Title	**RateMyProfessors.com: Students' Paradise, Professors' Nightmare?**
Course	QTM 385 Practical Approaches to Data Science with Text
Authors	Colleen Su, BS in Applied Mathematics and Statistics, kexin.su@emory.edu
	Derick Yang, BS in Quantitative Sciences, yuan.yang1996@gmail.com

Abstract

Since RateMyProfessor.com (RMP) started in 1999, it has grown to become one of the most used websites for students to review their professors. However, its objectivity and accuracy have always been discussed among professors and researchers. Instructors have complained that they got extreme comments from students that are inconsistent with their actual performances, while scholars have shown that RMP reviews and rating provide valid measurement of student learning. In this research, we will scrape the ratings and reviews of 12 schools' mathematical department professors from RMP website and analyze the evaluation scores and student comments to test our four hypotheses: 1. Students come to RMP to either rave or rant about their professor. 2. Students' ratings reach a higher consensus when the professor has high or low teaching quality. 3. The level of difficulty of the class affects the overall quality rating of the professor. 4. Students' reviews of professors on RMP are different across universities with different ranking.

Intellectual Merit

In this research, we will write our own code to scrape the RMP website for the information that we need because we cannot find any usable API or programs that are already available online. To test our hypothesis, we will utilize natural language processing tools including TF·IDF and sentiment analysis as well as statistical methods such as hypothesis testing and regression analysis. We further incorporated school major ranking into our analysis to see school's reputation and academic resources impact on student reviews to their professors.

Broader Impact

Our research aims to explore the validity and credential of student comments posted on RMP website. We started our research by analyzing the student ratings of professors in mathematics departments across 12 different prestigious universities in America. If our research logic and methods succeed, we can expand our research scope to incorporate more student reviews of other universities and department in the US. Meanwhile, the same method can also be applied to analyze other similar rating websites involving student rating and comments. Hopefully, by delving into the data, correlation and distribution of student rating scores on RMP, our research can provide

constructive suggestions for university professors to deliver effective instructive performance to help students to learn their lessons more efficiently.

1. Introduction

An anonymous tenure-track professor in a flagship state university posted a curious case on Reddit, talking about the distinct difference between the scores she received from her university internal evaluation and her general reviews on RateMyProfessors.com (RMP). On the scale of 1 to 5, the professor received a consistent 4.7 to 4.9 scores on her internal university evaluations while getting an average mid-3 score on the rating website. She also mentioned that some quotes and reviews from the anonymous students on RMP were false, even containing sexist and racist extreme comments. Those biased slanders, the professor said, put a huge negative impact on her reputation and daily life. Some other cases include one tenured professor from the University of Wisconsin at Whitewater suing one student for putting defaming commentary on RMP.

These are two extreme cases we found on Reddit and Inside Higher Ed talking about educational instructors criticizing the credentials of students' comments on RMP website. It would be reasonable to infer that a significant number of students who made comments on RMP were students who obtained either satisfied grades or unsatisfied grades from the professors, so it is natural that some students would make comments on RMP either to rave the professors for satisfaction or rant the professors for catharsis. But still, some peer reviewed journals provide evidence to support the validity of the comments on the RMP website, which will be further elaborated in the Background section. For this research, we aim to testify the validity of student reviews on RMP by examining the polarity of student reviews of each professor, students' consensus of evaluating one professor and the relationship between the professor's class level of difficulty and overall quality score. We furthermore incorporated school major ranking into our analysis to see whether there is any apparent difference among students' reviews from different tier schools, an analysis method which have not been conducted by any similar review before. The reason we chose major ranking to differentiate schools is based on the fact that higher ranking schools potentially have higher academic and research quality, which we inferred could promote the professor ratings. Here are the four major hypotheses our research aims to examine:

- Students use Rate My Professor to either rave or rant about their professors, that is, students come to RMP to post either very positive or very negative comments.
- Students' ratings reach a higher consensus when the professor's teaching quality is in the higher third or in the lower third; ratings have more variance when the professor is mediocre.
- The level of difficulty of the class affects the overall quality rating of the professor. The easier the level of difficulty, the higher the professor's overall quality rating will be.
- Students' reviews of professors on RMP are different across universities. Professors from higher major ranking schools should receive higher ratings on RMP.

2. Background

Since RateMyProfessor.com (RMP) started in 1999, it has grown to become one of the most used websites for students to review their professors. Realizing its increasing importance, researchers have done plenty of researches around RMP for the past twenty years, and a lot of research findings highlighted the importance and effects of RMP reviews on students' perception of professors. Overall, students have positive experiences with RMP and regard it as a reliable resource of information for course selection (Coladarci and Kornfield, 2007). Students who have posted reviews on RMP trust the website even more and have stronger tendencies to check the website before registering for classes (Coladarci and Kornfield, 2007). Students who have exposed to positive RMP reviews before class report a better impression on the professor, are more engagement in the class, and they get higher grades than students who expose to negative ratings of the professor (Reber, Ridge, and Downs, 2017). Furthermore, even though there have been incidents that make people question the equity of RMP reviews, many journal articles have shown that RMP is more valid than we thought. By looking at the students' comments of the professor, researchers were able to classify good professors from poor professors with an accuracy of over 90% (Azab, Mihalcea, and Abernethy, 2016). Research findings have also shown that ratings from RMP actually provide an accurate measure of student learning (Otto, Sanford, and Ross, 2008). These researches all show the extensive usage and impact of RMP among students, thus emphasizes the importance of our research because the research result can be beneficial to a wide range of the population. Different from these studies that emphasize the influence of RMP scores and reviews on the students, our current research focuses on the reviews themselves. We will be studying the trend of the student reviews and ratings as well as the attitude of those comments. Furthermore, we will be looking at the potential differences in ratings and comment attitude based on the ranking of the institution.

