Looking for Gendered Differences in Instructor Evaluation Language

Abstract

Instructors can receive feedback on their course content and teaching methods from students who fill out subject evaluation forms. However, some studies have indicated there may be gender bias in course evaluations: male and female professors seem to be rated differently. We are interested in seeing if we can predict the gender of the professor being evaluated using learned features from student text responses. We used two datasets, one from an MIT sorority's collection of informal student evaluations for humanities classes and another from MIT's national honor society for computer science and electrical engineering. We trained various predictive models such as logistic regression, support vector machines, and k-nearest neighbors using text features of the student responses. We are unable to train a predictive model that performs better than our baseline of guessing the most common gender on the honor society data and have only made small improvements on the sorority data.[1]

Background

A number of previous studies have found differences in the way male and female instructors are evaluated. One study had students listen to a lecture read by the same person, whose voice was gender-neutral, and then fill out an evaluation form after being told the professor was young and male, young and female, old and male, or old and female. The study found students gave higher ratings to male professors, with young and male professors receiving the best ratings.[2] Another study found results that suggested that the criteria the professors were evaluated on and the stringency with which they were evaluated on that criteria may be different for men and women. For example, there seemed to be more expectation of female professors to provide interpersonal support.[3]

Corpus

Our corpus consisted of the student text responses for subject evaluations from two sources. A summary of the corpus statistics is below:

---

[1] The code for our project can be found here: https://github.mit.edu/bazuzi/806project

[2] Arbuckle, Julianne, and Benne D. Williams. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles* 49, no. 9 (2003): 507-16. http://link.springer.com/article/10.1023/A:1025832707002#.

[3] Bennett, Sheila K. "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*: 170-79. Accessed December 11, 2015. http://search.proquest.com/docview/614362046?accountid=12492.

	AXO	HKN
number of courses evaluated	154	6
average number of reviews per course	1.7	32.5
standard deviation of reviews per course	1.01	49
maximum number of reviews for a course	6	239
minimum number of reviews for a course	1	5
total number of instructors reviewed	152	29
number of male instructors reviewed	88 (~57.9%)	26 (~89.7%)
number of female instructors reviewed	64 (~42.1%)	3 (~10.3%)

AXO Data

Our first dataset comes from MIT's Alpha Chi Omega (AXO) sorority. Evaluations of humanities, arts, and social science (HASS) classes were written by AXO members on a voluntary basis. These comments came in the format of a class name and number attached to the instructor of the class, a free-text comment of no specific required content, a yes or no recommendation of the class for other members, and finally an attribution of the comment with the reviewer's name and the semester in which she took the class. A sample review is below:

> The class was really chill with some interesting reading assignments that were not too long. In-class discussions were lengthy, but very interesting in my opinion. The professor was very understanding of college students in general, and was pretty chill. The writing assignments were definitely manageable (considering that it was a CI-H). I enjoyed the class.
> *Recommend to others?* Yes [(year and name redacted)]

HKN Data

Our second dataset came from MIT's national honor society for computer science and electrical engineering, Eta Kappa Nu (HKN). HKN takes course evaluations written by students on a voluntary basis and compiles an Underground Guide, consisting of summaries of all courses in the electrical engineering and computer science (EECS) department. We analyzed a set of original responses from students answering a set of questions asking about the subject generally, lab assignments, readings, quizzes, grading, and suggestions for future iterations of the course, as well as comments about the instructor specifically. A sample instructor review is below (All gendered pronouns such as 'he' and 'his' were removed before analysis occurred.) :

> "often has very harsh comments, has very high expectations. his blackboard technique is subpar, but he doesnt use the blackboard very much. he mostly gives comments on project presentations which are given each week on a laptop/projector setup"

A course review is more extensive and includes a set of comments like this, answering a whole set of questions, and concatenated in the dataset:

> "The class is pretty much divided into three parts - random processes, state-space models, and dt/ct transformations. Mostly theory
> The staff seems to really care and put effort into making themselves available.
> It's an EE header. I haven't taken any other ones, but I'd be surprised if they were better than 011.
> They are useful. I don't really understand the time table - things covered in MW are on the pset due Thursday.
> Maybe better to have psets due Monday. The way it is, it decreases my already too little motivation to start psets early.
> The class notes are like 10 pages of notes with 30 pages of problems for each chapter. It'd be nice if their were more solved examples.
> They are pretty good. The staff is very lenient on mathematical errors on tests.
> I think it's fair, even though grades aren't out yet. PSets are graded 1to 3, and are pretty lenient. The staff says that upwarded trending grades are better than downward trending ones."

