Analysing RateMyProfessors Evaluations across Institutions, Disciplines, and Cultures: The Tell-tale Signs of a Good Professor

Mahmoud Azab, Rada Mihalcea, Jacob Abernethy

University of Michigan
2260 Hayward Street
Ann Arbor, MI 48109, USA
{mazab,mihalcea,jabernet}@umich.edu

Abstract. Can we tell a good professor from their students' comments? And are there differences between what is considered to be a good professor by different student groups? We use a large corpus of student evaluations collected from the RateMyProfessors website, covering different institutions, disciplines, and cultures, and perform several comparative experiments and analyses aimed to answer these two questions. Our results indicate that (1) we can reliably classify good professors from poor professors with an accuracy of over 90%, and (2) we can separate the evaluations made for good professors by different groups with accuracies in the range of 71-89%. Furthermore, a qualitative analysis performed using topic modeling highlights the aspects of interest for different student groups.

1 Introduction

Assessing teaching quality is a difficult and subjective task. Most if not all schools evaluate their professors by asking students to provide course feedback, which often consists of ratings as well as open-ended comments in response to several prompts. With few exceptions, this feedback is kept confidential and is shared with neither current nor prospective students. It is therefore not surprising that the Web 2.0 wave has brought several sites that encourage students to share their in-class experiences and the opinions they hold on the professors teaching their courses. Among these sites, the one that is by far the most popular is RateMyProfessors[1] (RMP), where students can anonymously rate different aspects of their professors (i.e., clarity, helpfulness, easiness), and also provide open-ended comments. The site currently has approximately 15 million evaluations for 1.4 million professors from 7,000 schools in the United States, Canada, and United Kingdom. Students appear to have confidence in the RMP ratings and there is evidence that they use the site to make academic decisions [5].

In this paper, we analyze the language used by students when discussing their professors. Using a large collection of 908,903 RMP comments collected

[1] http://www.ratemyprofessors.com/

for 71,404 professors from 33 different institutions, our study aims to answer the following two questions. First, can we use automatic text classification to distinguish between professors regarded as good vs. professors regarded as poor? After several feature selection experiments, we show that we can reliably separate good professors from poor professors with an accuracy of over 90%.

Second, and perhaps more interestingly, we ask whether there are differences between what characterizes a good professor across different groups. To answer this question, we focus exclusively on the good professors in our dataset, and specifically look for differences across disciplines (e.g., Sociology vs. Computer Science), across institutions (top-ranked vs. low-ranked schools), and across cultures (U.S. vs. Canada). We perform a quantitative analysis of these differences by performing automatic classification of good professor comments contributed by different groups using domain-independent features, and show that we can achieve classification accuracies in the range of 71-89%, suggesting that different students value different aspects of a good professor. To understand these differences, we use topic modeling to perform a qualitative analysis through comparisons between the distributions of several topics in the students comments. This analysis leads us to several interesting findings, e.g., computer science students appear to exhibit greater appreciation for a professor's clarity, while philosophy students are more concerned with readings and discussions, and so on.

2 Related Work

While there is no previous work that we are aware of in the field of natural language processing focusing on the analysis of RMP student evaluations, there are several studies in fields such as education and sociology. These studies confirmed the validity of RMP evaluations and found significant correlations between RMP rating scores and their corresponding scores in official student evaluations of teaching for professors from different schools [5, 19, 14, 6]. There are also studies on the intercorrelations among RMP rating scores. For instance, RMP overall quality score is highly correlated with the easiness and the physical attractiveness of the professor [8, 7]. Freng and Webber [9] also showed that attractiveness is responsible of 8% of the variance in the data.

The study that is closest to our, although not computational, is the one by Helterbran [10], who manually analyzed RMP comments for 283 instructors from three universities in Pennsylvania, and identified certain personal attitudes and instructional behaviors that are most beneficial to students, such as being knowledgeable and approachable. This study was limited in terms of the numbers and institutions studied, and did not have discipline and cultural diversity.

