What Makes A Good Course and Professor: Through The Lens Of RateMyProfessor Website

Abstract. What makes a good professor? Do the courses a professor teach and their difficulty levels influence a professor's overall evaluation? What topics do the students mostly care about when evaluating a professor and a course? What emotions are pervasive in student reviews? Do universities from different locations, and Carnegie classification levels exhibit different rating? Using a large compilation of the RateMyProfessor (RMP) dataset, we proceed to explore answers to these questions in this paper by performing a comprehensive data-driven analysis. We first explore the distribution of good professors for university locations and classifications and understand whether a professor's perceived difficulty level affects their RMP rating. Then, for a granular review level investigation, we analyze the relationship between a course's difficulty level and student ratings and investigate the main topics and emotions prevalent in student reviews.

Keywords: data analysis, rate my professor, social network analysis

1 Introduction

Evaluation of an instructor's teaching is challenging due to its subjective nature. Usually, universities provide students with an evaluation form that includes both numeral ratings to rate predetermined teaching qualities and text fields that ask for open-ended comments regarding the professor's teaching performance. This feedback collection from students is then consolidated to provide a generic assessment of a particular instructor's teaching quality by the university — which is usually never available to prospective or current students. To this aim, Rate My Professor(RMP) was introduced in 1999, where students can anonymously rate different aspects of the course and the instructor[25]. RMP is the largest online destination for professor ratings. The site includes 8,000+ schools, 1.7 million professors, and over 19 million ratings[25]. In [3], researchers found that 83 percent of surveyed students have visited RMP, 36 percent have rated an instructor on the Web site, and 71 percent have avoided taking an instructor based on their ratings. Furthermore, 47 percent of respondents believed that RMP ratings are more representative of instructors' performance than official student evaluations[3]. These findings suggest that students have confidence in RMP ratings and use the site to make academic decisions.

In this paper, we compile a dataset of $481,712$ reviews for $90,466$ across 72 universities in the US. We choose the university list carefully across two factors: Carnegie university classifications[24] (R1, R2, liberal arts) and US locations

(East, West, Midwest, and South). Then, we investigate the professor-level data to understand factors influencing a professor's overall rating in RMP. Next, we explore professors' rating distribution for different university locations, classifications, and difficulty levels. Finally, we analyze every review individually to answer the following questions: What sentiments/emotions do students usually express the most when reviewing an instructor? What factors about a particular course do the students deem essential? What topics do students talk about in their professor reviews? When do students usually post the reviews? Is there any relation between the difficulty of a course or the level of a course (for example, 162 is a level 1 course code, 260 is a level 2 course code) and the rating given?

We organize the rest of the paper as follows. In Section 2 we delineate the related works in this topic. Next, in Section 3, we describe our dataset collection methodology and possible ethical concerns. In section 4 and 5, we analyze the professor-level metadata and review-level metadata to understand better what makes a professor/course a good professor/course. Finally, in Section 6, we conclude the paper by summarizing the findings and exploring future research directions.

2 Related Works

RMP evaluations in the field of Sociology and Education. Previous researches have found significant correlations between official student evaluations and RMP evaluations of a professor [4, 5, 8, 15]. A more recent study [9] however, found that RMP evaluations possess an inherent negative bias. A professor's overall assessment also depends on the relative ease of the courses offered, and the professor's physical attractiveness [6, 7]. Studies have also shown that professors in Science and Engineering schools tend to have lower ratings than their Arts and Humanities counterparts [17]. Other factors impacting a professor's rating in RMP include grades received in the course [11] and gender [13, 17].

Computational analysis of RMP evaluations. Computational analyses of RMP evaluations have been a focus of some recent studies. One recent study [10] analyzed 283 instructors' evaluations from Pennsylvania to understand attributes that are valued by students when assessing a professor. The study revealed that being knowledgeable and approachable are two of the most sought-after traits in a good professor. In addition, studies have analyzed the inherent bias of RMP evaluations compared to the official evaluations [9]. In that study, the University of South Florida's official student evaluation was used as the ground truth data to compare the corresponding RMP evaluations. The study also investigated the influences of STEM courses and the genders of professors in their assessments. Upon comparing the official and RMP evaluations, the authors found that strong/negative opinions are pervasive in RMP. Professors who teach relatively complex topics are also on the receiving end of much harsher assessments on RMP. The authors posited that — to explain the negative bias of RMP — students who loved the professor's course usually express that emotion in emails, or in-person meetings [9]; whereas negative assessments are more done anonymously through RMP. The study failed to find any correlation between gender and RMP evaluations, contrary to [17] but did find that STEM professors tend to get lower ratings.

