1] `https://www.chalmers.se/en/education/your-studies/plan-and-conduct-your-studies/course-evaluation/`

should be written, which is then made available to relevant students and staff.

2.4.1 The Course Evaluation Forms

The actual course evaluation forms that are sent out to all students in a course and the student replies are after this point referred to as "course evaluations" [28]. There are some standard questions that are common for all course evaluations at Chalmers. However, examiners can also add questions to the course evaluations if they want more feedback on specific aspects of the course. As a standard, the course evaluations are sent out to all students registered to a course (both new and re-registrations). However, in the case that there are other students that should be able to answer, such as PhD-students or students from another school (i.e University of Gothenburg), they are added manually by the departments.

2.4.1.1 The Standard Questions

There are 12 groups of questions that are used in all course evaluation forms, and these are consistent through the years with the exception of small reformulations. One of these question groups, question 7 Working environment, was added in 2018. These new questions have been excluded from our work since there would be no data from before that year. The questions are divided into two main categories: questions with numerical answers and questions with free-text answers. In the course evaluations these are often presented together as one or more numerical answers together with a free text comment question relating to the same subject. On each numerical question, the students give a score between 1 and 5 on a Likert scale. In the study year 2021–22, the standard questions were formulated as follows, where the questions are separated into groups with free text questions followed by the corresponding numerical question or questions. Note that the headlines for each question group are added by us to make it easier to understand the theme of each question group. Each of the standard questions is either marked with (N) if the answer is numerical and (T) if the answer is in a text format.

1. Prerequisites
   (a) (N)"I had enough prior knowledge to be able to follow the course"
   (b) (T)"Comments (For example: Did the course start at an adequate level? Was it assumed that you had knowledge which you could not get from your previous studies? etc.)"
2. Learning outcomes
   (a) (N)"The learning outcomes (see course syllabus) clearly describe what I was expected to learn in the course"
   (b) (T)"Comments (For example: Should some learning outcome be clarified? In what way? etc.):"
3. Learning
   (a) (N)"The course structure (as divided into lectures, exercises, lab sessions, simulations etc.) is appropriate in order to reach the intended learning outcome of the course"
   (b) (N)"The teaching worked well"

   (c) (N)“The course literature (including other course material) supported the learning well”
   (d) (T)“Comments (For example: Do you think that something should be changed in the course structure? What, and in what way? What made the teaching work well or less well? Are there aspects of the teaching in this course which could be high-lighted as a good example for other courses? etc.):”
4. Assessment
   (a) (N)“The assessment (including all compulsory elements, exams, assignments etc.) tested whether I had reached the intended learning outcomes of the course"
   (b) (T)"Comments:”
5. Course administration
   (a) (N)“The course administration (information during the course, course memo, course homepage etc.) worked well”
   (b) (T)“Comments (For example: What are the main reasons for your rating of the course administration? Are there aspects of the course administration in this course which could be high-lighted as best practice for other courses? etc.):”
6. Workload
   (a) (N)“The course workload as related to the number of credits was..."
   (b) (T)"Comments (For example: What is the main reason behind your rating of the workload? Would the perceived workload have been lower if deadlines in the course would have been distributed in a different way? How many hours have you on average spent on the course per week? etc.):”
7. Working environment
   (a) (N)“The organization, content and teaching of this course have been designed and executed so that everyone can feel included, welcome and seen”
   (b) (T)“Comments:”
8. Overall impression
   (a) (N)“What is your overall impression of the course?”
   (b) (T)“Comments (For example: What are the main reasons for your overall impression of the course?):”
9. (T)“How has the interaction between students and teachers worked in this course?”
10. (T)“If the course has contained group activities (lab sessions, simulations, group work, projects, or other types of cooperation between students): How have group roles and cooperation between students worked in this course?”
11. (T)“What should be kept for the next round of this course?”
12. (T)“Is there anything that should be changed for the next round of this course, and if so: How?”

2.4.2 Student Answers

Under normal circumstances, the course evaluations are sent out on the Monday after the exam week [28]. Thus, the students get the course evaluations after the exams are held, but before knowing their results.
The students then have the possibility to answer for two weeks, with the exception of the last courses of the school year where they have until the first Wednesday after the start of the new semester. The students also receive two reminders to fill in their course evaluations. Once the course evaluation forms are filled in and the deadline for doing so has closed, the answers are filtered manually to remove offensive language and insults. They are also anonymized to ensure the privacy of all students. This is done using a standard procedure that should be used for all courses and all course evaluations. Our data is from after this filtering and doesn't contain any offensive language. The examiner gets access to the results from the course evaluations gathered [28]. The individual student answers are not made public.

2.4.3 Consequences of a Poor Evaluation Result

[28] further describes what happens when a course receives poor results in the evaluation. For any course where the average answer to question 8(a) in 2.4.1.1, the overall impression score, from the students is lower than three, the examiner needs to create a plan of action on how to mitigate any issues and improve the course for the next time the course is held. The plan should include both specific actions as well as what consequences these are expected to have on the course and its quality. This document should be created in accordance with the guidelines that are available to the examiners. It is worth noting, however, that while there is no other consequence to the examiner other than needing to create an action plan, consistently bad results in their course evaluations could reflect poorly on them.

2.5 Preparation of Data for Machine Learning

Data that is to be used for machine learning needs to be structured in a matrix-like format in order to be used [29, p. 13]. This means that while some data is already structured like that, other data such as for example text data will need to be represented in some way to follow that matrix format. One such way of representing text data is the *Bag of Words* model. There can be many different factors in that matrix structure that make up a basis for predictions on the data [30, pp. 228–229]. Such factors can be a word or for example, a person's height or age, depending on what type of data we have. Each such factor is often called a *feature*. That means that each data point is made up of a vector of features.

2.5.1 Bag of Words

Bag of Words, *BoW*, is a way to represent language data so that it can be used for machine learning [30, p. 675]. A BoW model has a dictionary of words used to represent the documents. In the simple case, each word in the document is

represented by a vector of binary features (0 or 1) representing whether or not a certain word in the dictionary is present in the document or not. As an example, say the model has a dictionary consisting of three words: *apple*, *yum*, and *pear*. If we then have a document "Apple! Yum!", it would be represented as the vector [1, 1, 0], as the words *apple* and *yum* are present, but not *pear*. What words are present in the dictionary would normally be selected from the words used in the training data. Often, the BoW models also include a count of how many times the word appears in the document but without regard to where in the document the word appears [31, 6.2.3.1 The Bag of Words Representation]. Still, there are some limitations to the BoW model [31, 6.2.3.7 Limitations of the Bag of Words Representation]. For example, a BoW model cannot handle misspellings of words and as it doesn't store any positional information of the words, a lot of meaning that comes from the structure of sentences is also lost in this representation.

2.5.2 Imbalanced Datasets

When working with a classification problem one may come across the issue of imbalanced datasets. Such a dataset has a much larger proportion of one of the classes compared to the other, called a majority class. The other, less common classes are called minority classes [32]. In binary classification, the majority class is often also called the negative class, and the minority class is often called the positive class [29, p. 213]. This notation of positive and negative does not imply that any class is better than the other, it is just simply notation. Having imbalanced classes is a relatively common issue in machine learning applications and can come up in many different domains, in everything from spam filters for emails to fraud detection in the banking industry. Two common ways to deal with this issue are downsampling and upsampling [33].

Downsampling the majority class means that random samples of the majority class are removed from the training data [33]. When we remove instances of the majority class, the data isn't flooded with those samples which gives the classifier a better chance to learn the minority class as well. Downsampling can be combined with upweighting, i.e. making the classifier give the majority class a higher weight when training [32]. That way, while it still only sees the same number of samples of the classes (or at least a lower proportion of the majority class than in the original data) the model is provided with a better representation of real conditions as the majority class in the training data is likely to be the majority class in real life. As we have fewer samples this saves disk space and when there are relatively more samples of the minority class, some classifiers converge faster.

Upsampling the minority class is another way of handling the issue of imbalanced classes [33]. It is in some ways relatively similar to downsampling, but instead of randomly removing samples, random samples of the minority class are duplicated and added to the dataset again. These samples are drawn with replacement, whereas the random sampling in downsampling would be done without replacement.

- Downsampling: Removing samples of the majority class
  - Upweighting: Adding additional weight to the remaining samples in the

majority class
- Upsampling: Copying and adding samples of the minority class

In this project, we have for simplicity chosen to only work with downsampling the data to handle the imbalanced classes.

2.5.3 Cross Validation

Cross-validation is a method to make better use of the data at hand [29, p. 227]. Instead of only training one model on one training set and evaluated on the remaining data, the data is divided into k different training and testing sets. Then k different models can be trained on each of these sets, yielding a more robust result compared to only training one model, where chance plays a bigger part in the final results. Note that all data is used in each of the k training-testing sets, it is only what data is used for training and what data is used for testing that will differ.

2.6 Machine Learning

Machine learning is a collection of models and algorithms that continuously learn and improve based on input data [34]. They are used to solve different prediction tasks based on historical data. An example is training a model to predict whether or not an image contains a cat or using previous data to predict today's price of a certain stock. Two common categories of machine learning tasks are classification and regression [35]. Simply put, classification seeks to categorize data points, while the point of regression is to find a certain value of an output variable based on the data.

2.6.1 Classification

A classification problem is a problem in which we seek to predict a certain label or group to assign to a data entry [29, pp. 16–17]. We should have a finite set of possible labels that can be assigned. One such example is when we want to decide what genre a movie belongs to (i.e. drama, thriller, etc.). Binary classification is a classification problem where the data entries are categorized into one of two classes. Binary classification is utilized in this thesis both when trying to predict the gender of the examiner, the two classes then being male or female, and when predicting if a course is gender balanced. There are many different classification models available, but in this project, we have chosen to work with a logistic regression model and a support vector machine-based classifier.

2.6.1.1 Logistic Regression

Logistic regression is a type of classification algorithm that bases the predictions on a probability score that makes the user able to determine how "sure" the model is of the classification [36]. Logistic regression falls under a category of linear regression methods that can be used to solve classification problems [29, pp. 290–291]. Instead of using the regression line to extrapolate and predict new samples as on the line, it is

used to divide the samples into classes. In the binary classification case, that means that some samples are above the line and thus be categorized belong to the positive class, while other samples are below the line, categorized as the negative class. The quality of this type of division of above and below the line will differ greatly with the dataset and what the differences of the classes look like. It is worth noting that there are both benefits and issues with the linearity of this type of model. For some datasets, a lot of samples will be misclassified as there are many outliers that sway the regression line to not fit the rest of the data well. There are also datasets that simply don't have any linear structure to how the classes are divided in the feature space and that need some non-linear model to obtain a good classification result. However, the simplicity of a linear model brings many benefits such as not being particularly prone to overfitting. In our work, we use the weight vector that these linear models produce in order to see what features are associated with what class in the prediction.
The foundation of logistic regression is a sigmoid (s-shaped) function called the logit function [29, pp. 292–295]. This function is defined in equation 2.1, where f is the logit function and c is a constant that determines the steepness of the slope of the transition between the top and bottom of the function. The value of the function goes between zero and one.

$$f(x) = \frac{1}{1 + e^{-cx}} \tag{2.1}$$

In logistic regression, the logit function is used to produce a probability that a certain data point belongs to a certain class. This is done by first fitting a linear function to the data and then using the logit function, and including a loss function to create a logistic regression classifier.
A good thing about logistic regression is that, for binary classification, it can be used to find the optimal linear division between the classes [29, pp. 295–296]. However, this model still has issues. For example, this type of classifier isn't great with imbalanced classes, as the penalization simply will not add up to make the classifier produce any predictions of the minority class (or at least not more than very few). This means that in cases with imbalanced classes, these need to be dealt with in some other way as described in section 2.5.2.

2.6.1.2 Support Vector Machine

The foundation for a support vector machine classifier is similar to that of a standard linear regression model, such as logistic regression [29, pp. 367–368]. But instead of simply finding the best line that divides the classes, it also seeks a margin on both sides of that line. In many ways, this is a natural extension of the line of thinking that is used in for example logistic regression. By maximizing the margin to the samples of each class, there should be fewer misclassified samples as the line should be a better separation than any other contender.
There are both linear and non-linear SVMs [29, p. 369], but in our project, we have only used linear SVMs so that the results from the different classification models can be interpreted in the same way. In the end, the support vector machine model is based on an optimization problem that can have more or less general constraints,

depending on how it will be used and the data set. While the foundation is to find a line and a margin around it between the two classes, this is not always possible. Sometimes finding a large margin, while leaving some samples misclassified is better and at other times it's a smaller margin with fewer or no misclassified samples.
The *support vector* part of the name comes from the data points (i.e. feature vectors) that touch "support" the margin lines [29, pp. 367–369]. Think about the SVM as the optimal division line, surrounded by two margin lines that are parallel to the division line and at the same distance from it. Then the support vectors are the data points that touch these margin lines. These support vectors should be at most twice the number of features and preferably much fewer than that. An important difference to other models we use, such as logistic regression, is that in a support vector machine, only the support vectors are part of defining the model. This is in contrast to the logistic regression model where all data points play a part in defining the separating line.

2.6.2 Regression

When solving a regression problem, we seek to predict a numerical quantity to assign to a data entry [29, pp. 16–17]. An example of a regression task is predicting the proportion of female students in a course. While there are many different regression models, we have chosen to work with linear regression and ridge regression.

2.6.2.1 Linear Regression

A linear regression model is a type of regression model that is both easy to create as well as interpret [29, pp. 267–271]. It is based on finding the best linear representation of some data points, which can then be used to predict a value for future data points. In the perfect example, all points would be across the same line and that line would then be the linear regression representation of the data. This is, however, seldom the case. Instead, one needs to use the residual error, i.e. the difference between the actual value and what is predicted by the model, the point on the line. This residual error can be defined as in equation 2.2, where y_i is the value from the data for sample i and $f(x_i)$ is the predicted value for sample i. The residual error is denoted by r_i.

$$r_i = y_i - f(x_i) \tag{2.2}$$

While there are other ways to define the error or distance from the predicted line, the residual error is highly related to the data points (features) which makes it useful for linear regression [29, pp. 267–271]. Looking at figure 2.1 the residual error is simply the vertical distance between the point and the line.
When performing linear regression the goal is to find a line that minimizes the squared error between the data points and the line [29, pp. 267–271]. The reasoning for using the squared residual error as opposed to just the residual error is that the sign of the error is irrelevant to determine how good an estimation is, only the size is relevant. However, this could also be solved by using for example the absolute value of the error. To create a linear regression model we want to find the line (or rather

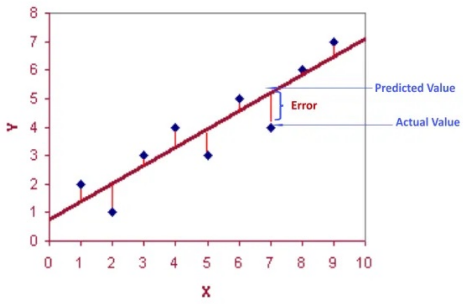

Figure 2.1: A simple visualization of linear regression and residual errors. Image from [37].

vector of coefficients, $\bar{w}$) that minimizes equation 2.4, using $f(x_i)$ from equation 2.3. Equation 2.4 calculates the mean squared difference between the predicted value $f(x_i)$ and the actual value y. Note that the coefficient vector has m coefficients in it, where $m-1$ is the number of features, and that each of the n data points has $m-1$ features and a target value y.

$$f(x_i) = w_0 + \sum_{j=1}^{m-1} w_j x_i \tag{2.3}$$

$$MSE_f(x, y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{2.4}$$

$$minimize\ MSE_f(x, y) \tag{2.5}$$

To solve equation 2.5 and obtain the weight vector for this linear regression, $\bar{w}$, is obtained by following equation 2.6.

$$\bar{w} = (A^T A)^{-1} b \tag{2.6}$$

In 2.6, a $n \times m$ matrix, A, is created by putting all the feature vectors, the x_i, of all the n data points in a matrix, and then padding it by adding a first column of ones [29, pp. 267–271]. b is simply the column of target values, the y_i. This comes from the notion that when we have the best possible line, no value in $\bar{w}$ can be changed to get a better line. Thus, the vector of errors $\overline{b - A\bar{w}}$ has to be orthogonal to all of the feature vectors, x_i, or there would be a way to improve the line.

2.6.2.2 Ridge Regression

Ridge regression is a linear regression model with an added regularization [29, pp. 286–287]. Regularization is a way to make sure that the model has small coefficients unless there is any strong reason for it, not to [38, pp. 109–112]. The reasoning for this is both to avoid the noise of strongly correlated features as well as to make the model have a good level of flexibility. There are two broader categories of regularization – explicit and implicit. Explicit means that we explicitly add to the cost function of the model and implicit can be done by some other means, such as

early stopping in the training, but not by adding to the objective function. Ridge regression is of the first category since a second set of terms is added to the objective function [29, pp. 286–287]. These new terms are dependent on the coefficients, rather than being dependent on the features of the data. That means that in order to make the coefficients larger, the model has to pay a penalty. This makes the loss function for ridge regression as in equation 2.7.

$$loss(w) = \frac{1}{2} MSE_f(x, y) + \lambda \sum_{j=1}^{m} w_j^2 \tag{2.7}$$

In this equation, w is the coefficient vector corresponding to the features, n is the number of samples, $y_i - f(x_i)$ is the residual error for sample i, $m-1$ is the number of features and finally, λ is a parameter used to tune the strength of the regularization. Note that in equation 2.7, the term $\lambda \sum_{j=1}^{m} w_j^2$ is the added regularization. The reasoning behind including $\frac{1}{2}$ in the loss function is simply for technical reasons [29, p. 279]. However, scaling the loss function for all samples has no actual effect on the results.

