

Abstract—Online reviews have a critical impact on e-commerce business. While many studies have been done on the various characteristics of online reviews, in this paper we present a preliminary experimental analysis on RateMyProfessors, a well-known review site that allows college students to post reviews and assign ratings to professors. We collected online review data from RateMyProfessors and compared them with assumptive ground truth. Our analysis suggests an evaluation bias where RateMyProfessors ratings tend to be more negative. About 76% of professors in general and 90% of professors teaching hard courses are negatively affected.

Index Terms—online reviews, evaluation bias, RateMyProfessors, big data analytics, data mining

I. INTRODUCTION

As a mass-collaborative way of evaluating product quality or performance, nowadays online reviews are ubiquitous in e-commerce such as Amazon, Home Depot and Best Buy. There are also numerous websites that collect and publish crowd-sourced online reviews about businesses, physicians, and college professors such as Yelp, Healthgrades and RateMyProfessors. Online reviews and ratings have a critical impact on e-commerce business. According to research from Nielsen, 73% of online respondents use reviews to make purchase decisions. According to a survey from marketing agency Fan & Fuel, 92% of customers are impacted and less likely to buy if online reviews are not available.

Many studies have been done on the fairness and other characteristics of online reviews. In this project, we focus and perform preliminary experimental analysis on RateMyProfessors (www.ratemyprofessors.com), a review site that allows college students to post reviews and ratings to professors and campuses of American, Canadian, and UK institutions. We collected review data from RateMyProfessors and compared them with assumptive ground truth so as to verify our observations and hypothesis that there could be a systematic evaluation bias in RateMyProfessors.

For each professor, we can define ground truth evaluation based on aggregating reviews/ratings from the entire population. Statistically, a population is a complete set of items that share at least one property in common that is the subject of a statistical analysis. In our case, a population can be considered as the entire group of students who have taken courses from the professor being evaluated. Based the law of large numbers [1], a truthful approximation to ground truth can be obtained if the sample size is large where each professor receives a large number of reviews/ratings. This way, the sample would well represent the distribution of various standards, preferences, opinions, and evaluations of the population.

Observations: We observe that in RateMyProfessors, on average each professor profile receives a small number of reviews, which form a small sample of the population that can easily suffer from statistical instability. As a consequence, the overall evaluation is less robust and is vulnerable to biased reviews, leading to significant deviation from ground truth.

We also observe that the small samples from RateMyProfessors are ill-distributed heavily representing students with strong/extreme opinions. As a consequence, the student ratings from RateMyProfessors would show high variance and form a J-shape [2] similar to many online review platforms. This observation can be explained by the fact that there is a certain degree of overhead/barrier in posting online reviews. To overcome this barrier, students tend to be motivated with strong opinions and feelings about the professors involved.

Hypothesis: We hypothesize that there is an evaluation bias in RateMyProfessors, where negative strong reviews disproportionately out-weight positive ones, leading to bias towards more negative overall evaluations compared to ground truth. This evaluation bias can possibly be explained as follows.

First of all, unlike the cases of reviewing products, restaurants or physicians where little to none face to face interaction exists between evaluators and evaluatees, students and professors communicate and interact in person on a daily basis. Strong positive feelings from students can find many channels to be expressed directly. For example, students would commonly show appreciation to professors in person, via emails, or by sending thank-you cards. By human nature and culture it is uncommon and unnecessary to express positive feelings anonymously. However, such transparent channels are impractical for strong negative feelings, making RateMyProfessors or similar sites a preferred venue for their release.

In addition, in the case of students reviewing professors, there is a stronger than usual conflict of interest. The fact that RateMyProfessors allows students to post ratings and...
comments after the semester ends provides an opportunity for some disgruntled and frustrated students to “seek revenge”. In social psychology, self-serving bias is a defense mechanism to protect self-esteem. It describes humans’ tendency to blame external forces (and feel unfair) when bad things happen and to give ourselves credit (and feel fair) when good things happen [3]. Such blame-shifting [4] is particularly common and intense in school settings where student performance can directly impact their career and is constantly judged by professors. In addition, professors teaching hard courses may face increased student frustration leading to increased blame-shifting and inaccuracy in RateMyProfessor ratings.