Baselines

We used two different baselines to compare with our predictors. First, we considered a mode classifier that will always predict whatever the majority gender in the training set was. Second, we considered a random classifier that predicts male with probability equal to the proportion of males in the training set. The results from our baseline classifiers can be seen below:

Classifier	Accuracy (AXO)	Accuracy (HKN)	F1 Score (AXO)	F1 Score (HKN)
Majority Baseline	0.5603	0.8571	0.7151	0.9231
Random Baseline	0.5078	0.7924	0.6700	0.8799

Methodology

Data processing

For each dataset, we constructed a script to extract comments and compile a data object for each instructor to capture the sets of comments and associated gender. We removed gendered pronouns from both sets of comments, and removed named references from AXO reviews when it became apparent that they were being identified as strongly weighted features.

The HKN data was a bit more structured than the AXO data, and so we were able to identify comments made for a course in general and for instructors specifically in the HKN data. We assigned course comments to the lecturer for a subject, in order to associate those comments with the most likely associated gender.

We then apportioned 50% of each dataset for training, 25% for test data, and 25% for a development set, used as training data for the bag of words vectorization and for parameter tuning for topic model vectorization.

Features

The first set of features we extracted from our data was done using a bag of words approach. We first acquired a list of stopwords, using the NLTK English corpora for stopwords, and after removing all stopwords from comments, used scikit-learn's CountVectorizer method.[4] From this, we were able to create feature vectors for each comment where the length of the feature vector was the size of the vocabulary in the training set, and indices in the feature vector that corresponded to words in the vocabulary were incremented by one each time that word appeared in the comment. In the AXO data, there was a vocabulary size of 1,923 words, while in the HKN data, the vocabulary size was 3,892.

The second set of features was determined using a topic modeling approach. Rather than considering only the words in students' text responses, we also wanted to consider what topics were used to evaluate the subjects and professor. Using the lda Python package,[5] we ran latent Dirichlet allocation (LDA) topic modeling with collapsed Gibbs sampling on the bag of words feature vectors to discover the topics associated with the dataset. The new feature vectors we used were the distribution of topics for each professor. We determined, using our validation set, that 7 topics gave us the best results.

[4] http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
[5] https://pypi.python.org/pypi/lda

Predictive models

We decided to do a survey of three various classifiers on our data. The three predictive models we used in an attempt to classify each instructor as male or female from their student text responses were: support vector machines (SVM),[6] logistic regression (LR),[7] and k-nearest neighbors (KNN).[8] To implement each of these models, we used scikit-learn's built-in implementations, more specifically to implement SVM, we used scikit-learn's svm.SVC class, to implement LR, we used scikit-learn's linear_model.LogisticRegression class, and to implement KNN, we used scikit-learn's KNeighborsClassifier class. For each model's hyperparameters, we used the default values provided by scikit-learn's implementation.