Also related to our work is research on opinion mining and sentiment analysis, which is a well-established area in natural language processing. It has been approached at different levels of granularity from document- to sentence- to phrase-level sentiment classification [20, 15, 11, 21, 1]. The nature of the examined data varied from online products and movie reviews to opinions posted on microblogs like Twitter [2]. These studies used different machine learning techniques for classification such as Naive Bayes and Support Vector Machines with different sets of features such as unigrams and bigrams. To the best of our knowledge, no previous work has tackled students' evaluations.

High-level classification of students opinions is not enough to understand what are the instructional behaviors that students care about the most. We found inspiration in recent work on topic modeling, which has been successfully used to extract personal values and behaviors from open-ended text [4], or to integrate expert reviews with opinions scattered over the Web in a semi-supervised approach [12].

3 Dataset

The study reported in this paper is based on a corpus compiled from the RMP site. Our goal was to build a dataset of professors and their evaluations from a diverse pool, covering institutions with different academic rankings, covering different countries, and also covering different disciplines.

The crawl, made during the summer of 2015, was started by specifying a list of 33 schools. When constructing this list, we considered the academic ranking of the schools according to the U.S. News ranking. We included 10 U.S. top-ranked public schools, such as the University of California Berkeley and the University of Michigan, 10 low-ranked public schools, as well as 4 additional U.S. public schools.

We also considered the country of each institution, and in addition to the 24 U.S. schools, we included 9 schools from Canada, such as the University of Toronto and University of Montreal.

We collected the records of every professor affiliated with each school, covering all 33 schools, which in aggregate provided a very diverse set of faculty disciplines. For each professor, we then collected the entire set of their students' ratings. Finally, we removed ratings that had the comment field left blank and also the professors who received no comments. The resulting dataset consists of 908,903 evaluations with textual comments for 71,404 professors from 33 schools. Table 1 shows the distribution of professors and comments in our dataset.

	Professors	Evaluations
U.S. top-ranked	21,119	245,553
U.S. low-ranked	15,631	195,728
Canada	19,672	313,868

Table 1. Statistics on the RMP dataset.

In addition to specifying the professor and the class, each evaluation includes an optional comment, as well as several attributes, such as helpfulness, clarity, and easiness scores. These attributes can have a value between [1, 5], where 1 is the worst score and 5 is the best score. Each evaluation also receives an overall classification of good, average or poor, determined by RMP based on the helpfulness and clarity scores. For each professor, overall helpfulness and

Overall	Helpfulness	Clarity	Easiness	Department	Comment
Good	4	5	4	Economics	Uses real world examples to make lectures more interesting. Clear and concise. Recommended.
Poor	1	2	1	Computer Science	Bad at explaining material, doesn't seem to care about individuals.
Good	5	3	2	Statistics	Statistics requires that you work for it, so be prepared to work for this.

Table 2. Sample RMP evaluations.

clarity scores are also calculated, as the average of all the helpfulness and clarity scores given to this professor by the students. Finally, RMP calculates the overall quality score of a professor as the average of her overall helpfulness and clarity scores. Table 2 shows examples of RMP evaluations.

In all our experiments, we use a random split of the dataset into training and test, consisting of 57,150 and 14,254 professors respectively. The comments are also split based on the professors they belong to. Therefore, a professor and her corresponding comments exist in either the training or the test set, but not in both. We do not balance the data because in our analyses we want to capture as many aspects and concerns in students' comments as possible. Balancing the data might result in a loss of important information.

4 Can We Tell a Good Professor?

Our first set of experiments is concerned with determining whether the textual comments from RMP can be used to automatically predict the overall classification of an individual comment or of a professor as either "good" or "poor" (see below for an explanation of these labels). This task is akin to that of sentiment analysis, in that we use the text of a comment to predict whether that comment is reflective of a "good" or a "poor" student evaluation (comment-level classification); or, we use the text of all the comments submitted for a professor to predict if that professor is rated as "good" or not (professor-level classification). These experiments, along with the feature selection discussed in Section 4.1, allow us to determine the words that have high predictive power in students' textual comments, which are necessary for our analyses to understand the characteristics of good professors.

