Analyzing Gender Bias in Student Evaluations

Andamlak Terkik (at3616@rit.edu)

Co-Advisors:
   Dr. Emily Prud'hommeaux
   Dr. Cecilia Ovesdotter Alm
   Dr. Christopher Homan
   Dr. Scott Franklin

A Project Submitted in Fulfillment of the Requirements for the Degree of
Master of Science in Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology

May, 2016
Dedicated to my parents.
Acknowledgments

I would like to greatly thank my advisors Dr. Emily Prud'hommeaux, Dr. Cecilia Ovesdotter Alm, Dr. Christopher Homan and Dr. Scott Franklin for providing me an interesting project (and data!) to work on. I’m also thankful for their valuable guidance and feedback along the way. I would like to thank Dr. Reynold Bailey for providing helpful supervision ensuring a smooth progression of the project. My academic advisor, Cindy, and Dr. Hans-Peter Bischof were always highly supportive and effective in their help throughout my 2 years at RIT-- thank you. Finally, I’m thankful to my fellow CS students for the friendship and support. I’ll miss those interesting discussions during late-night lab sessions.
Analyzing Gender Bias in Student Evaluations

Andamlak Terkik

Abstract

Sentiment analysis is a computational linguistics task that characterizes positive-negative tone expressed in language data. We apply sentiment analysis to analyze student comments spanning over eight years of College biology, chemistry, physics and math and explore differences in sentiment conveyed toward male and female faculty. We study differences in perceptions found in categories pertaining to instructor competence, organization/presentation, personality/helpfulness and overall satisfaction. We also use term ranking to identify words that are strongly associated with either gender.

1 Introduction

Students evaluations of teachers (SETs) are critical to instructors teaching careers. SETs affect teachers’ tenure and promotion decisions, compensations as well as teaching awards.

Sentiment analysis is the task of identifying positive and negative tone expressed in natural language text (Liu, 2015; Mohammad, 2016). Sentiment analysis methods can be roughly grouped into supervised or unsupervised methods. In supervised methods, sentiment classifiers are trained into supervised or unsupervised methods. In supervised methods, sentiment classifiers are trained using correctly labeled data. On the other hand, unsupervised sentiment analysis methods often work by aggregating individual sentiment scores of phrases and aggregating the results to give a final sentiment class (Turney, 2002). Lexicons containing the terms and associated subjectivity scores will affect the results on such methods. These methods are known to suffer from inaccuracies stemming from the their inability to model contextual information occurring in text.

In the first part of this paper, we present a supervised sentiment classifier model trained to classify the polarity of SET comments. This trained model is then applied onto a much larger set of the same (unlabeled) dataset to examine how sentiment varies across both genders.

2 Background

There is a lot of previous work investigating whether instructor gender affects the student commentaries and the extent to which it does.

MacNell, Driscoll & Hunt (2015) investigated the existence of gender bias in SETs by conducting an online college-level class. They ran an experiment by carefully disguising the genders of the instructors and collecting evaluations from the students at the end. The reported results indicate markedly lower scores to the perceived female instructor both in teaching effectiveness and interpersonal skills. This same perceived female instructor was able to achieve statistically significant higher ratings in a separate section of the same class where only the instructor’s gender identifiers were changed to be perceived as male. Conversely, a male instructor received significantly less favorable evaluations when he was perceived as a female in a separate section.

Over 22,000 student comments compiled from a French university were analyzed by Boring (2015). The analysis measures perceived teacher effectiveness from different dimensions such as class preparation, knowledge, enthusiasm, assessment criteria etc. The final analysis concludes that both student and instructor genders play a role in teaching evaluations. It found that male students are 30% more likely to give an excellent overall rating for satisfaction to men teachers in comparison to women teachers.

Young, Rush & Shaw (2009) examined the role
of student gender played in SET outcomes by running an experiment involving a total of 723 students from 5 different colleges. The student body were comprised from students having varying levels of education, gender and course characteristics. The experiment asked students to evaluate instructors on pedagogical, interpersonal and course characteristics. Results of the experiment showed that student gender also had a marked role in affecting SETs.

