Sentiment Analysis of Student Evaluations of Teaching

Heather Newman(✉) and David Joyner(✉)

Georgia Institute of Technology, Atlanta, GA, USA
{newman, david.joyner}@gatech.edu

Abstract. We used a sentiment analysis tool, VADER (Valence Aware Dictionary and sEntiment Reasoner), to analyze Student Evaluations of Teaching (SET) of a single course from three different sources: official evaluations, forum comments from another course, and an unofficial "reviews" site maintained by students. We compared the positive and negative valences of these sites; identified frequently-used key words in SET comments and determined the impact on positivity/negativity of comments that included them; and determined positive/negative values by question on the official course SET comments. Many universities use similar questions, which may make this research useful for those analyzing comments at other institutions. Previous published studies of sentiment analysis in SET settings are rare.

Keywords: Sentiment analysis · Student evaluation of teaching
Course evaluations · Natural language processing

1 Introduction

Student evaluations of teaching (SET) are an important part of universities' self-improvement programs, providing a viewpoint that may affect everything from professors' tenure case decisions to the structure of future semesters of those courses. SET typically include qualitative comments that may be difficult to present in a summary manner. Sentiment analysis, a form of natural language processing, attempts to assign a positive, negative or neutral valence or polarity to natural speech. We set out to determine whether sentiment analysis is a viable tool for analyzing evaluations.

1.1 Evaluation Sources

We analyzed evaluations of a single graduate-level online course of several hundred students over a period of two semesters. By limiting the evaluations to a single course, we were able to control for variability by instructor, semester, and course material. We analyzed three separate sources of SET: official course evaluations, consisting of a series of quantitative measurements followed by qualitative open-ended questions; informal peer evaluations from an unofficial online course evaluation site with quantitative and qualitative rankings and comments; and postings in another course, where students were asked to discuss specific classes they had taken.

C. Penstein Rosé et al. (Eds.): AIED 2018, LNAI 10948, pp. 246–250, 2018.
https://doi.org/10.1007/978-3-319-93846-2_45

1.2 VADER as a Sentiment Analysis Tool

After experiencing poor results with a standard SentiWordNet analysis [1], we turned to a more-sophisticated analytical tool for sentiment in informal postings. We found good results with VADER, the Valence Aware Dictionary for sEntiment Reasoning [2]. VADER not only analyzes individual word sentiment, but attempts to predict the normalized valence of positive or negative sentiment based on overall sentences, accounting for factors such as negation, punctuation or emoticon usage. It provided consistent analysis of SET comments, which are often written informally.

2 Related Work

Student evaluations may be flawed overall in how closely they track with the actual educational outcomes of a particular class; past studies have shown that positive evaluations may not correlate well with student learning, and that other factors may be in play [3–5]. Given that issue, the detailed sentiment analysis by topic discussed here may offer an option for instructors or institutions attempting to do a deeper dive into evaluations than just the summary ratings, by identifying classroom themes or components and students' positivity or negativity toward them.

Lim et al's [6] study on course evaluations did not focus on sentiment analysis per se, but a tangent: frequency analysis to determine key features of course evaluations. The applications of this work to the word groupings in our study seem very relevant. El-Halees' [7] analysis of comments to improve course evaluations comes the closest to approaching the subject of this study. The author conducted analysis identifying overall sentiment and features including teacher, exams, resources, etc., assigning sentiment to those features. He used NB, k-Nearest Neighbor and SVM methods.

3 Viability of Sentiment Analysis in SET

One method of determining viability for sentiment analysis in student evaluations of teaching comments using our datasets was to compare comments' sentiment polarity using a particular method with student-awarded quantitative scores. Just one set of evaluations included individually calibrated student comments and evaluation scores: the informal student website rankings. Official course evaluation survey data pre-summarized quantitative data, making it impossible to match scores with particular comments. The informal site included an overall rating for the entire class (based on a 1–5 scale) and comments on the class from the same individual.

3.1 Results

A clear trend can be seen between the average normalized compound sentiment scores for students awarding low quantitative scores versus those awarding high scores, suggesting that sentiment analysis does generally track with students' overall estimations of course value. A dip occurs at the "3" rating, possibly due to students' association of the mid-level “3” rating with “average,” as opposed to a true progression from 1 to 5. This supports the use of sentiment as an evaluative measure, especially in these edge cases, since it can be used to tease out the true expressed negative/positive valence of comments for those awarding an “average” score (Fig. 1).