II. DATA COLLECTION
A. RateMyProfessors Data
RateMyProfessors data were collected directly from the RateMyProfessors site using a Web crawler named ProfessorXBot (academicintegritygroup.com). Each professor profile is parsed and stored in a central database on Amazon Web Services. At the time of this study, 100% of the profiles (∼1.73M) on RateMyProfessors have been downloaded, processed, and stored. Parsing RateMyProfessors profile data is not straightforward because the site uses a complex dynamic Web application structure. We omit the technical details here due to the page limit.

To perform feature-specific analysis, we further augmented RateMyProfessors profiles with gender and discipline labels. In particular, we manually labeled each profile as STEM or Non-STEM. Gender label assignment was based on analysis of first names, where we used birth records in the United States from 1909 to 2018. Using this data, a probability score was derived based on total number of babies born per name and per gender. Any score above 92% was considered reliable to assign gender to a first name. Where this was not the case, the profile was researched manually to determine gender.

From the total of ∼1.73M profiles, 250K (each with ≥ 3 ratings) were extracted randomly for study. The data sample was further broken down based on gender (54% Male and 46% Female) and discipline (34% STEM and 66% Non-STEM).

B. Ground Truth Data
To establish a close approximation to ground truth, we used official college semester student evaluation data. All accredited colleges and universities mandate end-of-semester student course evaluations. These anonymous evaluations are conducted under strict guidelines and the participation rate is usually quite high. In the literature, such official college evaluations are considered truthful and accurate when a sufficiently large number of students participate [5], [6].

For the purpose of this study, we selected the University of South Florida as our data source (https://fair.usf.edu/EvaluationMart/), which makes their official student evaluation data public. 250 professors were randomly chosen with a selection criterion that each has at least 2 reviews from the corresponding RateMyProfessor profile. For each professor, all evaluation data from all semesters were acquired manually and compiled. For these 250 professors, the average total number of official student evaluations is 468, whereas their corresponding average number of total ratings from RateMyProfessors is 15.9.

The ground truth data sample was also augmented with gender and discipline labels in the same manner as the RateMyProfessor data described in Section II-A.

III. DATA ANALYSIS
A. RateMyProfessors vs Ground Truth
Table I shows statistics of RateMyProfessors (RMP) data in comparison with those of Ground Truth (GT) data with varying difficulty levels. The difficulty scores are extracted from RateMyProfessors reflecting student perceptions of difficulty on the courses being reviewed. Each professor’s RMT rating and GT rating are obtained by averaging all of his ratings from RMP data and GT data respectively. RMP rating and GT rating columns in Table I indicate average RMP rating and average GT rating for all the professor profiles in the specified category (row). From the table we can see that overall RMP rating (3.82) is significantly lower than GT rating (4.25) with a 10% decrease, where 69% of professors are negatively affected. This analysis shows that RateMyProfessors has a systematic evaluation bias towards negativity.

In addition, with the increase of difficulty level, RMP rating drops much faster than GT rating. For the difficulty level of [4, 5], RMP rating has a shocking 34% decrease over GT rating and 90% of professors are negatively affected. This analysis shows that professors teaching hard courses receive much more mistreatment from RateMyProfessors.

The standard deviation for RMP ratings of the 250 professors is much higher at 0.94 compared to that of GT ratings at 0.49. This analysis shows that strong/extreme opinions are over represented in RateMyProfessors.