While most previous studies focused on analysis of numerical scores, this project explores sentiment analysis of SETs. The learning model utilizes such features as unigrams, bigrams and sentiment scores based on term counting. A supervised approach for sentiment classification is shown to be more effective when compared to other unsupervised or document clustering approaches.

3 Sentiment Analysis

3.1 Dataset

The small-labeled set is labeled with satisfaction levels conveyed in the comments and is composed of student responses reflecting on their instructor’s effectiveness. All of the responses are responding to the prompt Comment on the instructor’s strength and weaknesses. While the set medium-labeled also contains responses to the same prompt, the comments are unlabeled. On the other hand, the large-unlabeled set contains responses to multiple distinct set of prompts and the comments are not labeled with satisfaction levels.

The small-labeled data is composed of 2076 comments written by students. These comments were annotated in terms of 5 levels of satisfactions conveyed in the text. In increasing orders of level of satisfaction, the options were very satisfying, somewhat satisfying, neither satisfying nor dissatisfying, somewhat dissatisfying and very dissatisfying.

We trained a supervised sentiment classifier model on the small-labeled data and use the trained model to predict the affect for the comments in medium-unlabeled. Before applying it on the medium-unlabeled set, the trained classifier is evaluated on a held-out subset taken from small-unlabeled. The third dataset, large-unlabeled contains student comments responding to multiple types questions not found in the two smaller sets. Therefore, this large-unlabeled set is used for computing analytics and identifying terms strongly associated with either of the genders.

A randomly chosen 15% subset of small-unlabeled was held-out as final test data. Table 1 presents statistics on the remaining 85% of the dataset. Analyzing the length of comments revealed that student comments for men are, on average, 10% longer than comments for women.

3.2 Inter-rater agreement

Annotating the sentiment of a given text can sometimes be subjective. Since supervised sentiment analysis relies on human annotation, it’s useful to measure the reliability of the ratings. Therefore, three independent annotators, including the person who rated the entire small-labeled set, were asked to rate a subset of 100 comments using the previously described rating scheme for affect.

The annotated set was then analyzed by computing Fleiss’ kappa (Randolph, 2005). Fleiss’ kappa is a variant of Cohen’s kappa (Byrt et al.,

<table>
<thead>
<tr>
<th>Data set</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>small-labeled</td>
<td>2,056</td>
</tr>
<tr>
<td>medium-unlabeled</td>
<td>15,896</td>
</tr>
<tr>
<td>large-unlabeled</td>
<td>107,855</td>
</tr>
</tbody>
</table>

Table 1: Three comment sets used. One small-labeled set & two larger, unlabeled sets

![Figure 1: Distribution of 5 satisfaction levels](image)
that assesses agreement among more than two raters. The kappa scores were computed for two groupings. For the first group, each of the five satisfaction classes are taken independently. For the second group, the two extreme classes on each side of the range were merged which results in a 3-class rating scheme. Naturally, the three-class agreement scores exhibit less ambiguity and thus raise the agreement scores.

Table 3 shows the Fleiss’ Kappa and overlap percentage of the results for both 5-class and 3-class groupings. Both of these scores fall in the range of 0.4 to 0.6 which indicates a moderate agreement according to a commonly cited scale of kappa interpretation. It’s also worth noting that inter-rater agreement scores such as Kappa score greatly depend on the task since it’s usually easier to agree on some tasks such as Part-of-speech tagging than more subjective tasks such as sentiment classification (Landis and Koch, 1977).

3.3 Classification Tasks

The sentiment analysis part was divided into two different tasks of varying levels of difficulty.

1) Extremes task: classifying extreme classes by differentiating between very dissatisfied and very satisfied

2) Merged task: classifying merged classes by differentiating between (very dissatisfied together with somewhat dissatisfied) and (somewhat satisfied together with very satisfied)

3.4 Classifier Architecture

The classifier is designed as a two-tier system sometimes referred to as Ensemble via stacking. The first tier is composed of five classifiers trained using unigram, bigram and trigram features. The ensemble of classifiers is composed of a Random Forests model (with 80 trees and maximum depth value of 100), a Multinomial Naive Bayes model, a Bernoulli Naive Bayes and a Logistic Regression model. The class likelihood predictions obtained from these 4 models are then combined together with sentiment term scores to train a final classifier at the second tier. A Random Forest model with similar parameters as the one in tier one is also used to train the final classifier. Figure 3 shows the architecture of the classifier. For implementation, the scikit-learn machine learning library (Pedregosa et al., 2011) was used.

