Analyzing Student and Instructor Comments using NLP

Zack Butler
zjb@cs.rit.edu
Rochester Inst. of Tech.
Rochester, NY, USA

Ivona Bezáková
ib@cs.rit.edu
Rochester Inst. of Tech.
Rochester, NY, USA

Shaoxuan Xu
cx2336@rit.edu
Rochester Inst. of Tech.
Rochester, NY, USA

Angelina Brilliantova
lb9849@rit.edu
Rochester Inst. of Tech.
Rochester, NY, USA

ABSTRACT

We report on our experience using common natural language processing (NLP) tools to analyze two vastly different data sets of free-form responses collected during a study of assignments in introductory computing courses. Our first data set consists of typically short comments left by hundreds of students on assignment surveys. Our second data set is comprised of semi-structured individual interviews of eight instructors of up to an hour long each. We collected the data across several years as part of our investigation of the use of pencil puzzles as a context for introductory computer science. In an earlier work, we manually analyzed a fraction of the student comments (all data collected until that point), using grounded theory. The results were illuminating, but the process was very time consuming, consisting of manual assignment of a small number of codes to each comment. In this work, we investigate the usability of common NLP tools to speed up the process for the entire data set of student comments. We also applied these tools to the instructor interviews. The NLP tools do not appear to be effective to create the code base, but, once the code base was determined, they performed the actual coding (assignment of codes to each student comment) promisingly well. For the long-form instructor interviews, the situation was much more challenging, due to the wide-ranging nature of semi-structured interviews, interleaving discussion topics, and elements of natural speech. We report on the lessons learned while automatically analyzing these complex data sets.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**.

KEYWORDS

Instructor interviews, Student survey comments, Manual qualitative analysis, Natural language processing

ACM Reference Format:
Zack Butler, Ivona Bezáková, Shaoxuan Xu, and Angelina Brilliantova. 2024. Analyzing Student and Instructor Comments using NLP. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE 2024), March 20–23, 2024, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3626253.3635593

SIGCSE 2024, March 20–23, 2024, Portland, OR, USA

ACM ISBN 979-8-4007-0424-6/24/03.
https://doi.org/10.1145/3626253.3635593

1 INTRODUCTION AND BACKGROUND

Grounded theory [3] is a well-established technique for qualitative analysis of free-form responses. It consists of the creation of a code base — a small set of strings capturing the most common themes in the data set. Then, a small number of codes are associated with each comment or a part of the free-form text (for example, a paragraph or a sentence in an interview). Having thus coded the free-form responses, standard statistical approaches can be applied to study the occurrences and co-occurrences of individual codes, determining correlations between the themes and sentiments in the responses.

The manual creation of the code base (often done using triangulation, when multiple researchers compare their proposed code bases and arrive at a mutually agreeable compromise), and, even more so, the manual assignment of codes to individual comments, paragraphs, or sentences, is very time consuming. We investigate the use of common NLP tools to help automate the process.

We collected our data set over several years, and across several institutions. We studied the use of *pencil puzzles* (puzzles such as sudoku that are designed to be solved by people, whether with a pencil or an app) as a context for introductory CS assignments. We initially collected data (student surveys including Likert-scale questions and open-ended comment boxes, plus grade data) at our home institution. We analyzed the Likert-scale data, and found that students reacted positively to these assignments, and that their experience was independent of their prior exposure to computing or of their gender [1]. We also performed qualitative analysis using grounded theory on the open-ended comments [2]. Notably, this work required manual coding of over 1000 individual comments.

As a followup to the initial study, we recruited 10 collaborating instructors (at institutions ranging from R1 schools to small liberal-arts schools to an international campus of our own university), who delivered puzzle-based assignments in 14 different course sections and allowed us to survey their students. Throughout this second study, we also conducted semi-structured interviews with the collaborating instructors and some of their teaching assistants, reflecting on how they valued the use of puzzle-based assignments and their adoption into their courses. We used grounded theory to manually code the interviews and the student comments in this

larger data set. Recent advances in large language models may allow some of this coding and analysis to be automated, and since we have the human coding available for comparison, we find ourselves in a strong position to evaluate the efficacy of such approaches.

2 RELATED WORK

Natural language processing (NLP) research has achieved admirable progress in recent years. However, the use of these tools in the context of education research has been limited thus far. A survey from 2021 [7] summarizes studies using NLP to analyze the sentiments of students' feedback. The authors identified almost a hundred publications, though only a fraction uses student feedback in the form of survey responses (other sources of student feedback included blog posts, forums, and feedback provided on online education and research platforms). Almost all of this work examines only whether students' comments are generally positive or negative, but we wish to examine whether similar techniques can distill the multitude of thoughts given at the end of a survey. In addition to sentiment analysis, we found papers discussing automated approaches to obtain various statistical analyses such as identifying important keywords and correlations between them, e.g. [5, 8], but we have not found publications aiming for (semi-)automated and more detailed analysis of student feedback.