There have also been researches that are interested in the same hypotheses as we do. A study done by Bleske-Rechek and Michels in 2010 analyzed the self-report data on the use of RMP from 208 students at a regional public university and the RMP reviews of the professors from that university. They were also interested in whether the student comments are polarized or neutral. Their results show that students ratings have a near-normal distribution rather than bimodal, which means that most students' ratings are neutral. This study is different from our current study because it analyzed the self-reported survey-like data that they gathered from the students to study their motives of posting on RMP, while our research analyzes RMP reviews using sentiment analysis to see if the reviews are actually polarized regardless of the students' motive. Also, this article focuses on only one university; for our paper, we will also analyze the difference in student comments between different universities and see if the ranking of the college has an impact on the polarity of the reviews.

Student consensus on ratings has also been studied by other researchers. Bleske-Rechek and Fritsch' study in 2011 studies the consensus of student ratings to see whether reviews from RMP are reliable. They found that variance in students' ratings about a given instructor was similar across the number of raters, with 10 raters showing the same degree of consensus as 50 or more raters. Furthermore, students showed the most consensus about instructors who were among the top third of the distribution in quality, and this effect occurred even among instructors rated as the most difficult. They thus concluded that RMP reviews are more reliable than we thought. For our research, we want to extend their research findings to see if this consensus also exists for professors that have ratings in the lower third since it is possible that students post bad reviews just because they got a poor grade. This way we can use the student consensus to see if the professor truly has low teaching quality, or does some student just give bad reviews because of their own poor performance.

Our third hypothesis also has previous evidence. Research done by Otto, Sanford, and Ross in 2008 found that the easiness rating on RMP (on the scale of 1 to 5, with 5 being very easy) is positively correlated with clarity and helpfulness rating, which can have an impact on the overall quality rating. However, since this study was published 10 years ago, the RMP website has changed a lot from then. They took out the clarity and helpfulness rating and changed the "easiness" rating to "level of difficulty" (1 to 5, with 5 being difficult). Thus, for our research, we can delve into the relationship between the level of difficulty and the overall quality score to test this result. We will also examine how the correlation changes when the ranking of the university changes. Potentially, we would expect that correlation are weaker for colleges that are ranked higher because the students are smarter and can be more appreciative of harder courses.

For our fourth hypothesis, we could not find a peer-reviewed journal article that is pertained to the exact topic that we want to investigate. However, In Pusser et al.'s article *University Rankings in Critical Perspective*, they mentioned that one critical aspect of university ranking across the globe is the academic reputation and faculty quality. Specifically, US News ranking particularly values the reputational survey of each university, which measures the faculties' research productivity. Thus, it is intuitive to assume that schools with higher ranking have better faculty members. Therefore, for our research, we expect that universities with higher major ranking would have better professor in that field, thus they would have higher ratings on RMP.

3. Research Process

3.1 Data Collection

In order to effectively compare the difference among student reviews from different tier schools, we selected 12 American universities based on US News' best mathematics program ranking (2018) and focused on analyzing the student reviews on RMP of these 12 schools' mathematics

department professors. Among these 12 schools, five schools are ranked in top 20, which are assigned as top tier schools in our dataset. Four schools are ranked between 20 and 50, which are assigned as middle tier schools and the rest of three, ranked beyond 50, are assigned as low tier schools. Because different school has different number of student reviews for their math professors and the variance is significant, to keep the number of professors from all three tiers approximately the same, there is a small variation in the numbers of schools in each tier.

On RMP, each professor has one specific ID and each professor's profile has a hidden html layer that stores all the student evaluation data and comments. We manually parsed down 738 professors' ID from RMP and used python to get access to these 738 hidden layers by inputting professor's ID into html link and further parsed down all the evaluation data and comments. Our analysis puts emphasis on three data variables: overall quality score with range between 1 to 5 with 5 being best quality and 1 being lowest quality; level of difficulty with range between 1 to 5 with 1 being easiest and 5 being hardest, and the number of comments each professor receives. Exhibit 1 in appendix shows more detailed explanation of every level of difficulty and overall score quality. By analyzing each student review, we created another gender variable that indicates the gender of the professor. We used python to loop through all the reviews of a professor and count the appearance times of gender pronouns. If the appearance times of male pronouns is larger than that of the female pronouns for one professor, we assigned that professor as male and vice versa. Exhibit 2 shows the pseudocode we used to do the gender filter. For the reviews with gender assigned as 'N/A', we further manually analyzed the student comments and searched the professor's personal information on school website to confirm their gender. In our dataset, we assigned 'male' as 1 and 'female' as 0. After we got the professor name, school, gender, overall quality score, level of difficulty score and student comment, we built all the information into panel data format for further regression analysis. Exhibit 3 shows the detailed structure of our dataset.

3.2 Methods

Regression

The dataset consists 7,782 reviews and the number of reviews received by a single professor varies from 2 to 125 (we eliminated the professors who received only one review out of the dataset). About 90% of our professors received reviews below 20 and 16 professors received reviews above 50. In order to eliminate the unbalanced weight issue of our panel data, we calculated the average overall quality score, average level of difficulty score, average polarity score and the log ratio of number of reviews that each professor received. We successfully reduced our sample size from 7,782 to 738 and run OLS regression on average overall quality score over average level of difficulty, average polarity score, gender, number of reviews log ratio, gender and average level of difficulty interaction term, and schools (as dummy variables based on Emory University).