Random Forests are modeled by an ensemble of trees where each tree is trained by random sub-
set of the dataset (Breiman, 2001). The Random Forests model was trained and evaluated with the parameters set to 100 trees having a maximum depth of 80. In previous studies, both of the Naive Bayes models are known to perform well in similar sentiment classification tasks (Pang et al., 2002). Both the Multinomial Naive Bayes and the Bernoulli Naive Bayes models were set to use Laplacian smoothing with value 1.0 with uniform priors.

3.5 Pre-processing

Before extracting features for the sentiment classifier, text normalization was done by lowercasing all of the comments and removing special characters.

3.6 Term Negation

Another crucial step undertaken during pre-processing is proper handling of negation terms. For instance, phrases such as great at teaching need to be differentiated from negative sentiments such as not great at teaching. In this project, 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. For this, the method described by Vivek et al (2013) was used during feature extraction.

Term negation was responsible for an increase of about 4 percentage points in classification accuracy. 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 newer terms prefixed with not –. For instance, not great at teaching would be turned into not not great not at not teaching. After a negation marker is once set, it negates every word that follows it until a punctuation mark or another negation term is encountered.

3.7 Lexical Features

3.8 N-grams

Currently, ngram features are extracted by using raw frequency counts of unigrams, bigrams and trigrams from the comments. Bigrams and trigrams capture useful contextual features that unigrams are unable to model. For example, while terms such as intelligent and nice will be accurately identified by unigrams, important phrases such as extremely well, very kind can only be captured using bigrams. Using these text features, over 30,000 terms were extracted for every comment. All three sets of n-grams, from unigrams to trigrams, were used both separately as well as in combination and evaluated on each of the tasks.

3.9 Term Scoring

Sentiment term scores are obtained by computing the aggregate positive and negative scores for each comment. To compute these aggregate scores, the prior polarities of the terms are looked up from domain-independent lexicons. We utilize 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 are computed. This is done by looking up scores from each of the three lexicons. Therefore, a 6-valued (i.e. 3 dictionaries x 2 sentiments) feature vector is computed for each comment.

3.10 Current Results

Both the majority baseline and main ensemble classifier are evaluated on the two tasks described above. We measure the classification accuracy and F1-score (Powers, 2011). All of these metrics were computed following a 5-fold cross-validation on the development set. Results for both of the
Table 4: Classification accuracy and f1-score results for both tasks

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>RF Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremes task (n=670)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>86%</td>
<td>91%</td>
</tr>
<tr>
<td>F1-score</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Merged task (n=1360)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>69%</td>
<td>80%</td>
</tr>
<tr>
<td>F1-score</td>
<td>57%</td>
<td>79%</td>
</tr>
</tbody>
</table>

tasks are shown in table 4. As the table shows, the ensemble classifier greatly reduces the error rate for both tasks compared to the baseline one. This is particularly true for the more ambiguous merged task. This demonstrates how effective the combination of n-grams and lexicon-based term scoring features are for sentiment analysis.

4 Further Analysis

4.1 Satisfaction by Gender

For the small-labeled dataset, we computed the ratio of satisfied comments to dissatisfied class broken down by gender on small-labeled. This was also done for the larger medium-unlabeled dataset after using the trained classifier to predict the comments. The results are given in table 5.

While the table shows that male instructors receive slightly less favorable satisfaction ratings in the small-labeled set, this is possibly due to the small size of the set. This is confirmed by looking at the results on medium-unlabeled set in which the satisfaction levels are predicted to be equally distributed for both genders.

On a related analysis, a comment predicted as dissatisfied was found to be on average 2x longer than a satisfied comment. This was done on the medium-unlabeled set.