2.6.3 Evaluation

While there are a few formal ways to evaluate both classification and regression models, this task is not as trivial as it might seem [29, pp. 210–223]. Even if the resulting evaluation metrics might show good results, the model might still have issues with faulty implementation, only making certain types of errors or other issues. There is always the possibility that the model could yield better results, and by just looking at the metrics there is not really a way of knowing. One way to discover and mitigate this type of error is by manually looking at some examples of what predictions the model is making with what data. To deepen the understanding of the model it is also interesting to examine what it predicts for completely new data entries. A key element of evaluation is to determine whether the results are reasonable for this task and use human judgment to determine whether or not the model is performing well [39, pp. 62–64]. The evaluation metrics are often somewhat standardized for a certain field or task. When choosing a metric it should be consistent with what is commonly used, unless there is a certain reason for making an exception.

2.6.3.1 Classification Model Baselines

When creating a classification model, there needs to be some way of knowing if the model is performing well enough [29, pp. 210–211]. For this purpose, baseline models are used and the results from the actual model can be compared to that of the baseline. One such model would predict the labels at random, not at all based on the data entry itself. Another method is to always predict the most common label in the training data. When we do have some notion of how the data is distributed, this method is better than assigning random labels. Other baseline models can be predicted based on one feature of the data or using a model someone else has created. It is important to choose a baseline model that is realistic for the task at

hand, so as to not set the bar for making a successful model way too low. It is also important that the baseline is still realistic to beat.

2.6.3.2 Common Evaluation metrics for Classification

Some common evaluation metrics for classifiers are accuracy, precision, recall, and ROC AUC[2] [39, pp. 62–64]. Perhaps the most common and well-known metric for classification is accuracy, but there are several others that help determine the quality of a model. The accuracy describes how large a proportion of the samples that were classified correctly as defined in equation 2.8 [29, p. 215].

$$\text{accuracy} = \frac{\text{correctly classified}}{\text{All samples}} \tag{2.8}$$

While the accuracy score is simple and relatively easy to understand and make sense of, it isn't always the best metric to use [29, pp. 214–215]. Especially in the case of imbalanced classes, when one class has many more samples than the other, it will give a misguiding view of the quality of the classifier. One way to handle this when using accuracy as a metric is to calculate an accuracy score for each class separately. Sometimes, making one type of error is a much bigger problem than making another type of error [39, p. 62]. An example of that could be for example a cancer screening application, where missing a cancer patient could risk their life.
However, when evaluating classification models it is often important to examine the type of errors the model makes. For this, it can be interesting to create a confusion matrix [29, pp. 213–214]. In the case of two classes, binary classification gives us four categories to look at. The *positive* class is the minority class and the *negative* class is the majority class.

- True Positives, TP, are the number of instances of the positive class that are classified correctly
- False Positives, FP, are the number of instances of the negative classified incorrectly as positive
- True Negatives, TN, are the number of instances of the negative class that are classified correctly
- False Negatives, FN, are the number of instances of the positive classified incorrectly as negative

These numbers are then put in a matrix where one axis has positive and negative, and the other axis has true and false. This gives a simple overview of what types of errors the model is making. Refer to figure 2.2 for an example of a confusion matrix. This notation of using True/False Positives/Negatives makes the accuracy formula in 2.8 instead be 2.9.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.9}$$

Precision and recall are two metrics that can be used to avoid some of the issues with accuracy when it comes to imbalanced classes since they take into account the type of error that the model is making [39, pp. 62–63]. Precision looks at how large

[2]Area Under the Curve of the Receiver Operating Characteristic

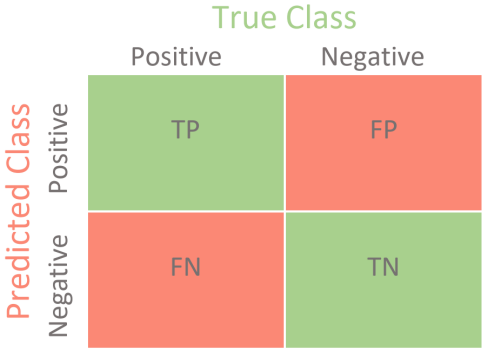

Figure 2.2: An example of a confusion matrix. Image from [40].

a proportion of the positives we predicted were actually positives, whereas recall instead looks at how large a proportion of the total number of positives we classified as positive. Using the counts of positives and negatives, the formula for precision is described in equation 2.10 and recall in equation 2.11 [29, pp. 215–216].

$$\text{precision} = \frac{TP}{TP + FP} \tag{2.10}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2.11}$$

When working with precision and recall, there is always a tradeoff between them and it is important to think about what we want for this specific application and what the scores mean [29, pp. 215–218]. One way to view it is that a high precision score means a careful decision, every time the classifier deems something positive it needs to be very certain. On the other hand, by labeling every instance as positive, a perfect recall score is achieved. But that doesn't make for a very good or interesting classification. A way to visualize the tradeoff between precision and recall is by producing a curve based on the confidence of the model [39, pp. 63–64]. This can prove very useful when comparing two models.
ROC is a curve that can be used to visualize how good the classifier is. On the y-axis we find the true positive rate [29, pp. 218–219] (which is the proportion of samples that were predicted positive that were classified correctly) [41] and on the x-axis, we have the false positive rate (proportion of samples labeled positive that were actually negative) [29, pp. 218–219]. The curve is obtained by varying the value of a threshold parameter. When the threshold parameter is low, more samples are assigned to the positive class [42]. One can think of it as the classifier having a low threshold to give classify as positive. As a simplification of this, a single number can be produced from the graph using the area under the curve [29, pp. 218–219]. This gives the ROC AUC score, which is perfect when it is 1, and the graph follows the upper left corner. However, this number should generally be above 0.5 which would be generated by selecting labels completely at random as can be seen in the baseline example in figure 2.3. This can be compared to the other curve in the same figure, which is closer to the upper left corner and thus better.

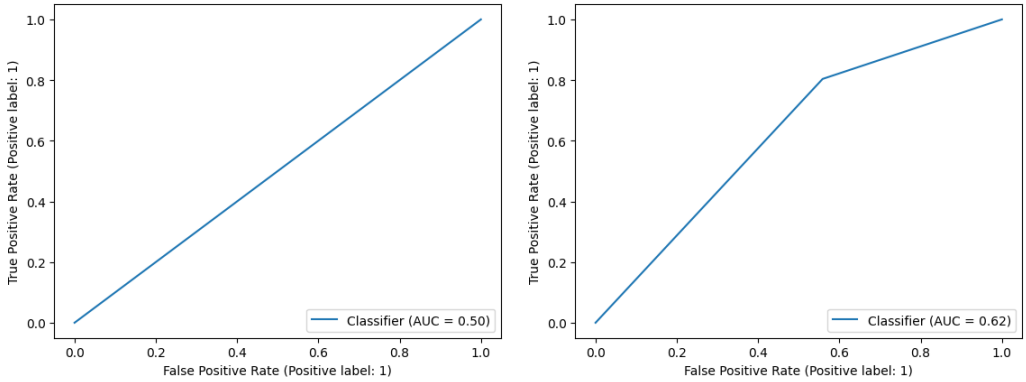

Figure 2.3: Examples of Roc curves from predicting the examiner gender on the English data. The left is the curve from the baseline classifier. The right is the best of the logistic regression classifiers trained on the same data. The better curve is the right as it is closer to the upper left corner, meaning there is more area under the curve.

2.6.3.3 Regression Model Baselines

When evaluating a regression model, i.e. a model used to predict a value from a set of features, there are a few different methods. One simple baseline model that can be used always predicts the mean or median from the training data [29, p. 212]. For non-linear models, training a simple linear regression model as a baseline can prove useful. In a problem where one seeks to predict a value at a certain point in time (such as the example of the price of a stock on a certain day), simply outputting the value from another day, such as the previous day can make a very good baseline that is hard to beat in many cases. Any baseline model should be both realistic to beat, while still posing some sort of challenge so as to make a meaningful target.

2.6.3.4 Common Evaluation metrics for Regression

A lot of the metrics used to evaluate regression models are based on the distance between the predicted value and the actual value, the error. Some of these metrics are mean squared error (MSE), root mean squared error (RMSE)[29, p. 223], and mean absolute error (MAE) [43]. One of the more common metrics used to evaluate regression models is the MSE, mean squared error, which is based on summing squaring the errors, and then averaging them. This can have issues in data with a lot of outliers since they are going to have a too large impact on the score. The formula for the MSE is found in equation 2.12 [29, p. 223] where the error for a certain sample, i, is denoted ϵ_i.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \tag{2.12}$$

For simpler interpretation, one might be interested in using the RMSE, root mean squared error, as a metric instead [29, p. 223]. It is calculated as the root of the MSE. This value is easier to interpret since it isn't squared and thus is on in the same

order of magnitude as the predicted and target values. Still, this metric is based on the MSE and does not mitigate any of the issues that MSE has with outliers.
The MAE, mean absolute error, is calculated by averaging the absolute value of all the errors as in equation 2.13 [43].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid \epsilon_i \mid \tag{2.13}$$

A good thing about the MAE is that it keeps the unit of the original data, making it relatively easy to interpret. It is also important, however, that this means that this measure cannot be compared between data of different scales, similarly to MSE and RMSE [43]. A metric that can be used without sensitivity to scaling is the mean absolute percentage error as this is designed to handle relative errors well [44].

3
Methods

3.1 Research Questions

In order to guide the work of finding any gender bias or other issues related to gender in the course evaluation data, we have produced four research questions that we seek to answer. We have used these as a foundation to aid in finding different interesting aspects and factors that may or may not play a part in how the courses are rated and what the students write in the evaluations. These questions can be seen as a concretization and a way to help find the solution to what is described in the problem description in the introduction of this thesis.

RQ 1 Is there a difference in the course grade (overall impression) depending on
- **a** the examiner's gender?
- **b** the examiner's gender and the gender balance of the students
- **c** the examiner's gender and students' grades?
- **d** the examiner's gender and the teaching language in the course?

RQ 2 Do students use other words in their comments depending on
- **a** the examiner's gender?
- **b** their own gender?

RQ 3 Do courses with a female examiner have a more even gender balance?

RQ 4 Do students in courses with a more even gender balance get higher grades?

3.2 Description of the Data

The data that this project is the basis for this project comes from four different sources, which we have then combined in order to be able to work with the data smoothly. The first source contains information about what examiners are responsible for which courses, what course code the course has, what year the course was held in, and some additional identifiers. The second data source comes from *Statistiska Centralbyrån* and contains Swedish name statistics from 1999 to 2020. It contains all names that are held by more than 10 people in Sweden. This data also contains information about what type of name it is (ie. first name or surname) and the legal gender of those who carry it. The third dataset holds information about student grades in the courses from 1979 to 2021. This set has information about how many students have what grade, what their legal gender is, and whether this is the first time they are registered in the course or if they are re-registered. The fourth data source is perhaps the main one, as it contains the actual course evaluations. This set contains approximately nine thousand course evaluations gathered from Chalmers

University between 2013 and the fall of 2021. For each course in the data, we have the exact formulations of all the questions that are asked in the evaluation forms. This means that we have the standard questions, but also the questions that the examiners have added. We then have all of the students' answers to these questions, both numerical answers and text comments. The numerical answers are on a Likert scale from one to five. The course evaluation process is described in section 2.4, and the text comments have been prepared as described in section 2.4.2. Note that the questions aren't mandatory so a student may only have answered a few of the questions and left some blank.

3.2.1 Standard Questions

While there are a lot of different questions that vary from course to course and year to year, there are 22 questions that are consistent through all years and courses, while the exact formulation might differ slightly the essence of the questions remains. However, we have saved the exact question formulation for all of the years we have course evaluation data from. We refer the reader to section 2.4.1.1 for the questions used. These are the only questions we have used in our work as their consistency makes comparison a lot easier.

3.2.2 Course Evaluations

When referring to the course evaluations in this thesis, we generally refer to the student answers to the standard questions in the course evaluations that are sent out to the students in a course. This data consists of both numerical answers, i.e. questions in the form of "How would you rate your overall impression of the course?" and text comments that the students can leave that are often connected to a numerical answer in the evaluation. When training the classifiers we generally use textual comments, and the numerical answers are part of what we base the statistical analysis on. The numerical answers that the students leave are on a Likert scale between one and five. It is worth noting that no question is mandatory for a student to answer, and not all students send in answers to the course evaluations at all. We do not have any information about what students have answered the course evaluations or who gave what answers. We only have the numbers and text comments, and what questions they are associated with.

3.2.3 Data Preparation

The data at hand was initially separated into several files with different information using different formats and we structured it into one single JSON file for simplicity and ease of use. This job was rather time-consuming, as the datasets needed to be combined in such a way that we got the most interesting information from all sets and kept correctness as to which instance of a course we had information about. We also cleaned the data quite extensively. Part of this work was also to find a way to represent the most interesting information about each course in an accessible way. This turned out to be slightly different for the different tasks so in order to simplify

when working on the machine learning tasks, we created additional files that only store the text comments and target variables.

3.2.4 Final Structure

When working with the data, we created a JSON-structure to make it easier to work with. This is an example of how a course was structured.

```
"MetaView": {
    "Course Code": "ank123",
    "Study Period": "lp2",
    "Year": 2022/2023,
    "Course Name": "ducks and where to find them",
    "Gender of Examiner": "M",
    "Examiner": ["Anki Duckling"],
    "Female Grades": {"3": 12, "4": 20, "5": 12, "-": 5},
    "Male Grades": {"3": 10, "4": 15, "5": 10, "-": 5},
    "Students": 87,
    "Answers": 63,
    "Female Proportion": 0.34523809523809523,
    "Language": "SWE"
},
"QuestionView": {
    "Learning outcomes" :
        ["Comments", "What did you learn"],
    "Overall impression" :
        ["Comments",
        "What is your overall impression of the course?"],
    ...
},
"AnswerView" : [
    {"Learning outcomes" :
        ["i learned a lot!", 5],
     "Overall impression" :
        ["great course! loved it!", 5]
     ...
    },
    {"Learning outcomes" :
        ["i only learned about ducks", 1],
     "Overall impression" :
        ["could have been better", 2],
     ...
    },
    ...
]
```

3.2.5 Examiner Genders

Originally, the data only contained the names of the examiners, but not their gender. This was something we had to annotate the data with in order to use it. To annotate the data we used parts of a script that predicts gender from a name, based on data från *Statistiska CentralByrån*.[1] However, this name data is not complete and left us with 40 first names that needed to be annotated manually. We did this by looking up the examiner on the Chalmers webpage and based our annotation on how we perceived their gender expression. As we ideally would use what gender the students perceive their examiner as, rather than a ground truth, our perception of their gender should be a reasonable proxy.

Some courses in the dataset are led by multiple examiners, and in some cases, the examiners are of mixed genders. For this project, we chose to exclude the courses with examiners of mixed genders but kept the courses where all examiners had the same gender. These courses were then annotated with that gender.

It is important to note that while we do have access to information about who is the examiner in a course, we do not, in fact, know who was actually present and teaching in that course. This is not always the same person as the examiner. The examiner of a course is always a professor. And as previously stated, the number of women in academia can be seen as a pyramid, with very few women at the top positions and more in the lower ranking positions (such as ph.D. students and post-docs). So, while a woman might actually be teaching a course, the listed examiner might still be a man. This could impact our results in a negative way, as we would ideally want to know the gender of the teacher as opposed to the examiner.

3.3 Data Selection

In order to really work with the data and avoid issues the evaluations of some courses needed to be removed. First of all, any courses where the course evaluations were not available in a spreadsheet format were removed as they would require a lot more work for cleaning. Other files were wrongly formatted or had other similar issues or were simply "broken" and got removed for that reason. We also made a decision to exclude evaluations from any courses where the examiners had different genders, as these courses would make the prediction tasks very difficult to manage. Finally, we set up some selection criteria stated in section 3.3.1 to avoid "noise" and outliers in the data. Please refer to table 3.1 for how many evaluations remained after each step.

[1] The script used was provided by our supervisor Peter Ljunglöf.

Selection of Course Evaluations		
Reason for removal	Removed	Remaining
Raw course evaluation data	0	9438
File format	1148	8290
Formatting and missing language data	37	8253
Mixed genders of examiners	1176	7077
Final data selection	2545	4532

Table 3.1: The reasons for removing course evaluations from the data and how many were removed after each step in the preprocessing.

3.3.1 Selection of Courses

While we have some data about a lot of courses, we do not have all the interesting data about all of those course instances. Thus, to select what course evaluations to use in our experiments, we started by determining what information we needed about each course instance to answer the research questions we had. All course evaluations selected needed to have the student answers, as well as information about examiner gender and student grades. We also wanted to exclude some of the outliers and other courses that may confuse the data and chose to add some additional criteria for the course selection. We decided only to include evaluations of courses with more than 25 students registered. In order to make reasonable comparisons between the courses, we also only included evaluations of courses with more than 10% female students. Also, note that we previously removed all courses that have multiple examiners where the examiners are of mixed genders. These selection criteria resulted in a final selection of 4473 courses. These are the course evaluations used in almost all experiments but with the exception of experiments centered around students' grades.

3.3.2 Student Grades

When working with the statistics of the grades that the students received in the courses, we made the choice to only use courses with the numerical grades (3–5 with 'U' as failure). This is because there is no good way to translate the other grading system of "G", "VG" and "U" into a fair number since that grading system is completely different from the 3–5 system. So in order to have some consistency we simply decided to exclude courses using any other grading system than the numerical grades. A failed course is counted as a zero grade, while the numbers keep their value. This selection of 3991 courses is consistent for all statistics that require the students' grades as part of the data and is a smaller subset of the total courses that are used for the other experiments.
For all grades, we have data about how many students with male or female as their legal gender have that grade. We do, however, not have any information about what individual student received what grade and cannot connect any student's grade to a specific answer in the course evaluations. In addition to the legal genders of the

students, we have information about how many of the students receiving a grade are registered to the course for the first time or if they are re-registered.

3.4 Statistical Analysis of the Data

When calculating descriptive statistics such as means or standard deviations for the dataset, we started by calculating a mean for each course. A couple of reasons that this approach can be problematic is that the number of students in the courses varies, and the number of students that have answered the survey varies a lot from course to course. This means that the courses with few students might have a larger impact on some statistics, such as the mean, as we then only calculate the mean of course averages and not that of student answers. However, this is an easier approach to implement and can in some cases be more meaningful. For example, it is more interesting to compare averages per course of overall impression, as we are interested in the overall impression per course and not per student.