$$OverallQualityScore = \beta_0 + \beta_1 * avgLof + \beta_2 * Gender + \beta_3 * avgLof : Gender + \beta_4 * log(avgComment) + \beta_5 * avgPol + \sigma_1 * CMU + \sigma_2 * Cornell + \sigma_3 * Stanford + \sigma_4 * Tufts + \sigma_5 * Tulane + \sigma_6 * UCLA + \sigma_7 * Chicago + \sigma_8 * NotreDame + \sigma_9 * USC + \sigma_{10} * WashingtonSt.Louis + \sigma_{11} * Yale$$

Sentiment Analysis

For sentiment analysis, we used Textblob, a Python library that is commonly used for processing textual data. Textblob has a "sentiment" property where after passing in a text, it returned a Namedtuple of the form "Sentiment (polarity, subjectivity)". The polarity score is a float within the range [-1.0, 1.0] where 0 means neutral, 1 means very positive and -1 means very negative. The subjectivity is a float within the range [0.0, 1.0] where 0.0 indicates very objective and 1.0 indicates very subjective. There are also two sub-methods sentiment.polarity() and sentiment.subjectivity() where we can access these two values separately. We used TextBlob to go over each review and created another two columns to store both polarity and subjectivity values for our panel dataset. To test the accuracy of the sentiment analysis, we randomly chose 100 comments from all the comments and manually rate their polarity, and compared those with the polarity score that are given by Textblob, and the result shows that Textblob's sentiment analysis has an accuracy of 84% on our data.

For our research, we define that the polarity of [-1, -0.3) indicates negative tone, [-0.3, 0.3] indicates neutral tone, and (0.3, 1] indicates positive tone. We also assigned three categories to the overall quality variable with (1, 2.5) being poor quality, [2.5, 4) being average quality, and [4, 5) being awesome quality. We decided that those who gave an awesome overall quality score and leave comments with positive polarity are considered as "rave", and those who rate poor over quality score and leave comments with negative polarity are considered as "rant", and others are considered as neutral. If the proportion of comments that are either "rave" or "rant" are greater than 50% of all comments, it would confirm our hypothesis that students come to rate my professor to either rave or rant about their professor.

We also noticed that RMP has a default comment, "No comment", when the student did not leave an actual comment, which gets 0 for both polarity and subjectivity score. Therefore, to avoid the potential impact this pile-up at 0 might have to our analysis, we eliminated those comments and then plotted the distribution of polarity and subjectivity separately for each college tier. Then, we performed ANOVA test and Tukey comparison to see if there are significant differences among the three tiers.

One-way ANOVA & Tukey Multiple Pairwise Comparison

The one-way analysis of variance (ANOVA) is often used to determine whether there are any statistically significant differences between the means of two or more independent groups. However, ANOVA can only tell us if there is a significant difference between any of the two groups, it cannot tell us which two groups has significant difference. Therefore, we use Tukey's

multiple pairwise comparison. It returns a confidence interval and an adjusted p-value for the difference of each two groups, and we can determine if the difference is significant based on the confidence interval. If the confidence interval does not include zero, it means the difference in mean is statistically significant. For our research, the combination of these two methods is crucial because we have a lot of categorical variables such as college tiers and average overall quality score, and these methods allow us to identify if there is a significant difference between the means of the groups that we are interested in.

Distribution of Standard Deviation

To test our second hypothesis on student consensus, we first calculated the average overall quality score for each professor, and categorized them into awesome quality, average quality, and poor quality based on their average overall quality score using criteria that we mentioned earlier. Then, we calculated the standard deviation of each professor's overall quality score and plotted each professor's standard deviation based on their quality categories. To test whether there is a significant difference between these three groups, we used the combination of one-way ANOVA test and Tukey comparison to get the result. If our hypothesis is correct, we would observe that the standard deviation is significantly lower for the awesome and poor quality group than for the average quality group.

TF-IDF and Vector Space

To analyze the US News ranking impact on student reviews, in other words, to analyze the similarity of student reviews across different university math department professors, we managed to approach the problem from two aspects. First, we tried to analyze the student comment similarity through bag of words and cosine similarity methods and second, we tried to analyze the evaluation scores, like review number, average overall quality score, average polarity score, through calculating the Euclidean distance among the vectors we built for each school.

For student comment analysis, we stored all the student comments from the same university into one txt file and split all the comments into tokens (words and special characters). We further counted both the appearance frequency of each token in each txt file and the appearance frequency of each token across all the txt documents to calculate the TF-IDF score of each token. We selected eight schools to run the cosine similarity test. The evaluation data analysis involved building vectors to summarize the evaluation score of each school. We constructed an eight dimension vector for every school and each vector consists of the log ratio of number mathematics professors that students have reviewed on RMP, the log ratio of the average number of reviews professors received, the average overall quality score, the overall quality score standard deviation, average level of difficulty, the level of difficulty standard deviation, the average polarity score and the polarity score standard deviation. The figure below shows our vector structure for each school.

$$[log(Numberof Professors), log(AvgNumberof Reviews), AvgOverallScore, StdOverallScore, AvgLevelof Difficulty, StdLevelof Difficulty, AvgPolarityScore, StdPolarityScore]$$

We further calculated the Euclidean distance between every two vectors, built a 12x12 matrix and plotted a heat map to see whether there is any cluster effect due to major ranking. The way we calculated Euclidean distance follows the equation below in which **p** and **q** represent any two vectors from our sample schools.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

4. Result

4.1 Regression

Several things are apparent from the parameter output (Table 1). In this OLS equation, average level of difficulty and average polarity score show statistically significance as both p values way out of the rejection region of 5% significance level. Although Gender variable shows less statistical significance, it still has greater practical effect on average overall quality score in terms of its coefficient. There is no much to tell from the school dummy variables; however, the majority of the schools that we selected show relatively lower average quality score than Emory University does, which could be simply attributed to our sample selection. Furthermore, an optimal 53.4% adjusted r-squared proves the validity of our regression model.

