Analyzing Gender Bias in Student Evaluations

Andamlak Terkik[1], **Emily Prud'hommeaux**[2], **Cecilia O. Alm**[2],
Christopher M. Homan[1], **Scott Franklin**[3]

[1]Golisano College of Computing & Information Sciences, Rochester Institute of Technology
[2]College of Liberal Arts, Rochester Institute of Technology
[3]College of Science, Rochester Institute of Technology
{†at3616|†emilypx|†coagla|§cmh|†svfsps}@{†rit.edu|§cs.rit.edu}

Abstract

University students in the United States are routinely asked to provide feedback on the quality of the instruction they have received. Such feedback is widely used by university administrators to evaluate teaching ability, despite growing evidence that students assign lower numerical scores to women and people of color, regardless of the actual quality of instruction. In this paper, we analyze students' written comments on faculty evaluation forms spanning eight years and five STEM disciplines in order to determine whether open-ended comments reflect these same biases. First, we apply sentiment analysis techniques to the corpus of comments to determine the overall affect of each comment. We then use this information, in combination with other features, to explore whether there is bias in how students describe their instructors. We show that while the gender of the evaluated instructor does not seem to affect students' expressed level of overall satisfaction with their instruction, it does strongly influence the language that they use to describe their instructors and their experience in class.

1 Introduction

Student evaluations of teachers (SETs), in which students are asked to provide their assessment of the quality of instruction they have received in a particular course, have been in use for over a century. At the end of a course, students are given forms, in paper or electronic format, containing a series of questions about the course and instructor, some requiring Likert-type scale responses and others seeking free text responses. SETs have increasingly become the de facto standard for evaluating university-level teaching performance (Centra and Gaubatz, 2000). The impact of these surveys on faculty is enormous, as they affect tenure, promotion, and compensation decisions.

Despite playing an outsized role in assessing teaching effectiveness, SETs have numerous shortcomings as tools for this task. Students, for example, are understandably often not well equipped to determine an instructor's knowledge of the subject matter, and they can be unreliable judges of how well they have mastered material, one important and generally accepted measure of teaching effectiveness. Perhaps more troublingly, several studies have demonstrated biases, whether conscious or unconscious, in students' evaluations of women instructors and instructors of color.

Most previous research in this area has focused on numerical (more precisely, *ordinal categorical*) ratings of various qualities related to teaching effectiveness. In this paper, we instead focus on computational analysis of the text responses to open-ended questions found in SETs. In the first part of this paper, we present a supervised instructor satisfaction classifier trained to identify the satisfaction polarity of SET comments. Next, we apply this model to a much larger dataset to examine how satisfaction varies across both genders. Finally, we analyze the patterns of word choice associated with each gender in order to explore how students' language changes according to the gender of the instructor they are evaluating.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 868–876, Osaka, Japan, December 11-17 2016.

2 Background

Several previous studies have found evidence for some sort of gender bias in SETs, although the results have been somewhat mixed and inconclusive. Female instructors reportedly receive lower overall numerical ratings than their male counterparts, particularly when male students evaluate them (Basow and Silberg, 1987; Basow, 1995; Young et al., 2009; Boring, 2015). At the same time, women whose personality and teaching style conform to expected gender stereotypes (e.g., warmth, helpfulness, accessibility) tend to receive higher marks overall, regardless of the gender of the evaluating students (Bennett, 1982; Kierstead et al., 1988). Gender bias also seems to vary according to the subject matter, level, and department of the course being taught (Basow, 1995; Centra and Gaubatz, 2000).

The rise of online instruction has provided a useful mechanism for identifying gender bias under more controlled conditions. MacNell et al. (2015) recently investigated the presence of gender bias in SETs of an online college-level social science class. In their experiment, two instructors, one male and one female, each taught two sections of the same course in an online setting: one in which the students were led to believe their instructor was a woman, the other in which they believed their instructor to be a man. Students in the sections with the *perceived* female instructors gave the instructors significantly lower scores in six areas, including the overall rating, than the students in the sections with the *perceived* male instructors. The differences in scores assigned to the instructors were not significantly different, however, across *actual* gender. Remarkably, it was noted that the instructor with the highest ratings was the female instructor who was perceived to be a man, while the instructor with the lowest ratings was the male instructor who was perceived to be a woman.