To represent the text, we extract features consisting of unigrams, bigrams, and a mix of unigrams and bigrams. Each instance in our dataset (whether an individual comment or a professor) is thus represented as a feature vector encoding the counts of the n-grams in the representation.

In addition to raw n-gram features, we also experiment with the use of sentiment/emotion lexical resources. Specifically, we use the following lexicons: OpinionFinder [21], which includes 2,570 words labeled as positive and 4,581 words as negative; a subset of WordNet Affect [18], with 1,128 words grouped into six basic emotions: anger, disgust, fear, joy, sadness and surprise; and General Inquirer [17], with 29,090 words mapped to 96 categories. We first filter the input text based on these lexicons by removing words that do not exist in the lexical resources and then generate unigram and bigram features from this filtered text.

To identify the most distinctive lexical features in the students' comments, we use feature selection, as described below. The features are then used in a multi-nomial Naive Bayes classifier; we also ran experiments using a Support Vector Machine classifier, but its performance was significantly below that of the multi-nomial Naive Bayes.

Note that all our experiments exclusively rely on the text in the comments, and are not making use of the other attributes available on the RMP site (helpfulness, clarity, easiness) in any ways.

4.1 Feature Selection

We experiment with two feature selection methods to identify the most useful features for our task. The methods are compared by using five-fold cross-validation on training data, and the best method is selected and applied on the test set.[2]

The first feature selection method is linear regression which, for each feature, uses uni-variate linear regression tests to compute the correlation between a target class and the data.

The second one is chi-square, which measures the degree of dependence between two stochastic variables: in our case, for each feature, we determine if there is a significant difference between the observed and expected frequencies in one or more target classes. For each feature selection method, we use their scores to rank the features, and keep the top K-percent features for the classification.

4.2 Comment-level Classification

In this initial experiment, we classify the individual comments as either "good" or "poor." We use the RMP overall *quality* rating, which is associated with each comment and can have one of the following values: good, average or poor. We only consider comments that are labeled as good or poor, and ignore those labeled as average.

To determine the training and test datasets, we use the random split mentioned in Section 3, ensuring that all the comments belonging to a professor are either in training or in test. Table 3 shows the distribution of the good and poor comments in the data. As seen in this table, the distribution is similar in both training and test, with 74% of the comments being labeled as good.

In order to tune the classifier and select the best set of features, we use five-fold cross-validation on the training data, and compare the accuracies obtained with the two feature selection methods (linear regression and chi-square) and different features (unigrams, bigrams, unigrams+bigrams, unigrams+bigrams prefiltered based on the lexicons). Fig. 1 shows the average accuracy obtained in this

[2] The feature selection methods and the machine learning algorithms used in this study have been implemented in Python using the Sci-kit Learn machine learning library [16]. We use a maximum document frequency of 0.5 and lowercased text. We also experimented with stemming but it was found to degrade performance.

cross-validation experiments on the training data for the top-K percentile of the features with an incremental step of size 2. The best accuracy is achieved using the top 18% of the mixed raw unigrams+bigrams features, ranked according to the chi-square test. Interestingly, the features based on the sentiment/emotion lexicons do not perform as well as the raw features, which may suggest that student comments are different from the opinions/reviews previously used in sentiment analysis research. We use these top 18% features to train and test our final classifier. Tables 4 and 5 show that our classifier achieves significantly higher accuracy, precision, recall, and f-score than a majority class baseline.

	Training	Test
Comment-level		
Good	471,566	117,816
Poor	165,593	40,631
Professor-level		
Good	36,958	9,265
Poor	8,615	2,152

Table 3. The distribution of the training and test data in the comment- and professor-level classification experiment.

4.3 Professor-level Classification

In a second experiment, instead of classifying individual comments, we now classify professors as either “good” or “poor.” To represent a professor, we use all the comments submitted for that professor. To label a professor as good or poor, we use the overall score field that is calculated by RMP for each professor. We consider a professor with an overall rating score of $\geq$3.5 as good, and a professor with an overall rating score of $\leq$2.5 as poor.