3 Dataset Description

In this section we give an overview of the RMP dataset and discuss the ethical concerns with our dataset collection.

Region	Classification	Universities
East	R1	Harvard, Temple University, University of Maryland
	R2	University of Rhode Island, Masachusetts Boston, Georgia Southern
	Liberal Arts	Colgate, Bucknell, Washington Lee
Midwest	R1	Michigan State, Iowa State, University of Chicago
	R2	Missouri University, Illinois State, Illinois Institute
	Liberal Arts	Grand valley State, Augustana, Dennison
South	R1	Texas Tech, Houston, Florida University
	R2	Florida Institute, Louisiana Tech, Texas Southwestern
	Liberal Arts	Austin College, Eckerd College, Ave Maria University
West	R1	Arizona State, Colorado State, University of Arizona
	R2	University of Idaho, Utah state, Colorado School of Mines
	Liberal Arts	Evergreen State, Harvey Mudd, Lewis Clark

Table 1: Examples of universities' RMP data collected

3.1 Dataset Statistics

For data collection, we consider a total of 72 universities across four geographic regions in the US (east, west, south, midwest; 18 for each) and three Carnegie classifications of higher institutions [24] (R1, R2, liberal arts; 24 for each). We leverage python's selenium [21] and beautiful soup [22] to crawl the RMP website during September, 2021—collecting all evaluations of all the professors associated with each university. Table 1 shows some examples of names of universities for each category we consider for our study. We collect two levels of metadata from RMP: professor level and evaluation level. For each professor, we collect the affiliated university, professor's overall rating, associated department/faculty/school name, and the total number of student evaluations. For each assessment, we collect the course-id, difficulty and quality rating given, evaluation comment, number of people who found the review helpful or not, and associated keyphrase-tags such as 'would-take-again', 'skip-class-you-wont-pass' etc. In total, we collect $481,712$ evaluations of 90466 professors. Overall rating of a professor, difficulty, and quality score for the assessments can have a value between [1,5]; 1 being the lowest and 5 being the highest rating. RMP calculates a professor's overall rating by averaging the quality rating given for each evaluation for that particular professor. Table 2 shows the overall distribution of professors and evaluations in the dataset.

Region	Classification	Professors	Evaluations
East	R1	13580	57578
	R2	9765	52764
	Liberal Arts	1304	11907
Midwest	R1	14638	47403
	R2	5196	28317
	Liberal Arts	2865	18582
South	R1	18522	86385
	R2	5152	42863
	Liberal Arts	3281	18467
West	R1	11923	78290
	R2	6830	25695
	Liberal Arts	1328	13461

Table 2: RMP Dataset Statistics

3.2 Ethical Considerations

Crawling and analyzing social media data entails serious ethical implications. However, in this research, we only concern ourselves with publicly available RMP website data and refrain from interacting with human subjects. Therefore, our study did not fall under the purview of IRB human subjects research. Moreover, we only perform and publish aggregated analyses of the collected data—conforming to the standard ethics guidelines to protect users [16].

4 Professor Level Analysis

Figure 1(R) shows an example of a professor's RMP page. It shows the professor's overall rating, the total number of ratings given, difficulty level, and top tags associated with the professor, extracted from the student's reviews. In this section, we try to analyze the professor-level data collected — specified in section 3.1 — to mine insights into what makes a good professor. To this end, we first investigate the overall rating distribution of the 90466 professors' ratings collected in our dataset. Figure 1 (L) shows the Cumulative Distribution Function (CDF) distribution of all the professors' ratings. As the figure shows, in our dataset, a significant portion of professors (40%) have ratings that are above 4.5. Moreover, less than 20% professors have less than ratings of less than 2.5. This paper follows the good and bad professor classifications specified in [1], labeling professors with ratings less than 2.5 as bad and more than 3.5 as good professors. According to this classification, from Figure 1, we can see that, in our dataset, *we have less than 20% bad professors and more than 60% good professors.*

