Analyzing Correlations Between Student Evaluations and Course Grades

First Author
Kriti Varghese: kvarghese@ucdavis.edu
Qui Le: qnle@ucdavis.edu
Nausheen Sujela: nsujela@ucdavis.edu

Abstract

It has been acknowledged that there might be a correlation between the average grade in a class and the number of negative reviews the professor has on Rate My Professors. After obtaining a corpus of over 70,000 student reviews of various instructors from Rate My Professors, we studied the correlation between the positivity of the reviews and the average grade to see if this was true. Our findings led us to conclude that there is no correlation between negative reviews and lower average grades. The most common grades for both negatively reviewed professors and positively reviewed professors were A and A-.

1 Credits

The code and analysis has been submitted in a zip file. The link to the training and testing sets are here:

https://drive.google.com/open?id=1YX1JGixTGRoTi48D0QBMVWBGTjrjo6QU

Kriti Varghese: ReviewTraining.py, ReviewTest.py, rmp_test.csv
Qui Le: average_grade.py, average_grade.csv
Nausheen Sujela: rmp_scraper.py, rmp_dump.csv
All: Results_Analysis_Graphs.xlsx, this report

2 Introduction

It is widely conjectured among students and teachers alike that the worse the average grade in a class, the more likely a professor is to receive negative reviews on RateMyProfessors.com. We aim to determine empirically with an extensive amount of data whether or not this is true.

After assembling a large corpus of RMP reviews ($>$ 70,000 individual student reviews for 5000 professors), we have analyzed the relationship between the average grade assigned by a professor and the predominant sentiment expressed by student reviewers for that professor. For our purposes, we have established two sentiment categories: positive and negative.

We chose to investigate this question for our final project due to widespread student reliance on RMP for selecting classes and instructors. We believe it is critical to assess the factors that influence student evaluations. This allows educators (professors, lecturers, and advisors) and perhaps even interested students to understand RMP behavior and how student grades correlate with reviews.

3 Data Collection

We searched extensively for an existing RMP dataset. Unfortunately, we could not find a single dataset to suit our needs. As a result, we built a script to scrape 70,000+ reviews from 5000 professors. The details of this script and the resulting dataset are discussed below.

3.1 The Script

The script utilized the BeautifulSoup Python library for extracting text from RateMyProfessors.com and outputting over 70,000 RMP reviews into a CSV file called rmp_dump.csv. We were able to extract the reviews as well as additional information corresponding to each review: the professor, the university, the grade received, the ”would take again” field (which specifies whether the student would take the professor/class again) as well as the student quality and difficulty ratings.

We used the student's response to ”Would Take Again” as well as their overall quality rating (which is on a scale of 1-5) to annotate the review as ”positive” or ”negative.” If the student marked ”Would Take Again” as ”Yes” *or* if they listed a quality rating of $>=$ 3, we marked the review as ”positive.” If neither of these conditions were met, we marked the review as ”negative.”

3.2 The Dataset

The training dataset scraped from RateMyProfessors.com contains the following columns: University Name, Professor Name, Average Rating (of the Professor), Review, Grade Received, Quality, Difficulty, Would Take Again, and Annotation. Each row of the dataset corresponded to a single review.

Ratings
The ”Average Rating” is located at the top of each professor’s dedicated RMP page. As its name suggests, it is an average of all student quality ratings that have been submitted.

These student quality ratings (and the resulting Average Rating) range from 1-5, with 1 being the poorest score and 5 being the highest.

Reviewers also have the option of leaving a ”Difficulty Score” to denote how challenging they found the professor and class to be. This score ranges from 1-5 with 1 being the easiest and 5 being the most difficult.

Grade Received
The ”Grade Received” ranges from the usual A-F and includes other options as well, such as ”Not sure yet” and ”N/A.”

Would Take Again
The ”Would Take Again” field represents whether or not a student enjoyed a professor’s instruction enough to retake them. Students can select ”Yes”, ”No”, or ”N/A.”