As before, we use the training/test split described in Section 3. Table 3 shows the distribution of professors labeled as good/poor in the data. Once again, the numbers indicate that the class distribution is similar in training and test, with 81% of the professors being labeled as good. We use the same approach as in the comment-level experiment to tune the parameters of this classifier, and run five-fold cross validation experiments on the training data. Fig. 2 shows the average accuracy for different methods using the top-K percentile of the features with an incremental step of size 2. The best accuracy is achieved using the top 4% of the unigrams+bigrams features with a chi-square test. This suggests that there are words that are not included in the lexical resources that can distinguish good from poor professors. We use this setting to train our final classifier, and evaluate it on the test data. The final result, shown in Tables 4 and 5, indicates that we can reliably distinguish between good and poor professors, with an accuracy, precison, recall, and f-score significantly higher than the majority class baseline.

Not surprisingly, the accuracy obtained in the professor-level classification is higher than the one obtained by the comment-level classifier. Although the num-

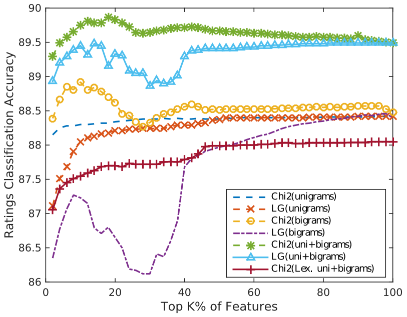

Fig. 1. The performance of different feature selection methods using different top-K lexical features (comment-level)

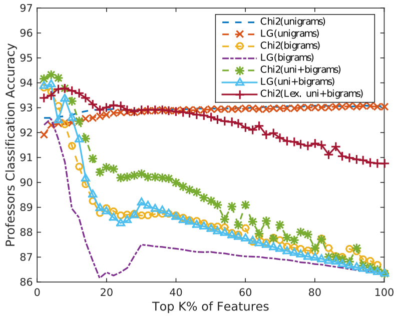

Fig. 2. The performance of different feature selection methods using different top-K lexical features (professor-level)

	Majority class	Multinomial Naive Bayes
Comment-level	74.35%	90.09%
Professor-level	81.15%	94.14%

Table 4. Comment- and Professor-level classification accuracy on test data.

ber of training instances is larger in the comment-level classifier, the professor-level classifier benefits from more data available for each instance, and also from a higher baseline.

To provide some insight into the features that play a significant role in the classification, Table 6 lists the top ten features for each class obtained from the professor-level classifier, ranked in reverse order of their chi-square weight. The Naive Bayes probability (i.e., P(feature|good), P(feature|poor)) was used to determine the class that each feature "belongs" to.

5 Can We Tell the Group Behind the Comments of a Good Professor?

The results presented in the previous section have shown that we can accurately classify a comment or a professor as either good or poor based on student language. While this is an interesting result in itself, we are also interested in finding whether there are differences between what is regarded as a good professor by different groups.

If we condition on professor quality, all else being equal, how well can we determine other particular factors of the faculty member in question, such as the rank of their institution, their discipline, or the country in which they teach? Our answers to these questions provide some insight into the complex attribute-specific components that determine the perception of professor quality. For in-

	Majority class			Multinomial Naive Bayes		
	Precision	Recall	F-score	Precision	Recall	F-score
Comment-level						
Good	74.36	100	85.29	91.44	95.62	93.48
Poor	–	–	–	85.36	74.04	79.30
Professor-level						
Good	81.15	100	89.59	95.84	96.99	96.41
Poor	–	–	–	86.33	81.88	84.04

Table 5. Comment- and Professor-level precision, recall and f-score of each class on test data.

Rating	Top features
Good	interesting, best, awesome, fun, funny, helpful, amazing, great teacher, great professor, highly
Poor	worst, avoid, horrible, terrible, teach, worst professor, worst teacher, useless, does, costs

Table 6. Top ten features associated with professors with a good/poor rating.

stance, are there differences between good professors in Canada vs. U.S.? Or good professors in Computer Science vs. Sociology?

