Detecting Objectifying Language in Online Professor Reviews

Angie Waller and Kyle Gorman
Graduate Center, City University of New York

Abstract

Student reviews often make reference to professors' physical appearances. Until recently RateMyProfessors.com, the website of this study's focus, used a design feature to encourage a "hot or not" rating of college professors. In the wake of recent #MeToo and #TimesUp movements, social awareness of the inappropriateness of these reviews has grown; however, objectifying comments remain and continue to be posted in this online context. We describe two supervised text classifiers for detecting objectifying commentary in professor reviews. We then ensemble these classifiers and use the resulting model to track objectifying commentary at scale. We measure correlations between objectifying commentary, changes to the review website interface, and teacher gender across a ten-year period.

1 Introduction

Natural language processing techniques have long been used to study subjectivity and sentiment in media and product reviews. In this study, we employ these technologies to study objectifying language in reviews of professors using archival data from RateMyProfessors.com (RMP). Detecting such language is difficult because it is somewhat rare, making up a small part of a small proportion of reviews (Davison and Price, 2009), and references to physical appearance show enormous linguistic variation (discussed in Section 2.2), making them difficult to detect accurately using simple text features.

This study provides insights into bias in professor reviews and their interaction with the design of the web user interface. We propose two models—a chunk tagger and a document classifier—used to build an ensemble to detect objectifying reviews at scale. This approach could be applied to many other domains where noisy user-generated reviews may contain harassment or exhibit harm.

We focus on the RMP website because it has been active for over twenty years, giving us ample data to study trends across time. The website has long been associated with students commenting on their professors' appearances (Lagorio, 2006) and has been the subject of many prior studies on bias in course reviews. Recent changes to the website interface allow us to consider how text reviews may have been influenced by its design feature for rating professor "hotness".

1.1 Prior work

We look to previous work on bias in professor reviews, effects of interface design on internet discourse, and detecting subjectivity and opinions in online reviews.

1.1.1 Bias and student reviews

Prior studies address bias among students' reviews of teachers. Freng and Webber (2009) find a positive correlation between "hotness" and quality scores of professors on RMP, accounting for 8% of variance. Chang and McKeown (2019) report gendered differences in students' descriptions of computer science professors on RMP which is also reflected in visualizations by Schmidt (2015) showing that words like *genius* are more frequently attributed to male professors and words like *nurturing* to female professors. This is supported in work by Boring et al. (2016) and Boring (2017) where in-class reviews show higher ratings for leadership skills among male professors and "being warm" among female professors. Noting that perceptions of "easiness" predict overall ratings, Davison and Price (2009) recommend an RMP interface change replacing the site's "easiness" rating with better-defined terms such as "amount learned".

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 171–180

1.1.2 Interface design and online discourse

The interaction between interface design and online discourse is a central focus in computer-mediated discourse analysis (Herring, 2004; Herring and Androutsopoulos, 2015) and critical technocultural discourse analysis (Brock, 2018). Both consider not only how people express themselves in online environments, but also how elements like interface design of a website shape people into "users", affecting how they express themselves. Because we are interested in the relationship between attractiveness commentary and interface design, we constrain this study to the RMP website and its interface elements, including its professor rating form (Appendix, Figure 3), featuring the "hot or not" chili pepper rating.

Interface design is also considered quantitatively, and at scale, in company-led user experience studies. For example, Facebook found that by curating users' News Feeds to positive or negative posts, they influenced the emotional tenor of the users' own posts (Kramer et al., 2014). NextDoor, a popular neighborhood classifieds website, made their web form for reporting suspicious activity more detailed and inadvertently arduous, successfully decreasing suspicious activity posts and therefore decreasing posts with racial profiling (Hempel, 2017). In an effort to combat online harassment, Twitter introduced interface elements to warn users before posting tweets with inflammatory language (Statt, 2020). These interventions suggest that small interface changes may produce measurable effects in online discourse.

1.1.3 Subjectivity in workplace reviews

Like all genres of review, professor reviews interweave subjective and objective statements (e.g., "the class was poorly attended"). We consider commentary on a professor's physical attractiveness, which we refer to as objectifying or attractiveness commentary, to be subjective content.