	Std.	Error t	value P	r(>|t|)	
(Intercept)	4.46339	0.33279	13.412	< 2e-16	***
avgLof	-0.42715	0.0968	-4.413	1.18E-05	***
Gender	-0.47082	0.32749	-1.438	0.151	
avgLof:Gender	0.13789	0.1015	1.359	0.1747	
log(avgComment)	-0.0179	0.03017	-0.593	0.5533	
avgPol	3.27284	0.15815	20.695	< 2e-16	***
CMU	0.03125	0.13061	0.239	0.811	
Cornell	0.03041	0.10583	0.287	0.7739	
Stanford	-0.10362	0.15032	-0.689	0.4908	
Tufts	0.14211	0.12609	1.127	0.2601	
Tulane	-0.13135	0.11178	-1.175	0.2404	
UCLA	-0.10417	0.12242	-0.851	0.3951	
Chicago	-0.04314	0.15129	-0.285	0.7756	
Notre Dame	0.16522	0.18174	0.909	0.3636	
USC	-0.23469	0.10989	-2.136	0.033	*
Washington St.Louis	-0.2414	0.13056	-1.849	0.0649	.
Yale	-0.39549	0.24132	-1.639	0.1017	

Residual Standard Error	0.7094	Adjusted R-Squared	0.534
Degrees of Freedom	721	F Statistic	53.78
R-Squared	0.5441	P-Value	2.20E-16

Table 1

Our research puts great emphasis on analyzing the polarity score and level of difficulty to examine the first and third hypothesis separately. The statistical significance generated by our regression

model indicates the legitimacy of our research logic of using these two variables to test the hypothesis. Furthermore, a 0.534 adjusted r-squared may prove that only a medium portion of data can be fitted into a linear regression, which may potentially disprove the fact that most of students come to RMP either to rave or rant. If the first hypothesis is the very fact, the distribution of overall quality scores will cluster on the two sides of the axis and lead to regression model with very low standard error and higher adjusted r-squared. Meanwhile, we worried that the number of reviews might put significant impact on the average overall quality score obtained by one professor, but it turns out that the log(comment) variable is not significant and a small coefficient implies that the variable has small practical impact on dependent variable. In order to visualize review size impact on overall quality scores, we plotted the distribution graph of the overall quality scores of professors who have received more than 50 reviews on RMP. Among all the professors, 16 professors have over 50 reviews. The overall quality score distribution can cluster to one side or two sides, but clustering doesn't necessarily imply that students come to either rave or rant the professor. One thing is that positive or negative comment can be purely objective. As the number of reviews for a single professor increases, the portion of rave and rant phenomenon significantly decreases. The relationship between the overall score distribution and number of reviews for a single professor is purely random. It can be uniformly distributed, cluster to either side or cluster to two sides. The graph below shows the overall quality scores distribution for these 16 professors.

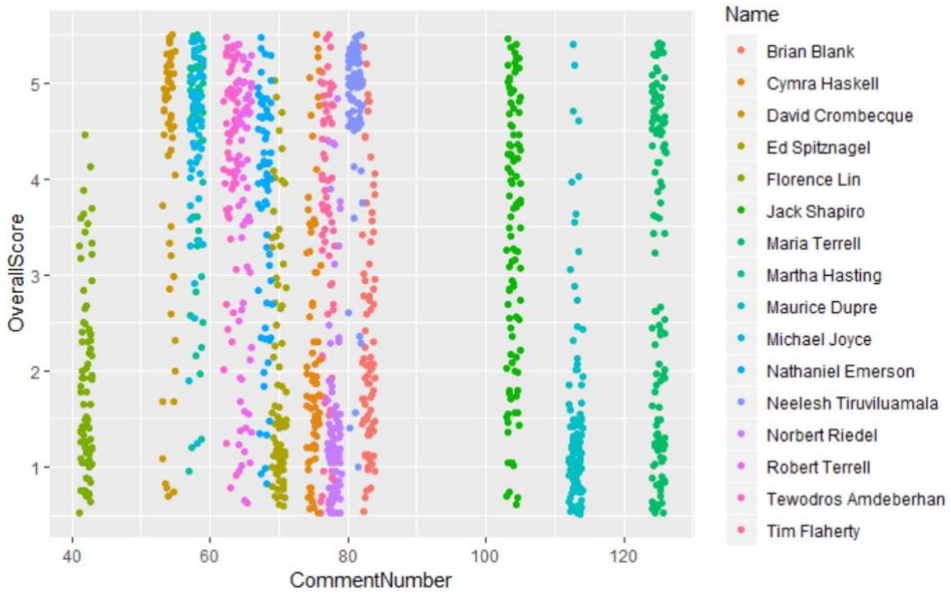

Figure 1

Based on the figure 1, we can see that there are about four professors who have students reviews clustering on the low level of overall quality score. If the professor has been known for giving poor instruction performance on RMP, why so many students still flowed into his or her class and gave negative evaluation. One possible explanation is that those classes are core classes that

students need to take to fulfill the major and usually those core classes are mainly instructed by the professor.

In conclusion, there is no apparent relationship between the number of comments and the overall quality score distribution for a single professor and probably this is why we obtain a insignificant t-score for log(Comment) variable.

4.2 Comment Polarity

As mentioned, we define reviews that give both awesome overall score and leave positive comments as rave, and those who give poor overall scores and leave negative comments as rant. From table 2, we can see that the sum of the rave and rant categories is about 30% of all comments, which is less than 50%. This means that our hypothesis that students either rave or rant about their professor on RMP is rejected by the statistics. Student's reviews are actually mostly objective.