In these experiments, we specifically focus on the “good” professors in our dataset, with an overall rating of 3.5 or higher similar to RMP criteria. We perform three different analyses: (1) cross-culture, where we separate good professors from U.S. schools vs. good professors at schools in Canada; (2) cross-institution, where we classify good professors from top-ranked vs. low-ranked public U.S. schools, according to the U.S. News ranking; and (3) cross-discipline, where we try to see if there are differences between good professors in different disciplines. For this third analysis, we work with three pairs of disciplines that are unrelated (Sociology vs. Computer Science), (Philosophy vs. Physics), and (Fine Arts vs. Biological Sciences); and one pair that is somewhat related (Management vs. Business Administration).

To create the experimental datasets for these analyses, we use the original training and test sets described in Section 3, and filter for the group of interest. For instance, to obtain the training dataset for Canada, we extract all the good professors from the large training dataset that are affiliated with a Canadian institution, and so forth. For the discipline datasets, we determine the discipline of the professor using the department name that the professor is affiliated with. Table 7 shows the number of good professors in our dataset, broken down for each of the groups mentioned above.

One difficulty with the classification of such groups is the presence of confounding factors: while our main goal is to identify differences between these groups in terms of what they appreciate in a good professor, the groups are also distinct because of culture-, institution-, or discipline- specific words. For example, the word “programming” is more likely to appear in comments made about Computer Science professors than in comments on Biology professors. Similarly, French words are more likely to be used in comments on professors at schools in Canada than in comments on professors at schools in the U.S. In

	Training	Test
Cultures		
Canada	9,463	2,395
U.S.	27,495	6,870
Institutions		
Low-ranked	8,139	2,059
Top-ranked	11,261	2,884
Disciplines		
Biological Sciences	203	49
Business Administration	122	29
Computer Science	674	182
Fine Arts	372	79
Management	236	45
Philosophy	793	195
Physics	539	141
Sociology	872	229

Table 7. Number of good professors in different groups.

order to disallow the classifier to use such words in the classification process, we impose on all these group classifiers the same set of features, consisting of the top 500 unigram features reversely sorted according to their chi-square weight obtained from the good vs. poor professor experiments, described in Section 4. Moreover, we manually revised these features, removing by hand all culture-, institution-, or discipline-specific words, to ensure that the feature set includes only general attribute words, e.g. “good,” “humorous,” or “knowledgeable.” We also normalized the words that are spelled differently in both Canada and the US, e.g. “favorite” and “favourite”.

Group Pair	Majority Class	Multinomial Naive Bayes
Canada vs. U.S.	74.15%	89.49%
Top- vs. low-ranked	58.35%	74.71%
Philosophy vs. Physics	58.03%	82.14%
Biological Sciences vs. Fine Arts	61.72%	89.06%
Sociology vs. Computer Science	55.72%	84.43%
Business Administration vs. Management	60.81%	71.62%

Table 8. Classification accuracy for different groups.

Table 8 shows the classification accuracy that our classification models achieve for each experiment. Table 9 shows the precision, recall, and f-score for each group in each classification experiment. The classification accuracies between these groups are statistically significant except for Business Administration vs. Management. Thus, it seems that the differences between the comments of different groups changes according to the (dis)similarity of the two disciplines they represent. These results indicate that the groups writing comments about good professors can be separated with an accuracy significantly higher than the base-

Group Pair	Majority class			Multinomial Naive Bayes		
	Precision	Recall	F-score	Precision	Recall	F-score
U.S.	74.15	100	85.16	91.91	94.1	92.99
Canada	–	–	–	81.85	76.24	78.95
Top-ranked	58.35	100	73.69	77.88	79.13	78.50
Low-ranked	–	–	–	70.09	68.53	69.30
Sociology	55.72	100	71.56	84.23	88.65	86.38
Computer Science	–	–	–	84.71	79.12	81.82
Philosophy	58.04	100	73.45	83.92	85.64	84.77
Physics	–	–	–	79.56	77.3	78.42
Fine Arts	61.72	100	76.33	90.12	92.41	91.25
Biological Studies	–	–	–	87.23	83.67	85.42
Business Administration	60.81	100	75.63	75.63	37.93	51.16
Management	–	–	–	70.00	93.33	80.00

Table 9. Precision, recall and f-score for each group.

line, which, given that the features used in the classification do not include any group-specific words, suggest that there are indeed differences between what is considered to be a good professor by different groups. For additional insight into these differences, Table 10 shows the top ten features for each group, according to their chi-square weight.