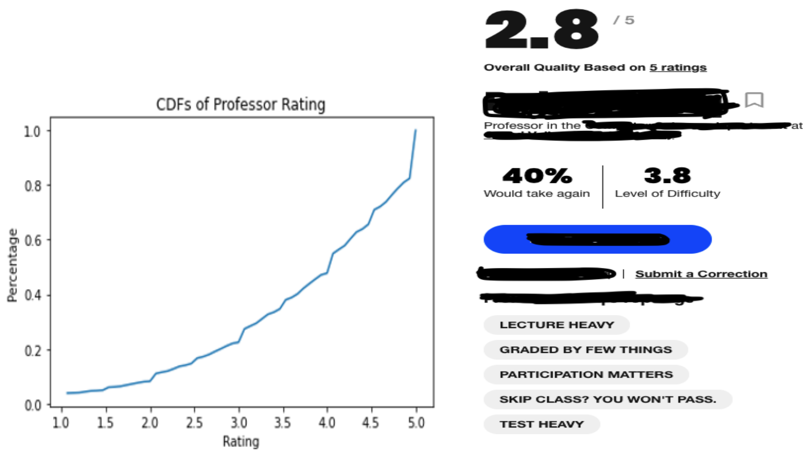


Fig. 1: (L) CDF of professor ratings (R) Example of a professor rating in Rate My Professor

Uni	% of Good Professors	% of Bad Professors
East	62	14.93
MidWest	60.39	15.60
South	63.74	13.75
West	61.70	14.61
R1	61.94	15.04
R2	60.26	15.79
Liberal Arts	**66.44**	**11.31**

Table 3: Percentage of good and bad professors based on location and carnegie classification respectively

Table 3 shows the percentage of good and bad professors across different regions and different university classifications separately. It is seen from the table that *good and bad professor distribution for universities across different regions and types usually follows the same percentage distribution* as the overall professor rating distribution in figure 1 — around 60% and less than 20% good and bad professors respectively. However, *Liberal Arts universities have the highest and lowest number of good and bad professors*, respectively: 66.44%, and 11.31%, according to the table. To further understand the distribution of good and bad

Region	Classification	% of Good	% of Bad
East	R1	62.22	15.09
	R2	60.04	15.73
	Liberal Arts	**71.30**	**7.7**
Midwest	R1	60.05	15.48
	R2	60.63	16.08
	Liberal Arts	61.61	15.28
South	R1	63.17	14.15
	R2	59.76	16.21
	Liberal Arts	66.34	14.3
West	R1	61.97	15.29
	R2	59.82	14.74
	Liberal Arts	**68.36**	**8.15**

Table 4: Percentage of good and bad professors based on location and carnegie classification combined

professors, Table 4 presents the numbers with location and university classification together. We find that, first, in the midwest region, R1, R2, and liberal arts universities have almost similar good and bad professors distribution, unlike others where Liberal Arts universities have a significantly higher number of good professors than R1 and R2 universities. Second, the East and West region's liberal arts universities have a significantly high percentage of good professors compared to others, 71.30% and 68.36% respectively.

We now identify the top tags associated with the good and bad professors. Table 5 shows the top 10 tags associated with the good and bad professors in our RMP dataset. According to our data, *human qualities such as "Caring", "Respected" and "hilarious" is more pervasive among all good professors*, accounting for 62.63%, 50.94%, and 33.81% respectively. Moreover, "Giving good feedback", "Amazing Lectures" and "Accessibility outside class" is also more frequent. These findings reveal that students appreciate "teaching" qualities such as quality of lectures, good feedback process, and overall helpfulness and human qualities such as caring for the students and making the lecture enjoyable. On the other hand, bad professors have frequent tags such as tough grader, lecture heavy, lots of homework, and test heavy.

