# Use Tags and Comments to Understand Gender Difference in the Online Evaluation from RateMyProfessors.com

**Ting Liu**[1,2], **Yuwei Chen**[1], **Logan Brandt**[1]
[1]Computer Science, Siena College, Loudonville, NY, USA
[2]Data Science Program, Siena College, Loudonville, NY, USA

**Abstract –** *In this paper, we presented a novel method to study the difference of student ratings towards female and male instructors. We collected all female and male instructors (over 10,000) who have more than 10 ratings with long comments (>= 5 words). The analysis on instructors' overall quality and tags distribution suggests that students like male instructors more than female instructors. After adding discipline as another dimension, the students from different discipline can have very different rating behaviors. Even though female instructors only have better rating in Math, but they achieved similar performance in some other science disciplines, such as Chemistry and Physics. One the other hand, female instructors rated much lower comparing with male instructor in some liberal arts disciplines, such as English and Music, and business-related disciplines. We are working on comments and will report the results in the near future. With the results, we hope to find out the reason(s) behind the difference.*

**Keywords:** Gender, Evaluation, Tag, Discipline, Comment

**Type: Short Paper**

## 1 Introduction

Students Evaluation of Teaching (SET) has been widely used as an evaluation scheme in higher education since 1920s [19]. At the end of a semester, students will be asked to anonymously answer a list of carefully designed questions (in Likert scale) from the evaluation with their comments [1]. The evaluations in turn will be, aggregated, studied, and used for teaching quality improvement as well as promotion and tenure decisions. SET evaluations usually were administered by educational institutions, but in recent decades, online evaluation websites like RateMyProfessors (RMP) provide students a public place to share their opinions that can be used as reference for other students.

### 1.1 About RateMyProfessors

Since 1999, RMP has collected more than 19 million reviews, 1.7 million instructors' at over 7500 higher education institutions across the United States, Canada, and United Kingdom. This is so far the largest online student evaluation system that offers much bigger dataset freely comparing with those from institutions, which usually are inaccessible because of confidential information. In addition, researches ([4], [8], [10], [11]) showed that peer recommendations about courses and professors are very important criteria for students to decide what course to take. Therefore, RMP becomes student favorite website that they can refer, which is also supported by the surveys from ([6], [24]), which showed that 80% college students either 'almost always' or 'sometime' visit the site when choosing courses. Considering both the size of the data and its strong impact on students' decisions on picking courses and instructors, it is worth well for educators and researchers to dive in deeply for analysis.

Students on RMP rate their Professors on three criteria: helpfulness, clarity, and easiness using 5-point scale rating (1-5)[1] on RMP. The average of helpfulness and clarity scores become professors' overall quality scores. In 2016, RMP replaced helpfulness and clarity with overall rating. Therefore, our data don't have helpfulness and clarity ratings since we collected our data in 2018 Spring. RMP also requires students input comments (maximum 350 characters) to be more specific about their ratings. In addition, students can choose up to 3 tags, such as "CARING", "GIVE GOOD FEEDBACK", from a list of predefined 20 tags.

To answer the question whether students' reviews from RMP are validate, Silva [22] and Timmerman [26] compared ratings by students on RMP with their SET's evaluations and found there are strong correlation between them consistent.