Wiebe et al. (2001) discuss a method of labeling spans of subjective text within news corpora so that opinion phrases, even those that occur infrequently, can be detected using collocation clues. To determine review sentiment, Pang and Lee (2004) automatically segment movie reviews into subjective and objective portions, discarding the objective portions before attempting to determine the overall sentiment. Here we are also interested primarily in a subjective portion of reviews, but whereas subjectivity is an expected feature in other genres of reviews, comments about a professor's "hotness" may constitute workplace harassment (Flaherty, 2018) among other harms. To our knowledge, this is the first study to target objectifying commentary in professor reviews and its relationship to website design.

1.2 Our approach

Our classification scheme is tailored for a low-resource setting with a limited amount of labeled data. The goal is to construct classifiers which achieve sufficient accuracy to allow extrapolation to a much larger set of unlabeled reviews. To achieve our goal of analyzing large-scale trends, we train two models for identifying objectifying commentary in RMP reviews: (1) a token-level chunk tagger similar to those used for named entity recognition; and (2) a review-level text classifier similar to those used for document classification. Unlike the chunk tagger, the document classifier can take into account more variety in features and account for attractiveness commentary that occurs multiple times in a document. Multiple spans can "gang up", allowing them to be more easily detected at document level. In contrast, the chunk tagger considers objectifying language as a highly-local phenomena and is therefore more able to detect attractiveness commentary in the context of longer reviews covering a range of topics.

We then build ensembles of these models. We anticipate that ensembling will be useful because we hypothesize that the two classifiers' patterns of errors will be only weakly correlated (van Halteren et al., 1998), and because labeled data is limited, high-variance, and class-imbalanced (Brill and Wu, 1998) for this task.

2 Data

For this study, anonymous RMP reviews of professors were scraped on two occasions.[1] The first scrape, in July 2018, paired textual data with the professor's "hotness" rating, defined by the number of times a student rated the professor as "hot" minus the number times they rated them as "not hot" (Felton et al., 2008). In the web interface, the names of professors receiving positive attractiveness scores are marked with a chili pepper emoji (see Appendix, Figure 5). The second scrape, in August 2019, targeted a broader set of regions and

[1] Scraping was seeded using a list of professors and their chili pepper scores (`http://morph.io/chrisguags`).

schools. Test data was drawn from this latter data set, which was also used for trend analysis.

By this latter date, the chili pepper emoji had been removed from the website in response to public criticism (McLaughlin, 2018), so it was no longer possible to extract hotness scores. In addition to text, both scrapes also collected the names of professors, student-reported quality and difficulty scores (averaged by professor, on a five-point scale), subject area, and the name of the school. See Table 8 and Table 9 in the Appendix for the full list of schools.

2.1 Defining objectifying language

We define objectifying or attractiveness commentary as reviews that describe a professor's physical appearance, demeanor, clothing style, or resemblance. In contrast to prior work (e.g., Felton et al., 2008), we also include language disparaging a professor's appearance. Although previous work has considered objectifying comments in limited RMP datasets (Davison and Price, 2009; Kindred and Mohammed, 2017), there are no previous annotation guidelines to follow for labeling these expressions. Kindred and Mohammed (2017) find out of 788 RMP ratings in their sample, only 3.6% describe teacher attractiveness. Given the low frequency of these reviews and their informal qualities, creating instructions that cover attractiveness commentary in all of its variations is not possible. We acknowledge some reviews like ones described in Section 2.3 will be more subjective than others.

2.2 Review characteristics

RMP reviews contain stylistic flourishes common to online discourse: slang and non-standard language, typographical errors, expressive punctuation and capitalization, and emoticons. The examples below are fragments from 30-to-50-word reviews representing attractiveness commentary. See Figure 4 in the Appendix for additional examples in screen-capture format.

- *Everyone LOOOOOVES sexy Jeff!*
- *..he doesn't assume students understand complex stuff like other math teachers do. Plus, hello, HOT!*
- *He's also pretty cute which helps. :)*
- *..when he talked about vector space he almost saw my O-face.*

2.3 Fuzzy samples

This section describes reviews that pose challenges in labeling attractiveness spans and the process behind how distinctions are made. Annotators were instructed that, when in doubt, reviews that imply romantic interest, or lack thereof, are considered objectifying.