Next, we analyze how much the difficulty level of a professor influences their overall rating. We extract the difficulty rating of a professor - as shown in Figure 1(R) - from the RMP website. The difficulty scores of the professors range from 1 to 5. We prescribe, in total, five levels of difficulty- L1, L2, L3, L4, and L5. Table 6 denotes the percentage of good and bad professors across five difficulty levels, along with the ranges for the difficulty levels. It is apparent that a *disproportionate percentage of L1, L2, and L3 professors are good professors — 79.95%, 82.71%, and 72.50%, respectively.* On the other hand, only 20.31 percent of L5 professors were labeled as good professors — ascertaining the fact that the difficulty levels of a professor play a significant role in the final evaluation of the professor, corroborating previous studies [9].

Tags	% of Good Professors	Tags	% of Bad Professors
Caring	62.63	Tough Grader	78.26
Gives Good Feedback	60.61	Skip Class? You won't pass	51.58
Respected	50.94	Lecture Heavy	46.96
Accessible outside class	40.16	Lots of Homework	46.23
Amazing Lectures	39.47	Test Heavy	40.57
Participation Matters	35.19	Get Ready to Read	40.44
Hilarious	33.81	Graded by Few Things	33.9
Clear Grading Criteria	31.28	Participation Matters	22.73
Inspirational	28.93	Group Projects	17.12
Get Ready to Read	27.50	Beware of Pop Quizzes	13.60

Table 5: Top tags for good and bad professors

Difficulty Level	Total	% of Good Professors	% of Bad Professors
L1 (0-1)	2928	79.95	10
L2 (1-2)	11827	82.71	4.6
L3 (2-3)	24347	72.50	7.3
L4 (3-4)	22039	49.69	17.98
L5 (4-5)	6438	20.31	51.77

Table 6: Percentage of good and bad professors based on Difficulty Level

5 Review Level Analysis

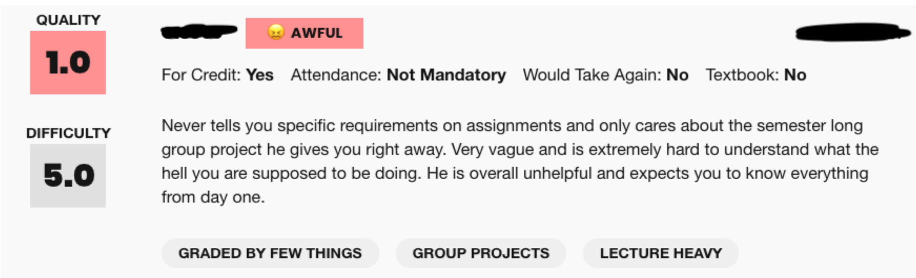


Fig. 2: Example of a student review in RMP

In this section, we investigate the 481,712 reviews in our dataset — mentioned in Section 3 — to understand better what drives a student to evaluate a particular course as a good one. Figure 2 depicts an example of a student review for a course in RMP. In a review, the student can assign two scores, from 1 to 5, for two attributes — quality and difficulty of the course. In addition to the review text, the students have two ways to provide additional information to a review. First, at the top end, the student can input information about the course in general; examples include if the course was "taken for credit" or "if attendance was mandatory". At the bottom, the student can also select relevant tags for the course; examples include "group projects", which means this course had a group project requirement.

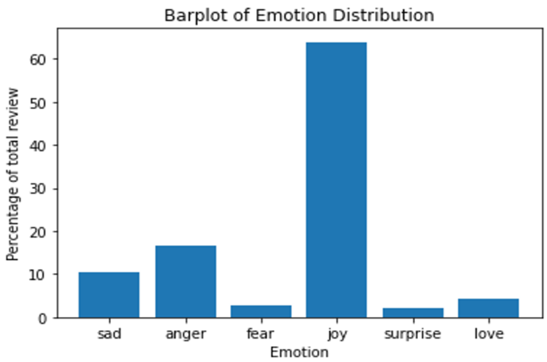


Fig. 3: Emotion Distribution of Student Reviews in RMP

We now explore the distribution of emotions and sentiments expressed in the student reviews. For this task, we employ DistilBERT, a distilled version of BERT, which is smaller, faster, cheaper, and lighter. DistilBERT model is inspired by the Knowledge distillation approach — a compression technique to train a small model to reproduce a larger model's behavior [18]. Using this technique reduces the size of a BERT model by 40% and faster by 60% faster [18] while keeping 97% of its language capabilities. DistilBERT has been one of the most efficient and accurate models for recognizing emotions in texts [14], [12]. The model was trained with Twitter emotion dataset [19] that was tagged with six major human emotions: anger, fear, joy, love, sadness, and surprise [20]. Let $E \in \{anger, fear, joy, love, sadness, surprise\}$, the set of six emotions. The DistilBERT emotion model outputs, for every review text $r \in R$ in our dataset, — R is the set of all review texts — a probability score $S_r(e)$ such that:

$$\sum_{e \in E} S_r(e) = 1 \tag{1}$$

We assign an review r, an emotion label e_{final}, where,

$$S_r(e_{final}) > \forall_{e \in E - \{e_{final}\}} S_r(e) \tag{2}$$

Figure 3 shows the distribution of the six aforementioned emotion labels in our RMP student reviews dataset. As it can be seen from the figure, *"joy" is the most pervasive emotion, accounting for more than* 60% *of the student reviews in our dataset. We also see those negative emotions such as "anger", "sadness", and* *"fear" account for less than* 40% *of the student reviews.* Next, for a more granular investigation, we analyze the outlier good and bad student reviews based on emotion scores calculated in Equations 1 and 2. Let $E_{pos} \in \{joy, love, surprise\}$ and $E_{neg} \in \{fear, anger, sadness\}$ be the set of positive and negative emotions. We define, for a list of real numbers L and a real number v, an outlier function $OF(L, v)$ such that:

$$OF(L, v) = \begin{cases} 1, & \text{if } v \geq L_{95thpercentile} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Finally, we call S^e as the set of a particular emotion's probability scores across all reviews in our dataset, as calculated by DistilBERT, such that:

$$S^e = \{S_r(e) \mid \forall r \in R\} \tag{4}$$

Now we proceed to define good and bad outliers, *GO* and *BO* respectively, as follows:

$$\begin{aligned} GO &= \{r \mid \sum_{\forall e \in E_{pos}} OF(S^e, S_r(e)) \geq 1\}, \forall r \in R \\ BO &= \{r \mid \sum_{\forall e \in E_{neg}} OF(S^e, S_r(e)) \geq 1\}, \forall r \in R \end{aligned} \tag{5}$$

Now, we take the *GO* and *BO* reviews and list the frequently occurring tags to explore the topics that make students exhibit such extreme emotions — both positive and negative — in their reviews. Table 7 lists the top tags and the percentage of outlier reviews they appear in. From the list, we see that *human attributes such as care, respect, hilarity, helpfulness, and accessibility, together with good feedback-giving, are prevalent in the positive emotion outliers.* On the contrary, the *negative review outliers are mostly related to course-works, such as grading, lecture heaviness, homework and tests, mandatory attendance, and group projects.*

Next, we put under the microscope the relation between the rating of a review with several factors associated with the course. First, we explore the influence of mandatory attendance and textbook requirements with a review's given rating. Figure 8a and 8b shows the distribution of mandatory-attendance and textbook-requirements for the five possible ratings. The figure shows that the distribution of attendance and textbook requirements remains fairly constant across the five ratings. This shows that *mandatory attendance or a textbook requirement for a course does not necessarily affect the course's assessment.*

We now look at the rating distribution on two dimensions: the difficulty dimensions for the course as specified by the student in the review (Figure 2) and the course level (as specified by the course code, for example, cs320 is a level 3 course, cs557 is a level 5 grad course). Note that, from now on, we define a good review as having ratings of 4 and 5, a bad review as having a rating of 1 and 2, and neutral as rating 3. Figure 8c demonstrates that *there is absolutely no real differentiation between the ratings given by the students and the associated*