3.4.1 Student grades

When working with statistical questions related to student grades, we use the smaller subset of the data of 3991 courses, where all courses with a pass/fail grade are removed. For all grades, we have data about how many students with male or female as their legal gender have that grade. We cannot connect a student's grade to a specific answer in the course evaluations. In addition to the legal genders of the students, we have information about how many of the students receiving a grade are registered to the course for the first time or if they are re-registered. The grades are averaged for each course separately and then these course averages are combined for the global average. The main reason for this is that we are comparing courses and the information related to them (such as average student grades) as opposed to comparing individual students.

3.4.2 Teaching Language and Course Evaluations

In the final data selection, we have 2601 course evaluations for courses taught in Swedish and 1931 courses in English. We were unable to obtain language data for all courses, and for this reason, some course evaluations were removed during the formatting stage. So the data we have contains 2191 evaluations of courses in Swedish with a male examiner and 410 in Swedish with a female examiner. For the English courses, there are 1575 with a male examiner and 348 with a female examiner.

3.4.3 Overall Impression

One way that we use to measure the students' experiences and impressions of the courses is by using their answers to the overall impression question, from the course evaluation questions in section 2.4.1.1. This question is divided into one numerical question and a free text comment that the student can leave. In the numerical

question, the students can give a rating ranging from 1 to 5, with 0 as the worst rating and 5 as the best rating. Students don't have to answer any question, which means some student answers appear as blank in the data. For overall impression questions, there were 32 such empty student answers. These non-answers were removed to avoid technical problems when working with the experiments related to the overall impression.

3.5 Examiner Gender Prediction

One part of the project consisted of using text data to predict the gender of the examiner. This has a couple of purposes. Firstly, it is interesting to see whether any of the machine learning methods we use for prediction are able to find some connection between the data at hand and the examiner's gender. So rather than purely trying to solve the classification task, we are interested to see if there are any specific features (i.e. words and part of speech) that the classifiers associate with male or female examiners. We use the weight vectors from the classifiers to produce lists of the ten most important features used in the predictions.

How good the classifier is at predicting the examiner's gender is also interesting, since that means that there is some difference in how the students write about the examiners and courses depending on the examiner's gender. A better classifier will have found a clearer connection or pattern.

The main data source is the text comments that the students have written. While these comments are anonymous, we do have access to other information, such as the gender balance of the students in the course. When processing the comments, we use two variations, single-comment where each individual comment is seen as a document and multi-comment where all comments in a course evaluation together are seen as a document. In both cases, the corpus consists of all the text comments in all course evaluations.

We compare a few different datasets, one with just all of the text comments, one where we have undersampled courses with male examiners, and one where we have included part-of-speech tags. The number of training instances in each dataset are presented in table 3.2.

Multi Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	2080
Undersampled	SWE	654
Included POS-tags	SWE	2080
Full text	ENG	1538
Undersampled	ENG	566
Included POS-tags	ENG	1538
Single Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	215124
Undersampled	SWE	72268
Included POS-tags	SWE	215124
Full text	ENG	129027
Undersampled	ENG	40098
Included POS-tags	ENG	129027

Table 3.2: The different training sets used in the examiner gender prediction task and their sizes.

3.5.1 Undersampling

In this classification task, we are dealing with a case of *imbalanced classes* as most of the courses have male examiners. This can lead to several problems down the line and can really affect the results when using some classifiers [33]. In order to avoid these issues we have chosen to include one dataset where we have undersampled the courses with male examiners, but also to compare these results to those where we use all the available data. When undersampling we remove random data instances, which means that the particular choice can affect the end result. We remove courses with male examiners until we have an equal number of courses with male and female examiners.

3.5.2 Preprocessing

In order to use the text data in classification it needs to be prepared. We first do this by using our own preprocessor in order to remove punctuation, pronouns, and names from the text.[2]
There are more features that are interesting than just the text comments themselves. One such example is the part of speech for the words in the text comments. In order to extract these we use the NLP library Spacy and extract the lemmas and parts of speech for the words in the text comments.[3] These lists are then concatenated and

[2]The parts of the code in the preprocessor related to names were provided by our supervisor Peter Ljunglöf.
[3]`https://spacy.io/`.

fed through a bag-of-words model. Another part of preparing the text comments was to decide what was to be classified, which was done in two separate ways, multi-comment prediction, and single-comment prediction.

3.5.2.1 Multi Comment Prediction

For multi-comment prediction, each sample that was to be classified consisted of all the text comment answers that students had made in regard to a specific course instance. These comments were all concatenated and fed through the bag-of-words model to create the data for the classifier to use. We then both train and evaluate the models on all comments from a course. That means that each row in the data corresponds to one instance of a course.

3.5.2.2 Single Comment Prediction

To perform single comment prediction, the comments were instead used separately and each comment was on its own fed through the bag-of-words model. This means that we both train and evaluate the classifiers on single comments as opposed to all comments from a course bunched together. In the single comment prediction each row in the data corresponds to a single comment from a course, meaning that one instance of a course can correspond to several data entries.

3.5.3 Classification

In order to find different patterns and see how well we could possibly classify what gender the examiner has we have tried and evaluated a few different methods for classification. When choosing methods, we focused on using classifiers where we could identify what features were important in the classification process so we could determine if there are any words, parts of speech, or other features that seemed to have greater importance than the others. The classification models we use are logistic regression and support vector machine. These models were chosen as the main goal is to find what features correlate to what class, and for that reason, a linear model gives an easily interpretable coefficient vector.

3.5.4 Cross Validation

In order to obtain more robust and reliable results we ran the experiments five times. To do this, we used five-fold cross-validation. That means we separated the data into five different training and testing sets using all of the data, meaning all training sets contained 80% of the data and the testing sets 20%. This enabled us to avoid some of the issues that can come when selecting the training and test set. Five separate classifiers were then trained and evaluated on their own training-testing set. For the final results of each experiment, we calculated an average of all the scores from the classifiers as well as the results from the best of the classifiers. To obtain the “best” classifier we compared their ROC AUC score.

3.5.5 Evaluation

In order to evaluate the quality of the classifiers we use a few different methods and scores. Perhaps the most obvious one is using the accuracy score. But due to having such imbalanced classes it is especially interesting to look at other evaluation methods, as a classifier could easily get high accuracy by simply labeling all samples as male. The other metrics we calculate are precision, recall, and ROC AUC. We also compute confusion matrices for the classifiers in order to evaluate what type of errors they make. This can help us see if courses with female examiners are consistently classified as male or vice versa. This gives a better image of what the classifiers actually produce.

3.5.5.1 Baseline

As a baseline, we use a *Dummy Classifier* from *Sklearn* that is trained to classify each sample as male, as this is the majority class. The reasoning behind using this as a baseline instead of simply outputting random or something else is that the classes are imbalanced. There are far more male than female examiners, and thus a classifier that only outputs "male" will still get a pretty good accuracy score. However, the other measures (such as precision) might not be as high for the baseline compared to the other classifiers. By using that classifier as a baseline, we know we will want our actual classifiers to perform better than that.

3.6 Comment Author Gender Prediction

Since the data is anonymized, we cannot use the author's gender directly. However, we do have access to the gender balance among the students in the course and can base our predictions on that. That is based on the assumption that the genders of the students that write course evaluations follow the genders of the students in the course as a whole. This is probably not completely true, but since we don't have access to the gold standard labels (i.e. the actual comment author gender) we need to work with some proxy. When working with this, we tried two different approaches, treating it as a regression problem and as a classification problem.

3.6.1 Regression

To do comment author gender prediction as a regression problem, we train a regression model on the comment data, using the proportion of female students as the target value. To evaluate the quality of the regression models trained on the different data sets look at some different metrics for evaluating regression models as well as the coefficient vector to determine what features have been important to predict a high or low proportion of female students respectively. The models were trained on four different data sets: English data with and without part-of-speech tags and Swedish data with and without part-of-speech tags. Using cross-validation, the models were trained on five different versions of these datasets to create more robust results and to some extent mitigate the impacts of chance. The number of training instances in each dataset are presented in table 3.3.

Multi Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	2083
Included POS-tags	SWE	2083
Full text	ENG	1544
Included POS-tags	ENG	1544
Single Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	215218
Included POS-tags	SWE	215218
Full text	ENG	129259
Included POS-tags	ENG	129259

Table 3.3: The different training sets used in the course gender balance regression task.

3.6.1.1 Preprocessing

The preprocessing for the course gender balance regression was done in a similar fashion to the preprocessing for examiner gender prediction, but with a couple of differences. Firstly, the target variable, in this case, is the proportion of female students in the course. This was relatively easy to extract since we had already calculated it (based on the course grades that are separated by gender) and made it readily available for each course. For the regression task, we also didn't have to do any undersampling as there are no classes, and thus no majority or minority class. For the actual text data, we performed the regressions in two different ways, multi-comment prediction, and single-comment prediction. While we don't have access to the author gender of each comment, it is still interesting to compare if the models perform better or worse on single versus multiple comments as a data entry.

3.6.1.2 Multi-comment Prediction

For the multi-comment prediction, all student comments in a course were put together and fed through a bag-of-words model. When including the part-of-speech tags, these were concatenated with the comments before feeding them to the BoW model. The bag-of-words representation of the comments (or comments and tags) is then fed forward to the regression model.

3.6.1.3 Single Comment Prediction

When using singular comments for the predictions, one comment is treated as a data entry. It is fed through the bag of words model, either with the part-of-speech tags for the words concatenated with it or on its own. This is then what is passed to the regression model.

3.6.1.4 Regression Models

For the actual regression, we use two linear models, one classical least squares linear regression, and a ridge model. Since the goal of the regression is rather to see if the models can capture some difference (i.e. bias) in how people write depending on their gender as opposed to creating an exceptional regression model, no hyperparameter tuning was performed.

3.6.1.5 Evaluation and Baselines

To evaluate the regression models we calculate four different evaluation metrics that can be compared to the results from the baseline regressor. These metrics are mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and maximum error (Max err). The baseline regressor always outputs the mean from the training set. The reason for choosing the mean rather than some more complicated value is mainly the simplicity, but also that there aren't that many outliers in the data, most courses have a relatively small portion of female students.

3.6.2 Classification

Solving the comment author's gender prediction problem as a classification problem meant we decided and tried to predict whether there was a balance among the student genders or not. Due to the nature of the data (most courses have a vast majority of male students), this in practice meant examining which of the courses had more than 60% male students and which had fewer.

Multi Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	2083
Undersampled	SWE	1542
Included POS-tags	SWE	2083
Full text	ENG	1544
Undersampled	ENG	886
Included POS-tags	ENG	1544
Single Comment Data Sets		
Dataset	Language	Number of training instances
Full text	SWE	215218
Undersampled	SWE	156690
Included POS-tags	SWE	215218
Full text	ENG	129259
Undersampled	ENG	70146
Included POS-tags	ENG	129259

Table 3.4: The different training-testing sets used in the course gender balance classification task.

3.6.2.1 Preprocessing

The preprocessing for this classification task was very similar to the preprocessing for the other task. Stopwords, such as gendered pronouns or names were removed from all comments and the comments are tokenized and fed through a bag-of-words model. This is done in two separate ways, just as for the other experiments, using single-comment and multi-comment prediction. For the single comment prediction, each comment is its own data entry and is classified separately. When using multi-comment prediction instead each course, i.e. all comments in a course, are all concatenated into one string and treated as a data entry and classified together. In the case of using the part-of-speech tags, these are concatenated to either the lemmas of the words in the comment or the group of comments before being fed through the bag of words model.

3.6.2.2 Annotation and labels

For this task we didn't have any pre-annotated data, so we annotated the data based on the proportion of female students in the course. We set a threshold of 0.4, annotating all courses with a proportion above that as balanced and all courses below as unbalanced. While there might still be some courses present in the data sets with a very high proportion of female students that thus aren't technically "balanced", we decided that these are few enough for it to not make too much of an impact on the results. Only concerning ourselves with a high or low proportion of female students also makes it easier to compare the results to the results we get from the regression task as those are just an estimated proportion of female students.

3.6.2.3 Undersampling

In this classification task, we are also dealing with imbalanced classes as most courses have a majority of male students. To mitigate this, we have performed undersampling as described in section 3.5.1, with the difference that the majority class, in this case, is courses with an unbalanced gender balance.

3.6.2.4 Classification Models

To solve the classification task we train two different linear classification models. One logistic regression model and one support vector machine classifier. These were trained on the different datasets and then evaluated.

3.6.2.5 Evaluation and Baselines

To evaluate the classification models, we run a five-fold cross-validation and choose the best model based on the roc auc score. We then compute the evaluation metrics for this model as well as the average over all of the five models. The reasoning for this is that we are interested in seeing if a classifier could find some difference in how students write depending on their gender, and thus even if it is based on some randomness, the best model might give some interesting information to that regard. However, it is also interesting to examine the average of the models as this

gives a more fair result and helps in having randomness play too large a part in the final result. The evaluation metrics used are the same as for the examiner gender classifiers, i.e. accuracy, precision, recall, ROC AUC as well as computing confusion matrices to understand what types of errors the classifiers make. These results can be compared to that of the baseline classifier, which classifies all samples as the majority class, which in this case is "unbalanced", i.e. having more than 60% male students in the course.

4
Results

4.1 Statistical Analysis of the Data

4.1.1 Gender Distributions

There are 5158 courses in the final data selection, of which 983 have female examiners and 4175 have male examiners. This means that 19% of the courses have female examiners. Thus, the vast majority of courses have male examiners. Among the students, 33% are female and 77% are male.

4.1.2 Student Grades

In **RQ 1**c we seek to find if there is any difference in the overall impression of the courses depending on the students' grades. Grades are also an important factor in **RQ 4**, examining whether students in courses with a more even gender balance get higher grades or not. To answer these questions we need to understand the grades better and make a deep dive into statistics related to the grades and genders of examiners and students.
For all students, the average grade is 3.23 with a standard deviation of 0.8. For female students, the average is 3.26 across all courses and the average grade for male students is 3.19. The difference between the average grades for the genders is then 0.06. These results are shown in figure 4.1.

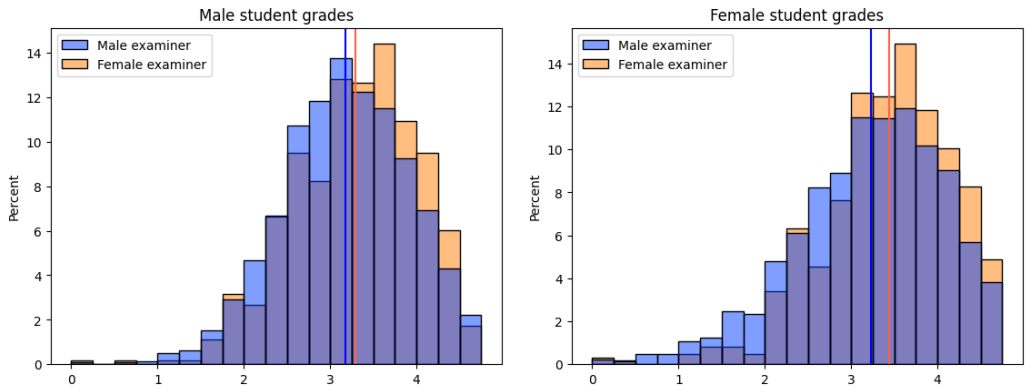

Figure 4.1: Histograms of the average per course grades for male and female students.

Female examiners on average give higher grades to students. Female students with

female examiners receive an average grade of 3.44, while male students with the same examiners get an average grade of 3.29. With male examiners, female students get an average grade of 3.22 and male students of 3.18. Thus, a female examiner gives their female students an average grade that is 0.15 higher compared to the male students. For male examiners, the difference is 0.05 in favor of female students. This is visualized in figure 4.2.

Receiver of grades	Person who grades: M	Person who grades: F
M	3.18	3.29
F	3.22	3.44

Figure 4.2: A visualization of who gives what students what grades.

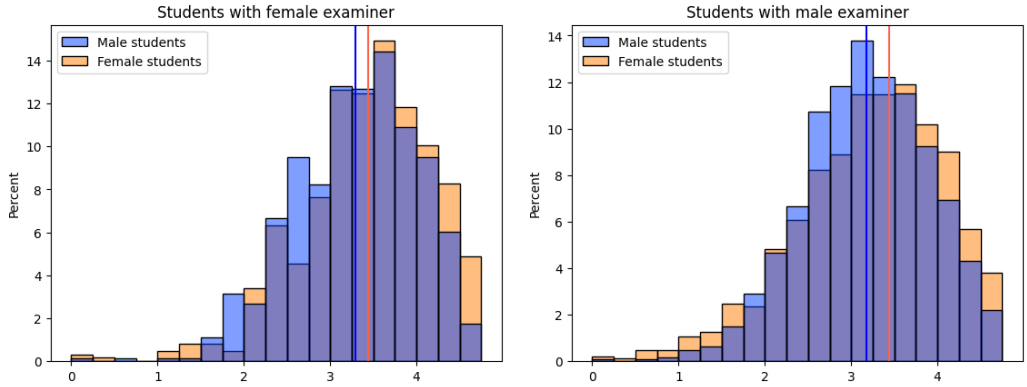

Figure 4.3: Histograms of the average per course grades for female students in orange and male students in blue. The courses with female examiners on the left and those with male examiners on the right.

Female students receive higher grades than male students from both male and female examiners, but especially from female examiners. The average grade that female students receive is 0.22 higher than what they receive from male examiners. The same number for male students is 0.12. The difference can be seen clearly in figure 4.3.

To further examine the data and what grades the students receive, we plot the student grades separated by the teaching language of the course. The difference can be seen in figure 4.4, students in English courses have an average grade of 3.43 and the students in Swedish courses have 3.05, meaning there is an average grade difference of 0.38 in favor of the students in English courses.