6 What are the Tell-tale Signs of a Good Professor?

The results of the experiments described in the previous section show clear differences between what is considered to be a good professor by different groups. However, the numbers by themselves do not say much about what the actual differences are. In order to gain a better understanding of what each group looks for in a good professor, we use topic modeling to determine the main topics of interest in the students comments, and consequently compare the distribution of these topics in different groups.

To perform topic modeling, we use the Latent Dirichlet Allocation (LDA) implementation provided in Mallet (a machine learning for language toolkit) [13], applied on the professor-level representation of the data. LDA is a generative model that in our case considers each professor as a mixture of a small number of topics, and assumes that each word in this professor's data are associated with one of the topics [3]. Consistent with the analyses in the previous section, aiming at identifying differences among good professors as regarded by different groups, we extract ten topics using the data corresponding to the "good" professors. Table 11 shows these topics, along with several sample words for each topic.

Starting with these ten topics, we determine their distribution in each of the groups considered in the previous section. Figures 3, 4, 5, 6, 7 and 8 show these distributions, leading to interesting findings.[3] For instance, students in Canada

[3] In each of these figures, the topic distributions for a group add up to 100% (e.g., the blue/dark and yellow/light columns in Fig. 3 each add up to 100%).

Group	Top ten features
Canada	prof, marker, profs, notes, textbook, fair, excellent, clear, approachable, best
U.S.	homework, credit, grader, book, papers, interesting, extra, guides, material, reading
Top-ranked	lecturer, office, ta, readings, clear, reading, interesting, engaging, fair, slides
Low-ranked	attendance, credit, help, extra, gives, work, study, notes, willing, book
Sociology	readings, reading, papers, paper, study, discussion, essay, attendance, loved, passionate
Computer Science	homework, comments, teach, guy, excellent, office, time, help, explains, mistakes
Philosophy	papers, readings, reading, essays, paper, essay, marker, discussion, discussions, boring
Physics	homework, problems, exams, curve, help, accent, office, book, solutions, extra
Biological Sciences	notes, exams, material, questions, prof, clear, study, understand, fair, textbook
Fine Arts	work, nice, inspiring, comments, does, help, awesome, teaching, best, little
Management	paper, boring, book, papers, excellent, essay, kept, teachers, dr, instructor
Business Admin	prof, arrogant, curve, fair, extremely, lecturer, clear, engaging, approachable, definitely

Table 10. Top ten features associated with good professors rated by different groups

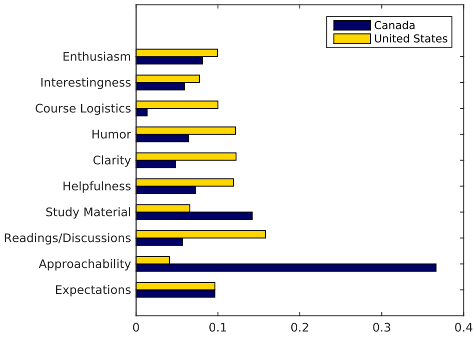

Fig. 3. Top topic distribution among good professors from U.S. schools vs. good professors from Canadian schools

seem to be more concerned with Approachability and Study Materials, whereas students from U.S. schools appear to talk more about Readings/Discussions and Clarity (Fig. 3). Students at top- and low-ranked U.S. public schools appear to be concerned with similar aspects of their good professors, with a somehow higher interest for Readings/Discussions and Clarity among students in top-ranked institutions, and more interest in Course Logistics among students in low-ranked schools (Fig. 4).