[1] 5 - Awesome, 4 - Good, 3 - Average, 2 - Poor, 1 - Awful

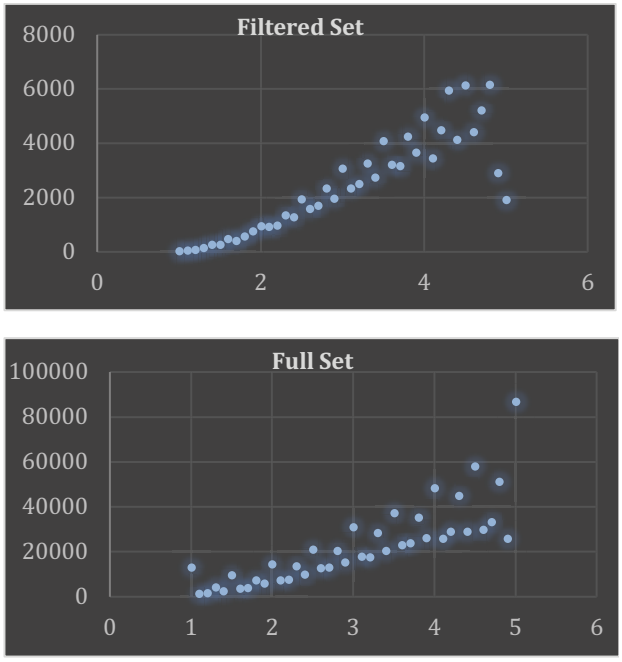


Figure 1. Rating Distribution of Full Set and Filtered Set

Otto et. al [15] even argued that reviews from RPM may a useful supplement to traditional SETs.

### 1.2 Gender issue

One big uncertain question arose from prior research on SET evaluation is whether instructor's gender can affect the evaluation and if it does matter, how strong it is and what is the reason behind? Some research ([7], [14], [24], [20]) can't find statistically significant difference in rating of overall quality. However, some other research has very different conclusions. Sandler et. al [20] found that male instructor are evaluated more positively than female instructor. Basow ([2]) found that the interaction between professor gender, student gender, and discipline has more significantly impact on the effect of gender than those researches only focusing on instructors' gender. Their results showed men's overall quality/effectiveness is 0.07/0.24 higher than women's. Even though the limitations of the RMP dataset wouldn't be able to involve students' gender as part of research, the recent research ([21]) on 190,000 instructors[2] pulled from RMP reveals the similar conclusion, "When looking at RateMyProfessors as a whole, the effect of a professor's gender on rating criteria is small but statistically significant".

To understand gender issue from different aspects, we propose a novel idea that use tags and students' comments as new dimensions for gender analysis. We employ natural language process technique to extract both topics from students comments and also sentiment opinions (positive/negative/neutral) towards the topics. Our goal is to find out 1) whether students' comments can reveal the reason(s) of gender issues in students' ratings. 2)Whether students from different disciplines have different rating patterns and how they affect students' evaluation towards instructors in different genders.

## 2 Data

We finished our data collection in 2018 spring. Our crawler pulled 1.5 million instructors who have at least one rating. Some of the ratings have either no comment or very short comments, from which no useful information can be extracted. Therefore, we create a filter that only keeps the instructors who have at least one comment with a minimum length of 5 words. We arbitrarily set words threshold to 5 because according to our observations, students usually use more than 5 words (including 5) to form a complete sentence. Our assumption is that students who are willing to put long comments (at least one full sentence) are more "serious" about their ratings and therefore such ratings are more reliable.

### 2.1 Gender classification

The method in which these professors were classified for gender is through the use of the comments. Each comment will be searched through to find pronouns or gender indicators that would associate that professor with a gender. For example, keywords such as Mr., he, him, or guy would be associated a male instructor while keywords such as she, her, or Mrs. would be associated with a female instructor. A counter for both genders will be populated for every pronoun we found in the professor's comments. These counters will then be compared to each other to classify the professor into a gender. This process of filtering found 399K female instructors and 538K male instructors from the raw data. This filter can achieve high accuracy when instructors have multiple long comments since it is high possible either pronoun or gender indicators will be used. It is also very unlikely that the opposite genders pronouns will be mentioned in the professor's comments by studying a small set with 125 random selected professors. In addition, the data set we're working on has instructors with at least 10 long comments (explained in next section). Therefore, we have strong confidence that this filter finds a good set of female and male instructors.