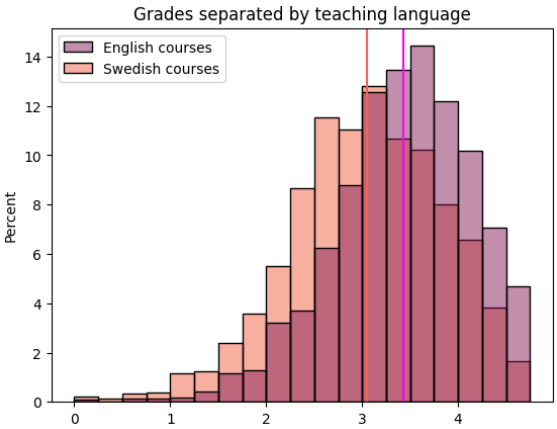

Figure 4.4: Student grades separated by teaching language in the course.

4.1.3 Overall Impression of Courses

The overall impression scores that are given by the students in the course evaluations, as seen in figure 4.5, are relatively similar for male and female examiners. Across all courses, the mean overall impression rating is 3.83, for courses held by a male examiner the result is 3.85 and for female lead courses it's 3.75. The courses held by female examiners receive an overall rating that is 0.10 lower than their male peers and 0.08 lower than the mean.

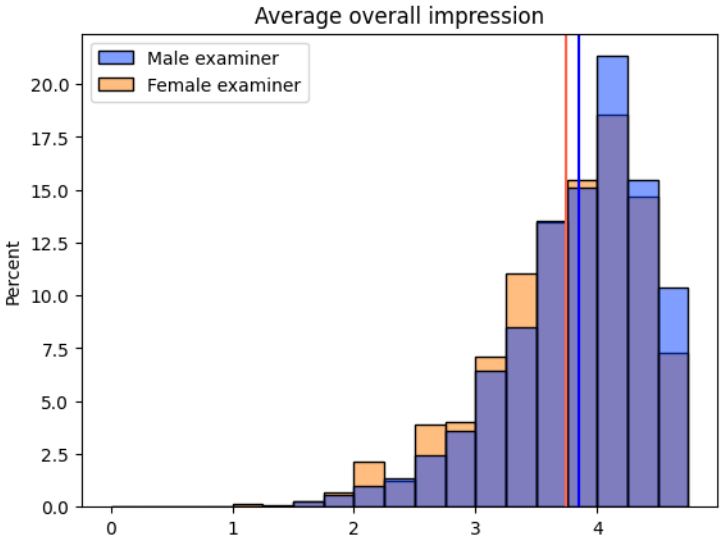

Figure 4.5: Histograms of the average per course overall impression rating by students, separated by examiner gender.

4.1.3.1 Overall impression and teaching language

To answer **RQ 1d** we examine if there is a difference in the overall impression depending on the teaching language and the examiner's gender. This question is twofold, and we have already uncovered that there seems to be some difference from

the examiner gender, as can be seen in figure 4.5. Courses with male examiners tend to get higher overall impression scores, so is there a difference between Swedish and English courses?
We found no significant difference in the performance between Swedish and English courses, with an average overall impression score of 3.84 for English courses compared to the Swedish score of 3.83. Looking at figure 4.6, however, we can see that there are fewer English courses at the top of the distribution, compared to the Swedish courses at the top of the Swedish distribution. The same holds for the very bottom of the distributions as well, where there are more Swedish courses.

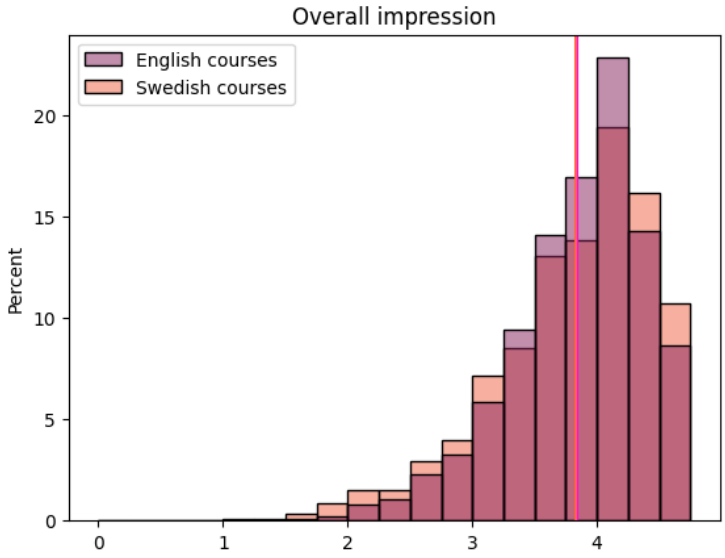

Figure 4.6: The overall impression for courses in Swedish and English.

How do the gender differences compare over the languages? For the Swedish courses, the average overall score for female lead courses is 3.71 and for the male we have 3.85, making it a difference of 0.13. When it comes to English courses the difference is smaller, 0.06, for their courses with female examiners getting 3.79 and male lead courses getting 3.85.

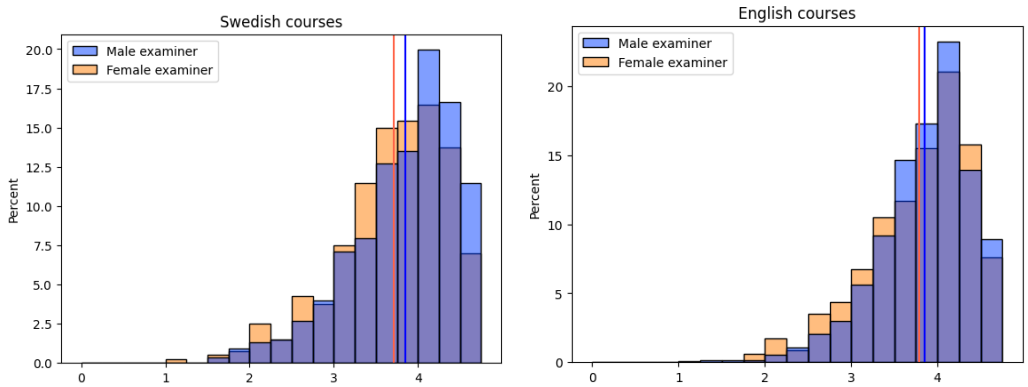

Figure 4.7: Overall impression for courses with male and female examiners, separated by teaching language.

In figure 4.7, it is clear that male examiners are most common at the top of the distribution in the Swedish data, and female examiners are more dominant in the bottom part. For the English data, the pattern isn't quite as clear, and the differences between the genders aren't as big.

The relationship between overall impression, teaching language, and examiner gender can also be viewed as in figure 4.8.

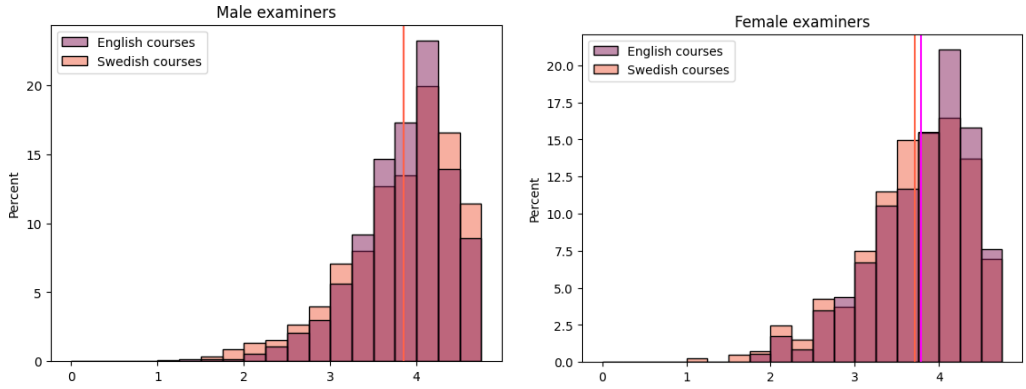

Figure 4.8: Overall impression for courses separated by teaching language.

For courses with female examiners, the mean for Swedish courses is 3.71 and for English courses it's 3.79, making the difference 0.08. We can see in figure 4.8 that for courses with female examiners, the English courses consistently outperform the Swedish courses in the top part of the distribution. The pattern is not as clear for courses with male examiners, as the male examiners in Swedish courses get both the top and bottom overall impression scores, while the examiners in the English courses are more common in the middle part of the distribution.

When it comes to courses with male examiners there is no significant difference in the overall impression due to teaching language. Both the English and the Swedish courses have an average rating of 3.85. Interestingly, for courses with male examiners, the difference in overall impression between languages is much smaller than for the courses with female examiners.

4.1.3.2 Proportion of female students and examiner gender

In our data, it is clear that courses with female examiners have a larger proportion of female students compared to their male colleagues. This is not only evident in the mean of the data, 0.38 for female examiners and 0.32 for male examiners. But also in how it is distributed in the courses. As can be seen in figure 4.9, while the lowest and highest values observed (excluding outliers) are similar for both genders, all quartiles are placed at significantly higher values for the female examiners. The outliers in the data are data points that are far enough away from the rest of the data to be deemed noise by the plotting software. However, the outliers for the female examiners are higher than the ones for the male examiners.

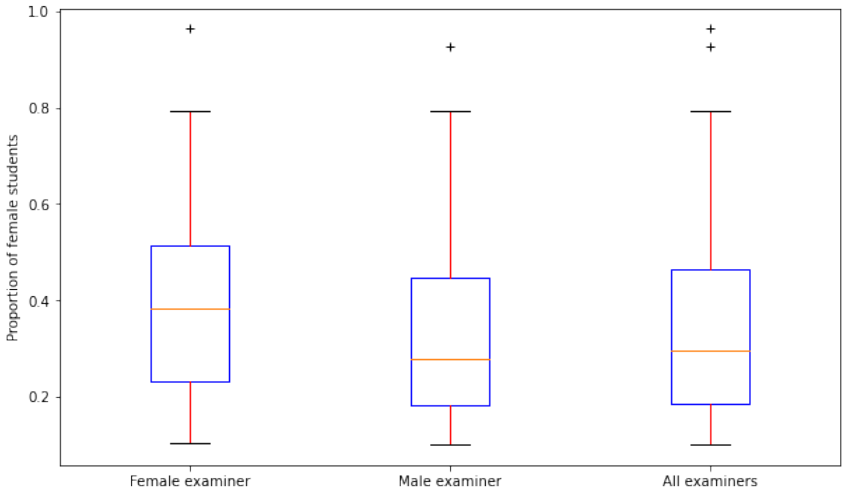

Figure 4.9: The proportion of female students separated by examiner gender.

However, there does not seem to be any clear relation between the proportion of female students and the overall impression of the course. Examining figure 4.10, we can see that there seems to be a higher density of orange data points (i.e. female examiner) compared to blue data points higher up on the graph. This is hardly surprising as we already determined that there seems to be a higher proportion of female students in courses with female examiners. In terms of higher and lower overall impressions, however, there are no clear patterns. Using only the naked eye, we can identify some potential clusters of courses with one type of examiner, but no linear relation or similar.

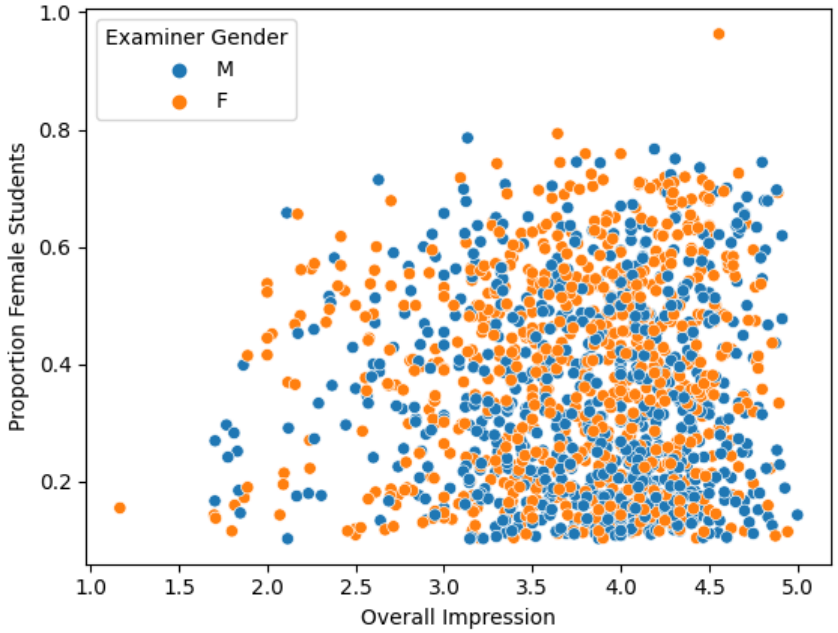

Figure 4.10: A visualization of the proportion of female students and the overall impression of the course on a downsampled version of the data.

4.1.3.3 Overall Impression and Student Grades

To answer **RQ1c** we wanted to explore the data in terms of examiner gender, overall impression, and student grades.

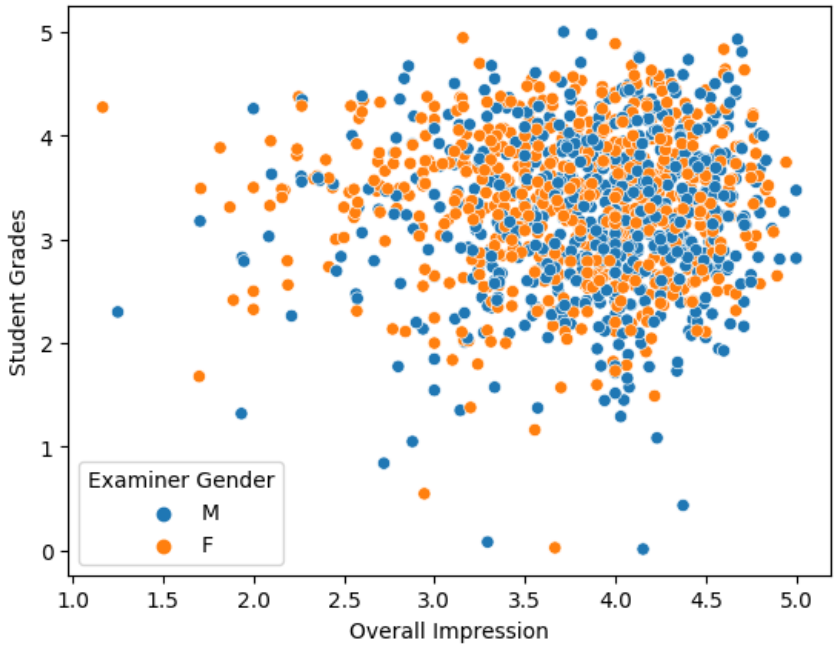

Figure 4.11: A visualization of students' grades and the overall impression of the course on a downsampled version of the data.

In figure 4.11, we cannot find any clear pattern in the data. Most data points are gathered a bit from the upper right corner, indicating that students with high grades give a relatively high overall impression score. There does not seem to be any clear relationship between examiner gender and the other two variables in this case.

4.1.4 Proportion of Female Students and Student Grades

To answer **RQ 4**, we examine the data in terms of average student grades and the proportion of female students in a course.

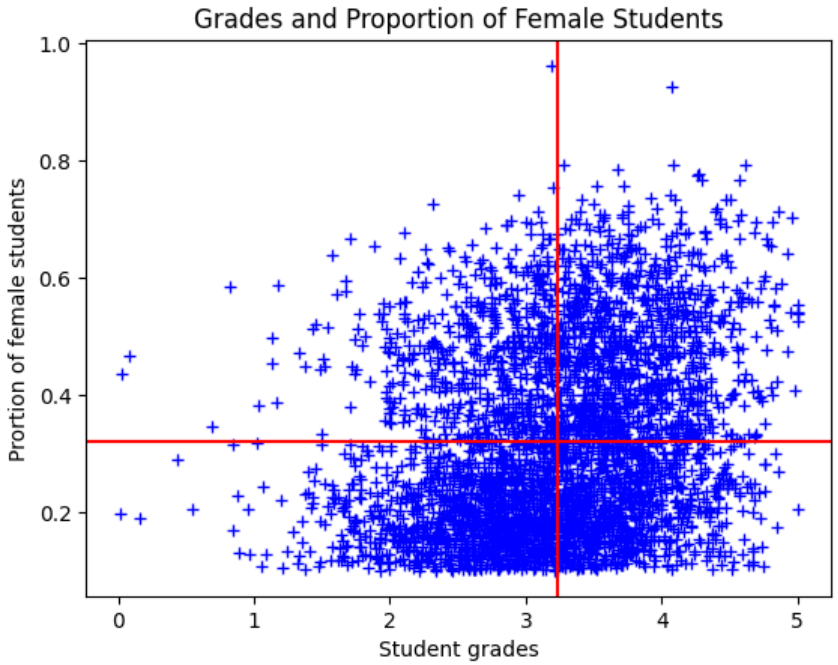

Figure 4.12: A scatter plot of the proportion of female students and the grades that students receive in a course. With the average grade marked as a vertical line and the average proportion of female students marked as a vertical line.

In figure 4.12 we visualize the relationship between these variables using a scatter plot. There seems to be some positive relationship between them (i.e. students in courses with a higher proportion of female students tend to get higher grades), but the relation isn't completely clear, and there is a lot of noise in the data.

4.2 Examiner Gender Prediction

To answer **RQ 2b** regarding if students use different words in their course evaluations depending on the examiner's gender, we try predicting the examiner's gender from the comments that the students have left in the evaluations. The main purpose of the examiner's gender classification is not really to solve the classification task of predicting the examiner's gender from the text comments. Instead, we are examining whether or not the classifiers can solve the task sufficiently, which would mean that there is some difference in how students write about their courses and examiners based on their gender. For this reason, the main results are focused on what features the classifiers have deemed important rather than the classification accuracy or precision/recall. Still, a better classifier should yield more reliable features as it has made a better prediction.

4.2.1 Baseline

The results from the baseline classifiers can be seen in table 4.1. To read the baseline values for all of the calculated metrics, refer to A.1.1.

Baseline		
Result	Eng	Swe
Accuracy	0.80	0.84
ROC AUC	0.50	0.50

Table 4.1: The baseline metrics for Swedish and English data.