This previous work on gender bias in SETs has relied primarily on students' numerical ratings of their instructors for a variety of qualities related to their teaching effectiveness. In our research, we focus on the open-ended text comments that students are sometimes allowed to provide in order to elaborate on their numerical ratings. We rely on techniques and algorithms typically used for the NLP task of sentiment analysis, in which a variety of linguistic features are used to identify positive and negative tone expressed in natural language text (Mohammad, 2016; Liu, 2012). Our approaches to the task of identifying the degree of student satisfaction in their instructors, as expressed in their written comments, are inspired by, though distinct from, seminal work by Turney and colleagues (2002; 2003). Although our methods and features are not independently novel, the application of these methods in combination to the task of analyzing comments in SETs and the framing of the task itself constitute new and important contributions to the fields of NLP, gender studies, and education theory.

3 Data

SETs from a period of eight years were collected from a variety of undergraduate courses in math, physics, statistics, biology, and chemistry offered at a four-year, degree-granting institution in the United States. STEM courses were chosen because of national focus on gender disparities in physical sciences and the potential to eventually explore differences between scientific fields that do show such disparities (physics, mathematics) and those that do not (biology). Focusing on STEM also allows us to sample the majority of students, as the introductory courses in these fields are often service courses required for computing and engineering majors and can be used to meet distribution requirements for students in the humanities and social sciences.

We divide the SETs into three groups, which we call *small-labeled*, *medium-unlabeled*, and *large-unlabeled*. Each item in the first two sets contains one student response to the prompt: "*Comment on the instructor's strength and weaknesses*." The *large-unlabeled* set contains responses to multiple distinct prompts for comments on specific qualities associated with teaching effectiveness, including helpfulness, materials, organization, and presentation. Table 1 shows the the number of comments in the three groups.

The *small-labeled* dataset is the full set of "strengths and weaknesses" comments for two introductory statistics courses taught by multiple professors over several semesters, manually labeled by an undergraduate research assistant. We instructed this annotator to rate the level of satisfaction expressed in the comments. The options were *very satisfied*, *somewhat satisfied*, *neither satisfied nor dissatisfied*, *somewhat dissatisfied* and *very dissatisfied*. The comments in the remaining two datasets were not manually

Data set	Size
small-labeled	2,076
medium-unlabeled	15,896
large-unlabeled	107,855

Table 1: The three comment sets used in this study.

	Fleiss' Kappa	%Overlap
5 classes	0.49	0.61
3 classes	0.58	0.73

Table 2: Agreement scores for all annotators.

labeled. We trained and tested our classifier, discussed in detail below, on the *small-labeled* data, and then used it to predict the satisfaction level of the comments in the *medium-unlabeled* set. The *large-unlabeled* set was used for computing analytics and identifying terms strongly associated with either gender.

3.1 Inter-rater Agreement of Manual Annotations

Annotating the satisfaction level of a given text is an inherently subjective task. Since our classifier, described below, is trained on these annotations, it is important to estimate their reliability. Three individuals (co-authors), including the research assistant who rated the entire *small-labeled* set, annotated a subset of 100 comments using the previously described scheme. All of the annotations were performed after anonymizing the text and replacing all gendered titles, nouns, and pronouns with their equivalent gender neutral placeholders as explained below in Section 3.2. We then analyzed this newly annotated set by computing Fleiss' kappa (Randolph, 2009), using an online tool (Geertzen, 2012). Fleiss' kappa is a variant of Cohen's kappa (Byrt et al., 1993) that measures agreement among more than two raters.

The kappa scores were computed for two groupings. In the first grouping, each of the five satisfaction classes is considered independently. In the second, the two extreme classes on each side of the range were merged, yielding a 3-class rating scheme. The 3-class agreement scores exhibit less ambiguity, resulting in improved agreement.

Table 2 shows the Fleiss' kappa and overlap percentage of the results for both the 5-class and 3-class groupings. Both of these scores fall in the range of 0.4 to 0.6, which indicates moderate agreement under most interpretations of kappa in the psychology literature (Landis and Koch, 1977). It is also worth noting that inter-rater agreement scores such as the kappa score greatly depend on the level of subjectivity inherent in the task itself.

3.2 Data Preprocessing

Before extracting features for the satisfaction classifier, we anonymized the text by replacing all first and last names with a placeholder, merged all gendered pronouns (e.g., both *he* and *she* became *he/she*), replaced all words referring to a particular gender with a gender-neutral equivalent (e.g., *guy* and *lady* became *person*), downcased, and removed special characters. Another crucial step undertaken during preprocessing was the handling of negation terms. For instance, phrases such as *great at teaching* were differentiated from negative sentiments such as *not great at teaching*. This was achieved by applying the negation term *not* to each term that follows it until a special character or other negation term is encountered, using the method described by Narayanan et al. (2013).