Flirtation but no attractiveness commentary Examples where the review may be flirtatious but not directly describing professor appearance present a grey area.

- *Damn, I love that man.* **None**

Referring to a professor as "that man" borders on objectifying, but without additional context it is not considered attractiveness commentary.

- *I love him so much, I would totally marry him if I could.* **Obj.**

However, we consider references to marriage or dating the professor like the above example to be objectifying commentary. We decide this because the element of fantasy in samples like these is taken to be indicative of an attraction to the professor.

- *He is a math god!* **None**

Reviews that compare the professor to a deity are also difficult to distinguish. If the focus could be the professor's expertise, the review is not considered attractiveness commentary.

Accents The most common challenging examples refer to the professor's voice or accent. These types of reviews primarily fall into two categories: (1) the accent is sexy, charming, or appealing; and (2) denoting professors who are non-native English speakers described as difficult to understand. Reviews in the latter category can be considered denigrating of the professor but are not necessarily attractiveness commentary. We consider the intent of the student. If the comment is personally derogatory, such as "horrible accent", it is considered objectifying. The following examples illustrate these distinctions:

- *And he's British, such a charmer! Love his accent!* **Obj.**
- *He has the cutest accent.* **Obj.**
- *His accent was difficult to understand.* **None**

	He	is	CUT	for	a	Stanford	professor
word-lower	he	is	cut	for	a	stanford	professor
lemma	he	is	cut	for	a	stanford	professor
pos	PRON	AUX	NOUN	NOUN	DET	PROPN	NOUN
has-hot	false	false	false	false	false	false	false
next-word	is	CUT	for	a	stanford	professor	[END]
next-pos	AUX	NOUN	ADP	DET	PROPN	NOUN	[END]
prev-word	[START]	He	is	cut	for	a	stanford
prev-pos	[START]	PRON	AUX	NOUN	ADP	DET	PROPN
prev-iob	O	O	O	B	O	O	O
all-caps	false	false	true	false	false	false	false
prev-all-caps	false	false	false	true	false	false	false
next-all-caps	false	true	false	false	false	false	false

Table 1: Example feature vector for chunk tagger using review snippet commenting on professor's physique.

3 Methods

We implement classification techniques with unique strengths for capturing the qualities and contexts of objectifying comments. The first, a chunk tagger, represents a bottom-up strategy, whereas the second, a document classifier, uses top-down processing and a richer feature set.

3.1 Chunk tagger

Because discussion of a professor's attractiveness may only be a small portion of any given review, we annotate spans of tokens which refer to attractiveness. These labels can then be automatically propagated from spans to the document level. That is, if a review contains any spans tagged as containing objectifying language, the whole review is labeled objectifying. We employ a chunk tagger customized to identify these spans within reviews. During preprocessing, labeled data is tagged for part of speech (POS) using the `spaCy` tagger (Honnibal and Montani, 2017). Text spans that refer to attractiveness are tagged using the CoNLL-2003 IOB format (Tjong Kim Sang and De Meulder, 2003). The chunker is built using the `nltk.chunk` library (Bird et al., 2009, ch. 7); it uses a multinomial logistic regression classifier and a greedy left-to-right decoding strategy.

Attractiveness features In addition to token features, we develop a dictionary of words describing attractiveness (see Appendix, Table 10); these are matched using regular expressions so alternative spellings (e.g., *hooooottt*, *hottttttt*) are also captured. See Table 1 for an example token feature vector.

3.2 Document classifier

We also develop a model that can take advantage of features extracted from the entire review. The document classifier is built using a linear-kernel support vector machine classifier from `sklearn` (Pedregosa et al., 2011). The primary features used are term frequency-inverse document frequency weighted unigrams and bigrams. Several other types of features, described below, are used to improve classifier accuracy.