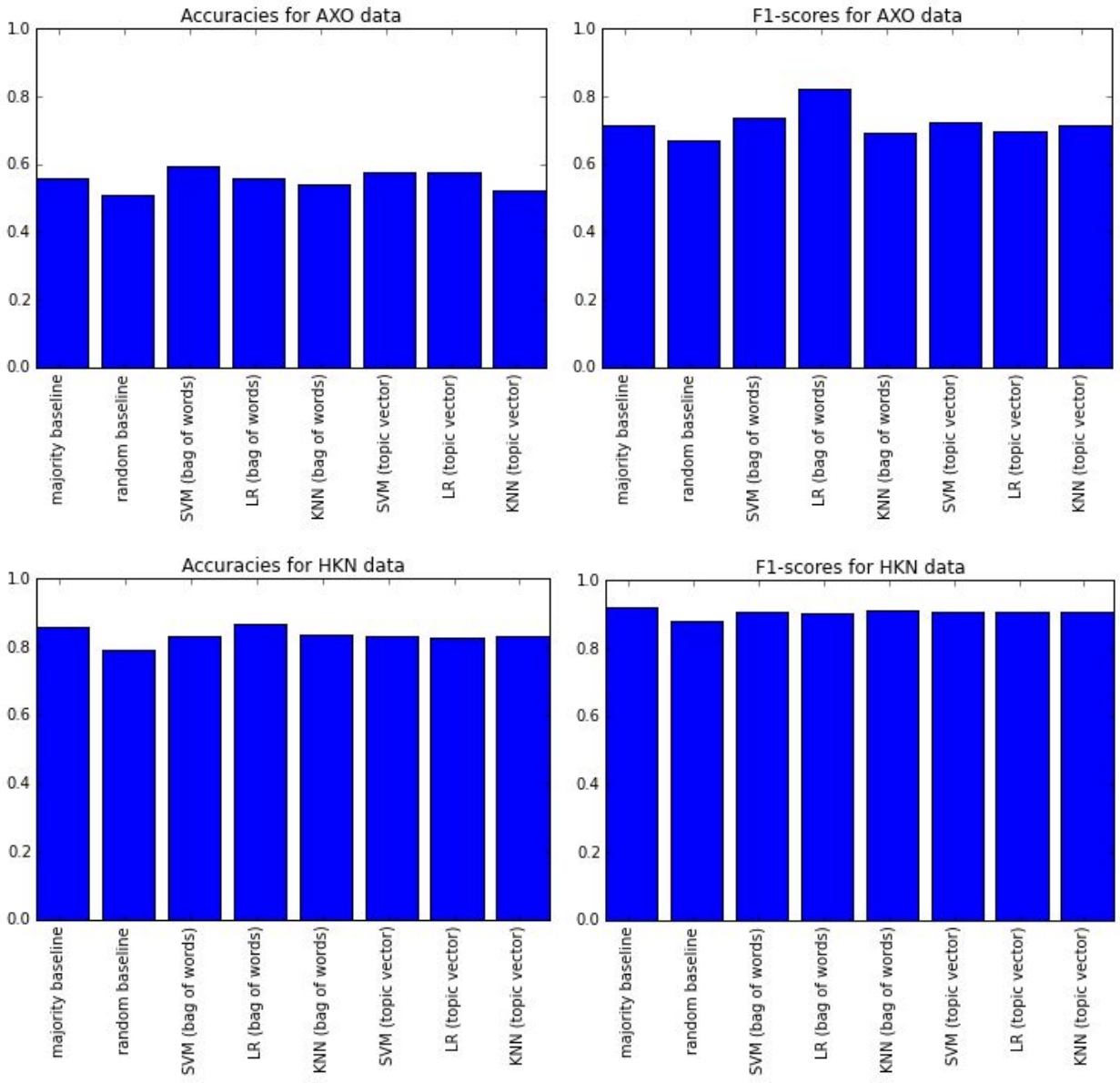
Results

Quality of predictive models

Unfortunately, our predictive models did not perform significantly better than our baselines. Any differences we saw in accuracy and F1 score were small. A summary of our results can be seen in the table and charts below:

Features	Model	Accuracy (AXO)	Accuracy (HKN)	F1 Score (AXO)	F1 Score (HKN)
Bag of words	SVM	0.5938	0.8295	0.7373	0.9067
Bag of words	LR	0.5573	0.8682	0.8224	0.9051
Bag of words	KNN	0.5413	0.8372	0.6943	0.9114
Topic vector	SVM	0.5776	0.8295	0.7223	0.9068
Topic vector	LR	0.5764	0.8287	0.6982	0.9068
Topic vector	KNN	0.5244	0.8295	0.7163	0.9068

[6] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[7] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[8] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html