The negation routine works by first detecting the negation markers *not* or *n't*. Whenever these markers are encountered, words that follow them are transformed into new terms prefixed with *not_*. After a negation marker is set, it negates every word that follows until a punctuation mark or another negation term is encountered. For instance, *not great at teaching* would be turned into *not not_great not_at not_teaching*. Term negation was responsible for an increase of about 4 percentage points in classification accuracy.

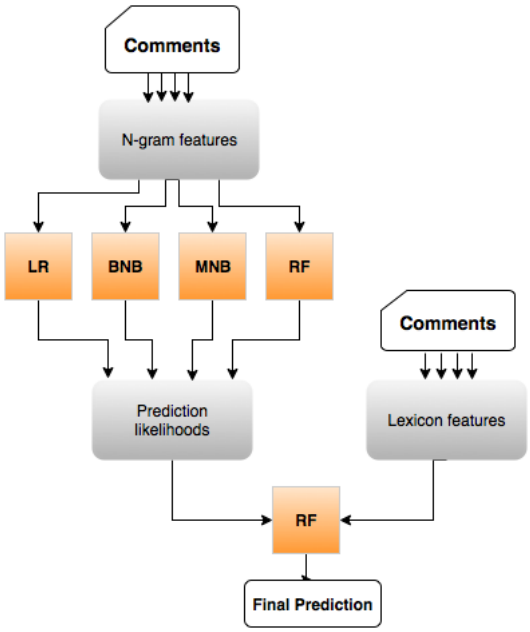

Figure 1: Architecture of the satisfaction classifier.

4 Method

4.1 Lexical Features

We extracted two types of lexical features from the data: n-gram features and sentiment term scores. Unigrams, bigrams, and trigrams served as features for the first tier of the classifier. Bigrams and trigrams can model useful local contextual features that unigrams are unable to model. For example, while unigrams features would be sufficient to capture single-word terms such as *intelligent* and *nice*, higher-order n-grams are required to capture the composite meaning found in phrases such as *extremely well* or *hardly ever available*. Over 30,000 n-grams were extracted from the dataset, resulting in a feature vector of this length for each comment.

Sentiment term scores were obtained by computing the aggregate positive and negative scores for each comment. To compute these aggregate scores, the prior polarities of the terms were determined using domain-independent lexicons. We relied on three general-purpose sentiment lexicons: the MPQA lexicon (Wilson et al., 2005), the NRC emotion lexicon (Mohammad and Turney, 2010; Mohammad and Turney, 2013), and Bing Liu's opinion lexicon (Liu, 2012). For each comment, the aggregate raw positive and negative term scores were computed from the scores from each of the three lexicons. Therefore, a 6-valued (i.e., 3 dictionaries x 2 sentiments) feature vector was computed for each comment.

4.2 Classifier Architecture

The classifier was designed as a two-tier system called *stacked generalization* (Wolpert, 1992), illustrated in Figure 1. The first tier comprises four classifiers: a random forest model, a multinomial naive Bayes model, a Bernoulli naive Bayes model, and a logistic regression model, each trained on unigram, bigram, and trigram features. The class likelihood predictions obtained from these four models, along with sentiment term scores, were then used to train a final classifier in the second tier. We used for this final classifier a random forest with parameters similar to those in the first tier.

Both the multinomial and the Bernoulli naive Bayes models have performed well in previous sentiment classification tasks (Pang et al., 2002) similar to ours. With both models we used Laplacian smoothing (i.e., $\alpha = 1.0$) with uniform priors. We trained and evaluated the random forest with 100 trees having a maximum depth of 80. The framework was implemented using the scikit-learn machine learning library (Pedregosa et al., 2011).

Total comments	2076
Total lexicon types	73802
Total tokens	80970
Type/token ratio	0.9
Avg. comment length	206 characters

Table 3: Descriptive statistics for the *small-unlabeled* data set.