Formality Impressionistically, RMP reviews that discuss teacher appearance tend to be less formal than those that focus on the quality of instruction. To capture this distinction, we use features proposed by Pavlick and Tetreault (2016) to measure textual formality. These include average word and sentence length, the ratio of nouns to verbs, and the proportion of words over 4 characters. We also add one-hot features for the use of non-standard punctuation and capitalization. Finally, we also extract features tracking the use of titles such as *Dr.*, *Professor*, *Mrs.*, and *Mr.*

Gender We extract professor gender by tracking third-person singular pronouns (e.g., *he*, *his*, *she*, *her*) in reviews; gender-non-specific pronouns like *they* and neo-pronouns like *ze* were not present in

	Reviews	Tokens	Words
Labeled	4,050	12,209	139,091
Unlabeled	358,970	71,700	15m

Table 2: Summary statistics for datasets.

the labeled data and therefore not tracked. We also do not track gender of the reviewers as all reviews are submitted anonymously.

Subjectivity Davison and Price (2009) and Ritter (2008) argue that student reviews largely follow a transactional consumerist discourse similar to customer service reviews. We hypothesize that this would be reflected in the ratio of first-person to third-person pronouns; a greater proportion of first-person pronouns may indicate a review about personal opinions and feelings (*consumerist*) rather than instruction. We also reuse the attractiveness dictionary regular expression patterns from the chunk tagger, expanding this to include common idioms such as *easy on the eyes* and *good looking*. Additionally, each review is scored for its sentiment and subjectivity using the `textblob`[2] sentiment classifier.

Style We consider features measuring the use of text properties characteristic of internet discourse, including the use of emoticons, repeated exclamation points, and words in all uppercase letters.

3.3 Feature ablation

For the document classifier, a feature ablation study on the development data (Appendix, Table 11) shows accuracy scores rely on the custom dictionaries for "hotness" including pattern matching for idiomatic expressions. However, omitting formality and stylistic features does not impact performance.

3.4 Ensembling

After training the chunk tagger and document classifier on the labeled data, a simple document-level ensemble of these models is applied to the unlabeled data. Since there are only two weak classifiers available, we use two forms of voting: in ensemble 1 we consider reviews with disagreement as non-objectifying reviews; in ensemble 2, we completely discard reviews when the classifiers disagree to achieve higher accuracy.

4 Results

4.1 Experiment setup

The labeled data of 4,050 reviews is randomly split into training (80%) and development sets (20%), the latter used for feature ablation (Appendix, Table 11). During annotation, professors labeled "hot" were deliberately oversampled. Review and token counts can be found in Table 2.

[2] `https://textblob.readthedocs.org`

Chunk tagger	Doc. classifier	
	Targeted	None
Targeted	8,573	9,858
None	4,295	336,242

Table 3: Confusion matrix for chunk tagger and document classifier models; Targeted: reviews which contain attractiveness commentary.

4.2 Annotation

To estimate interannotator agreement, a subset of the labeled data was independently labeled by a second annotator, a graduate student in linguistics, according to the authors' guidelines. This gave a span-level Cohen's $\varkappa = .785$ and a document-level $\varkappa = .801$; both correspond to "substantial" agreement according to the Landis and Koch (1977) qualitative guidelines.

4.3 Performance

After applying the chunk tagger and document classifier to the unlabeled data, we find the classifiers disagree on 4.1% of the reviews (see Table 3). This is roughly what one might expect given the overall low proportion of true positive samples. We determine accuracy by creating a test set from 600 of these reviews. This set includes reviews with classifier agreement on 150 documents predicted to contain, and 150 documents predicted not to contain, objectifying language. We also sample 300 reviews in which the chunk tagger and document classifier disagree.[3] These 600 samples are randomly sorted and then adjudicated by a human judge to create the test set.

The results for the chunk tagger and document classifier are shown in Table 4. As can be seen, both classifiers have relatively high accuracy but significantly lower precision and recall. Table 5

[3] We oversample from recent date ranges to better capture any new trends in reviews.

Classifier	Prec.	Rec.	$F1$	Acc.	$\varkappa$
Chunk tag.	.42	.21	.28	.89	.23
Doc. class.	.44	.23	.30	.93	.26

Table 4: Weak classifier results.