### 2.2 Gender classification

Some of prior research ([8], [24], [21]) built their data set by adding instructors with minimum of 20 ratings. For our dataset, we set two thresholds, 10 minimum ratings and each rating contains comment at least 5 words. After the thresholds applied, our data size reduces to 46,602 female instructors and 54,253 male instructors. Figure 1 shows rating distribution charts of full and filter sets. The numbers of instructor with rating 1 (awful) and 5 (awesome) are abnormally high comparing their neighbors, which means

[2] Only instructors have at least 20 ratings were chosen

Table 1. Overall Instructors ratings' distributions based on gender

| | Good (3.5 - 5) | Average (2.5 - 3.4) | Poor (1 - 2.4) |
|---|---|---|---|
| # of Femal Instructors | 31107 | 11235 | 4260 |
| # of Male Instructors | 37416 | 12452 | 4385 |
| % of Femal Instructors | 67% | 24% | 9% |
| % of Male Instructors | 69% | 23% | 8% |

that without enough number of ratings' support, the instructor's overall rating is not reliable. Therefore, all analysis reported in this paper focus on filtered dataset unless otherwise stated.

# 3 Gender study

The overall quality of an instructor on RMP is the average of the ratings from all students to this instructor. RMP breaks down instructors' quality into three categories: Good (3.5 to 5), Average (2.5 to 3.4), and Poor (1 to 2.4). Table 1 shows numbers and percentages of female and male instructors fall into the three categories. The majority part of instructors from both genders falls into good category and only 2% more male instructors are evaluated as good instructor comparing with female instructors. The average scores of overall qualities from male and female instructor are 3.84 and 3.79, which are higher comparing with those (Male: 3.71, Female: 3.66) reported in [21]. It could suggest that students giving long comments intent to rate their instructors higher that those without writing comments.

## 3.1 Top Ranked Tags

RMP allows students to pick up to 3 tags, which are most applicable to the instructor, for one rating. We are very curious what tags instructors are getting and how they distribute in the instructors' rating categories. To do so, we calculated the percentage of instructors having each tag given in the three categories. Table 2 displays 5 most frequent tags assigned to Male and Female instructors. Female and Male Average and Poor tags are almost identical

Table 2 Top 5 tags from each rating category of both Male and Female instructors

| Gender and Rating | Tag | Percent |
|---|---|---|
| | GIVES GOOD FEEDBACK | 12% |
| | CARING | 10% |
| Female Good | RESPECTED | 9% |
| | PARTICIPATION MATTERS | 8% |
| | TOUGH GRADER | 7% |
| | TOUGH GRADER | 19% |
| | SKIP CLASS? YOU WON'T PASS. | 11% |
| Female Average | GET READY TO READ | 10% |
| | LOTS OF HOMEWORK | 9% |
| | PARTICIPATION MATTERS | 8% |
| | TOUGH GRADER | 26% |
| | SKIP CLASS? YOU WON'T PASS. | 12% |
| Female Poor | GET READY TO READ | 12% |
| | LOTS OF HOMEWORK | 11% |
| | LECTURE HEAVY | 8% |
| | RESPECTED | 10% |
| | GIVES GOOD FEEDBACK | 10% |
| Male Good | CARING | 9% |
| | HILARIOUS | 8% |
| | SKIP CLASS? YOU WON'T PASS. | 7% |
| | TOUGH GRADER | 19% |
| | SKIP CLASS? YOU WON'T PASS. | 11% |
| Male Average | GET READY TO READ | 10% |
| | LOTS OF HOMEWORK | 9% |
| | LECTURE HEAVY | 8% |
| | TOUGH GRADER | 27% |
| | SKIP CLASS? YOU WON'T PASS. | 11% |
| Male Poor | GET READY TO READ | 11% |
| | LOTS OF HOMEWORK | 11% |
| | LECTURE HEAVY | 10% |