Topic	Sample words
Approachability	prof, fair, clear, helpful, teaching, approachable, nice, organized, extremely, friendly, super, amazing
Clarity	understand, hard, homework, office, material, clear, helpful, problems, explains, accent, questions, extremely
Course Logistics	book, study, boring, extra, nice, credit, lot, hard, attendance, make, fine, attention, pay, mandatory
Enthusiasm	teaching, passionate, awesome, enthusiastic, professors, loves, cares, wonderful, fantastic, passion
Expectations	hard, work, time, lot, comments, tough, expects, worst, stuff, avoid, horrible, classes
Helpfulness	helpful, nice, recommend, cares, super, understanding, kind, extremely, effort, sweet, friendly, approachable
Humor	guy, funny, fun, awesome, cool, entertaining, humor, hilarious, jokes, stories, love, hot, enjoyable
Interestingness	interesting, material, recommend, lecturer, engaging, classes, knowledgeable, enjoyed, loved, topics
Readings/ Discussions	readings, papers, writing, ta, interesting, discussions, grader, essays, boring, books, participation
Study Material	exams, notes, questions, material, textbook, hard, slides, study, answer, clear, tricky, attend, long, understand

Table 11. Ten main topics addressed in students comments, along with sample words.

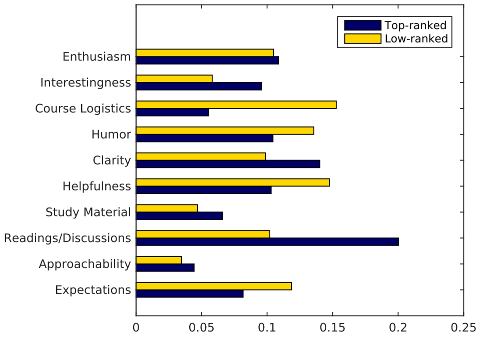

Fig. 4. Top topic distribution among good professors from top-ranked vs. low-ranked U.S. public schools

There are also differences among the aspects of interest for different disciplines. Sociology students talk more about Readings/Discussions, whereas Computer Science students focus more on Clarity (Fig. 5). A similar difference is observed between Philosophy and Physics (Fig. 6). Fine Arts students are more concerned with the Enthusiasm of their professors and tend to talk more about the Expectations of their classes; on the other hand, Biological Sciences students primarily talk about Course Logistics and Study Materials (Fig. 7). Finally, although Management and Business Administration are related disciplines, we note differences with Management students showing higher interest in Course

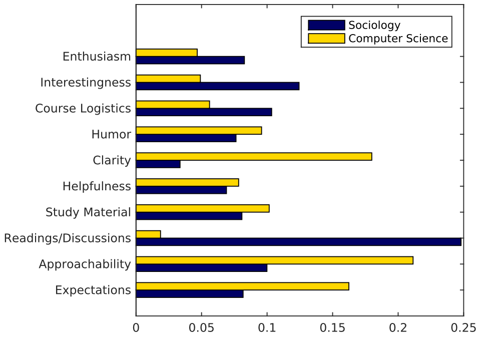

Fig. 5. Top topic distribution among good professors from Sociology vs. Computer Science

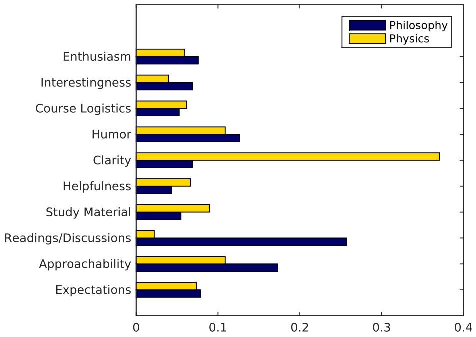

Fig. 6. Top topic distribution among good professors from Philosophy vs. Physics

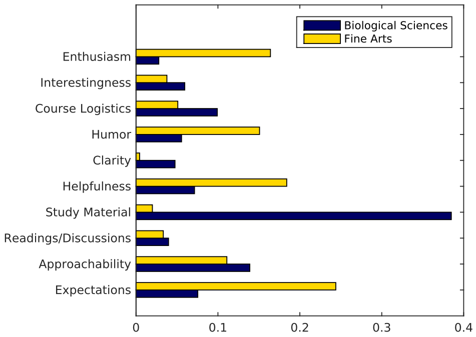

Fig. 7. Top topic distribution among good professors from Biological Sciences vs. Fine Arts

Logistics, and Business Management students talking more about Approachability and Enthusiasm (Fig. 8).