Review	Chunk tagger	Doc. class.
Not a great teacher (in fact pretty awful) but she's looking GOOD.	FN	TP
he is now bald, but he still has the look ;)	FN	TP
His classes are worthwhile because he's a good teacher, but mostly because he has the most awesome accent in the world. Rawr.	FN	TP
*the WORST ****ING TEACHER EVER. WORST CLASS, WORST PERSON. NOT PROFESSIONAL AT ANYTHING, DOES NOT KNOW PHYSICS FROM THE HOLE IN HIS ASS. AVOID!*	FP	TN
My experience with this professor was awful. He wasn't helpful and I ended up learning everything on my own without his help. I should have just stared at the wall rather than wasting my time in this class. He did not BUMP my grade up!	FP	TN
Probably the BEST Org Chem prof out of all the ones I've had. His slides are actually notes, not just pictures with lines on the side for you to write on. The exam is based onthe notes, but you also need to read the book. Def didn't mind looking at him for 1 hour 25 mins either.	TP	FN
not a bad prof. has a nice smile. class discussions were pretty interesting. grades are based on ur attendance, and ur blog entries (they are not hard, but be careful, cuz her way of grading is kinda picky). overall, not a hard class. kinda interesting. take it if u want, but if u cant stand reading don't. TONS of reading.	TP	FN
Jason's a fantastic section leader–some of the best classes I've had here were in section for this class. Plus, he knows his stuff, is super eloquent, and kicks ass in suits (just sayin'). I will say that he can come off as cold and intimidating at first, but he actually cares and is really willing to help you.	TP	FN
I can't understand her heavy accent. I found her subject boring.	TN	FP
Jenny is an extraordinary professor- she truly cares about how you do in her class, and does her best to help you in whatever fashion she can.	TN	FP
Very easy going. Knows what he is doing from lived experience. The power-points are very good. You can skip class and just follow along on the slides and get the idea of things (although probably not a good grade). Hot daughter.	FP	FP
worst prof ever, and i really mean that. she's not even a professor, just some plant biologist hired as a lecturer. she is completely inept as a lab manager and universally hated by the students. oh, and very not hot	TP	TP

Table 5: Examples with classifier disagreement; reviews have been modified to reflect their original format while protecting the identity of professors.

shows example reviews where the classifiers disagree. The chunk tagger performs better in reviews with higher word counts. In some cases, the chunk tagger avoids false positives of the document classifier where keywords from custom dictionaries appear but are in a context that is not objectifying. The document classifier performed well on lower word count reviews and where words from custom dictionaries and regular expression patterns are present. In Table 6, we show the same results for the two ensembles; note that results for ensemble 2 do not include the 300 samples from the test data on which the two weak models disagree.

We see that both ensemble classifiers achieve greatly improved results compared to either the chunk tagger or the document classifier alone, and as expected, error can be further reduced in ensemble 2 by discarding data on which they disagree.

We conclude that ensemble methods are effective for detecting objectifying commentary in student reviews in the face of unbalanced data. In what follows, the ensemble 2 classifier is used to to analyze trends in attractiveness commentary on 344,815 reviews.

5 Analysis

Building on previous RMP research studying bias in student reviews, we continue this inquiry focusing on how attractiveness commentary is distributed based on teacher gender, and quality and