with only exception that Female Average has "Participation Matters" but Male Average has "Lecture Heavy". Another interesting finding is that the frequency of "Tough Grader" is significantly higher than other tags, which indicate that "Grading procedure" is one of the most concerned topics by students and instructors who get this tag can be easily downgraded. For "Female Good" and "Male Good", the top ranked tags and their orders are quite different, which could mean that female and male instructors are good at different teaching styles both preferred by students. For example, 12% good female instructors have "Gives Good Feedback" while 10% good male instructors having this tag. Another example, 8% of good male instructors have "Hilarious" but only 5% of good female instructors have this tag (ranks 13). In

| | Languages | CompSci | Chemistry | Anthropology | Economics | Mathmatics | Physics | Phsycology |
|---|---|---|---|---|---|---|---|---|
| # of Good Female Instructors | 75.2% | 54.0% | 56.5% | 69.3% | 55.8% | 63.7% | 48.4% | 76.3% |
| # of Average Female Instructors | 18.3% | 28.5% | 29.4% | 21.6% | 31.7% | 25.6% | 33.1% | 18.6% |
| # of Poor Female Instructors | 6.5% | 17.5% | 14.1% | 9.1% | 12.5% | 10.7% | 18.5% | 5.1% |
| # of Good Male Instructors | 78.6% | 56.9% | 55.4% | 72.2% | 56.6% | 60.5% | 50.4% | 77.4% |
| # of Average Male Instructors | 17.8% | 28.7% | 31.1% | 21.1% | 31.2% | 27.0% | 33.1% | 18.3% |
| # of Poor Male Instructors | 3.7% | 14.5% | 13.5% | 6.6% | 12.2% | 12.5% | 16.4% | 4.3% |

Table 3.1 Instructor ratings' distributions in top 20 most popular disciplines 1

| | Management | Political Science | Biology | Business | Communications | Education | English | History | Music | Philosophy | Sociology | Accounting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # of Good Female Instructors | 61.9% | 67.2% | 61.9% | 63.6% | 71.3% | 63.9% | 70.5% | 67.3% | 63.7% | 67.3% | 68.1% | 59.8% |
| # of Average Female Instructors | 28.7% | 26.0% | 27.9% | 25.9% | 20.6% | 24.3% | 21.2% | 25.7% | 27.1% | 25.1% | 24.1% | 30.9% |
| # of Poor Female Instructors | 9.4% | 6.7% | 10.2% | 10.6% | 8.1% | 11.8% | 8.3% | 7.0% | 9.2% | 7.6% | 7.8% | 9.4% |
| # of Good Male Instructors | 67.9% | 76.5% | 66.6% | 69.4% | 77.2% | 77.6% | 77.5% | 76.2% | 76.7% | 72.5% | 72.8% | 65.3% |
| # of Average Male Instructors | 23.6% | 18.7% | 26.1% | 21.6% | 18.6% | 16.8% | 17.4% | 18.9% | 19.8% | 22.5% | 20.4% | 23.9% |
| # of Poor Male Instructors | 8.5% | 4.8% | 7.3% | 9.0% | 8.0% | 5.6% | 5.1% | 5.0% | 3.5% | 5.0% | 6.8% | 10.7% |

Table 3.2 Instructor ratings' distributions in top 20 most popular disciplines 2

addition, "Tough Grader" ranks 5th in female good instructors' tag list. This can be an evidence for the claim that women are evaluated first as a woman, then as a professor ([3], [23]) since these two tags are somewhat against woman's nature.

### 3.2 Ratings in Different Disciplines

To study the difference between female and male instructors' ratings, we clustered the instructors according to their disciplines. To improve the clustering accuracy, we built a list of variant names for each discipline, e.g. for Physics, we have Physics Department, Department of Physics, the Physics Astronomy Department, etc. For this paper, we focused on the top 20 ranked disciplines with most instructors because of page limits. Table 3.1 displays all disciplines that either female instructors did better or no significant difference (less than 4%) comparing with male instructors. Surprisingly, Math is only the discipline that Female instructors did better and in other science department, like Chemistry, Physics, and Computer Science, female instructors actually have similar performance. On the other hand, female instructors get significant lower ratings (Table 3.2) from some liberal arts disciplines such as English (7% difference for Good category), History (8.9% difference), and Music (13%). This finding came out a contradict conclusion comparing with prior research ([2]), which claim women receive less evaluation ratings in traditionally masculine disciplines (e.g., Physics) but more evaluation ratings in traditionally feminine disciplines (e.g., English).