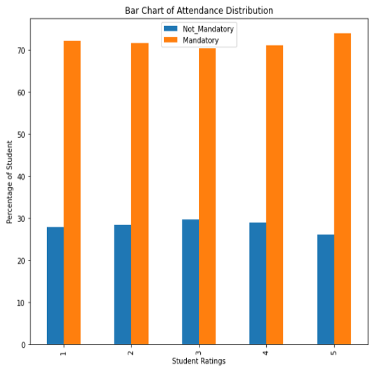
GO Tags	% of GO reviews they appear	BO Tags	% of BO reviews they appear
Caring	9.9	Tough Grader	9.2
Gives Good Feedback	7.77	Skip Class? You won't pass	7.0
Respected	6.5	Lecture Heavy	6.1
Accessible outside class	5.4	Lots of Homework	5.6
Amazing Lectures	5.3	Get Ready to Read	5.45
Hilarious	5.1	Graded by Few Things	4.45
Inspirational	4.42	Test Heavy	4.36
Participation matters	4.3	Participation Matters	4.03
Skip class? You won't pass	3.6	Group Projects	4.03

Table 7: Top tags for good and bad outliers

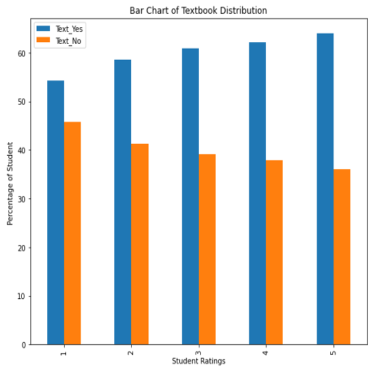
course's level. However, from Figure 8d, it is evident that the difficulty of a course does play a very significant role in a student's assessment of a course. We see that out of all the difficulty level 1 reviews, almost 78% of them have ratings of 4 or more. For difficulty levels 2 and 3, they are more than 75% and 65% respectively. On the contrary, courses with difficulty level 5 have almost 60% of the reviews as neutral and less than 20% as good. To add granularity to this investigation, we plot the same for all the difficulty levels, but this time for the *GO* and *BO* reviews defined earlier. Figure 8e shows the distribution of *GO* and *BO* reviews across the difficulty levels of the courses. It is clear from the figure that high difficulty courses (4 and 5) tend to have a high percentage of *BO* reviews, with percentages rising as much as 80% of all the outliers. These findings solidify the claim that the *difficulty level of a course certainly plays a part in RMP assessments, with higher difficulty level courses are highly likely to get negative reviews.*

In order to understand the main topics students talk about in their reviews, we employ Latent Dirichlet Allocation (LDA)[2]. LDA is a generative model that, in our case, considers each review as a mixture of a small number of topics and assumes that each word in this review's text is associated with one of the topics. Upon LDA analyses, we list the top few topics for good reviews (rating 4 and 5), bad reviews (rating 1 and 2), and five different course levels in Table 9. It can be seen from Table 9 that *topics such as tests and grades being tough, as well as lectures and materials being boring, are the main topics for bad reviews. On the other hand, topics that dominate the good reviews are interesting materials, funny lectures, and the relative easiness of the course.*

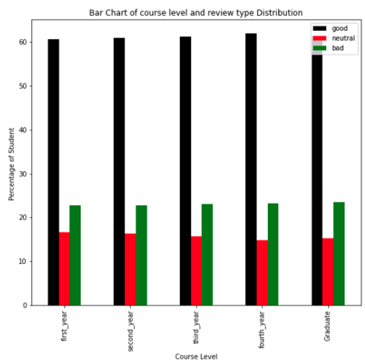
Finally, we investigate when students usually post their reviews on RMP. We do this to check if bad reviews come in early during a course when the students who are struggling usually drop out. Figure 4 shows our dataset's count of good and bad reviews from 2003. Interestingly, in our dataset, all the peaks of good and bad reviews occur in May and December — understandably at the end of the semesters for the US-based universities in our dataset. Therefore, from the



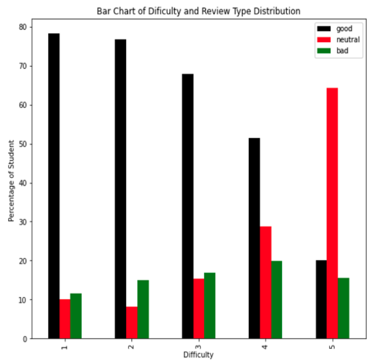
(a)