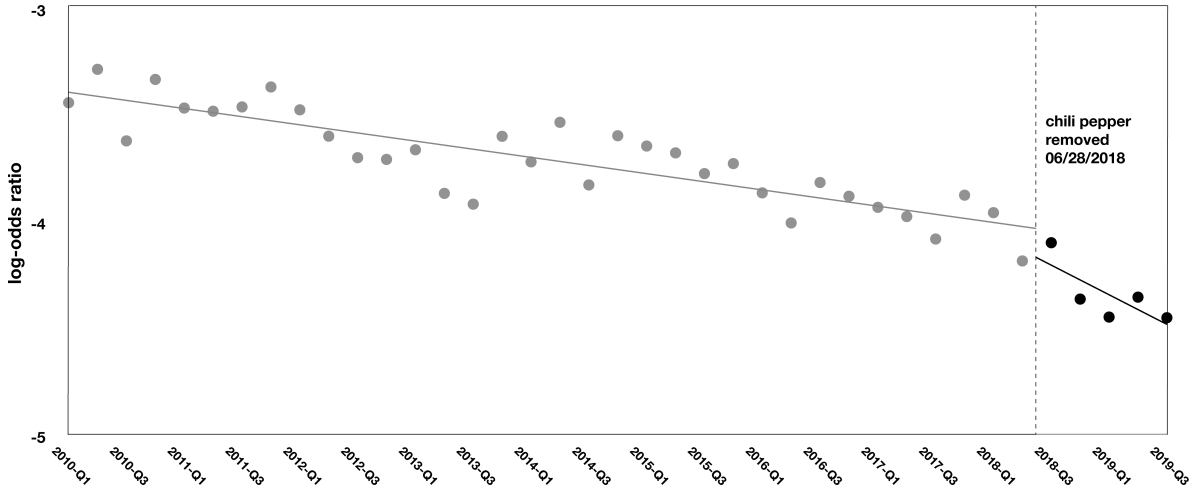

Figure 1: Log-odds of attractiveness commentary in reviews from 2010 to August 2019.

Classifier	Prec.	Rec.	$F1$	Acc.	$\varkappa$
Ensemble 1	.72	.44	.55	.93	.50
Ensemble 2	.72	1.00	.84	.99	.83

Table 6: Ensemble classifier results.

difficulty scores. We then focus on a logistic regression analysis using generalized estimating equations (GEE) to determine if there was a decrease in attractiveness commentary following the removal of the chili pepper feature from the web interface.

5.1 Teacher gender

Our dataset contains 39.7% female professors (11,192) compared to 60.2% male professors (16,967). This proportion is similar to those found in U.S. higher education where women make up only 31% of full-time faculty (Kelly, 2019). Since this breakdown leads to more reviews for male professors overall, we consider each professor and whether or not they have at least one objectifying comment. In our dataset, 21.0% of male professors have at least one attractiveness review compared to 18.4% of female professors. We find in a chi-square test of independence that this difference is significant ($\chi^2 = 17.75, p < .01$). In contrast, Rosen (2018) found that women were more likely to have the chili pepper rating (27.8%) than men (22.7%). We believe this difference could be attributed to the distinction between the low effort act of clicking "hot" on the review form versus actually writing commentary on the teacher's appearance. Also, unlike chili pepper ratings, our counts include reviews with negative commentary.

5.2 Logistic regression

We deploy logistic generalized estimating equations (GEE; Liang and Zeger 1986), an extension of the generalized linear modeling that takes into account the correlation between observations. A logistic GEE accommodates the unequal number of earlier observations across professors and conditions as well as the variation in review activity volume over quarterly time intervals. This is optimal for the noise in the dataset and allows utilization of the entire collection of reviews. The final model parameters are determined by the best goodness-of-fit score computed using the full log quasi-likelihood function. School size and tuition did not have significant outcomes in the results and were discarded. The final model includes presence or absence of the chili pepper interface feature, teacher quality and difficulty scores, and professor gender. Time is input as an interval covariate by quarter, while chili pepper condition is a binary factor; final parameters and their outcomes are given in Table 7.

Chili pepper and time interval First, we focus on our primary question concerning the proportion of objectifying comments and the removal of the chili pepper. We observe a downward trend over the time period prior to the interface change; however, the log-odds of attractiveness commentary after the chili pepper was removed on June 28, 2018 is lower than the time variable can account for

	Estimate (log-odds)	Std. err.	Wald χ^2	$p(\chi^2)$
(Intercept)	−3.111	.143	476.18	$< .001$
pepperAbsent	−.428	.136	9.93	.002
timeInQuarters	−.020	.002	79.44	$< .001$
difficultyHigh	−.075	.022	11.49	$< .001$
qualityHigh	.051	.026	3.76	.053
genderFemale	−.528	.174	9.19	.002
qualityHigh:genderFemale	.097	.043	5.09	.024

Table 7: GEE model parameter estimates with attractiveness commentary as dependent variable. The intercept represents pepperPresent, timeInQuarters = 0, difficultyLow, qualityLow, genderMale. $N = 344,815$.

alone (see Table 7). Our analysis finds a significant effect of time and condition (with vs. without the chili pepper). These findings support our hypothesis: RMP's removal of the chili pepper coincides with a decline in reviews mentioning professor attractiveness.