To find more reasons about the difference, we also looked into the 5 top tags from the 20 disciplines, but because of space limits, we are not able to display them, but we can put them as appendix later on if paper get accepted. By comparing with tags by Math students and by Music students (Music is picked because it has biggest difference between Male Good and Female Good ratings percentage), we can see some different evaluation behavior. Good music and math instructors have identical top 5 tags. The top 5 tags from average Female instructors in Math have two positive tags, "Caring" and "Clear Grading Criteria", but no positive tags for average music instructors. There could be two reasons, 1) the culture from Math discipline is friendlier to female instructor, 2) The teaching style of Match female instructors is friendlier and more acceptable by students. In the next step, we will look into the comments to see if any pattern(s) or model(s) can be found to explain such difference.

## 4 Conclusion and Future works

In this paper, we reported a novel research that introduces tags in analyzing the difference of students' ratings towards male and female instructors across disciplines. The top 5 Tags collected from overall data suggest that students like male instructors in a different way comparing with their ratings to female instructors, for example, the most favorable tag, "giving good feedback", from good female instructor indicates that students really love this teaching method. On the other hand, we found that students' evaluation behavior from different disciplines can be very different. The comments in the rating could reveal the reasons behind it, which is our ongoing research. We currently focus on topic extraction and target-based sentiment analysis.

In recent years, more and more researchers applied sentiment analysis technique to extract students' opinions out of the comments ([12], [13]). Their research focuses on predicting the overall positive/negative opinions of the comments. Some other research ([17]) also clusters the opinion words (such as "good", "nice", "difficult", …) to visualize the student feelings. However, students already expressed their overall ratings clearly through scaled questions. Such predictions won't add extra benefit for evaluation analysis. Therefore, our interest is to extract anchors from student's comments that can explain their ratings. For example, one student gave "average" rating to the professor. By looking at the comment (Figure 2), we can see that this student liked the professor's practice exam, but didn't like the teaching style. This comment contains at least two different opinions toward three topics. If we can successfully extract the information, it will help us understand student evaluations much deeper.

For topic extraction, we are using dependency parser (from Stanford University) and for sentiment detection, we decide to adapt the approach using tree structured LSTM model ([25]) since it beats the performance of state of art systems. With the extracted topics and their sentiments, we will cluster them according to their semantics. We hope to find good explanations by looking into these clusters and report them in the near future.

## 5 References

[1] Philip C. Abrami, Sylvia d'Apollonia, and Steven Rosenfield. 2007. "The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not." In The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, edited by Raymond P. Perry and John C. Smart, 385–456. Dordrecht: Springer Netherlands.

*His practice exams are extremely helpful. … The class is extremely strict and boring.*

Figure 2 Comment Examples

[2] Susan A. Basow. 1995. Student evaluations of college professors: When gender matters. Journal of Educational Psychology 87: 656–65.

[3] Susan A. Basow. 2000. Best and worst professors: Gender patterns in students' choices. Sex Roles, 43: 407–17.

[4] Michael J. Beatty &Christopher J. Zahn 1990. Are student ratings of communication instructors due to "easy" grading practices? An analysis of teacher credibility and student-reported performance levels. Communication Education, 39, 275–291.

[5] Sheila Bennett. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. Journal of Educational Psychology, 74, 170–179.

[6] Elizabeth Davison and Jammie Price. 2009. "How Do We Rate? An Evaluation of Online Student Evaluations." Assessment & Evaluation in Higher Education 34 (1): 51–65.

[7] Kenneth A. Feldman. 1993. College students' views of male and female college teachers, part I: Evidence from students' evaluations of their classroom teachers. Research in Higher Education 34: 151211.