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>small-labeled</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfied</td>
<td>74%</td>
<td>62%</td>
</tr>
<tr>
<td>Dissatisfied</td>
<td>26%</td>
<td>38%</td>
</tr>
<tr>
<td>medium-unlabeled</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfied</td>
<td>94%</td>
<td>94%</td>
</tr>
<tr>
<td>Dissatisfied</td>
<td>6%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Table 5: Affect distribution broken down by gender for small-labeled and medium-unlabeled

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

<table>
<thead>
<tr>
<th></th>
<th>F (per 1000)</th>
<th>M (per 1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professor</td>
<td>167</td>
<td>209</td>
</tr>
<tr>
<td>&lt;Last name&gt;</td>
<td>139</td>
<td>151</td>
</tr>
<tr>
<td>Dr</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Teacher</td>
<td>95</td>
<td>79</td>
</tr>
<tr>
<td>Instructor</td>
<td>172</td>
<td>172</td>
</tr>
<tr>
<td>&lt;First name&gt;</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>

4.2 Gendered Language

To understand how word usage differs for both genders, words were ranked based on the strength of their co-occurrence with each gender. Mutual information (MI) (Church and Hanks, 1990) was used for measuring term-gender co-occurrence. The top 200 words as measured by MI score were selected and categorized into two groups based on the words’ semantic function. The first group contained terms of address while the second group contained words used to describe a teacher or the course.

After the words and groups were identified, raw frequency counts broken down by gender were computed for each word. Standard text normalizing steps such as special character removal, text lowercasing, negation handling and normalizing possessive noun forms (For example, transforming the Professor’s notes to the professor notes) were undertaken beforehand. The analysis was done in the large-unlabeled set. We report the occurrence count of each term per 1000 comments after adjusting the number of comments for each gender.

The results show that students are significantly more likely to address their male instructors with more professional or formal titles such as Professor or refer to them by their last names. Conversely, female teachers are more likely to be simply addressed as teacher or instructor followed by their first names. It’s crucial to keep in mind that the comments were compiled from an institution having a faculty with roughly similar distributions of professional qualifications for both genders. Therefore, this finding demonstrates a bias towards using more regarded titles and similar addressing mechanisms for male instructors.

On the other hand, the results from the second category confirms previous claims that women faculty are more likely to be regarded as wonder-
Table 7: Gender differences in keywords used to describe men vs. women faculty. Keywords describing instructor/course per 1000 comments adjusted by number of comments for both genders.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>F</th>
<th>M</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazing</td>
<td>32</td>
<td>18</td>
<td>128%</td>
</tr>
<tr>
<td>Love[d]</td>
<td>59</td>
<td>32</td>
<td>84%</td>
</tr>
<tr>
<td>Wonderful</td>
<td>28</td>
<td>12</td>
<td>57%</td>
</tr>
<tr>
<td>Organized</td>
<td>243</td>
<td>178</td>
<td>37%</td>
</tr>
<tr>
<td>Willing</td>
<td>114</td>
<td>88</td>
<td>30%</td>
</tr>
<tr>
<td>Helpful</td>
<td>454</td>
<td>402</td>
<td>13%</td>
</tr>
<tr>
<td>Tangent[s]</td>
<td>3</td>
<td>16</td>
<td>400%</td>
</tr>
<tr>
<td>Funny</td>
<td>4</td>
<td>14</td>
<td>250%</td>
</tr>
<tr>
<td>Knowledgeable</td>
<td>21</td>
<td>33</td>
<td>57%</td>
</tr>
<tr>
<td>Interesting</td>
<td>68</td>
<td>92</td>
<td>35%</td>
</tr>
<tr>
<td>Understanding</td>
<td>110</td>
<td>126</td>
<td>15%</td>
</tr>
</tbody>
</table>

The classifier was then used to analyze how sentiment satisfaction differs across genders and the relation between comment length and satisfaction level. We found sentiment is equally distributed for both genders. Additionally, male faculty are more likely to go on tangential topics, based on the comments analyzed.

5 Conclusion

In this project, we developed a reliable sentiment classifier for student evaluation comments. The classifier was then used to analyze how sentiment satisfaction differs across genders and the relation between comment length and satisfaction level. We found sentiment is equally distributed for both male and female instructors. Future work on this regard can improve on the classifier by using more higher-level features such as semantic roles and parts-of-speech.

On the other hand, analysis of terms more strongly associated with gender showed that women are more likely to be perceived as willing, organized and helpful while men faculty are perceived as funny and knowledgeable.

References