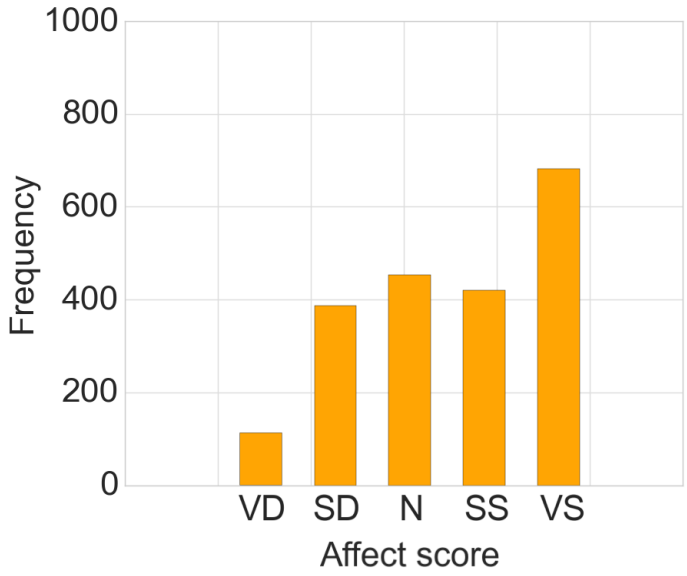

Figure 2: Distribution of 5 satisfaction levels in the *small-labeled* data set.

4.3 Results

We considered two classification tasks.

1. The *extremes* task: distinguishing *very dissatisfied* from *very satisfied*

2. The *merged* task: distinguishing (*very dissatisfied* or *somewhat dissatisfied*) from (*somewhat satisfied* or *very satisfied*)

In each case, we tested the classifier by running 10-fold cross validation on the *small-labeled* data. Table 3 presents statistics compiled from those tests. Figure 2 shows the distribution of satisfaction labels.

We evaluate a majority-class baseline and a Radial Basis Function (RBF) kernel-based SVM with penalty term $C = 2.0$, along with the main ensemble classifier. The SVM classifier utilized the same feature set as the ensemble one. However, instead of a two-tier architecture, the SVM directly combines the n-gram features with sentiment term score features. We report the classification accuracy and F1-scores in Table 4.

We see that our ensemble classifier yields large improvements over the majority-class baseline. This is especially true for the more challenging *merged* task. It also outperforms the SVM classifier by a notable margin for both tasks. Given that inter-annotator reliability on the 3-class task was just under $\kappa = .6$, achieving classification accuracy of 81% is impressive. Although there is room for improvement, these results demonstrate the efficacy of our framework.

	Baseline	SVM	RF Ensemble
Extremes task (n=795)			
Accuracy	86%	86%	**91%**
F1-score	80%	80%	**91%**
Merged task (n=1602)			
Accuracy	69%	79%	**81%**
F1-score	57%	77%	**81%**

Table 4: Classification accuracy and F1-score for both tasks. Boldfacing marks performance increases.

	Women	Men
small-labeled		
Satisfied	74%	62%
Dissatisfied	26%	38%
medium-unlabeled		
Satisfied	94%	94%
Dissatisfied	6%	6%

Table 5: Affect distribution broken down by gender. The top half shows breakdown of affect as annotated in the *small-labeled* set. The bottom half shows breakdown of affect in *medium-unlabeled* as predicted by the ensemble classifier.

5 Further Analysis

5.1 Satisfaction by Gender

For the *small-labeled* dataset, we computed the ratio of manually labeled *satisfied* to *dissatisfied* comments by gender. For the *medium-unlabeled* dataset, we computed this ratio using the satisfaction values from our classifier. Table 5 shows the results of these comparisons.

In the manually annotated dataset, male instructors receive slightly less favorable satisfaction ratings, in contrast to previous work reporting higher numerical ratings for male instructors. We note that this discrepancy is unlikely to be related to the gender of the students themselves since men were somewhat over-represented in the student body of the university from which the SETs were gathered. Rather, it seems that students were more satisfied with the instruction that they received from female instructors in the two introductory statistics courses from which these comments were drawn. We note, however, that satisfaction is a relatively subjective concept, and so may be influenced by the annotator's perception.

This difference in satisfaction disappears in the *medium-unlabeled* set, where the classifier predicts satisfaction levels to be equally distributed between genders. This could be an artifact of the larger size of the dataset, the broader range of course subject matter and level, the larger number of instructors, or simply the behavior of the classifier itself. In any case, our results do not seem to provide evidence for the presence of gender bias in students evaluations of teaching effectiveness. This does not preclude, however, the possibility of differences in students' language use according to the gender of the instructor being evaluated.

5.2 Gendered Language

In order to understand how word usage differs by the gender of the rated instructor, we first normalized and lemmatized the *large-unlabeled* set to account for morphological variation and abbreviation. We then ranked words based on their strength of co-occurrence, in terms of mutual information (MI) (Church and Hanks, 1990), with each gender. We selected the top 200 words from this ranking and sorted them into two groups based on their semantic functions. The first group contains terms used to *address* or *refer* to the instructor, and the second contains words *describing* an instructor or a student's experiences. We report the occurrence count of each term per 1000 comments after adjusting for the number of comments for each gender.