		Polarity vs. Overall Quality		
		[-1,-0.3) **Negative Polarity**	[-0.3,0.3] **Objective**	(0.3,1] **Positive Polarity**
[4.0, 5]	**Awesome**	0.33%	29.38%	24.43%
[2.5,4)	**Average**	0.72%	12.57%	2.58%
[0,2.5)	**Poor**	5.50%	22.94%	1.55%

Table 2

To further analyze out data to ensure the result, we looked at the polarity and subjectivity scores based on the college tier. If the result confirms our hypothesis, we would expect to see a bimodal distribution indicating pile-up at negative and positive polarity. However, as figure 2 and figure 3 shown, the polarity and subjectivity distribution for all three tiers are normal distribution which further rejects our first hypothesis.

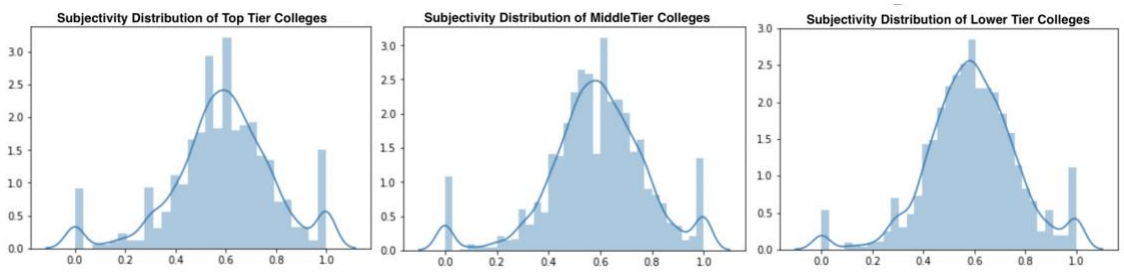

Figure 2

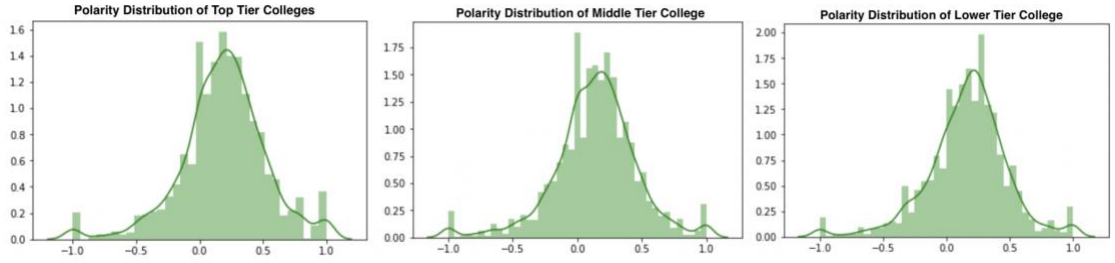

Figure 3

Even though our hypothesis is rejected, we still conducted the ANOVA test on both polarity and subjectivity to see if there is a significant difference between the mean of each group. We found that there is no significant difference between the subjectivity of the three tiers (Table 3), but the difference in polarity between all the tiers are significant (Table 4). From the Tukey comparison, we get that the top tier students wrote the most positive comments, and middle tier students wrote the most negative comments among the three groups (Table 4).

One-way ANOVA for Subjectivity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RMP_adj$Tier_cat	2	0.01	0.00435	0.117	8.90E-01
Residuals	7478	278.02	0.03718		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

Table 3

One-way ANOVA for Polarity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RMP_adj$Tier_cat	2	3.4	1.679	16.07	1.09E-07 ***
Residuals	7478	781.3	0.1045		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

Tukey Comparison (95% Confidence Interval)

	diff	lwr	upr	p adj
mid-top	-0.052466	-0.074497093	-0.030435	0.0000001
low-top	-0.024171	-0.047077194	-0.001265	0.0357283
low-mid	0.0282949	0.007845791	0.048744	0.0033975

Table 4

4.3 Student Consensus

Figure 3 displays the distribution of standard deviation for the three average overall quality score categories. From the graph, we can see that there are pile-ups at zero for both awesome quality and

poor quality professors, which means that the standard deviation of their overall quality rating is relatively small. To test whether the difference is statistically significant, we again used one-way ANOVA and Tukey comparison. The results of these two tests are shown in table 5, we see that there is a significant difference in mean between all three groups based on confidence interval and p-value. The average standard deviation for awesome quality professor is the lowest among the three groups, poor quality professor is the second lowest, and average quality professor has the highest average standard deviation among the three groups, which confirms our second hypothesis.

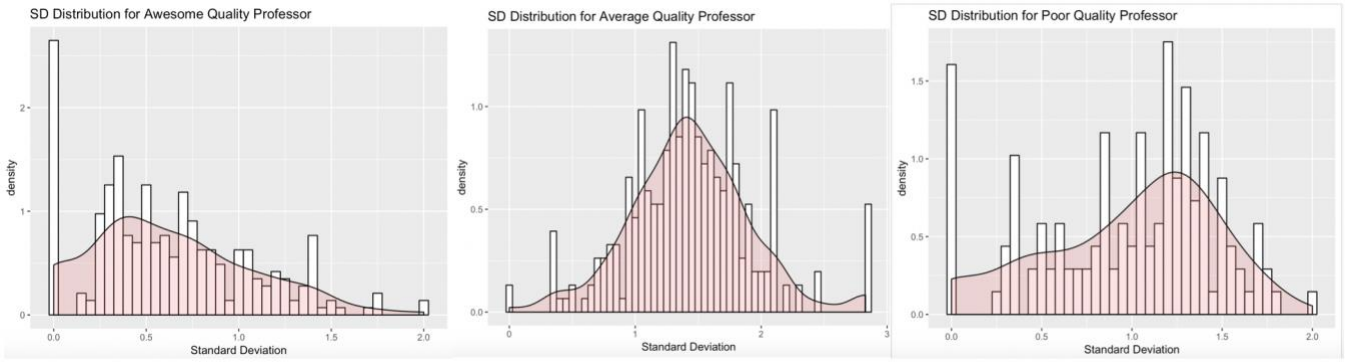

Figure 4

One-way ANOVA for SD					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RMP$Quality	2	95.22	47.61	224.1	<2e-16 ***
Residuals	717	152.36	0.21		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '

Tukey Comparison (95% Confidence Interval)

	diff	lwr	upr	p adj
Avrg-Poor	0.2792132	0.1697169	0.38871	0
Awesome-Poor	-0.5337153	-0.6475149	-0.419916	0
Awesome-Avrg	-0.8129285	-0.9034598	-0.722397	0

Table 5

4.4 Overall Quality vs. Level of Difficulty

To test our third hypothesis, we plotted the joint distribution of overall quality score and level of difficulty score (denoted as Lod in figure 4) with respect to the three tiers using the Python Seaborn Package. From figure 4, we can see that for top tier universities, the correlation is the weakest, with correlation coefficient -0.36. As we go down the tier, the correlation become strong, with lower tier has the strong correlation coefficient of -0.46. This confirms our third hypothesis. Furthermore, when we look at professors with level of difficulty of 5, we can see that a large portion of students from top tier colleges give high overall quality scores; however, as the school ranking decreases, less and less student gives high overall quality score for hard courses.

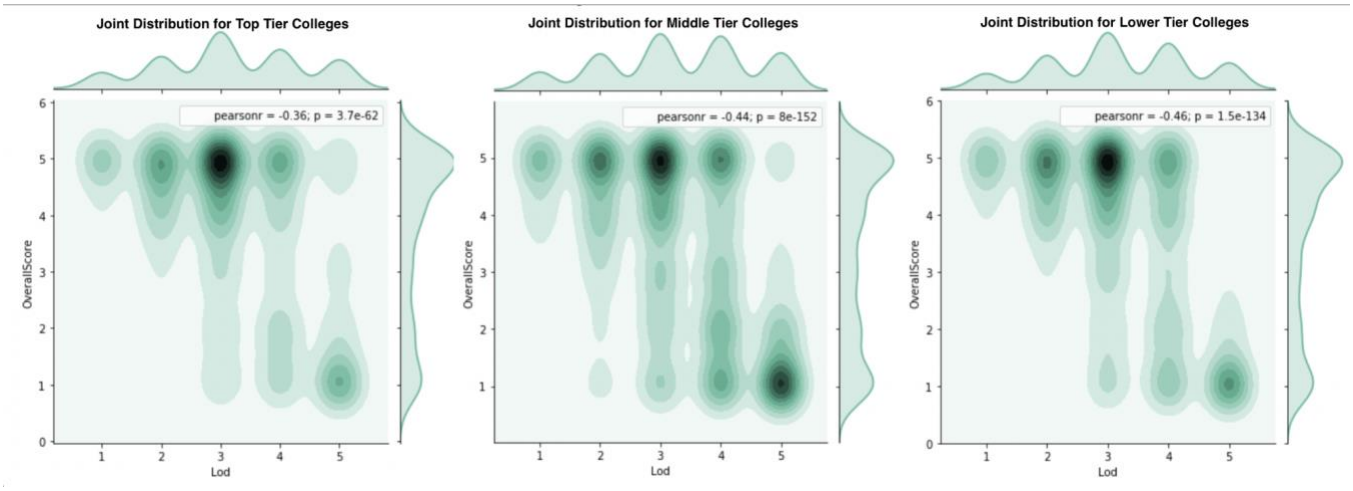

Figure 5

4.5 Student Review and School Ranking

We selected eight schools from our sample to run the cosine similarity test. Exhibit 3 shows us the student comment similarity between every two schools. The larger the ratio, the more similar the comments are. Similarly, we also built one matrix to calculate the Euclidean distance of the student evaluation score vector across these 12 schools. Opposite to comment similarity, the larger the Euclidean distance, the less similar the reviews are. Exhibit 4 shows the detailed data of our Euclidean distance matrix and we further plotted both matrix into heatmaps for better visualization.

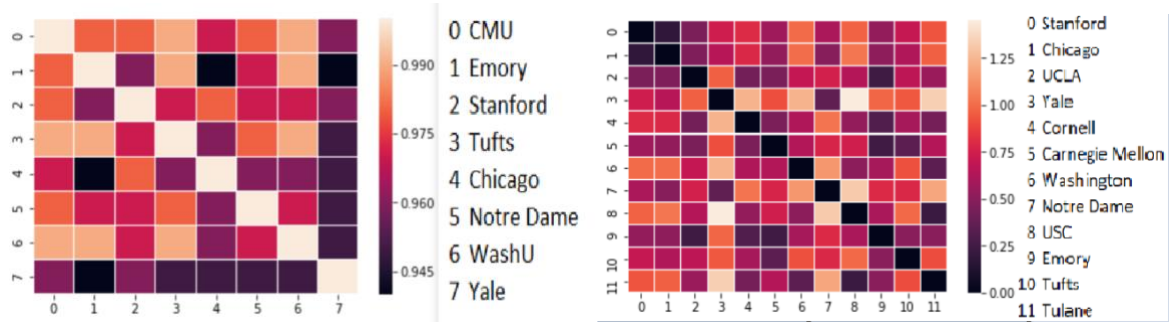

Figure 6

Based on the cosine similarity feedback, the majority of the similarity ratios are above 95%, which implies there is almost no difference among student comments across different schools. Even if Yale University shows us a lower cosine similarity comparing to other peers, it is caused by the small sample size and the contents of student comments are highly similar in these eight schools which we have run the cosine similarity test. In terms of evaluation score similarity, we calculated the Euclidean distance of every two score vectors and it turns out that there no distinct pattern on the Euclidean distance heatmap. The colors (figure 6 right), which represents the similarity distance between every two schools, are randomly dispatched without any dark and light clustering pattern effect shared by the schools belonging to the same tier.