Annotation
How we annotated each review is discussed above in Section 2.1

4 Sentiment Analysis of Reviews

The code for the sentiment analysis of the reviews was done using the Naive Bayes model. Openpyxl was used after converting the csv dump to an excel spreadsheet, to retrieve the data. Then, using a similar code to HW2, We created the model. We used Openpyxl again to access the data in the test set and then created a file for each professor that would contain that professor’s reviews. We then ran the Naive Bayes model on each file to figure out if the professor’s reviews were overall positive or negative and put the results in another excel sheet using XlsxWriter, a python package that allows you to write excel workbooks more efficiently than Openpyxl (which is more efficient for reading data).

This model accurately predicted overall negatively-reviewed professors but didn't accurately predict positive reviews. This is probably because the training set contained more negative reviews than positive reviews, which could either be a characteristic of reviews on Rate My Professors or an issue with the annotation of the training set (more likely the latter). As there were more negative reviews, there were more words classified as negative so during testing, more words would be predicted to be negative and, by extension, more reviews would be predicted to be negative.

The accuracy of this model was 93%. The precision, recall and F1 score for the positive classification were 0. For the negative classification, the precision, recall and F1 score were 0.976, 0.954 and 0.964 respectively. As mentioned before, the reason for this disparity is most likely the disparity in the amount of negative and positive reviews in the training set.

We ran this model without add one smoothing as there were no words that gave a probability of 0. We expected this as most of the reviews on the website are by college students who would all have a similar vocabulary. In the future, manually annotating the training set would improve the evaluation of positive reviews.

5 Average Grade Analysis

One of the key things for the task of finding the average grade from all the reviews of each professor is to establish a grade scale. This will be essential for us to do the translation between letter grades and estimate percentage points, and create data graphs in Excel for further analysis.

One unexpected difficulty was encountered in understanding how CSV File Reading and Writing module of Python work generally. Extracting grades of students who left reviews for one particular professor and storing them into a plain text was also a difficult process to figure out.

After getting all grades for each professor, we also see that there are many reviews that did not provide the grades for the professor. Thus, calculating the average grade for such a professor would run into division-by-zero error. For a lot of professors, all the reviews do not provide the grades received in the class, and it shows in our data with very high number of N/A. The large number of N/A is particularly not helpful in determining the correlation between average grade and sentiments that are expressed in reviews, and it is a big reason why we could not confirm our initial hypothesis which is that there is a correlation between the sentiment of a professor's reviews and average grades received by that professor's students.

6 Results and Future Work

After obtaining data about reviews and average grades, we do not see a strong correlation between grades and either positive or negative reviews. In both categories, the most common grades are A and A-, except for an unexpectedly high number of N/A. Moreover, having a large amount of negative reviews without having a comparable amount of positive reviews was not beneficial in training our classifier to predict the correct sentiment for each review.

In possible extensions of this project, we would like to look more at the reviews and see how students' online use of language is different from in-person conversations.

We are interested to know how students address professors when they are online. Also, by studying the language of online reviews using concordance, we would want to learn about what adjectives students use to describe their experiences with class materials, office hours, exams, projects, and assignments. We would attempt to tag the corpus with parts of speech for that purpose using POS tagger, such as the one provided by Natural Language ToolKit.

If we were to make a second attempt at this project, we would pay closer attention to the number of positive reviews and the number of negative reviews we have in our training dataset. In our current training set, the majority of reviews are negative. This may have influenced the correlation analysis as we had too few data points for positive reviews.

While we were able to gauge a sufficient amount of information about students that leave negative reviews, we did not enjoy the same luxury regarding those leaving positive ones. In a second attempt, we would ensure a greater number of positive reviews in order to conduct a fairer analysis.

We would also try to discard reviews with grades listed as ”N/A” or ”Not sure yet” as these reviews provide no useful information in answering our question about the correlation between grades and review sentiment.

7 Acknowledgements

We would like to express our gratitude towards everyone who helped make it possible to complete this project. We want to especially acknowledge our instructor, Professor Kenji Sagae, for his constant support.