[8] James Felton, John B. Mitchell, and Michael Stinson. 2004. Web-Based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. Assessment & Evaluation in Higher Education, 29, 91–108.

[9] James Felton, Peter T. Koper, John B. Mitchell, and Michael Stinson. 2008. Attractiveness, easiness, and other issues: Student evaluations of professors on RateMyProfessors.com. Evaluation in Higher Education 33: 4561.

[10] Roxanna Harlow. 2003. "Race doesn't matter, but. . . ": The effect of race on professors' experiences and emotion management in the undergraduate college classroom. Social Psychology Quarterly, 66, 348 – 363.

[11] Katherine Grace Hendrix. 1997. Student perceptions of verbal and nonverbal cues leading to images of black and white professor credibility. Howard Journal of Communications, 8, 251–273.

[12] Jagtap, B. and V. Dhotre, (2014) SVM & HMM based hybrid approach of sentiment analysis for teacher feedback assessment International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. 3, no. 3, pp. 229–232, 2014.

[13] Kaewyong, P., Sukprasert, A., Salim, N., and Phang, A., The possibility of students' comments automatic interpret using lexicon based sentiment analysis to teacher evaluation. Conference: The 3rd International Conference on Artificial Intelligence and Computer Science 2015, At Penang, MALAYSIA, 2015.

[14] JoAnn Miller and Marilyn Chamberlin. 2000. Women are teachers, men are professors: A study of student perceptions. Teaching Sociology 28: 28398

[15] Lincoln Mullen. 2015. Gender: Predict Gender from Names Using Historical Data. https://github.com/ropensci/gender.

[16] James Otto, Douglas A. Sanford Jr, and Douglas N. Ross. 2008. Does Ratemyprofessor.com Really Rate My Professor? Assessment & Evaluation in Higher Education 33 (4): 355–368.

[17] Rajput, Q., S. Haider, and S. Ghani (2016) Lexicon-Based Sentiment Analysis of Teachers' Evaluation Applied Computational Intelligence and Soft Computing Volume 2016, Article ID 2385429, 12 pages 2016

[18] Landon D. Reid. 2010. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. Journal of Diversity in Higher Education, 3, 137-152.

[19] Remmers, H. H. (1927). The Purdue Rating Scale for Instructors. Educational Administration and Supervision, 6, 399-406.

[20] Bernice Resnick Sandler, Lisa A. Silverberg, and Roberta M. Hall. 1996. The chilly classroom climate: A guide to improve the education of women. Washington, DC: National Association for Women in Education.

[21] Andrew S Rosen. 2018. Correlations, Trends, and Potential Biases among Publicly Accessible Web-Based Student Evaluations of Teaching. Assessment & Evaluation in Higher Education, v43 n1 p31-44 2018

[22] Kathleen M. Silva, Francisco J. Silva, Megan A. Quinn, Jill N. Draper, Kimberly R. Cover, Alison A. Munoff. 2008. Rate my professor: Online evaluations of psychology instructors. Teaching of Psychology, 35, 71– 80.

[23] Joey Sprague and Kelley Massoni. 2005. Student evaluations and gendered expectations: What we can't count can hurt us. Sex Roles, 53: 779–93.

[24] Jenny M. Stuber, Amanda Watsonb, Adam Carlea and Kristin Staggsa. (2009). Gender expectations and on-line evaluations of teaching: Evidence from RateMyProfessors.com. Teaching in Higher Education, 14, 387–399. doi:10.1080/13562510903050137

[25] Tai, K. S., R. Socher, and C. D. Manning. 2015 Improved semantic representations from tree-structured

long short-term memory networks. In Proceedings of 53rd annual meeting of the Association for Computational Linguistics (ACL).

[26] Thomas Timmerman. 2008. On the validity of RateMyProfessors.com. Journal of Education for Business, 84, 55– 61.