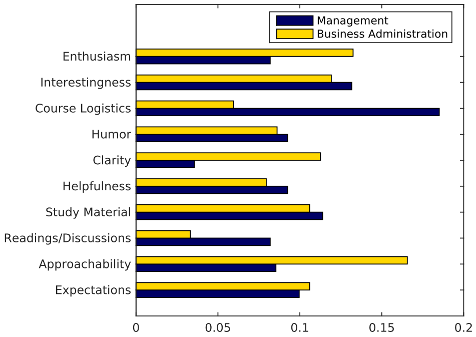

Fig. 8. Top topic distribution among good professors from Management vs. Business Adiminstration

7 Conclusion

In this paper, we explored a novel text processing application, targeting an analysis of the language used by students when evaluating their professors. Research work in the field of computational linguistics is typically divided into algorithms, data, and applications; our work falls under the applications category. We constructed a new dataset of 908,903 evaluations collected for 71,404 professors from 33 different institutions, covering different disciplines, different institutions, and two different cultures. We showed that we can reliably distinguish between good professors and poor professors with an accuracy of over 90%, by relying exclusively on the language of the students comments. Moreover, we performed experiments to determine if there are differences between what is regarded as a good professor by different student groups, and showed that we can separate between the comments made by students from different institutions, disciplines, or cultures, with accuracies in the range of 71-89%. Using topic modeling, we were able to identify the main aspects of interest in student evaluations, and highlighted the differences between the aspects appreciated more by different student groups.

We believe these results are interesting in themselves, as they clearly show differences in what is regarded as a good professor by different groups. Our findings can also be useful to professors, by enabling them to identify the aspects that matter to their students, so that they can improve the overall teaching quality.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation award #1344257, by grant #48503 from the John Templeton Foundation, and by the Michigan Institute for Data Science. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Michigan Institute for Data Science.

References

1. Agarwal, A., Biadsy, F., Mckeown, K.R.: Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. pp. 24–32. Association for Computational Linguistics (2009)
2. Bermingham, A., Smeaton, A.F.: Classifying sentiment in microblogs: Is brevity an advantage? In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1833–1836. ACM (2010)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022 (2003)
4. Boyd, R.L., Wilson, S.R., Pennebaker, J.W., Kosinski, M., Stillwell, D.J., Mihalcea, R.: Values in words: Using language to evaluate and understand personal values. In: Ninth International AAAI Conference on Web and Social Media (2015)
5. 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)
6. Coladarci, T., Kornfield, I.: Ratemyprofessors.com versus formal in-class student evaluations of teaching. Practical Assessment, Research & Evaluation 12(6), 1–15 (2007)
7. 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)
8. 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)
9. Freng, S., Webber, D.: Turning up the heat on online teaching evaluations: Does hotness matter? Teaching of Psychology 36(3), 189–193 (2009)
10. Helterbran, V.R.: The ideal professor: Student perceptions of effective instructor practices, attitudes, and skills. Education 129(1), 125 (2008)
11. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics. p. 1367. Association for Computational Linguistics (2004)
12. Lu, Y., Zhai, C.: Opinion integration through semi-supervised topic modeling. In: Proceedings of the 17th international conference on World Wide Web. pp. 121–130. ACM (2008)
13. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
14. 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)
15. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics. p. 271. Association for Computational Linguistics (2004)

16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
17. Stone, P.J., Dunphy, D.C.: The General Inquirer (1966)
18. Strapparava, C., Valitutti, A.: Wordnet affect: an affective extension of wordnet. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation. pp. 1083–1086 (2004)
19. Timmerman, T.: On the validity of ratemyprofessors.com. Journal of Education for Business 84(1), 55–61 (2008)
20. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 417–424. Association for Computational Linguistics (2002)
21. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. pp. 347–354. Association for Computational Linguistics (2005)