Analyzing Gender Bias in Student Evaluations
Andamlak Terkik at3616@rit.edu
Advisors: Emily Prud’hommeaux, Cecilia Ovesdotter Alm, Christopher Homan, Scott Franklin
Rochester Institute of Technology

MOTIVATION & GOALS

• Teaching evaluations are one metric used for promotion and tenure decisions, despite evidence of gender bias

• Previous work has shown students assign lower scores in their teaching evaluations of instructors they believe to be women [1]

• Can comments’ sentiment be predicted using NLP?

• Can gender bias be detected using NLP methods?

GENDER ANONYMIZED DATA

<table>
<thead>
<tr>
<th>Data set</th>
<th>Phase</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>small-labeled</td>
<td></td>
<td>2,045</td>
<td>Labeled set, one question</td>
</tr>
<tr>
<td>medium-unlabeled</td>
<td></td>
<td>15,896</td>
<td>Unlabeled, one question</td>
</tr>
<tr>
<td>large-unlabeled</td>
<td></td>
<td>107,855</td>
<td>Unlabeled, multiple questions</td>
</tr>
</tbody>
</table>

Table 1: Three comment sets used. One small-labeled set & two larger, unlabeled sets to motivate the applied experimental design

• Distribution of satisfaction levels on the labeled dataset

Fig 1: Satisfaction level of small-labeled data skew towards neutral to positive level. VD - Very Dissatisfied, SD - Somewhat Dissatisfied, N - Neutral, SS - Somewhat Satisfied, VS - Very Satisfied.

• Comments are on average 10% longer for male instructors in large-unlabeled set

• Text normalization (removing punctuation, lowercasing)

• Negation handling by transforming phrases followed by no, not, n’t
  Example: not great at teaching becomes not not_great not_at not teaching

FEATURES

• N-gram word frequencies

• Aggregated positive and negative sentiment scores for each comment based on the MPQA, Bing Liu's and NRC EmoLex sentiment lexicons [2]

SENTIMENT CLASSIFIER

• A binary sentiment classifier was trained on small-labeled to predict satisfaction levels

• Stacking ensemble architecture: We train 4 distinct models on n-gram features, then combine their predicted probabilities with term scoring features into a joint classifier

APPLYING CLASSIFIER

• Ran the merged classifier on the medium-unlabeled set

• Sentiment/satisfaction is evenly distributed across genders in medium-unlabeled

• The median dissatisfied comment is 2x longer than the median satisfied comment in medium-unlabeled

GENDERED LANGUAGE

• Using mutual information score, the top 200 most correlated terms for each gender were selected and categorized

• Gender differences in some terms of address used to refer to women vs. men faculty

CLASSIFIER RESULTS

• Two tasks were evaluated on the small-labeled set: classifying extreme classes (VD or VS) and classifying merged classes (VD + SD or SS + VS)

• Ensemble classifier outperforms the baseline on both tasks. Particularly, on the merged task

VALUES

<table>
<thead>
<tr>
<th>Phrase</th>
<th>F</th>
<th>M</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professor</td>
<td>488</td>
<td>610</td>
<td>25%</td>
</tr>
<tr>
<td>Teacher</td>
<td>276</td>
<td>221</td>
<td>25%</td>
</tr>
<tr>
<td>&lt;Last name&gt;</td>
<td>408</td>
<td>442</td>
<td>8%</td>
</tr>
<tr>
<td>&lt;First name&gt;</td>
<td>288</td>
<td>54</td>
<td>2%</td>
</tr>
<tr>
<td>Dr</td>
<td>185</td>
<td>188</td>
<td>2%</td>
</tr>
<tr>
<td>Instructor</td>
<td>490</td>
<td>481</td>
<td>2%</td>
</tr>
</tbody>
</table>

Table 2: Terms of address used to refer to faculty. Term frequency per 1000 comments adjusted by number of comments for both genders

CONTRIBUTIONS

• Built successful sentiment classifier for evaluation comments

• Showed difference in language usage when referring to instructors

• Provided evidence that negation handling improves classifier accuracy in evaluation data

SELECT REFERENCES


http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Bing Liu’s Opinion Lexicon: https://www.cs.uic.edu/~liub/FLSS/sentiment-analysis.html#Lexicon

NRC EmoLex: http://nrextomhanns.com/WebPages/NRC-EmoLex-Emotion.htm