5. Discussion

In this research, we centered around four core hypotheses. For our first hypothesis, we try to use sentiment analysis to prove the hypothesis that most students post on rate my professor to either rave or rant about their professor. However, the research result rejects our hypothesis; we see that most student comments have moderate tone and there is no significant difference between the subjectivity of student comments across schools with different ranking. This result is coherent with Bleske-Rechek and Michels' research where they also found a normal distribution when analysing the students' reviews, indicating no significant extreme sentiments. However, with further investigation using ANOVA test and Tukey comparison, we found that there are significant polarity differences with respect to school ranking. Student comments from top tier schools are more positive in tone than the other two tiers. What surprises us is that comments from lower tier also has significantly higher polarity than comments from middle tier. One potential explanation of this finding might be the "bronze medal affect" where gold medalists and bronze medalists are both happier than silver medalist because silver medalists are more likely to think about what "might have been". In our case, it is possible that since the ranking of middle tier colleges is not too far from the top tier colleges, students from middle tier schools might expect themselves to get educations that are the same as top-ranked schools. Therefore, when they get professors that does not meet their expectation, they would be disappointed and leave worse ratings for that professor. Whereas for students in lower tier universities, they are more likely to be content with the resources that they have, so they tend to leave higher rating and better comments.

For our second hypothesis, we expect that student's ratings will reach higher consensus when the professor's teaching quality is in the top third or the bottom third, there will be more variation in ratings for average quality professors. The result confirms our hypothesis, we found that top quality professors has the lowest variance in their ratings, poor quality professor has the second lowest, and average quality has the highest. This finding also confirms previous research done by Bleske-Rechek and Fritsch in 2011, and we further extend their research finding from only top quality to the top third and bottom third.

Thirdly, we tested our third hypothesis to see whether there is a correlation between level of difficulty and overall quality. We found that there is negative correlation between the two among all three college tiers, which means that professors that teach more difficult classes tend to get lower overall quality scores. As the school ranking goes down, the correlation becomes even stronger. Also, we noticed that when only looking at professors with the highest level of difficulty, more students from top tier give high overall rating, while almost none of the students in bottom tier colleges give high overall rating. There might be two possible implication of this result. First, it might be that students from top colleges are different from students from bottom colleges in that top students are more appreciative of hard classes, and they might expect to have hard classes in top colleges. Second possibility is that the professors in top colleges are better in quality, so that

they make hard classes more interesting and not so hard to have. Even though these two are possible explanation of this research finding, it requires further research to identify which on is more prominent, and whether there are other explanations.

Although our first hypothesis found that the polarity mean between different tiers is significantly different, if we grouped all the comments together and summarized all the score information into vector data, the difference among three tiers actually started to fade away. After testing our fourth hypothesis, we found that all the school student comments are significantly similar across different schools and the US News major ranking basically has no impact on the evaluation students gave to their professors. Although the evaluation scores may provide more apparent difference than the comments do, student reviews to their professors do not depend on either the school reputation or the research and academic resources, but hugely depend on the professor him or herself. Probably, this conclusion confirms another common sense: a good research professor doesn't necessarily make him or herself a good instructor.

6. Limitations and Improvement

Even though we tried to perfect the research process as much as possible, there are still some parts that can be modified for future research in order to get optimum result. First of all, when doing sentiment analysis, we noticed that Textblob's tends to underestimate the tone of a sentence, which means that what human think is a completely positive sentence, Textblob might only give it a polarity of 0.5 or 0.6. This indicates that if we are able to develop or find a sentiment analysis tool that more accurately determines the tone of the text, we may get more significant result and even other results that are significant.

Furthermore, our research focuses on analyzing students who have posted reviews on RMP; however, without survey as comparison, we couldn't tell whether students who posted reviews on RMP are different from students who didn't. Thus, based on the current evidence we have, it is hard to provide solid fact to either prove or disprove survivorship bias. Furthermore, because our research only collected student reviews of mathematics department professors, we want to see whether the patterns we found in our analysis also apply to other student reviews of different subjects. For further improvement, our research needs to incorporate more sample reviews from other major department. If the research sample is large and diversified enough, we would like to see the common patterns share by both high rating and low rating professors to see what are the apparent criteria that students use to do professor rating.

One more innovative method we propose to expand current research methods on RMP is that instead of focusing on analyzing the reviews belonged to one professor, we can analyze the reviews posted by one student account across different time periods. It is possible that some students may

consistently give low scores no matter what class he or she takes, which could be a possible evidence that the student deliberately used RMP to rant professor. However, this method requires more complicated parsing method to track down each user's reviews across different professor's profile.

Citation

Azab, Mahmoud, et al. “Analysing RateMyProfessors Evaluations Across Institutions, Disciplines, and Cultures: The Tell-Tale Signs of a Good Professor.” *Lecture Notes in Computer Science Social Informatics*, vol. 10046, 23 Oct. 2016, pp. 438–453., doi:10.1007/978-3-319-47880-7_27.

Bleske-Rechek, April, and Kelsey Michels . “RateMyProfessors.com: Testing Assumptions about Student Use and Misuse.” *Practical Assessment, Research & Evaluation*, vol. 15, no. 5, May 2010.

Bleske-Rechek, April, and Amber Fritsch. “Student Consensus on RateMyProfessors.com.” *Practical Assessment, Research and Evaluation*, vol. 16, no. 18, Nov. 2011, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.646.8431&rep=rep1&type=pdf.