The baseline classifier always predicts the label of the majority class, which in our case is male. As can be seen in table 4.1, the accuracy for the baseline classifiers are relatively high: 0.80 and 0.84 for English and Swedish respectively. This means that roughly 80% and 84% of the courses have a male examiner. Though it should be noted that the ROC AUC score is very low at 0.50, which means that the ROC curve is a straight line on the diagonal. This indicates that the classifier makes a lot of errors misclassifying courses with female examiners - which is hardly surprising as all of them are misclassified by the baseline classifier.

4.2.2 Metrics

The resulting accuracy scores from the best classifiers trained on the different datasets can be found in table 4.2 with the best score for each language marked in bold. All of the calculated metrics can be found in the appendix in A.1.2.1.

Examiner Gender Prediction Accuracy					
Swedish					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.83	0.77	0.83	0.76	0.80
Undersampled	0.59	0.61	0.60	0.57	0.80
Part of Speech	**0.84**	0.79	0.80	0.77	0.80
English					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	**0.85**	0.76	**0.85**	0.74	0.84
Undersampled	0.55	0.62	0.64	0.62	0.84
Part of Speech	**0.85**	0.75	0.73	0.76	0.84

Table 4.2: The accuracy scores from the best classifiers when predicting the examiner gender on the different datasets using the different classification models.

When training the different classification models on the Swedish data, the best accuracy score of 0.84 was obtained by training a logistic regression model on the data with added part-of-speech tags and separated into single comments. The worst accuracy of 0.57 came from training a support vector classifier on the undersampled data, lumping all comments in a course together. The baseline accuracy for Swedish data is 0.84, meaning that almost all of them performed worse than the baseline, except for the best classifier that performed on the same level as the baseline.

For the English data, three classifiers shared the same best score of 0.85. Again the logistic regression model with part-of-speech tags trained on single comments, but this time it was tied with the logistic regression model trained on the full text as single comments and the support vector classifier trained on the full-text single comments. The worst score of 0.55 was obtained by training a logistic regression model on the undersampled dataset as single comments. The baseline accuracy for English data is 0.80, meaning that almost all of the classifiers performed worse than baseline, except the three best classifiers which all had an accuracy that was 0.05 points above it.
Altogether, for both languages, classifying single comments yielded higher accuracy scores, although there were some exceptions to this pattern. While the logistic regression models yielded more of the best scores, overall the accuracy scores were pretty similar between the two classification model types. Comparing the different datasets, we see that the accuracy scores for the undersampled data are significantly lower than the others, across all models and languages. For single-comment classification, the full-text data performed better than adding the part-of-speech tags in most cases, but for multi-comment classification the opposite was true. Still, for both multi and single-comment classification, the differences between the full-text data and the data with added part-of-speech tags were very small.

Examiner Gender Prediction ROC AUC					
Swedish					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.50	0.54	0.50	0.61	0.50
Undersampled	0.54	0.57	**0.62**	0.59	0.50
Part of Speech	0.50	0.60	0.59	0.60	0.50
English					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.50	0.64	0.50	0.60	0.50
Undersampled	0.57	0.65	0.63	**0.65**	0.50
Part of Speech	0.50	0.62	0.62	0.62	0.50

Table 4.3: The ROC AUC scores from the best classifiers when predicting the examiner gender on the different datasets using the different classification models.

However, it is important to note that the classification accuracy doesn't tell the entire story, especially when dealing with an imbalanced dataset. We have also included the ROC AUC scores in table 4.3, to make a more nuanced comparison of the classifiers.
Evaluating the models on the Swedish data, the best ROC AUC score of 0.62 was obtained by training a support vector classifier on the undersampled single-comment data. The worst result of 0.5 came from training a logistic regression model on single-comment full-text data and the same model trained on single-comment data with part-of-speech tags, as well as training a support vector classifier on single-comment full-text data.

For the English data, the results are relatively similar. But the best result of 0.65 came from training both classification models on the undersampled multi-comment data. The worst result is again 0.5 and comes from training the same models on the same datasets as the worst results from the Swedish data.
An interesting difference between the ROC AUC scores and the accuracy scores is that for ROC AUC, training the models on the undersampled dataset yields higher scores - while training on those sets gives a lower accuracy. This could indicate that the undersampling does work and makes the classifiers correctly classify more courses with female examiners.
The ROC AUC scores are in general relatively low, and for some of the classifiers they are even at the level of the baseline at 0.5 (for both Swedish and English). While the goal for this thesis wasn't to create and tune the best classifier possible for predicting examiner gender, it is worth keeping in mind what classifiers got higher ROC AUC scores when comparing important features.

4.2.3 Important Features

The main part of this task is finding what features are important for predicting the examiners' genders. We have produced lists of the ten most important features for each classifier to predict a course as having a female or male examiner. These lists can be found in the appendix, in A.2.1 and A.2.2 respectively. It is interesting to note that in none of these lists of important features for predicting the examiner's gender, we find any part-of-speech tags.

Important features for the Best Classifiers			
ROC AUC		Accuracy	
Female	Male	Female	Male
begrepp	kurslitteraturen	kemi	programmering
mest	snabbt	matlab	kod
tydlig	lika	diskussion	praktisk
slutet	ihop	bevis	tentorna
lagom	började	presentation	labbarna
examinator	dom	text	absolut
ens	framförallt	dra	tentame
tar	intressant	repetition	tavlan
upplevde	bygga	diskutera	tillgänglig
konstigt	gärna	skön	duggorna

Table 4.4: A comparison of the most important words for predicting that a course has a male or female examiner for the single-comment support vector classifier trained on undersampled data on the left and the single-comment logistic regression model trained on the data with the added part-of-speech tags on the right. The classifiers are trained on Swedish data.

When training the classifiers on the Swedish data, the best classifier in terms of accuracy was the logistic regression model trained on the data with the added part-of-speech tags in a single comment format. The best in terms of ROC AUC score, however, is the support vector machine model trained on the undersampled dataset in a single-comment format. The ten most important features for the predictions for these models are presented in table 4.4.

Important features for the Best Classifiers			
ROC AUC		Accuracy	
Female	Male	Female	Male
your	help	okay	around
communication	page	activities	several
helpful	tutorials	lack	code
required	computer	low	assistants
something	helped	years	solving
writing	jag	spend	using
learned	det	focused	complete
team	last	improved	fast
three	yes	attend	ones
workshop	clearly	issues	believe

Table 4.5: A comparison of the most important words for predicting that a course has a male or female examiner for the support vector machine model, trained on undersampled data in the multi-comment format on the left, and the single-comment full-text data on the right. The classifiers are trained on English data.

For the English data, more classifiers shared the best accuracy score. However, using more of the digits in the initial precision leaves us with the single-comment support vector classifier trained on the full-text data as the best classifier in terms of accuracy. Do note, however, that this choice is more to just select one of the three best and that this classifier is not significantly better than the others in any way. The best ROC AUC score for the English data is obtained by training the support vector machine model on the undersampled data in the multi-comment format, however, note that there is no significant difference between this classifier and the logistic regression model trained on the same dataset. This model was also simply chosen to choose since it had a slightly better result than the other. The features that were important for these two best classifiers in terms of accuracy and ROC AUC, can be found in table 4.5.

To further analyze the results, we have used the lists of important features for all classifiers, and counted in how many lists each feature appears. This is to find what features have been most important across all classification models and datasets, as this should mean that they carry a higher relevance to detecting gender bias.

Important features for Prediction in the Swedish Data			
Female Examiner	Occurences	Male Examiner	Occurences
ännu	7	praktisk	4
skriva	5	labbarna	3
skönt	4	faktisk	3
matlab	3	absolut	3
tar	3	handledare	3
flesta	3	-	3
projekt	3	-	3

Table 4.6: A comparison of the most important words for predicting that a course has a male or female examiner and the number of times that word appears among the ten most important features across all classifiers and datasets for the Swedish data. Each word can appear at most twelve times.

The most common words for predicting a course has a male or female examiner can be found in table 4.6. The words for predicting a female examiner seem to appear in more word lists than the words to predict a male examiner. None of these words are part-of-speech tags, which is hardly surprising as only four of the datasets contain them.

Important features for Prediction in the English Data			
Female Examiner	Occurences	Male Examiner	Occurences
open	5	computer	11
stressful	5	code	9
works	4	tutorials	6
low	3	assistants	4
except	3	page	4
guest	3	matlab	3
paper	3	programming	3
mandatory	3	lab	3
based	3	complete	3
another	3	-	-

Table 4.7: A comparison of the most important words for predicting that a course has a male or female examiner and the number of times that word appears among the ten most important features across all classifiers and datasets for the English data. Each word can appear at most twelve times.

In the English data, the features for predicting a male examiner seem to have more significance. Several of these words are related to computers and programming. There are no words directly related to any subject in the words for predicting a female examiner.

4.3 Comment Author Gender Prediction

When predicting the author's gender we used two separate approaches, treating this as a regression problem and as a classification problem. While the task at hand is to predict the gender of the author of a comment, we have to use the proportion of female students in a course as a proxy as the comments are left anonymously. In tables 4.8 and 4.9 we have gathered the most important features to predict that a course has a low or high proportion of female students, using both the regression and classification models. These lists have been created by counting in how many of the lists of important features for the comment author prediction tasks each word appears.

Important features for Prediction in the English Data			
High proportion	Occurences	Low proportion	Occurences
text	6	tas	8
individual	4	assistants	5
process	4	ta	5
system	4	process	5
tas	4	stressful	5
programming	4	problems	4
ta	4	programming	4
code	4	code	4

Table 4.8: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression and classification models and datasets for the English data. Each word can appear at most twenty times.

In the English data in table 4.8, there are a lot of words related to teaching assistants that are used for predicting both a high and low proportion of female students. The word "ta" is most likely the commonly used TA, meaning Teaching Assistant. The plural form "tas", i.e. TAs, is also present in the lists. Interestingly, the two lists for high and low proportion have several words in common, which could indicate that there is not a lot of difference to how students write depending on the proportion of female students in a course or that the methods aren't good for measuring it. Note that the words in tables 4.18 and 4.17 come from both the regression models and classification models. This could mean that one type of model has learned that a word is important for predicting one thing, while the other type of model has learned the opposite.

Important features for Prediction in the Swedish Data			
High proportion	Occurences	Low proportion	Occurences
kritik	8	kritik	6
tydliga	4	labbar	4
därför	4	projekt	4
labbarna	4	pm	4
labbar	4	programm	4
labb	4	jättebra	4
betyg	4	tillbaka	4
labben	4	-	4
personligen	4	-	

Table 4.9: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression and classification models and datasets for the Swedish data. Each word can appear at most twenty times.

In the Swedish data in table 4.9, there are a lot of words related to labs to predict a higher proportion of female students. The word "kritik" (criticism) is also at the top of both of the lists.

4.3.1 Regression

The regression approach means that we predict the proportion of female students in a course directly. For this task we have used two regression models, a linear regression model and a ridge regression model. These are then trained and evaluated on both single comments and all comments together, both with and without part-of-speech tags.

4.3.1.1 Baseline

The results from the baseline model that always predicts the mean of the data can be found in table A.2. To read the baseline values for all of the calculated metrics, refer to A.1.1.

Baseline		
Result	Eng	Swe
MAE	0.18	0.14

Table 4.10: The results from the baseline dummy regressors trained on Swedish and English data.

The mean absolute error is slightly higher for the English data, at 0.18 compared to 0.14 for the Swedish data. However, we don't have information about the sign of

the error, so we don't know if it has predicted a higher or lower proportion of female students than there are.

4.3.1.2 Metrics

To compare the different regressors we have calculated several different metrics that are all available in the tables in the appendix, at A.1.2.2. However, we have also included the mean absolute error for all of the classifiers, which can be found in table 4.11.

Gender Balance Prediction MAE					
Swedish					
Model	Ridge Regression		Linear Regression		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.14	**0.13**	0.14	**0.13**	0.14
Part of Speech	0.14	**0.13**	0.14	**0.13**	0.14
English					
Model	Ridge Regression		Linear Regression		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.13	0.13	0.13	0.13	0.18
Part of Speech	0.13	**0.12**	0.13	**0.12**	0.18

Table 4.11: The Mean Absolute Error for the best regressors in predicting the gender balance in a course.

For the MAE score for the regression task, the regressors performed very similarly to each other. For the Swedish data, the best score is 0.14 and the worst score is 0.13. The linear regression and the ridge regression model trained on any dataset in the single comment format yield the best score and training in the multi-comment format gives the worst score. The regressors trained on the English data perform slightly worse, with a best score of 0.13 and a worst score of 0.12. Here the worst score comes from training the models on the data with added part-of-speech tags in the multi-comment format and training on all the other datasets yields the best score.

4.3.1.3 Important Features

The main part of this task is examining what features are important to predict a higher or lower portion of female students in a course. To do this, a list of the ten most important features to predict a low or high proportion respectively has been collected from each regressor. These features can be found in the appendix, in A.2.4 and A.2.3 respectively. However, we have also counted in how many of these lists each feature occurs, to create a table per language of the most important features, tables 4.12 and 4.13. As all regressors had such similar MAE scores, we decided not to include a table for the most important features for the best regressor.

Important features for Prediction in the English Data			
High proportion	Occurences	Low proportion	Occurences
system	4	process	5
tas	4	stressful	5
programming	4	around	3
ta	4	included	3
code	4	background	3
effort	3	previous	3
completely	3	making	3
-	-	feels	3
-	-	final	3
-	-	works	3
-	-	look	3

Table 4.12: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression models and datasets for the English data. Each word can appear at most eight times.

Some features are more important when predicting the proportion of female students in a course. In table 4.8 we can identify the most important words for predicting a high proportion of female students in the English courses. It can be interesting to note that while there are no part-of-speech tags present in the ten most important features for either language.

Important features for Prediction in the Swedish Data			
High proportion	Occurences	Low proportion	Occurences
labbarna	4	kritik	6
labbar	4	projekt	4
labb	4	pm	4
betyg	4	programm	4
labben	4	jättebra	4
personligen	3	tillbaka	4
-	-	examinatorn	3

Table 4.13: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression models and datasets for the Swedish data. Each word can appear at most eight times.

The most important features for the Swedish data are listed in table 4.13. Several of the words listed to predict a high proportion of female students are different variations of the word “labb” (lab) or the plural form “labbar” (labs), all of the different variations are present in half of the lists.

4.3.2 Classification

Two separate models were used for the classification task, a logistic regression model and a Support Vector Machine based classification model. These are trained on all datasets, i.e. full text, undersampled, and with added part-of-speech tags, all in both a multi-and single-comment fashion. We have set a threshold of the proportion of female students at above 0.4 as high and below as low.

4.3.2.1 Baseline

The baseline accuracy score and ROC AUC can be found in table 4.14. The baseline classifier simply classifies all samples as the majority class, which in this case is having a low proportion of female students. To read the baseline values for all of the calculated metrics, refer to A.1.1.

Baseline		
Result	Eng	Swe
Accuracy	0.71	0.65
ROC AUC	0.50	0.50

Table 4.14: The baseline metrics for author gender prediction classification for Swedish and English data.

The baseline accuracies for this task are lower than they are for examiner gender prediction. For the Swedish data, the baseline accuracy is down at 0.65, compared to the slightly higher English baseline at 0.71. Both ROC AUC scores are at 0.50 since the baseline classifier misclassifies all courses with a high proportion of female students.

4.3.2.2 Metrics

While the results from all of the gathered metrics are available in the appendix at A.1.2.3, we have included the accuracy score in table 4.15 and the ROC AUC score in table 4.16 for all of the classifiers. These two scores together give a reasonably nuanced, easy-to-understand picture of the results. However, as the model simply took too long to train, we were unable to obtain results from the support vector machine-based classifiers for the multi-comment data with added part-of-speech tags as well as all of the data in the single-comment format.

Gender Balance Prediction Accuracy					
Swedish					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.64	0.74	-	0.73	0.65
Undersampled	0.59	0.71	-	0.69	0.65
Part of Speech	0.65	**0.77**	-	-	0.65
English					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.73	0.77	-	0.77	0.71
Undersampled	0.59	0.72	-	0.71	0.71
Part of Speech	0.73	**0.79**	-	-	0.71

Table 4.15: The accuracy scores from the best classifiers when predicting the gender balance in a course, for the different datasets and using the different classification models.

The best accuracy of 0.77 for this task using the Swedish data was obtained by training a logistic regression model on the data with added part-of-speech tags in the multi-comment format. This result is a bit above the baseline accuracy, which is at 0.65 for the Swedish data. The worst accuracy of 0.59 comes from training the logistic regression model on the undersampled dataset in the single-comment format.

For the classifiers trained on English data, the best accuracy is 0.79. This came from training a logistic regression model on the data with added part-of-speech tags in the multi-comment format. While the data is different, the same model trained on the same variation of the data performed best in terms of accuracy for both languages. The baseline for the English data is a little bit higher at 0.71, compared to the baseline for the Swedish data. However, the best classifier also performs a little bit better for the English data compared to the Swedish. The worst accuracy is again at 0.59 for the English data and is obtained in the same way as for the Swedish data, by training the logistic regression model on the undersampled dataset in the single-comment format.

For both languages, the part-of-speech and full-text data sets perform better compared to the undersampled data. However, this difference is much larger for the single-comment format than the multi-comment format. While we do not have results for all variations of the support vector machine classifier, the performance between the models is similar for the results that we do have.

Gender Balance Prediction ROC AUC					
Swedish					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.52	0.71	-	0.71	0.50
Undersampled	0.56	0.70	-	0.69	0.50
Part of Speech	0.53	**0.74**	-	-	0.50
English					
Model	Logistic Regression		Support Vector Classifier		
Dataset	Single	Multi	Single	Multi	Baseline
Full Text	0.52	0.74	-	0.75	0.50
Undersampled	0.59	0.72	-	0.71	0.50
Part of Speech	0.52	**0.76**	-	-	0.50

Table 4.16: The ROC AUC scores from the best classifiers when predicting the gender balance in a course, for the different datasets and using the different classification models.