Heavily weighted features

Even though our predictors either show little or no improvement to our baselines, we see some interesting results when we look at the heavily weighted features of the models. Starting with the AXO data, we picked out the 20 most heavily weighted features and reduced this list to consist only of adjectives. Qualitatively, we notice that many of the adjectives used to describe female instructors are comments on their personality, more so than their teaching style or mastery of content. For instance, we see “energetic”, “helpful”, and “friendly” used to describe female instructors. Meanwhile, for men, we see some adjectives used to describe their personality (“awesome”, “cool”, “fun”), but also adjectives that describe their class and mastery of the material (“easy”, “hard”, “knowledgeable”). Although, we do not see this same distinction when

looking at the HKN data, perhaps this can be explained by the lack of data available on women instructors in the data set. Nevertheless, we think that the differences in adjectives describing men and women indicate at least the possibility of gender bias in student evaluations.

Male (AXO)	**Female (AXO)**	**Male (HKN)**	**Female (HKN)**
Knowledgeable	Super	Clear	Suggestions
Awesome	Amazing	Good	Prompt
Cool	Gender	Helpful	Informed
Fun	Energetic	Interesting	Learning
Flexible	Helpful	Available	Replying
Good	Simple	Liked	Prepared
Easy	Nice	Understand	Responded
Weird	Enjoyable	Friendly	Comfortable
Hard	Friendly	Worth	Feedback

Discussion

We were not able to effectively classify the gender of the instructor based upon language in the comments at a significantly better rate than our baselines. We saw only small improvements over the baselines for some tests. We believe this may be a result of the data we were able to procure, rather than a definitive statement of lack of gendered difference in instructor evaluation language.

We believe that all-female authorship and public reviewer attribution of the AXO comments may have contributed to the lack of discernible bias in that dataset, and that additional data is needed to confirm results.

We were not allowed access to very much data for our HKN Underground Guide dataset, and very few female instructors were present in the limited sample, a feature which is exacerbated by the gender imbalance in the EECS faculty in general. Significantly more data in this set is necessary to make conclusive judgements.

In addition to our quantitative results, we performed a qualitative analysis of the features that our bag-of-words models weighted most heavily for each gender in each dataset. The differences in

features identified for male and female instructors in the AXO dataset are fascinating, and indicate that there may in fact be a discernible difference in reviewer language given more data to analyze.

Future Work

Larger dataset

The data shared with us was rather limited and exploration of much larger and more varied data is certainly necessary. Data from multiple schools, multiple departments, more instructors, and more reviews per instructor would all be useful to explore the presence of language differences and the scope of the differences beyond a single population of instructors and reviewers.

Gender of the reviewer/audience

One interesting feature of our AXO dataset was that all reviewers and the primary intended audience members were female, and we had no information available about the gender of HKN reviewers. It would be interesting to explore how the gender of the reviewer, the the subject of the review, and the intended audience (when limited) each separately, and in combination, affect the language used in the review.

Evaluations for different subjects

Each of our datasets focused on a subset of MIT classes. AXO reviews focused on a selection of Humanities, Arts, and Social Science classes, while HKN reviews focused on a selection of classes in Electrical Engineering and Computer Science. The relative gender imbalances in different academic fields as well as differences in general thought patterns in different fields of study might each impact the way that students in certain majors review instructors in each field. Exploring reviews across subjects and in comparison to the aggregation of many fields could shed light on this question.

Attributed vs. anonymized evaluations

Many subject evaluations are anonymously submitted, but our AXO reviews were publicly attributed by name and graduation year, and the HKN data we procured includes the usernames of students that responded, though not associated directly with their responses. It would certainly be enlightening to explore differences in language for instructors of different genders when reviewers do and do not have an expectation of privacy. It is possible that reviewers put forth greater effort to be unbiased when they know their comments can be traced back to them. The anonymous nature of evaluations is often considered crucial to their usefulness, precisely because reviewers are more likely to share their true, unfiltered thoughts when they are unidentifiable.

Correlation with numerical rankings

Instructors are often rated numerically on their general quality of instruction, and it may be interesting to explore how the amount of difference in language or the type of difference in language might change when looking at instructors that are all ranked highly or all ranked poorly.