Coladarci, Theodore, and Irv Kornfield. “RateMyProfessors.com versus Formal in-Class Student Evaluations of Teaching .” *Practical Assessment, Research & Evaluation*, vol. 12, no. 6, May 2007, pareonline.net/pdf/v12n6.pdf.

Flaherty, Colleen. “Rating or Defaming?” *Inside Higher Ed*, Inside Higher Ed, 23 May 2014, www.insidehighered.com/news/2014/05/23/professor-sues-student-over-his-online-reviews-her-course.

Otto, James, et al. “Does Ratemyprofessor.com Really Rate My Professor?” *Assessment & Evaluation in Higher Education*, vol. 33, no. 4, Aug. 2018, pp. 355–368.

profthrowaway. “How Is RateMyProfessor Not Slander? Is It Possible to Get My Reviews Taken down?” *Reddit*, www.reddit.com/r/legaladvice/comments/7vchj4/usa_how_is_ratemyprofessor_not_slander_is_it/.

Pusser, Brian, and Simon Marginson. “University Rankings in Critical Perspective.” *The Journal of Higher Education*, vol. 84, no. 4, 2013, pp. 544–568., doi:10.1353/jhe.2013.0022.

Reber, Jeffrey S., et al. “Perceptual and Behavioral Effects of Expectations Formed by Exposure to Positive or Negative Ratemyprofessors.com Evaluations.” *Cogent Psychology*, vol. 4, no. 1, 12 June 2017, doi:10.1080/23311908.2017.1338324.

Appendix

Exhibit 1. RMP Rating Explanation

Overall Quality Score	Explanation
[4.0,5.0]	Awesome
[2.5,4.0)	Mediocre
[0,2.5)	Poor

Level of Difficulty	Explanation
1	Show up & Pass
2	Easy A
3	The usual
4	Makes you work for it
5	The hardest thing I have ever taken

Exhibit 2. Gender Filter Pseudocode

1: **procedure** GENDER˙FUNCTION:
2: **assign** variables m=0 and f=0
3: **for** each comment of one professor:
4: **if** *term like 'He' or 'he' or 'His' or 'his' or 'him' or 'guy' or 'dude' appears in comment*, **then**:
5: m variable +1
6: **if** *term like 'She' or 'she' or 'Her' or 'her' or 'lady' appears in comment,* **then**:
7: f variable +1
8: **end if**
9: **end for**
8: **if** *m is larger than f* **then**: the professor is assigned as 'male'
9: **if** *f is larger than m* **then**: the professor is assigned as 'female'
10: **Otherwise**, the professor is assigned as N/A
11: **end if**
12: **end procedure**

Exhibit 3. Dataset Overview

Total Number of Reviews	7782
Max Number of Reviews of One Professor	125
Number of Professors	738
Male	610
Female	128

College	Reviews	Professors	Average Number of Reviews	US News. Math Ranking
Stanford University	153	34	5	2
University of Chicago	153	33	5	6
UCLA	365	68	5	7
Yale University	42	10	4	9
Cornell University	1330	130	10	13
Carnegie Mellon University	665	52	13	32
Washington University of St.Louis	924	54	17	34
University of Notre Dame	68	20	3	39
USC	1472	110	13	44
Emory University	601	69	9	55
Tufts University	777	60	13	74
Tulane University	1210	98	12	74

Exhibit 4. Cosine Similarity Matrix

	CMU	Emory	Stanford	Tufts	Chicago	Notre Dame	WashU	Yale
CMU	1	0.98	0.98	0.99	0.97	0.98	0.99	0.96
Emory	0.98	1	0.96	0.99	0.94	0.97	0.99	0.94
Stanford	0.98	0.96	1	0.97	0.98	0.97	0.97	0.96
Tufts	0.99	0.99	0.97	1	0.96	0.98	0.99	0.95
Chicago	0.97	0.94	0.98	0.96	1	0.96	0.96	0.95
Notre Dame	0.98	0.97	0.97	0.98	0.96	1	0.97	0.95
WashU	0.99	0.99	0.97	0.99	0.96	0.97	1	0.95
Yale	0.96	0.94	0.96	0.95	0.95	0.95	0.95	1

Exhibit 5. Euclidean Distance Matrix

	Stanford	Chicago	UCLA	Yale	Cornell	CMU	Washington	Notre Dame	USC	Emory	Tufts	Tulane
Stanford	0	0.181	0.456	0.74	0.802	0.577	0.998	0.617	0.969	0.534	0.706	0.93
Chicago	0.181	0	0.466	0.656	0.79	0.531	1.007	0.494	1.033	0.521	0.63	0.965
UCLA	0.456	0.466	0	0.968	0.427	0.453	0.701	0.756	0.645	0.241	0.675	0.557
Yale	0.74	0.656	0.968	0	1.24	0.914	1.238	0.362	1.45	0.985	0.941	1.342
Cornell	0.802	0.79	0.427	1.24	0	0.455	0.629	1.027	0.538	0.306	0.6	0.432
CMU	0.577	0.531	0.453	0.914	0.455	0	0.638	0.775	0.744	0.244	0.349	0.643
WashU	0.998	1.007	0.701	1.238	0.629	0.638	0	1.139	0.511	0.597	0.922	0.35
Notre Dame	0.617	0.494	0.756	0.362	1.027	0.775	1.139	0	1.32	0.797	0.791	1.192
USC	0.969	1.033	0.645	1.45	0.538	0.744	0.511	1.32	0	0.607	0.996	0.222
Emory	0.534	0.521	0.241	0.985	0.306	0.244	0.597	0.797	0.607	0	0.5	0.5
Tufts	0.706	0.63	0.675	0.941	0.6	0.349	0.922	0.791	0.996	0.5	0	0.908
Tulane	0.93	0.965	0.557	1.342	0.432	0.643	0.35	1.192	0.222	0.5	0.908	0