For the Swedish data, as can be seen in table 4.16, the best ROC AUC score of 0.74 is obtained by training a logistic regression model on the data with added part-of-speech tags in the multi-comment format. This is well above the baseline of 0.50. The worst score comes from training a logistic regression model on the full-text data in the single comment format, at 0.52, that score is barely above the baseline.

The same model trained on the same version of the data, i.e. the logistic regression model on the data with added part-of-speech tags in the multi-comment format, produces the best ROC AUC score for the English data as well. This score of 0.76 is marginally higher than it is for the Swedish data. The worst score for the English data is the same as for the Swedish data at 0.52 and comes from the same model - the logistic regression model on the full-text data in the single comment format.

Altogether, for classifying whether a course has a high or low proportion of female students, the logistic regression model trained on multi-comment data with added part-of-speech tags performed the best - across both languages as well as across both of the metrics. However, it is important to note that we do not have access to all results as we were unable to train the support vector classifier on all of the datasets. One of these might have performed better or worse than the best and worst model at hand.

4.3.2.3 Important Features

The most important features have been gathered and listed as for the other tasks. These features are found in tables 4.13 and 4.18. However, the full lists of the ten most important features for each classifier for predicting a high and low portion of female students are found in the appendix, at A.2.3 and A.2.4 respectively.

Important features for Prediction in the Swedish Data			
High proportion	Occurences	Low proportion	Occurences
kritik	8	labbar	4
tydliga	4	informationen	3
därför	4	personligen	3
ligger	3	vilken	3
genomgång	3	använda	3
önskat	3	labbarna	3
projekt	3	labb	3
-	-	betyg	3

Table 4.17: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression models and datasets for the Swedish data. Each word can appear at most eight times.

Examining the Swedish data, we can see a couple of things. When predicting a low proportion of female students there are three variations of the word "labb" present. However, no word is present in more than half of the lists. There are no part-of-speech tags present in the most important features at all.

Important features for Prediction in the English Data			
High proportion	Occurences	Low proportion	Occurences
text	6	tas	7
individual	4	assistants	5
process	4	problems	4
feels	3	ta	4
important	3	far	3
almost	3	notes	3
zoom	3	theoretical	3
stressful	3	prepared	3
-	-	programming	3
-	-	code	3

Table 4.18: A comparison of the most important words for predicting that a course has a high or low proportion of female students and the number of times that word appears among the ten most important features across all regression models and datasets for the English data. Each word can appear at most eight times.

In the English features in table 4.18, we find that the word "tas" is present in almost all of the lists of important features (seven out of eight) to predict a low proportion of female students. This word is the pluralization of the appreciation "TA" or teaching assistant. Interestingly, the list of important features to predict a low proportion of female students also includes the words "assistants" and "ta" in the singular form. There are also a couple of words present related to code and programming. For

predicting a high proportion of female students there aren't any words related to teaching assistants or any subject. However, the word "text" is present in almost all of the lists.

Interestingly, the same classifiers performed best in terms of both ROC AUC and accuracy for this task. Also, it was trained on the same corresponding dataset for both languages. The best was a logistic regression model trained on the data with the added part-of-speech tags in a multi-comment format.

Important features for the Best Classifier	
High Proportion	Low Proportion
kritik	programmering
lyssna	labbarna
programm	kod
omfattande	labbar
skapa	labb
metod	pass
därför	betyg
INTJ	ja
ordentlig	kurshemsida
gammal	kursbok

Table 4.19: The ten most important features to predict that a course has a high and low proportion of female students for the logistic regression model trained on the Swedish data with added part-of-speech tags in the multi-comment format.

In the most important features for the best classifier to predict that a course has a low proportion of female students in the Swedish data in table 4.19, there are three variations of the word "labb" (lab). There are also two words related to programming. Several of the words are related to specific parts of the course and course administration, such as "labb", "kurshemsida", "kursbok", "pass" (which in this case is likely a lab session) and "betyg". Predicting a high proportion of female students, the words seem less about the course as such. While the misspelled "programm" (which may be due to the preprocessing) could be related to that. In this list we also find the part-of-speech tag *INTJ*, interjection.

Important features for the Best Classifier	
High Proportion	Low Proportion
area	tas
person	programming
expect	note
second	code
life	ta
provide	simulation
email	assistant
mostly	homework
almost	partner
miss	problem

Table 4.20: The ten most important features to predict that a course has a high and low proportion of female students for the logistic regression model trained on the English data with added part-of-speech tags in the multi-comment format.

When training the best classifier on the English data, see table 4.20, there are some similarities to the results for the Swedish data. For example, the words "programming" and "code" are present in the lists of important features to predict a low proportion of female students for both languages. In the English list are also some variations of the word TA (teaching assistant). Other words that are related to the course and course content are "simulation", "homework" and "partner". Again, in the list of features to predict a high proportion of female students there is not such a clear pattern and none of the words seem directly directed to the course. It is interesting to note that none of the important features are a part-of-speech tag, despite both classfiers being trained on the data with the added part-of-speech tags.

5

Discussion

5.1 Statistical Analysis of the Data

5.1.1 Student Grades

It is well worth noting, that while our results clearly indicate that female students have received consistently higher grades from both male and female examiners, this might have different reasons. One such reason is of course that female students on average perform better academically and thus are deserving of higher grades, on the other end of the spectrum is the reason for gender bias against male students or in favor of female students. Another thing that should be noted here is that male examiners have a smaller difference between the grades of female and male students, 0.05 as compared to 0.15 from the female examiners. It is very possible that this difference is related to what courses that examiners of different genders teach and what students attend those courses. Meaning that the subject of the course is likely a factor in both what grades students get as well as who is teaching and attending. There are two main explanations for this difference that would imply gender bias from the examiners: Either female examiners give their female students too high grades because they are biased in favor of them (or against male students) or male examiners give the female students too low grades because they are biased against them (or for male students). Had one group of students had consistent grades for both examiner genders, this distinction should be clear. However, interestingly, male examiners generally give students lower grades than their female colleagues do, even if the difference is smaller for male students. Female students get an average grade score that is 0.22 higher when they have a female examiner, and that same number is 0.12 for male students.

Still, it is important to note that a lot of the examinations in courses are anonymous at the time of grading, such as exams. While the examiner does know what student is receiving a final grade, they cannot stray too far from the student's results in the different parts of a course without raising suspicion. Then there are examination forms that are far from anonymous, such as group projects and lab sessions, where students may be in constant contact with the examiner and any inherent bias from them would be much clearer to the students. Since we do not have access to information about what courses have anonymous examinations and what courses don't, we cannot properly investigate whether the form of examination makes a difference. Another aspect of any potential bias is that it is not always clear and direct. Perhaps bias comes through in how the course is designed, fitting one group better than the other. It might also take hold of the culture of the course, making some

students uncomfortable and thus performing worse. Some groups may also be more comfortable in different types of examinations, such as group projects compared to sit-down exams, etc.
It is clear from our data that female examiners have given their students consistently higher grades compared to their male counterparts. One simple explanation to this could simply be that female examiners at Chalmers are on average better teachers and thus the students learn more and get higher grades. A related explanation could be that women take the teaching more seriously. However, both of these explanations should be reflected in the overall impression scores of the courses unless the students are biased, as courses by better or more dedicated teachers should be better. Another possible explanation could simply be that the subjects that are taught by female examiners are easier to obtain higher grades in. Meaning that the difference may really not be related to examiner gender at all, but rather the subjects that examiners of that gender choose to teach. However, we do not have sufficient data to properly study whether this is the case or not.

5.1.2 Overall Impression of Courses

Interestingly, even if female examiners give higher grades to their students, the courses that they teach receive lower ratings in the course evaluations. The difference in overall impression score isn't very large at about 0.10, compared to for example the results in [8] where they found that male instructors were rated as high as 14.2% – of a standard deviation higher than the female in one of their studied student groups. In their study, however, the evaluation questions they used were more directly related to the instructor, in contrast to giving a score to the course, which may perpetuate any bias further. Interestingly, when it came to more course-related questions, their findings where similar to ours, with a difference of around 0.1 on a Likert scale in favor of male teachers. While 0.1 isn't a great difference in score, it is important to note that this is in contrast to the consistently higher grades that the students receive in these courses. Higher grades should indicate a course of higher quality, at least in terms of what the student is learning unless the teacher that is giving the grades is biased in some way and the grade doesn't reflect what the student has learned at all. Still, the courses with female examiners are rated lower. If it is the case that the higher grades are due to a course of higher quality, these results could indicate a bias against female examiners from the students. On the other hand, if the overall impression scores are reflective of the quality of the course, then this indicates bias from the examiners in their grading process.
Another interesting aspect is that this difference is larger for examiners in courses taught in Swedish compared to English. It is also worth noting that the courses in English receive higher overall ratings in general, even if the difference is very small – 3.84 for English compared to 3.83 for Swedish.

5.1.2.1 How Does Teaching Language Affect Overall Impression?

For courses with female examiners, there seems to be a clear difference in overall impression for English and Swedish courses. The English courses receive clearly higher grades. This difference is not evident for male examiners where the results

for English and Swedish courses are very similar, even if the English courses perform marginally better.
Why are we seeing these patterns? Are the female examiners in English courses simply better teachers? If that is the case, then why don't we see this pattern for male teachers?
A factor when it comes to Chalmers courses and teaching language is the level at which the course is taught. Bachelor courses are typically taught in Swedish, while master's courses are taught in English. It could be related to master's courses having better quality than bachelor's courses. But then, we should still see similar results for male examiners. It is also possible that the students on a master's level are less biased compared to bachelor students. Here it is interesting to note that on the masters level, there are more international students compared to the bachelor courses, which might mean that international students are more favorably dispositioned in their opinion on women as examiners.
It is also interesting to note that there is a larger proportion of female examiners for English courses compared to Swedish courses. This could mean that the results for the Swedish data are driven by some examiners getting especially bad ratings compared to their colleagues. It is possible that when there are more female examiners, they are either better at teaching or are perceived as such by the students. If there is a bias against female teachers, this might be smaller if there are more female teachers, and students thus get used to being taught by women. As courses in English generally are on an advanced level at Chalmers, this could be related to more female examiners teaching these courses. The courses on the bachelor's level are taught in Swedish and are on more foundational subjects that are often shared as broader basic courses for several programs, one such subject is for example mathematics, where there might be more male examiners. The advanced courses, on the other hand, are more specified. Thus these courses are probably more related to the core subjects of the master's programs that take them. This all means that there might be a difference in what subjects are taught at a bachelor's and advanced level, rather than that the teaching language having an impact on the proportion of female examiners. It should also be noted that the courses on the more advanced level are often chosen specifically by the student and thus better tailored to the student's specific interests, which could explain better ratings as well.

5.1.2.2 How is the Overall Impression Affected by the Students' Genders?

Examining the relationship between the proportion of female students and the overall impression of a course in section 4.1.3.2, we can identify some potential clusters in the data. However, even after reducing the number of data points of the majority class to the same number as the minority class in order to get a clear picture, it isn't obvious how the overall impression and proportion of female students relate. Still, there could be some relation between the variables, even if the nature of it isn't clear. It is possible that using a clustering algorithm to make sense of the data could give a clearer picture. The gender of a student could be a factor in how they perceive a course and how they evaluate it, but this is a much more complex question and relates to many different factors of a student – such as prerequisites,

self-confidence, etc. Thus, even if a relationship between student gender and overall impression would be found in further analysis of the data, it would not be surprising if that relationship isn't completely clear.

5.1.2.3 How Do Student Grades Affect the Overall Impression?

In section 4.1.3.3 we explore the data in terms of the relationship between the average grade for the students in a course and the average overall impression score of the course. While we expected to see some linearity to the relationship, i.e. students with higher grades being happier with the course – this did not seem to be the case. However, the data is noisy and has several outliers that may cloud a relationship that is actually present. We can see indications of such a relationship for courses where students have an average student grade of around 4.0, where many courses are clustered around the same value on the overall impression axis. To explore these indications further, it would have been interesting to perform some form of clustering algorithm on the data to further explore it.

5.2 Student gender Distribution

When examining how the student genders are affected by the examiner's gender, we found a clear tendency that more female students enroll in classes with female examiners. We need to note that this can be due to a multitude of reasons, in particular, the subject being taught. While the courses at Chalmers are technical in nature, different areas still have different gender distributions among students and staff. So some areas just have a larger proportion of women. Still, while this type of reason may seem circumstantial, it is interesting to note why more women are attracted to some subjects. There might still be some bias about women's aptitude for the subject and whether or not they can find a career in that line of work, at play. Another thing that we need to address is that students don't always actively choose courses, and may not at all be aware of the examiner's gender before the course starts. Most courses are mandatory in programs, especially at the bachelor's level, and the students have only chosen their program, not the specific course. There are also compulsory elective courses, where the students choose from a set of courses and need to pick a certain number and finally, there are the elective courses that students choose as they like.

5.3 Words that are Important for Predicting the Gender of an Examiner

Almost all models for examiner gender prediction perform below the baseline in terms of accuracy on the Swedish data. This could indicate that there is no big difference in how students write about the courses in terms of examiner gender. However, in terms of AUC ROC, most of the classifiers performed above baseline. This, together with the relatively low accuracies could indicate that while the classifiers still make some errors, they are improving and are to some extent able to

classify the courses with female examiners correctly. Thus, the classifiers seem to find some correlation between examiner gender and the words that are used in the comments. Looking at what features are deemed most important by the classifiers, while we cannot find a clear pattern some words seem to matter a bit more for predicting a female examiner such as “ännu” and “skriva”. When it comes to predicting a male examiner, no words seem to be as indicative. One possible reason could be that we have a lot more training data for courses with male examiners, which makes each comment less important when training the classifier. Though, it is interesting to note that for the English data, the results are opposite. The words for predicting a male examiner are more frequently the same in the lists for the different classifiers. For example, as can be seen in table 4.7, the word computer has eleven occurrences in the lists of important features. Several of the words for predicting a male examiner in the English data are related to computers and programming. However, none of the words for predicting a female examiner seem related to the subject that is taught. One possible reason is that why there are male examiners in all fields, the percentage of male examiners is especially high in subjects within computer science. The female examiners, however, are fewer and spread out in a lot of different fields. Thus the contribution that field-related words have to the classification models might not be that significant. The words are more related to the course as such and course administration as opposed to the contents of the course. However, this could also be due to some bias in what students write, that they write more about administrational issues to female examiners and more related to the course contents to male examiners. Understanding exactly why this difference is, would need to be investigated further. Something that can potentially affect our results is that we do not have access to information about who has actually been teaching a course, only the official examiner. It is not uncommon for the teaching to be performed by someone other than the examiner – potentially someone with a different gender or academic position. In these cases, the person teaching is often junior compared to the examiner.

As we did not have access to a lot of data, we used all text comments in the predictions. However, it could be interesting to only use the answers to some of the questions to see if there is more bias to the answers to some of the questions. Specifically, the questions regarding the teaching, as these are more related to the examiner and how their performance is perceived by the students. It would also be interesting to train and evaluate the classifiers on different questions and compare the results between these.

5.3.1 Are Important Features a Good Measure for Bias in Text?

Is comparing the most important features of a prediction model trained on the text a good method for detecting bias against examiners in the comments? It can be argued that the similarities between the lists of important features for the classifiers rather point to how good they are at predicting the examiners’ genders rather than how clear the differences are in the data. Still, we have previously argued that a good classifier should have found some connection between the comments and the gender.

Looking at what features are important for the classifier trained on Swedish data with the best ROC AUC score in table 4.4, we cannot clearly identify any patterns in the words. In table 4.5 we see the most important features for the best classifiers trained on the English data. While the patterns in the English words are still not clear, it seems like more of the words related to the female examiners are centered around communication. Words like “helpful”, “communication” and “team”, may point to another way of teaching compared to some of the words used to predict a male examiner such as “tutorials”, “assistants”, “labbarna” and “tentorna” that indicate a more traditional approach to teaching and learning.
Some words like “helpful” or “tillgänglig” are not very clear since we do not have the whole context. We cannot say if the phrase were “the teacher is helpful” or “the teacher is not helpful” and thus we do not know if they are words with positive or negative connotations. One could make use of context clues or other knowledge to guess in what sense the word is meant. As an example, knowing that female teachers are often seen as approachable and spend more of their time helping colleagues and students [4], we can guess that the word “helpful” is meant in the sense that “the teacher is helpful” in the list of features to predict a female examiner. However, this method is far from fool-proof and risks inducing further bias from the interpreter of the words.

5.3.2 Potential Issues in Relation to Names

In our project, we don’t have information about the gender of the examiner directly but rather make an estimation based on their first name. In turn, these predictions for the examiners’ genders aren’t necessarily perfect. While the statistics that are the basis for the name-to-gender prediction are Swedish, far from all examiners are Swedish. An example of a name that could be wrongly assigned is “Andrea”, which is commonly used for girls and women in Sweden, but a typical man’s name in Italy. A person’s official name may also not correspond with their own gender identity, which may not even be something that they are displaying publicly. However, in the examiner gender prediction task, we are more interested in how the students perceive the examiner’s gender, as it is their comments we are using as data for the classification.

5.4 Predicting the Gender of the Author of a Comment

When examining whether male or female students write differently, we did not have information about the gender of the author of the individual comments. We did, however, know the proportion of female students registered in the course. Thus, any patterns we find may also be related to how students write when there is a higher or lower proportion of women in the group, rather than men and women writing in different ways. To truly know the reason behind this would require more data and possibly a deeper study of human behavior.
Classifying whether a course had a high or low proportion of female students was

the only time when we had a very clear answer as to which classifier or regressor performed the best, and it was the same for both languages. When looking at the most important features to predict that a course has a low proportion of female students for the best classifier in table 4.19, we find several variations of the word labb (lab). The other words also seem related to the course as such, such as "betyg" (grade) and "kurshemsida" (course page). All except the word "ja", which just means "yes". For the words to predict a high proportion of female students, the words seem less direcly related to the course and perhaps more related to communication during or about the course. We find words such as "lyssna" (listen) and "kritik" (criticism) at the top of the list. But also other words that may be related to the course but not as clearly as the words used to predict a low proportion of male students, such as "metod" (method), "skapa" (create) and "omfattande" (extensive). However, these words seem a bit less concrete compared to the words for predicting a lower portion of female students and may be more related to the teaching as opposed to the specific parts of the course (such as labs). This could indicate that women write more about inter-person communication instead of the specific parts of the course. In this list we also have the part-of-speech tag INTJ, meaning interjection, which could indicate that women use more interjections in their writing. Examining the English words for the best classifier in table 4.20 we find that none of the words seem related to the course as such. However, the words don't seem as clearly related to people and communication as their Swedish counterparts. Only the words "email" and "person" are directly related to that. Predicting a low proportion of female students we find several words for teaching assistants. The words are also more related to the course and its contents compared to the words used to predict a high female proportion. Words like "code", "partner" and "homework" are all more directly related to the course compared to "expect", "provide", "mostly" etc.