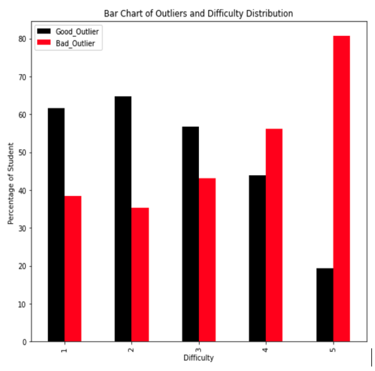
(b)



(c)



(d)



(e)

Table 8: Distribution of given ratings vs (a) mandatory attendance and (b) mandatory textbook

figure, we can safely claim that *students generally post their evaluations after the semester is over, irrespective of the review's sentiment.* Note that there is a significant fall in the review counts in 2020 and 2021. We find this consistent with enrollment declines due to Covid-19 in the US for 2020 and 2021 [23].

Type	Top Topics
Good Reviews	funny lectures, easy, amazing classes, interesting material, always help, extremely funny
Bad Reviews	boring lectures, worst grading, boring materials and exam, worst tests, hard tests and boring lectures, rude

Table 9: Top topics for good and bad reviews

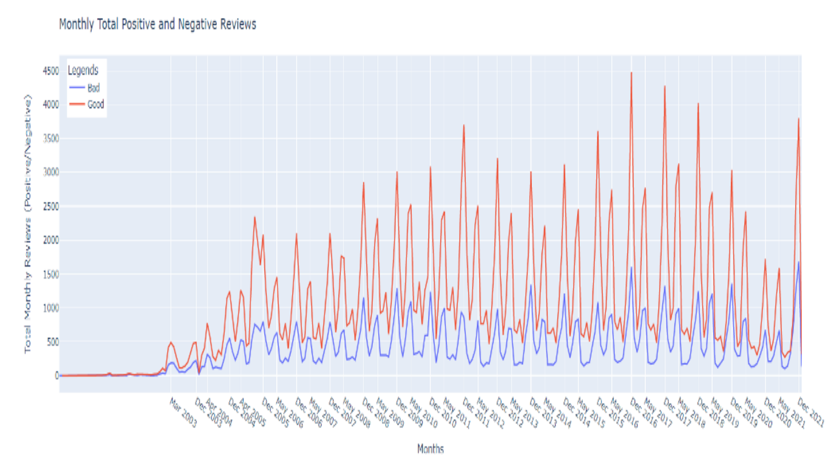


Fig. 4: Count of Reviews posted across time

6 Conclusion and Future Works

In this research, we perform a data-driven analysis to dissect the factors and qualities that make a course's and a professor's assessment positive in RMP. Analyses performed on the professor-level data in Section 4 revealed that a professor's perceived difficulty level negatively impacts their overall rating on RMP. We also find that in our dataset, good professors are more in number than bad professors. We also find that Liberal Arts universities have the highest and lowest number of good and bad professors, compared to others. Furthermore, we fail to find any relationship between the distribution of good and bad professors and the location or classification of a university. We then perform analyses on individual reviews posted by students on RMP. We found that, like the professors, the more difficult a course is, the higher the likelihood is of it getting a negative review on RMP. We also perform an emotion distribution analysis of RMP reviews, revealing that emotions such as joy in positive reviews and anger in negative reviews are more pervasive. We did not find any relationship between mandatory attendance, textbook requirement or course levels, and student ratings. Finally, both for professors and course reviews, we saw that students appreciate human qualities such as care, respect, hilarity, helpfulness, and accessibility, together with good feedback-giving; on the other hand, most of the complaints usually focus on the course-related factors such as lots of homework, strenuous tests and exams, and boring lectures. Finally, we see that irrespective of the review sentiments, RMP reviews almost always come at the end of the semester.

Our analyses pave the way for several interesting future research directions. First, it would be interesting to explore — analyzing our RMP dataset — the evolution of a professor through time. RMP currently scores a professor by averaging a professor's ratings through all the active years. The idea is to investigate if professors change how they teach a course by getting feedback from earlier reviews. Second, another research direction will be the implementation of a machine learning algorithm that incorporates all the students' reviews, the recentness of the reviews, the relative difficulty of the course, and reviews across other universities for similar levels of courses together to make a holistic prediction of a professor's actual rating.