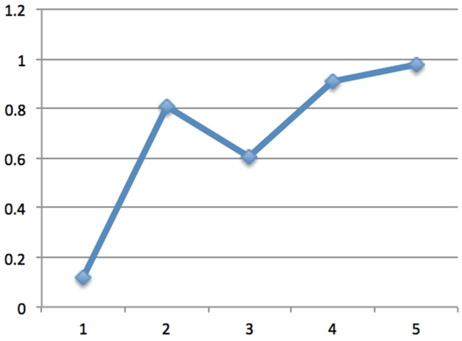

Fig. 1. Normalized compound VADER sentiment polarity of student comments expressed as a function of student quantitative ratings of the same course (1 low to 5 high). Normalized compound scores of sentiment range from −1 (completely negative) to +1 (completely positive.)

While evaluation comments by students overall tend to employ positive terms, the average for those who awarded 1 of 5 stars was 11.72% (.1172) in normalized VADER compound scores, while the average for those awarding 5 of 5 stars was close to universally positive, at 97.96% (.9796).

3.2 Differences in Evaluation Comment Sentiment Scores Based on Environment

All three sources scored roughly the same on compound ratings; however, the informal student website scores were slightly more likely to be negative (7.5%, versus 5.9% for the forum posts and 6.2% for official evaluation comments), while also being slightly more likely to include fewer positive comments (13.2%, versus 15% for forum posts and 15.5% for official evaluations.)

3.3 Scores for Particular Question Types

We tracked sentiment scores by question by term, since the course had changed. In general, positive questions asking about the course or instructor’s strengths received the lowest negative and highest positive scores; and vice versa for questions asking about weaknesses, supporting our methodology. For instance, in Fall 2016, “course best aspect” had an average normalized positive valence of 20.2% and negative of 0.5%; “course improvements” had a positive valence of 10.7% and negative of 7.7%.

3.4 Scores for Comments Including Frequently-Used Keywords

One of the key findings that could be helpful for evaluating parts of a course that students resonated with more or less strongly is the list of items (nouns) that students mention most frequently. We analyzed the most frequently used terms and averaged the sentiment polarity for comments using those terms.

We iterated through the comments themselves, identifying whether they contained one of these frequently-occurring nouns, and if so, adding it to a total score for that noun. The resulting totals were averaged to produce positive, negative and compound normalized VADER sentiment scores for each noun. For samples, see Tables 1 and 2.

Table 1. The top two frequently-used nouns with negative associations.

Word	Neutral	Negative	Positive	Compound
Feedback	0.812	0.091	0.097	0.9816
Questions	0.814	0.084	0.102	0.9996

Table 2. The top two positive terms for sentiment.

Word	Neutral	Negative	Positive	Compound
Interviews	0.763	0.051	0.185	0.9999
Idea	0.779	0.064	0.157	1

4 Discussion and Limitations

Sentiment analysis cannot provide a replacement for the content and contextual analysis done manually now. It may break down in environments where student comments are too short or factually phrased to provide consistent results. This analysis focused on a single course within a single degree program, and further study is needed to determine whether the results found here carry over to other types of classes (e.g., traditional in-person instruction) and other types of evaluative measures.

5 Conclusion

The potential for sentiment analysis as a tool for analyzing Student Evaluations of Teaching appears to be significant. It offers an additional summarization tool for "quick looks" at positive and negative factors within a single class. Use of frequently occurring keywords might help to identify where the course instructor was strong but particular materials were weak, or vice versa. Correlation between overall sentiment analysis scores for a review and overall scores awarded to a class appear to support the validity of sentiment analysis as a measurement.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 17–23 May 2010
2. Hutto, C.J., Gilbert, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM 2014)
3. Marsh, H.W., Roche, L.A.: Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias, and utility. Am. Psychol. **52**(11), 1187 (1997)
4. Uttl, B., White, C., Gonzalez, D.: Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. Stud. Educ. Eval. **54**, 22–42 (2017)
5. Boring, A., Ottoboni, K., Stark, P.B.: Student evaluations of teaching (mostly) do not measure teaching effectiveness. ScienceOpen Res. **10** (2016)
6. Lim, S.D., Lee, J., Park, H.S., Yu, J., Lee, K.Y., Sohn, I.S., Lee, R.: Experience and consideration on online course evaluation by medical students. Korean J. Med. Educ. **20**(4), 367–371 (2008). https://doi.org/10.3946/kjme.2008.20.4.367
7. El-Halees, A.: Mining opinions in user-generated contents to improve course evaluation. In: Zain, J.M., Wan Mohd, W.M., El-Qawasmeh, E. (eds.) ICSECS 2011. CCIS, vol. 180, pp. 107–115. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22191-0_9