For the regression task, the regressors performed very similarly to each other in terms of metrics, and we could not identify the "best" regressor. However, all regressors performed on par with or better than the baseline. The regressors trained on the English data had an MAE well below the baseline, and as should be remembered, when measuring error we want to have a low score. Aggregating the resulting important features from the different regressors in section 4.3.1.3, we find some interesting things. First, several of the words that are important to predict that a course has a high proportion of female students in the English data are related to code and programming. But the courses in subjects related to that have relatively few women enrolled. There are also several variations of the word "TA" (Teaching Assistant). This is interesting as these results seem like the opposite of what we found when exploring the important features in the classification task. It is also important to note that only two of the words in the regression lists occur in more than half of the lists: "process" and "stressful" to predict a low proportion of female students. For the Swedish data, most words indicating a high proportion of female students were different variations of the word "labb", which all appeared in half of the lists. This could indicate that women write about labs, but could also be due to that they are enrolled in courses with a lot of labs. The chemistry department has a relatively high proportion of female students on a bachelor's and master's level, and that could explain these results. No other clear patterns could be deduced from the

data.
Combining the results from the regression and classification solutions to this problem, we see something interesting. The two lists of the eight most important features to predict that a course has a high or low proportion of female students have five words in common – that is more than half of the words are in both lists. This could indicate that there men and women write similarly and use the same words. But it also leaves us with three words that are separate, "text", "individual" and "system" for women and "assistants", "stressful" and "problems" for men. It is also important to note that while we could not see any patterns in the results from the regression task, we did see some indications of different writing styles in the classification results. They indicate that men might write more about the matter directly at hand, such as the course in this case, while women broaden and write more about communication, etc. This would need to be investigated further to get a more definitive answer, but there do seem to be some indications from the resulting most important features in the classification task.

5.5 Issues with the Course Evaluation Data

Another aspect of the comments that is worth mentioning is that we only have access to the comments after they have been filtered to remove profanities and insults, but it would have been interesting to see how those words would have affected the results. It would also have been interesting to see how the comments that have been filtered out relate to gender, both in terms of who writes and who receives profanities.
In order to avoid getting features that are basically the same word in the results we could have lemmatized the texts beforehand. For example "labb" would only occur once in the important feature table, giving more room for other words and therefore would have given a more complete view of the words which had an impact on the tasks.
Whenever dealing with any type of voluntary data, such as course evaluations, we are dealing with self-selection bias [45]. This occurs since not all people are equally eager to participate in the survey, interview, or whatever the way of collecting data is. This leads to making people who have stronger opinions on the matter overrepresented in the data. For us in the course evaluations, that means that students who strongly liked or disliked the course are more likely to answer in the course evaluations.
It is also important to think about the course evaluation process as such when discussing any results extracted from this data. When evaluating a course, is the student actually evaluating the course or are they evaluating their own performance? Or are they in fact evaluating the examiner? While the purpose of the evaluations are to evaluate the course as such, separating these factors is not always easy for a student that is supposed to write their evaluation. They may have personal issues with a teacher (for example just don't like them) or be really disappointed in how poorly they performed in the course and "know" that they are getting a low grade. They may also like or dislike the subject, which in turn can affect how they value the course. It could be argued that at a university level, a student would be likely to enjoy what they are studying - but an engineering degree contains a lot of courses in different subjects and not everyone is bound to like everything, even if it's in some

way related to something they want to study.
Another issue regarding the course evaluations is related to the examiners. While the main purpose of the evaluations is to improve on the courses and catch any issues before they get too big, these evaluations could be seen as an evaluation of the examiner's performance in the teaching part of their work. This could then be used against the examiner when it comes to discussing promotions and salary raises. While this may not be the purpose of the evaluations and not even an active decision from the superior of the examiner, having access to the information may still affect their decisions and create an unconscious bias against the examiner.

5.6 Conclusion and Future Work

In conclusion, we refer back to the research questions in section 3.1. To answer **RQ1a**, we find that there seems to be some indication that female examiners do receive lower overall impression scores in the evaluation of their courses, compared to their male colleagues.
From our work, we have no distinct answer to **RQ1b**. While we can see indications of clusters in figure 4.10, this would need to be further examined. We suggest using some clustering techniques to explore the data and remove some noise to unveil a possible relationship between the variables. There could be some relationship between the variables that is not evident using only the naked eye.
When examining the data in terms of examiner gender, overall impression, and student grades to answer **RQ1c**, we do not find any clear indication of a connection between the variables.
To answer **RQ1d**, the difference between the data for the teaching languages is small but present. English courses consistently perform better in terms of overall impression. This difference is mainly driven by that female examiners get a higher overall impression score on their courses in English courses compared to the Swedish courses.
When examining the results from the examiner gender classification, we found no significantly clear patterns in the important features for either language. However, there are some indications of some words that seem related to another way of teaching between the sexes, but these word lists would need to be examined by someone more versed in gender studies to truly answer **RQ2a**. It is also important to note that none of the classifiers were able to get particularly good metrics, which could indicate that there are no clear differences in the comments depending on the examiner's gender. One of the patterns we did find was in the English data, where words related to computers often were important to predict a male examiner, while words used to predict a female examiner didn't seem related to any subject.
Answering **RQ2b**, the different prediction models seemed to perform better compared to their baseline, than the models used to predict the examiner gender did. Using the proportion of female students in the course as a proxy for the gender of the author of a comment, we found some patterns that are worth exploring further. The words that were associated with a female author seemed more related to people and communication, while the words associated with a male author were more related to the course and its content. However, these patterns were clearest for the

best classifiers but were consistent across both languages. Interestingly, there was no such pattern for the regression task at all. This is possibly due to not having enough training data.
RQ3 has a clear answer in terms of our data. As presented in section 4.1.3.2, courses with female examiners have a higher proportion of female students. In this data, this means that they have a more even gender balance, as most courses have a large majority of male students.
Properly answering **RQ4** would require further examination of the data. While we can see patterns in the data indicating that there is a relationship between the proportion of female students in a class and the average grade of the students, the exact nature of it isn't completely clear. To continue this work, outliers and noise would need to be examined and potentially removed to more clearly understand the connection.
Regarding all of the research questions, it would have been interesting to account for how the subject of a course affects the results. This could possibly be done by using the course codes to identify the subjects in a broad sense, or by gathering additional data. Other experiments that were lifted as ideas for this thesis, but that couldn't fit withing the time scope where to do a sentiment analysis on the text answers to the overall impression questions, with the numerical answers as an annotation. This training could be done on separate datasets for male and female examiners to possibly find indications of bias. Finally, we suggest as future work to perform the machine learning experiments done in this thesis with some improvements by training on the lemmas instead of words, using bi-grams or tri-grams and choosing to only use specific questions from the evaluations instead of all of them.

Bibliography

[1] S. Marken, *28% of Women in Academia Say Gender Limits Their Advancement*, Mar. 2022. [Online]. Available: https://news.gallup.com/opinion/gallup/391226/women-academia-say-gender-limits-advancement.aspx.

[2] G. Abramo, D. W. Aksnes, and C. A. D'Angelo, "Gender differences in research performance within and between countries: Italy vs norway", *Journal of Informetrics*, vol. 15, no. 2, p. 101 144, 2021, ISSN: 1751-1577. DOI: 10.1016/j.joi.2021.101144.

[3] O. Ajmal, C. Hemberg, M. Inkala, and A. Lundeteg, "Vetenskapsmannen, inte kvinna", Allbright, Apr. 2019. [Online]. Available: https://www.allbright.se/nyheter/2019/4/4/rapportslpp-vetenskapsmannen-inte-kvinna.

[4] B. J. Casad, J. E. Franks, C. E. Garasky, *et al.*, "Gender inequality in academia: Problems and solutions for women faculty in stem", *Journal of Neuroscience Research*, vol. 99, no. 1, pp. 13–23, 2021. DOI: 10.1002/jnr.24631.

[5] *Flest män i toppen på akademin - Tidningen Curie*, Aug. 2019. [Online]. Available: https://www.tidningencurie.se/nyheter/flest-man-i-toppen-pa-akademin.

[6] *Jämställdhets-och lika villkorsplan*, 2022. [Online]. Available: https://www.chalmers.se/en/about-chalmers/organisation-and-governance/equality/gender-and-equal-opportunities-plan/.

[7] *Genie - Gender Initiative for Excellence | Chalmers*. [Online]. Available: https://www.chalmers.se/om-chalmers/organisation-och-styrning/jamstalldhet/genie-gender-initiative-for-excellence/.

[8] F. Mengel, J. Sauermann, and U. Zölitz, "Gender Bias in Teaching Evaluations", *Journal of the European Economic Association*, vol. 17, no. 2, pp. 535–566, Feb. 2018, ISSN: 1542-4766. DOI: 10.1093/jeea/jvx057.

[9] A. Boring, "Gender biases in student evaluations of teaching", *Journal of Public Economics*, vol. 145, pp. 27–41, Jan. 2017, ISSN: 0047-2727. DOI: 10.1016/J.JPUBECO.2016.11.006.

[10] K. M. W. Mitchell and J. Martin, "Gender bias in student evaluations", *PS: Political Science & Politics*, vol. 51, no. 3, pp. 648–652, 2018. DOI: 10.1017/S104909651800001X.

[11] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services", *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019. DOI: 10.1145/3359246.

[12] B. N. Torgrimson and C. T. Minson, "Sex and gender: What is the difference?", *Journal of Applied Physiology*, vol. 99, no. 3, pp. 785–787, 2005. DOI: 10.1152/japplphysiol.00376.2005.

[13] Folkhälsomyndigheten. "Rätten till hälsa hur normer och strukturer inverkar på transpersoners upplevelser av sexuell hälsa". (Jun. 2016), [Online]. Available: https://www.folkhalsomyndigheten.se/publikationer-och-material/publikationsarkiv/r/ratten-till-halsa-hur-normer-och-strukturer-inverkar-pa-transpersoners-upplevelser-av-sexuell-halsa/.

[14] S. Biradar, B. Torgal, N. Hosamani, R. Bidarakundi, and S. Mudhol, "Age and gender detection system using Raspberry Pi", *International Journal of Computer Sciences and Engineering*, vol. 7, no. 6, pp. 14–18, Jun. 2019. DOI: 10.26438/ijcse/v7i6.1418.

[15] U. United Nations Development Programme, "2020 gender social norms index (GSNI)", New York, Mar. 2020. [Online]. Available: https://hdr.undp.org/content/2020-gender-social-norms-index-gsni.

[16] E. Larson. "New research: Diversity + inclusion = better decision making at work". (Sep. 2017), [Online]. Available: https://www.forbes.com/sites/eriklarson/2017/09/21/new-research-diversity-inclusion-better-decision-making-at-work/?sh=9de51344cbfa (visited on 12/12/2021).

[17] S. Sancier-Sultan, D. Garibian, and J. Sperling-Magro. "Taking the lead for inclusion | McKinsey". (Nov. 2019), [Online]. Available: https://www.mckinsey.com/featured-insights/gender-equality/taking-the-lead-for-inclusion (visited on 12/12/2021).

[18] T. A. Huston, "Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity?", *Seattle Journal for Social Justice*, vol. 4, no. 2, p. 34, 2006.

[19] V. Lavy and R. Megalokonomou, "Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study", National Bureau of Economic Research, Tech. Rep. 26021, Jun. 2019. DOI: 10.3386/w26021.

[20] M. Carlana, "Implicit stereotypes: Evidence from teachers' gender bias", *The Quarterly Journal of Economics*, vol. 134, no. 3, pp. 1163–1224, Mar. 2019, ISSN: 0033-5533. DOI: 10.1093/qje/qjz008.

[21] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality", *Journal of Personality and Social Psychology*, vol. 101, no. 1, pp. 109–128, Jul. 2011, ISSN: 00223514. DOI: 10.1037/A0022530.

[22] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text", *Digital Investigation*, vol. 8, no. 1, pp. 78–88, 2011, ISSN: 17422876. DOI: 10.1016/J.DIIN.2011.04.002.

[23] *Program och kurser | Chalmers*, Feb. 2023. [Online]. Available: https://www.chalmers.se/utbildning/program-och-kurser/.

[24] *Betyg | Chalmers*, Oct. 2022. [Online]. Available: https://www.chalmers.se/utbildning/dina-studier/planera-och-genomfora-studier/betyg/.

[25] *Before examination | Chalmers*, Nov. 2022. [Online]. Available: https://www.chalmers.se/en/education/your-studies/plan-and-conduct-your-studies/examinations-and-other-summative-assessments/before-examination/.

[26] University of Gothenburg, *Collaborate with us*, Oct. 2022. [Online]. Available: https://www.gu.se/en/about-the-university/collaborate-with-us.

[27] *Course evaluation | Chalmers studentportal*, Jun. 2022. [Online]. Available: https://www.chalmers.se/en/education/your-studies/plan-and-conduct-your-studies/course-evaluation/.

[28] *Kursvärdering processbeskrivning*, Aug. 2016. [Online]. Available: https://www.chalmers.se/en/education/your-studies/plan-and-conduct-your-studies/course-evaluation/#process-description.

[29] S. S. Skiena, *The data science design manual*. Cham: Springer, 2017, ISBN: 9783319554433.

[30] D. Jurafsky, J. H. Martin, and A. Kehler, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall, 2000, ISBN: 0130950696.

[31] SciKit Learn, *6.2. Feature extraction*. [Online]. Available: https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction.

[32] Google Developers, *Imbalanced Data*, Jul. 2022. [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data.

[33] *How to Handle Imbalanced Classes in Machine Learning*, Jul. 2022. [Online]. Available: https://elitedatascience.com/imbalanced-classes.

[34] *What is Machine Learning? | IBM*. [Online]. Available: https://www.ibm.com/topics/machine-learning.

[35] *Classification vs Regression in Machine Learning - GeeksforGeeks*. [Online]. Available: https://www.geeksforgeeks.org/ml-classification-vs-regression/.

[36] S. Bandgar, *Logistic Regression*. May 2021. [Online]. Available: https://medium.com/analytics-vidhya/logistic-regression-c5a6c047363e.

[37] S. Gupta, *RMSE: What does it mean?. Contributed by: Shweta Gupta | by Great Learning | Medium*, Apr. 2021. [Online]. Available: https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e.

[38] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. [Online]. Available: https://smlbook.org.

[39] H. Daumé III, *A Course in Machine Learning*, 0.99. 2022, ch. "Practical Issues". [Online]. Available: http://ciml.info/.

[40] *Weighting Confusion Matrices by Outcomes and Observations - Bryan Shalloway's Blog*. [Online]. Available: https://www.bryanshalloway.com/2020/12/08/weighting-classification-outcomes/.

[41] *True positive rate (TPR) – IBM Documentation*. [Online]. Available: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=overview-true-positive-rate-tpr.

[42] Google Developers, *Classification: ROC Curve and AUC*, Jul. 2022. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[43] S. Hiregoudar, *Ways to Evaluate Regression Models*, Aug. 2020. [Online]. Available: https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70.

[44] *3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 1.2.0 documentation*. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics.

[45] *What Is Self-Selection Bias? | Definition & Example*. [Online]. Available: https://www.scribbr.com/research-bias/self-selection-bias/.

A

Appendix 1

A.1 Prediction Metrics

A.1.1 Baselines

Baseline		
Result	Eng	Swe
Accuracy	0.80000	0.84069
Precision	0.80000	0.84069
Recall	1.00000	1.00000
ROC AUC	0.50000	0.50000

Table A.1: The baseline metrics for examiner gender prediction using Swedish and English data.

Baseline		
Result	Eng	Swe
MAE	0.17815	0.14417
MAPE	0.70292	0.55902
RMSE	0.20638	0.16861
Max error	0.40207	0.47738

Table A.2: The results from the baseline dummy regressors trained on Swedish and English data.

Baseline		
Result	Eng	Swe
Accuracy	0.80	0.84
Precision	0.80	0.84
Recall	1.00	1.00
ROC AUC	0.50	0.50

Table A.3: The baseline metrics for classifying the course gender balance using Swedish and English data.

A.1.2 Results

A.1.2.1 Examiner Gender

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.76967	0.61228	0.79079
Precision	0.85177	0.87107	0.87302
Recall	0.87900	0.63242	0.87900
ROC AUC	0.53588	0.56922	0.60215
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.83383	0.58813	0.83693
Precision	0.83406	0.84676	0.83724
Recall	0.99960	0.61460	0.99947
ROC AUC	0.50064	0.53703	0.50093

Table A.4: The resulting metrics from training and evaluating the best logistic regression model on the different Swedish datasets to predict the examiner's gender.

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.76104	0.62338	0.74545
Precision	0.85762	0.88626	0.84768
Recall	0.84091	0.60714	0.83117
ROC AUC	0.64123	0.64773	0.61688
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.84611	0.54636	0.84534
Precision	0.84695	0.87992	0.84607
Recall	0.99861	0.53648	0.99875
ROC AUC	0.50263	0.56840	0.50278

Table A.5: The resulting metrics from training and evaluating the logistic regression model on the different English datasets in the examiner gender prediction task.

SVC Swedish			
Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.76346	0.57198	0.76538
Precision	0.88193	0.88530	0.88517
Recall	0.83182	0.56393	0.83333
ROC AUC	0.60966	0.58919	0.60088
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.83297	0.59615	0.80000
Precision	0.83297	0.88298	0.88367
Recall	1.00000	0.58451	0.88367
ROC AUC	0.50000	0.61672	0.58567

Table A.6: The resulting metrics from training and evaluating the best support vector classifiers on Swedish data to predict the examiner's gender.