	F (per 1000)	M (per 1000)
Prof./professor	167	**209**
<Last name>	139	**151**
Dr.	75	**77**
teacher	**95**	79
instructor	172	172
<First name>	22	21

Table 6: Terms of address used to refer to faculty. Term frequency per 1000 comments adjusted by number of comments for both genders.

Word	F	M	Diff
amazing	**32**	18	128%
love(d)	**59**	32	84%
wonderful	**28**	12	57%
organized	**243**	178	37%
willing	**114**	88	30%
helpful	**454**	402	13%
tangent(s)	3	**16**	400%
funny	4	**14**	250%
knowledgeable	21	**33**	57%
interesting	68	**92**	35%
understanding	110	**126**	15%

Table 7: Gender differences in words used to describe men vs. women faculty. Values are per 1000 comments, adjusted by number of comments for each gender.

Table 6 shows that students are more likely to refer to their male instructors with the appropriate professional title (e.g., *Prof.*, *Dr.*) and by their last names. Conversely, female instructors are more likely to be referred to by their first names or descriptors that do not reflect their status as university professors (e.g., *the teacher* or *the instructor*). It is important to keep in mind that these comments were compiled from an institution having a faculty with roughly similar distributions of professional qualifications for both genders. These results therefore demonstrate an unwarranted bias toward more frequent use of prestigious titles and traditionally respectful forms of address for male instructors, regardless of their actual academic qualifications.

Table 7 shows the words with the most extreme differences in frequency according to instructor gender. Women were far more likely to be described with very positive but generic terms (*amazing, wonderful, loved*) than men. Perhaps more interestingly, students tended to describe women more often than men in terms of how the instructors impacted their learning experiences (*organized, willing, helpful*). Men, on the other hand, were recognized primarily for personal qualities (*funny, knowledgeable, interesting, understanding*) that may be independent of their ability to teach. The only negative term on this list, *tangent*, was also the term with the largest relative frequency difference between genders.

6 Conclusion

In this paper, we investigated the use of computational methods to analyze the language used in open-ended comments from student evaluations of teaching effectiveness in order to explore the possibility that gender bias exists in these evaluations. In contrast to previous research that relies on numerical ratings, our results fail to reveal differences by instructor gender in overall student satisfaction, as expressed in written comments. This results holds whether those satisfaction values are determined via direct human annotation or from machine learning models trained on the annotations. We do, however, observe real,

qualitative, gender-based differences in the language students use when providing written comments about their instructors.

Future work will follow several distinct but related paths. First, we will continue to develop our classifier using more complex features of language, such as those derived from semantic role labels or extracted from neural word embeddings or other vector space models. Secondly, we will explore the various student-assigned numerical ratings that accompany the text comments analyzed here. In particular, we hope to compare these ratings with our automatically generated satisfaction ratings in order to see the degree to which positive comments correlate strongly with high numerical ratings. As for gendered language, we plan expand our analysis by exploring whether certain syntactic structures are more strongly associated with either gender. Finally, we plan to investigate the various potential confounding factors in data, including subject matter, level, and instructor rank, as well as features of the students themselves, in order to shed light on the mixed and inconclusive evidence for gender bias described in the literature.

Acknowledgments

Many thanks to Ja'Nai Gray for providing manual annotations of the data used in this study and to the anonymous reviewers for their helpful feedback and suggestions.

References

Susan A. Basow and Nancy T. Silberg. 1987. Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, 79(3):308–314.

Susan A. Basow. 1995. Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, 87(4):656–665.

Sheila K. Bennett. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74(2):170–179.

Anne Boring. 2015. Gender Biases in Student Evaluations of Teachers. Technical report.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429.

John A. Centra and Noreen B. Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, pages 17–33.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Jeroen Geertzen. 2012. Inter-rater agreement with multiple raters and variables. https://nlp-ml.io/jg/software/ira/.

Diane Kierstead, Patti D'Agostino, and Heidi Dill. 1988. Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, 80(3):342–344.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Lillian MacNell, Adam Driscoll, and Andrea N. Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

Vivek Narayanan, Ishan Arora, and Arjun Bhatia. 2013. Fast and accurate sentiment classification using an enhanced naive Bayes model. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, pages 194–201. Springer.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pages 79–86.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Justus J. Randolph. 2009. Free-marginal multirater kappa (multirater κ_{free}): An alternative to Fleiss' fixed-marginal multirater kappa. *Advances in Data Analysis and Classification*, 4.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Peter D. Turney. 2002. 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*, pages 417–424.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Suzanne Young, Leslie Rush, and Dale Shaw. 2009. Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning*, 3(2):19.