Quality and teacher gender We compare the proportions for attractiveness commentary in relation to quality and difficulty rating scales (Figure 2). There is a significant interaction between teacher quality and gender, female professors rated high quality are significantly more likely to receive attractiveness commentary than male professors rated high quality (see Table 7). Difficulty was also a significant factor, the higher the difficulty score, the less likely the reviews for the professor will contain attractiveness commentary.

6 Discussion

While our work has focused on the text contents of reviews, our analysis of objectifying comments follows previous findings about biases of the original chili pepper rating, correlating with teacher gender, quality, and difficulty ratings. This is the first study to find a correlation between attractiveness commentary and the website interface.

More research is needed to understand the observed steady eight-year decline. As this was an observational study rather than a controlled experiment, there are many uncontrolled variables. For instance, we cannot compare attractiveness commentary by size of professor's class or attributes of the reviewer. We tried to estimate these factors with proxies such as university size, geographic area, and tuition amounts, but these only provide rough estimates and did not have significant effect on the presence of attractiveness commentary. McNeil (2020) reflects on how users' perceptions of anonymity have changed, from posting to online bulletin boards in the late 1990s, to present-day "sharing" on corporate-owned, heavily surveilled social network sites like Facebook. This turn from anonymity to self-awareness is observed by Marwick and boyd (2011) in their study of Twitter users. These users describe their own self-censoring behaviors by imagining their audiences to include not only friends but also parents and employers. The decline in attractiveness commentary on RMP may reflect broader internet trends, corresponding with internet users being more conscious of their perceived audience and realizing that true online anonymity is impossible.

7 Conclusion

We find that a small change to the RMP website, removal of the chili pepper rating, is associated with a lower likelihood of comments on professor attractiveness. Our experiments show that an ensemble of classifiers can accurately detect objectifying language in online professor reviews and can allow us to analyze trends in a large unlabeled dataset.

One area where classifiers disagreed was in the "fuzzy samples" such as accents and godliness discussed in Section 2.3. Breitfeller et al. (2019) describe similar challenges in classifying microaggressions and label themes within their dataset to better define these utterances. Based on our classifier's success in pulling out objectifying comments from large datasets, we can identify enough examples to consider labeling categories such as accent criticism and comments about unattractiveness. Finally, one could apply an active learning approach (Yarowsky, 1995) to label and train on examples where the classifiers disagreed.

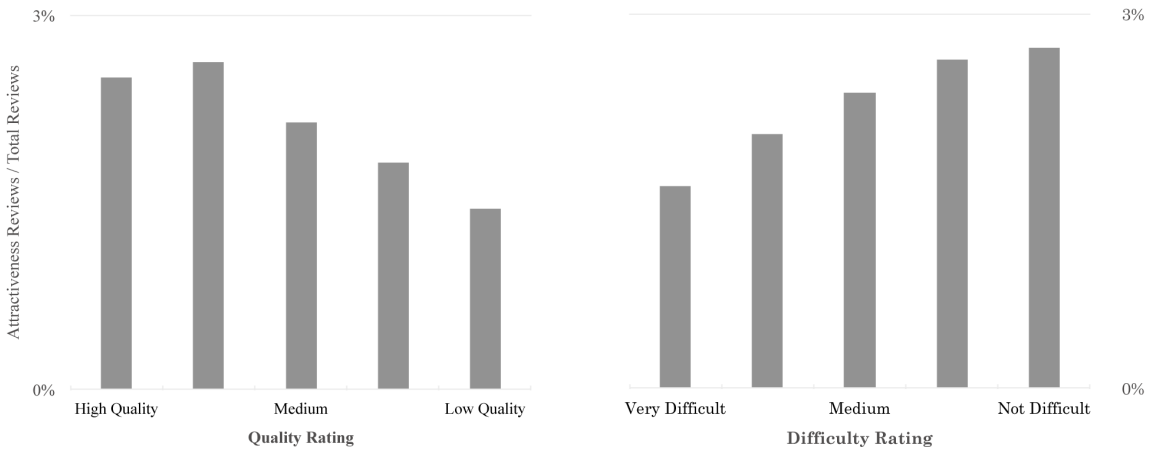

Figure 2: Proportion of reviews with attractiveness commentary by quality and difficulty ratings.