SVC English			
Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.74026	0.62078	0.76042
Precision	0.84818	0.88995	0.87171
Recall	0.82637	0.60194	0.83333
ROC AUC	0.60237	0.64966	0.62121
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.84627	0.64323	0.72727
Precision	0.84627	0.87336	0.86111
Recall	1.00000	0.64935	0.79233
ROC AUC	0.50000	0.63389	0.61839

Table A.7: The resulting metrics from training and evaluating the best support vector classifiers on English data to predict the examiner's gender.

A.1.2.2 Regression for Author Gender

Linear Regression English		
Multi Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.12650	0.12174
MAPE	0.54768	0.50471
RMSE	0.15772	0.15282
Max error	0.51269	0.61903
Single Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.12624	0.12675
MAPE	0.52229	0.52250
RMSE	0.15268	0.15316
Max error	0.69307	0.67442

Table A.8: The results from the best linear regression models trained on English data to predict the gender balance in a course.

Linear Regression Swedish		
Multi Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.13188	0.12932
MAPE	0.50359	0.48086
RMSE	0.16738	0.16105
Max error	1.09899	0.60443
Single Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.13681	0.13624
MAPE	0.52153	0.52202
RMSE	0.16110	0.16064
Max error	0.51452	0.87435

Table A.9: The results from the best linear regression models trained on Swedish data to predict the gender balance in a course.

Ridge Regression English		
Multi Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.12644	0.12165
MAPE	0.54741	0.50435
RMSE	0.15764	0.15271
Max error	0.51243	0.61870
Single Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.12624	0.12675
MAPE	0.52229	0.52250
RMSE	0.15268	0.15315
Max error	0.69306	0.67442

Table A.10: The results from the best ridge regression models trained on English data to predict the gender balance in a course.

Ridge Regression Swedish		
Multi Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.13184	0.12929
MAPE	0.50347	0.48073
RMSE	0.16733	0.16100
Max error	1.09808	0.60410
Single Comment Prediction		
Result	Full text	Included POS-tags
MAE	0.13681	0.13624
MAPE	0.52153	0.52202
RMSE	0.16110	0.16064
Max error	0.51444	0.87410

Table A.11: The results from the best of the ridge regression models trained on Swedish data to predict the gender balance in a course.

A.1.2.3 Classification of Author Gender

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.73896	0.70633	0.76583
Precision	0.77650	0.79866	0.84795
Recall	0.82371	0.71903	0.80556
ROC AUC	0.70873	0.70162	0.74129
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.64297	0.58628	0.64773
Precision	0.64779	0.68529	0.65247
Recall	0.96240	0.64772	0.95462
ROC AUC	0.52293	0.56314	0.53300

Table A.12: The resulting metrics from training and evaluating the best logistic regression model on the different Swedish datasets in the course gender balance classification task.

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.76943	0.72280	0.79275
Precision	0.88417	0.88085	0.87361
Recall	0.79514	0.72378	0.83630
ROC AUC	0.74451	0.72189	0.75624
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.73090	0.58787	0.73151
Precision	0.73585	0.79243	0.73631
Recall	0.98386	0.58864	0.98373
ROC AUC	0.51776	0.58722	0.51928

Table A.13: The resulting metrics from training and evaluating the best logistic regression model on the different English datasets in the course gender balance classification task.

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.72937	0.69482	-
Precision	0.76261	0.76871	-
Recall	0.80818	0.71293	-
ROC AUC	0.70704	0.68980	-
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	-	-	-
Precision	-	-	-
Recall	-	-	-
ROC AUC	-	-	-

Table A.14: The resulting metrics from training and evaluating the best support vector machine classifier on the different Swedish datasets in the course gender balance classification task.

Multi Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	0.77202	0.70984	-
Precision	0.89062	0.83408	-
Recall	0.79167	0.71264	-
ROC AUC	0.75298	0.70832	-
Single Comment Prediction			
Result	Full text	Undersampled	Included POS-tags
Accuracy	-	-	-
Precision	-	-	-
Recall	-	-	-
ROC AUC	-	-	-

Table A.15: The resulting metrics from training and evaluating the best support vector machine classifier on the different English datasets in the course gender balance classification task.

A.2 Important Features for Prediction

A.2.1 Predicting that a Course has a Female Examiner

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
ännu	hp	ännu
senare	tar	etc
materialet	ännu	lämna
arbeta	chalmers	förstod
skönt	låg	tempo
flesta	flesta	dra
etc	följa	särskilt
inlämningsuppgifter	laborationen	extrem
varandra	tiden	välja
bästa	mesta	lärande
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
intressanta	matlab	kemi
matlab	engelska	matlab
skriva	kritik	diskussion
projekt	skriva	bevis
veckor	intressanta	presentation
duggan	repetition	text
lärandemålen	upplägget	dra
senare	rapporten	repetition
skönt	projekt	diskutera
arbetet	veckan	skön

Table A.16: The ten most important features for predicting a sample as female for the best logistic regression model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
ännu	mail	ännu
princip	lösa	förstod
skönt	ännu	jämföra
arbeta	konstigt	laboratione
begrepp	all	rekommenderad
bästa	projekt	10
slutet	tillfällen	länge
tar	relevanta	välja
skriva	speciellt	redan
förstod	skriva	relevan
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
jämfört	begrepp	rekommenderad
sådär	mest	särskilt
nej	tydlig	lämna
skönt	slutet	ännu
flesta	lagom	jobb
kolla	examinator	personligen
samarbetet	ens	jämföra
kontakt	tar	förstod
båda	upplevde	skriva
rolig	konstigt	vidare

Table A.17: The ten most important features for predicting a sample as female for the support vector model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
stressful	based	structured
works	stressful	stressful
except	important	another
three	writing	correct
open	works	base
based	test	consider
feels	examples	until
giving	when	reason
change	period	leave
background	small	opportunity
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
guest	guest	article
low	lecturers	guest
articles	paper	calculation
lecturers	open	individual
paper	articles	activity
activities	mandatory	mandatory
discussions	low	paper
mandatory	works	text
open	discussions	team
individual	based	open

Table A.18: The ten most important features for predicting a sample as female for the logistic regression model trained and evaluated on English data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
improve	your	again
wrong	communication	appreciate
fair	helpful	administration
open	required	important
works	something	another
stressful	writing	stressful
another	learned	its
except	team	real
spend	three	correct
improved	workshop	text
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
okay	communication	certain
activities	wrong	base
lack	through	credit
low	made	opportunity
years	confusing	recommend
spend	extra	nothing
focused	improve	contribute
improved	their	extremely
attend	zoom	except
issues	sessions	actually

Table A.19: The ten most important features for predicting a sample as female for the support vector model trained and evaluated on English data.

A.2.2 Predicting that a Course has a Male Examiner

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
nytt	gått	kvar
förväntades	rolig	handledarna
givande	möjlighet	faktisk
hjälpte	där	pga
rolig	bort	praktisk
tänka	dessutom	nivån
däremot	heller	tentorna
klart	ger	ex
ger	dom	skapa
kurslitteraturen	la	lärorik
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
inlämningsuppgifterna	inlämningsuppgifterna	programmering
labbarna	labbarna	kod
tentamen	inlämningarna	praktisk
tentor	klart	tentorna
labb	går	labbarna
tavlan	labb	absolut
handledare	tentamen	tentame
bok	laborationerna	tavlan
inlämningarna	handledare	tillgänglig
betyg	aldrig	duggorna

Table A.20: The ten most important features for predicting a sample as male for the best logistic regression model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
förväntades	dålig	praktisk
nytt	gärna	absolut
tänka	intressant	krav
försöka	givande	faktisk
enda	hinna	alltså
onödigt	snabbt	pass
inlämning	alltid	lärorik
aldrig	inom	bygga
började	handledare	god
förkunskaper	tyvärr	förbättra
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
att	kurslitteraturen	praktisk
duggan	snabbt	pga
kurs	lika	behövt
stod	ihop	bild
lång	började	ex
övningsledare	dom	framförallt
kommer	framförallt	absolut
överlag	intressant	faktisk
på	just	bygga
chalmers	gärna	betydlig

Table A.21: The ten most important features for predicting a sample as male for the support vector model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
assistants	help	code
code	page	allow
tutorials	tutorials	computer
process	computer	explanation
computer	helped	suggest
clearer	jag	term
higher	det	believe
reports	last	someone
complete	yes	contain
follow	clearly	prepare
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
around	videos	code
several	computer	apply
code	ve	professor
assistants	does	computer
solving	went	okay
using	taken	explanation
complete	those	hade
fast	help	tool
ones	mostly	complete
believe	bad	page

Table A.22: The ten most important features for predicting a sample as male for the support vector model trained and evaluated on English data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
computer	tutorials	professor
harder	way	computer
tutorials	might	okay
code	det	page
perfect	over	code
those	problem	introduce
process	page	form
such	needs	perfect
learnt	exams	overall
clearer	computer	tool
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
matlab	assistants	computer
computer	computer	matlab
tutorials	matlab	code
code	labs	tutorial
assistants	tutorials	programming
labs	code	lab
programming	tas	assistant
tas	programming	simulation
lab	lab	partner
videos	notes	note

Table A.23: The ten most important features for predicting a sample as male for the logistic regression model trained and evaluated on English data.

A.2.3 Predicting that a Course has a High Proportion of Female Students

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
completely	minute
issues	effort
system	guess
presented	mean
prepared	although
using	far
effort	place
its	completely
members	without
lack	leave
Single Comment Prediction	
Full text	Including POS-tags
tas	tas
programming	programming
assistants	ta
labs	assistant
notes	note
ta	homework
code	partner
software	code
page	system
deadlines	lab

Table A.24: The ten most important features for the best English linear regression model predicting that a course has a proportion of female students to the higher end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
personligen	relativt
tempo	personligen
dugga	nej
överlag	vanlig
dom	direkt
extremt	främst
hand	fast
pga	hamna
stod	mål
enda	engelsk
Single Comment Prediction	
Full text	Including POS-tags
labbarna	labbarna
labbar	programmering
labb	labbar
laborationer	labb
kurshemsidan	kod
betyg	pass
labben	kurshemsida
laborationerna	labben
jämfört	betyg
problem	laboration

Table A.25: The ten most important features for the best linear regression model trained on Swedish data for predicting that a course has a proportion of female students to the higher end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
completely	minute
issues	effort
system	guess
presented	mean
prepared	although
using	far
effort	place
its	completely
members	without
lack	leave
Single Comment Prediction	
Full text	Including POS-tags
tas	tas
programming	programming
assistants	ta
labs	assistant
notes	note
ta	homework
code	partner
software	code
page	system
deadlines	lab

Table A.26: The ten most important features for the best ridge regression model trained on English data for predicting that a course has a proportion of female students to the higher end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
personligen	relativt
tempo	personligen
dugga	nej
överlag	vanlig
dom	direkt
extremt	främst
hand	fast
pga	hamna
stod	mål
enda	engelsk
Single Comment Prediction	
Full text	Including POS-tags
labbarna	labbarna
labbar	programmering
labb	labbar
laborationer	labb
kurshemsidan	kod
betyg	pass
labben	kurshemsida
laborationerna	labben
jämfört	betyg
problem	laboration

Table A.27: The ten most important features for the best ridge regression model trained on Swedish data for predicting that a course has a proportion of female students to the higher end of the spectrum.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
kritik	kritik	kritik
arbetsbelastningen	tydliga	lyssna
ligger	längre	programm
okej	otroligt	omfattande
skrev	genomgång	skapa
genomgång	arbete	metod
önskat	heller	därför
tydliga	klart	intj
därför	ligger	ordentlig
ihop	sent	gammal
Single Comment Prediction		
projekt	projekt	kemi
kritik	kritik	seminari
projektet	projektet	studiebesök
pm	upplevde	handledning
jättebra	klassen	kritik
zoom	jättebra	programm
matlab	strukturen	projekt
upplevde	matlab	projekten
klassen	sent	pm
kommunikation	önskat	stressig

Table A.28: The ten most important features for predicting a sample as having a higher proportion of female students for the best logistic regression model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
feels	feels	area
three	own	person
ones	important	expect
previous	others	second
try	person	life
important	individual	provide
almost	mycket	email
instead	text	mostly
text	background	almost
pass	complete	miss
Single Comment Prediction		
zoom	workshops	zoom
workshops	zoom	individual
individual	stressful	workshop
process	individual	text
presentations	lecturers	process
methods	process	reading
text	schedule	stressful
lecturers	text	final
stressful	final	online
schedule	seminars	method

Table A.29: The ten most important features for predicting a sample as having a higher proportion of female students for the logistic regression model trained and evaluated on English data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
önskat	tillfällen	-
kritik	ligger	-
arbetsbelastningen	tydliga	-
därför	kursens	-
tydliga	fort	-
välja	etc	-
strukturen	därför	-
genomgång	kritik	
kommunikation	stort	-
okej	upplägget	-

Table A.30: The ten most important features for predicting a sample as having a higher proportion of female students for the SVC model trained and evaluated on Swedish data. There are no words for single comment prediction as these experiments took too much time.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
feels	around	-
came	grading	-
three	took	-
almost	expected	-
text	might	-
ones	why	-
previous	pass	-
process	fact	-
important	several	-
possible	your	-

Table A.31: The ten most important features for predicting a sample as having a higher proportion of female students for the SVC model trained and evaluated on English data. Note that there are no words for single comment prediction as these experiments took too much time.

A.2.4 Predicting that a Course has a Low Proportion of Female Students

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
around	down
included	appreciate
background	text
process	early
previous	email
making	finish
feels	final
works	så
look	next
stressful	each
Single Comment Prediction	
Full text	Including POS-tags
around	workshop
included	zoom
background	seminar
process	reading
previous	stressful
making	guest
feels	final
works	article
look	process
stressful	individual

Table A.32: The ten most important features for the best linear regression model trained on English data for predicting that a course has a proportion of female students to the lower end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
tillbaka	nivån
arbetsbelastningen	bild
heller	programm
laborationen	ännu
kommunikation	tillbaka
kritik	metod
jättebra	plats
roligt	vela
examinatorn	strukturera
examinatorn	underlätta
Single Comment Prediction	
Full text	Including POS-tags
kritik	handledning
projekt	seminari
pm	kemi
önskat	studiebesök
klassen	kritik
kul	pm
upplevde	programm
jättebra	grupparbete
grupparbetet	stressig
skönt	projekt

Table A.33: The ten most important features for the best linear regression model trained on Swedish data for predicting that a course has a proportion of female students to the lower end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
around	down
included	appreciate
background	text
process	early
previous	email
making	finish
feels	final
works	så
look	next
stressful	each
Single Comment Prediction	
Full text	Including POS-tags
workshops	tas
zoom	programming
process	ta
seminars	assistant
workshop	note
lecturers	homework
presentations	partner
schedule	code
stressful	system
methods	lab

Table A.34: The ten most important features for the best ridge regression model trained on English data for predicting that a course has a proportion of female students to the lower end of the spectrum.

Important features	
Multi Comment Prediction	
Full text	Including POS-tags
tillbaka	nivån
arbetsbelastningen	bild
heller	programm
laborationen	ännu
kommunikation	tillbaka
kritik	metod
jättebra	plats
roligt	strukturera
examinatorn	vela
skrev	underlätta
Single Comment Prediction	
Full text	Including POS-tags
kritik	handledning
projekt	seminari
pm	kemi
önskat	studiebesök
klassen	kritik
kul	pm
upplevde	programm
jättebra	grupparbete
grupparbetet	stressig
skönt	projekt

Table A.35: The ten most important features for the best ridge regression model trained on Swedish data for predicting that a course has a proportion of female students to the lower end of the spectrum.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
överlag	hjälpte	pass
informationen	jämfört	tvung
personligen	labbar	enkel
mesta	båda	funka
instruktioner	informationen	nej
vilken	vilken	god
enda	överlag	hänga
stod	ställa	vanlig
särskilt	använda	koppla
högre	materialet	bevis
Single Comment Prediction		
Full text	Undersampled	Included POS-tags
labbarna	labbar	programmering
labbar	labbarna	labbarna
labb	labb	kod
laborationer	betyg	labbar
betyg	laborationer	labb
labben	kurshemsidan	pass
kurshemsidan	kurslitteraturen	betyg
inlämningsuppgifterna	kursboken	ja
kurslitteraturen	inlämningsuppgifterna	kurshemsida
nästan	problem	kursbok

Table A.36: The ten most important features for predicting a sample as having a lower proportion of female students for the logistic regression model trained and evaluated on Swedish data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
assistants	although	assistant
far	using	minute
tas	notes	tas
harder	spend	theoretical
prepared	solve	leave
come	problems	completely
bad	quality	please
needs	actually	engineering
similar	prepared	note
members	team	far
Single Comment Prediction		
tas	tas	tas
assistants	notes	programming
programming	assistants	note
notes	programming	code
ta	ta	ta
code	code	simulation
labs	labs	assistant
software	software	homework
problems	theoretical	partner
theoretical	problems	problem

Table A.37: The ten most important features for predicting a sample as having a lower proportion of female students for the logistic regression model trained and evaluated on English data.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
båda	enda	-
informationen	jämfört	-
vilken	extremt	-
tempo	använda	-
personligen	går	-
särskilt	högre	-
mot	konstigt	-
instruktioner	personer	-
använda	mot	-
sett	personligen	-

Table A.38: The ten most important features for predicting a sample as having a lower proportion of female students for the SVC model trained and evaluated on Swedish data. Note that there are no words for single comment prediction as these experiments took too much time.

Important features		
Multi Comment Prediction		
Full text	Undersampled	Included POS-tags
assistants	smaller	-
far	problems	-
harder	team	-
tas	need	-
prepared	probably	-
come	en	-
similar	ta	-
members	assistants	-
needs	actual	-
stuff	tas	-

Table A.39: The ten most important features for predicting a sample as having a lower proportion of female students for the SVC model trained and evaluated on English data. Note that there are no words for single comment prediction as these experiments took too much time.