Bibliography

[1] Azab, M., Mihalcea, R., Abernethy, J.: Analysing ratemyprofessors evaluations across institutions, disciplines, and cultures: The tell-tale signs of a good professor. In: International Conference on Social Informatics. pp. 438–453. Springer, Springer, Cham, Bellevue, WA, USA (2016)

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[3] Brown, M.J., Baillie, M., Fraser, S.: Rating ratemyprofessors. com: A comparison of online and official student evaluations of teaching. College Teaching **57**(2), 89–92 (2009)

[4] Brown, M.J., Baillie, M., Fraser, S.: Rating ratemyprofessors. com: A comparison of online and official student evaluations of teaching. College Teaching **57**(2), 89–92 (2009)

[5] Coladarci, T., Kornfield, I.: Ratemyprofessors com versus formal in-class student evaluations of teaching. Practical Assessment, Research, and Evaluation **12**(1), 6 (2007)

[6] Felton, J., Koper, P.T., Mitchell, J., Stinson, M.: Attractiveness, easiness and other issues: Student evaluations of professors on ratemyprofessors. com. Assessment & Evaluation in Higher Education **33**(1), 45–61 (2008)

[7] Felton*, J., Mitchell, J., Stinson, M.: Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. Assessment & Evaluation in Higher Education **29**(1), 91–108 (2004)

[8] Freng, S., Webber, D.: Turning up the heat on online teaching evaluations: Does “hotness” matter? Teaching of Psychology **36**(3), 189–193 (2009)

[9] Gao, B.J., Katrompas, A.: A preliminary experimental analysis on ratemyprofessors. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 5679–5681. IEEE, IEEE, Virtual (2020)

[10] Helterbran, V.: The ideal professor: Student perceptions of effective instructor practices, attitudes, and skills. Education **129** (01 2008)

[11] Magathan, H., Schreyer, S.: What factors affect web-based professor ratings? (2020)

[12] Mahmud, M.S., Bonny, A.J., Saha, U., Jahan, M., Tuna, Z.F., Al Marouf, A.: Sentiment analysis from user-generated reviews of ride-sharing mobile applications. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). pp. 738–744. IEEE, IEEE, Erode, India (2022)

[13] Mitchell, K.M., Martin, J.: Gender bias in student evaluations. PS: Political Science & Politics **51**(3), 648–652 (2018)

[14] Nimmi, K., Janet, B., Selvan, A.K., Sivakumaran, N.: Pre-trained ensemble model for identification of emotion during covid-19 based on emergency response support system dataset. Applied Soft Computing **122**, 108842 (2022)

[15] Otto, J., Sanford Jr, D.A., Ross, D.N.: Does ratemyprofessor. com really rate my professor? Assessment & Evaluation in Higher Education **33**(4), 355–368 (2008)
[16] Rivers, C.M., Lewis, B.L.: Ethical research standards in a world of big data. F1000Research **3**(38), 38 (2014)
[17] Rosen, A.S.: Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of ratemyprofessors. com data. Assessment & Evaluation in Higher Education **43**(1), 31–44 (2018)
[18] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR **abs/1910.01108** (2019), http://arxiv.org/abs/1910.01108
[19] Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized affect representations for emotion recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3687–3697. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1404, https://aclanthology.org/D18-1404
[20] Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized affect representations for emotion recognition. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3687–3697. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1404, https://www.aclweb.org/anthology/D18-1404
[21] Selenium: Selenium with Python. https://selenium-python.readthedocs.io/ (2022), [Online; accessed 12-August-2022]
[22] Soup, B.: Beautiful Soup Documentation. https://beautiful-soup-4.readthedocs.io/ (2022), [Online; accessed 12-August-2022]
[23] Welding, L.: U.S. College Enrollment Decline Statistics. https://www.bestcolleges.com/research/college-enrollment-decline/ (2022), [Online; accessed 11-August-2022]
[24] Wikipedia: Carnegie Classification of Institutions of Higher Education. http://en.wikipedia.org/w/index.php?title=Carnegie%20Classification%20of%20Institutions%20of%20Higher%20Education&oldid=1090284147 (2022), [Online; accessed 07-June-2022]
[25] Wikipedia: Rate My Professor. https://en.wikipedia.org/wiki/RateMyProfessors.com (2022), [Online; accessed 07-June-2022]