With further exploration it is hoped these techniques could be applied to detecting other forms of abusive language in online reviews. Insofar as the removal of the chili pepper feature correlated with a significant decrease in attractiveness commentary, we suggest that web interface design may positively influence online discourse. As the scope of the gig economy continues to expand and more workers find themselves evaluated by anonymous online reviews, we hope these findings will inspire future research around potential biases in online reviews based on gender, appearance, and the design of the online interface used.

8 Acknowledgments

We would like to thank Martin Chodorow for his guidance in statistical analysis and Deepali Advani for her assistance with data preparation. We appriciate Jonathan Butterick for helping with data collection. We would also like to acknowledge Sara Morini for her assistance with the data annotation, William Jordan for proofreading, and anonymous reviewers for their helpful feedback.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly.

Anne Boring. 2017. Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41.

Anne Boring, Kellie Ottoboni, and Philip B. Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1:1–11.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674.

Eric Brill and Jun Wu. 1998. Classifier combination for improved lexical disambiguation. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 191–195.

André Brock. 2018. Critical technocultural discourse analysis. *New Media & Society*, 20(3):1012–1030.

Serina Chang and Kathy McKeown. 2019. Automatically inferring gender associations from language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5746–5752.

Elizabeth Davison and Jammie Price. 2009. How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education*, 34(1):51–65.

James Felton, Peter T. Koper, John Mitchell, and Michael Stinson. 2008. Attractiveness, easiness and other issues: student evaluations of professors on ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 33(1):45–61.

Colleen Flaherty. 2018. Bye, bye, chili pepper: Rate My Professors ditches its chili pepper "hotness" quotient. Accessed 10/28/2018.

Scott Freng and David Webber. 2009. Turning up the heat on online teaching evaluations: Does "hotness" matter? *Teaching of Psychology*, 36(3):189–193.

Hans van Halteren, Jakub Zavrel, and Walter Daelemans. 1998. Improving data driven wordclass tagging by system combination. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 491–497.

Jessi Hempel. 2017. For Nextdoor, Eliminating Racism Is No Quick Fix. *Wired*.

Susan C. Herring. 2004. Computer-mediated discourse analysis. In Sasha Barab, Rob Kling, and James H.Editors Gray, editors, *Designing for Virtual Communities in the Service of Learning*, page 338–376. Cambridge University Press.

Susan C. Herring and Jannis Androutsopoulos. 2015. Computer-mediated discourse 2.0. In Deborah Tannen, Heidi E. Hamilton, and Deborah Schiffrin, editors, *The Handbook of Discourse Analysis*, pages 127–151. John Wiley & Sons.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Accessed 1/4/20.

Bridget Turner Kelly. 2019. Though more women are on college campuses, climbing the professor ladder remains a challenge.

Jeannette Kindred and Shaheed N. Mohammed. 2017. "he will crush you like an academic ninja!": Exploring teacher ratings on RateMyProfessors.com. *Journal of Computer-Mediated Communication*, 10(3).

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Christine Lagorio. 2006. Hot for teacher. *The Village Voice*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Alice E. Marwick and danah boyd. 2011. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

BethAnn McLaughlin. 2018. I killed the chili pepper on Rate My Professors. Accessed 1/4/2020.

Joanne McNeil. 2020. *Lurking: How a Person Became a User*. Macmillan.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271–278.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kelly Ritter. 2008. E-valuating learning: Rate My Professors and public rhetorics of pedagogy. *Rhetoric Review*, 27(3):259–280.

Andrew S. Rosen. 2018. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of ratemyprofessors.com data. *Assessment & Evaluation in Higher Education*, 43(1):31–44.

Ben Schmidt. 2015. Gendered language in teacher reviews. Accessed 7/13/19.

Nick Statt. 2020. Twitter tests a warning message that tells users to rethink offensive replies. *The Verge*.

TextBlob. 2018. TextBlob: simplified text processing. Accessed 1/8/2020.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Janyce Wiebe, Theresa Wilson, and Matthew Bell. 2001. Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.