3 METHODOLOGY

The various themes expressed in student comments in our initial data set included the puzzle concept, the student's learning experience, the details of the assignments, and the overall course. In each of these areas, there were many comments both positive and negative, and our codes reflected the combination of a theme and the accompanying sentiment. Once the codebook was developed, each comment (from both the initial study and the followup study) was assigned between one and three codes based on its content. For the interviews in the followup study, we used a similar approach, though here the coding is more free-form, as large parts of the interview may be irrelevant to the study topic and remain uncoded, while some codes may apply to a long section of the interview.

To perform supervised learning of comment codes, rather than developing new models, we chose to investigate how well common existing language models would be able to learn the coding. We initially chose the BERT model [4] as a well-known and effective deep-network model for general language understanding tasks, but also looked into less computationally intensive models such as FastText [6]. For the interview data, we also used BERT, but this data required significant pre-processing to turn the natural human conversation into a more regular form. We also used various data augmentation techniques on both data sets to increase the size and breadth of the training data.

Finally, we investigated clustering techniques on the survey comments to see if they could generate the initial set of comment codes. Here again we chose to explore a common technique, Latent Dirichlet Analysis (LDA), which is often used for topic modelling across a document corpus. LDA uses a probabilistic model that describes documents as collections of topics, where each topic is represented as a distribution over all words. The algorithm then uses an inference process to determine both the most likely distribution of topic clusters and the most likely distributions of the words in each cluster.

4 RESULTS

For the comment analysis, since each comment has up to three correct codes (as determined by the manual coding), instead of a binary accuracy measurement per comment, we measured success with precision and recall metrics. BERT was able to achieve a precision of 0.660 and recall of 0.750, and manual investigation showed that many "errors" assigned codes similar to the correct ones. (This often happens also with human coding, when different researchers pick up different nuances of the text. The final coding is then the result of a discussion between the researchers.) On the other hand, Fasttext was able to achieve only 0.53 precision and recall.

For the interviews, BERT performed extremely well at determining which parts of the interviews were relevant and which were not. Including the "uncoded" label, the precision of the model was 0.976 and recall was 0.966. However, if we look only at the portions of the interviews which had codes assigned in the ground truth, the accuracy was lower, though still somewhat reasonable, with a precision of 0.410 and recall of 0.667. As such, it appears that NLP techniques can be a meaningful time saver at least in terms of locating the most relevant portions of these long interviews.

On the other hand, clustering techniques appear unable to generate a meaningful set of codes for these survey comments. After manually examining the comments put together in each cluster, there appeared to be little consistency among them. We also compared the manually-assigned codes within each cluster, and likewise found that there was little correspondence between them. In short, it does not appear that this type of analysis is (yet) suitable for topic discovery amongst collections of short comments.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DUE-1821459. We thank Kimberly Fluet for her valuable guidance on qualitative analysis techniques.

REFERENCES

[1] Zack Butler, Ivona Bezáková, and Kimberly Fluet. 2017. Pencil Puzzles for Introductory Computer Science: An Experience- and Gender-Neutral Context. In *Proceedings of SIGCSE 2017*. 93–98. https://doi.org/10.1145/3017680.3017765

[2] Zack Butler, Ivona Bezáková, and Kimberly Fluet. 2018. Analyzing Rich Qualitative Data to Study Pencil-Puzzle-Based Assignments in CS1 and CS2. In *Proceedings of ITiSCE 2018*. Association for Computing Machinery, New York, NY, USA, 212–217. https://doi.org/10.1145/3197091.3197109

[3] K. Charmaz. 2014. *Constructing Grounded Theory*. SAGE Publications. https://books.google.com/books?id=y0ooAwAAQBAJ

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[5] Explorance 2016. Analyzing student comments in online course evaluations with Blue Text analytics. https://gatorevals.aa.ufl.edu/media/gatorevalsaaufledu/BTA_whitepaper_Analyzing-Student-Comments.pdf.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[7] Zenun Kastrati, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences* 11, 9 (2021). https://doi.org/10.3390/app11093986

[8] Anna Koufakou, Justin Gosselin, and Dahai Guo. 2016. Using data mining to extract knowledge from student evaluation comments in undergraduate courses. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 3138–3142. https://doi.org/10.1109/IJCNN.2016.7727599