<table>
<thead>
<tr>
<th>Difficulty level</th>
<th># Profiles</th>
<th>Average difficulty</th>
<th>RMP rating</th>
<th>GT rating</th>
<th>Rating drop</th>
<th>Affected professors</th>
<th>RMP stdev</th>
<th>GT stdev</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 2)</td>
<td>38</td>
<td>1.55</td>
<td>4.40</td>
<td>4.49</td>
<td>-2%</td>
<td>61%</td>
<td>0.56</td>
<td>0.38</td>
<td>0.61</td>
</tr>
<tr>
<td>[2, 3)</td>
<td>102</td>
<td>2.43</td>
<td>4.19</td>
<td>4.37</td>
<td>-4%</td>
<td>60%</td>
<td>0.64</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>[3, 4)</td>
<td>89</td>
<td>3.36</td>
<td>3.43</td>
<td>4.07</td>
<td>-16%</td>
<td>78%</td>
<td>0.97</td>
<td>0.51</td>
<td>0.66</td>
</tr>
<tr>
<td>[4, 5]</td>
<td>21</td>
<td>4.39</td>
<td>2.61</td>
<td>3.96</td>
<td>-34%</td>
<td>90%</td>
<td>0.84</td>
<td>0.61</td>
<td>0.30</td>
</tr>
<tr>
<td>[1, 5)</td>
<td>250</td>
<td>2.79</td>
<td>3.82</td>
<td>4.28</td>
<td>-10%</td>
<td>69%</td>
<td>0.94</td>
<td>0.49</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Correlations are computed between RMP ratings and GT ratings for professors in the specified categories, from which we can see that overall the two only have a moderate (not as high as one would expect assuming no evaluation bias) positive relationship. For the difficulty level of [4, 5], the correlation is fairly weak at only 0.3. This analysis is consistent with previous conclusions about the evaluation bias and professors teaching hard courses.

### Table II
**RateMyProfessors vs Ground Truth: Number of Ratings**

<table>
<thead>
<tr>
<th># Ratings</th>
<th># Profiles</th>
<th>RMP rating</th>
<th>GT rating</th>
<th>Rating drop</th>
<th>Affected professors</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 2</td>
<td>250</td>
<td>3.82</td>
<td>4.25</td>
<td>-10.0%</td>
<td>0.69%</td>
</tr>
<tr>
<td>≥ 4</td>
<td>204</td>
<td>3.79</td>
<td>4.24</td>
<td>-10.6%</td>
<td>0.72%</td>
</tr>
<tr>
<td>≥ 7</td>
<td>143</td>
<td>3.69</td>
<td>4.20</td>
<td>-12.1%</td>
<td>76%</td>
</tr>
<tr>
<td>≥ 15</td>
<td>72</td>
<td>3.70</td>
<td>4.23</td>
<td>-12.4%</td>
<td>79%</td>
</tr>
</tbody>
</table>

Table II shows statistics of RMP data and GT data with varying number of ratings from RateMyProfessors. From the table we can see that for slightly larger number of ratings (e.g., ≥7) where the RateMyProfessors samples are more self-robust, stabilized and truthful with less noise, RMP ratings are on average ~12% lower than GT ratings with more than 76% of professors negatively affected.

### Table III
**RateMyProfessors vs Ground Truth: Disciplines**

Table III shows statistics of RMP data and GT data for both STEM and Non-STEM disciplines with varying number of ratings. From the table we can see that STEM profiles suffer more mistreatment from RateMyProfessors with bigger rating drops. Also, RMP ratings for STEM profiles are less correlated with GT ratings. This conclusion is consistent with that from Table I, which makes sense as STEM courses are generally considered more difficult.

### B. RateMyProfessors 250K Data Sample

For additional insight, we also studied a relatively large sample from RateMyProfessors data containing 250K profiles that were randomly selected each having at least 3 reviews.

Table IV shows statistics of the 250K sample for various difficulty levels. From the table we can see that: (1) STEM courses are indeed considered difficult with an average difficulty score of 3.07 compared to 2.76 for non-STEM courses. (2) STEM profiles have a lower average rating of 3.63 and higher standard deviation of 0.99 compared to 3.88 and 0.9 for non-STEM profiles. (3) There is a moderate negative correlation between average ratings and difficulty levels across all categories, showing a similar conclusion to that from Table IV. (4) There is no significant difference between male and female profiles.

### References