EPiC Series in Built Environment

Volume 7, 2026, Pages 236–244

Proceedings of Associated Schools of Construction 62nd Annual International Conference

A Proposed Approach to Describe Online Construction Educator Perspectives beyond Word Cloud

Sandeep Langar[1], Rachel Mosier[2], Tulio Sulbaran[1], Sanjeev Adhikari[3],
[1]University of Texas at San Antonio, [2]Oklahoma State University, [3]Kennesaw State University

The advent of improved computer processing power and internet speed has allowed universities to increase the number of online courses. This capability to deliver classes online became very important during the recent pandemic to protect the well-being of educators and students when the gathering of individuals was restricted to try to slow down the spread of the virus. Although the delivery of online courses allowed the students to progress toward their degrees, concerns about this delivery method existed. Also, as the pandemic was getting under control and the universities were considering moving back to delivering courses in person, the educators also expressed concerns about the face-to-face classes. Thus, the focus of this paper is an approach to describe data collected from the educators regarding their concerns about course delivery methods during the transition to Online Learning during the pandemic. A qualitative research methodology was used for this research, where the data was collected using open-ended questions that allowed educators to fully describe their concerns without pre-established options (aka. Closed-end questions). The information collected was content-rich unstructured qualitative data, processed using text mining and sentiment analysis used in other areas but seldom used in construction education. The analysis using the proposed AI approach indicated equivocally that students are the foremost important consideration of the educators.

Keywords: Text mining, Course Delivery, Online Learning, Construction Education, Artificial Intelligence

Introduction

Natural Language Processing (NLP) is a machine-learning technology that allows computers to interpret, manipulate, and comprehend human language. This research used NLP techniques such as tokenization, stemming, and stop word removal to analyze qualitative data and identify the most common issues among educators for delivering courses in a face-to-face setting during the pandemic. Using an online survey, the research collected unstructured qualitative data and in-depth insights from educators. To ascertain the primary concerns and factors that educators have about traditional in person learning, especially considering exogenous events such as pandemics the data was examined using NPL techniques, such as text mining and sentiment analysis. However, these techniques do not typically include qualitative analysis.

While qualitative analysis is frequently used in social sciences (Ma and Fu 2020) and can help researchers identify various “dimensions” across a social system including the perception of the

respondents for particular situations and can help build theory through "*discovering patterns and connections*" (Fossey et al. 2002). Qualitative data often comprises information that is rich in context which is generated in form of words (written or recorded) which can be collected in numerous ways such as interviews direct observations, and others (Swanson and Holton 2009). Analyzing such data can face numerous challenges including it being time-consuming (Wolff et al. 2018, González Canché 2023). Responses to open-ended questions may be incomplete in thought and may not respond directly to the questions posed. The need to form cogent hypotheses from the data presented creates a need to find relationships within the data collected (Ragin 1987). Statistical methods typically used in quantitative analysis cannot be a substitute for qualitative analyses. While bias exists in any methodology (Sherratt and Leicht 2020), the open-ended question is intended to allow respondents to describe their perceptions freely. However, the questions themselves retain some amount of implicit bias, may provide an impetus towards a response, and may limit responses by the wording of the question itself (Sherratt and Leicht 2020).

Thematic or content analysis has long been used in qualitative analysis. This method for analyzing data can be arduous and time-consuming, involving the review of the same dataset in an iterative process with concepts or themes developed by identifying similar concepts from the responses (Braun and Clarke 2006). The process may be performed by multiple researchers independently, in order to limit internal biases (Braun and Clarke 2006). Weighting of individual responses may be necessary in order to determine the strength of a concept. Weighting should be identified early and set as a "rule" for the basis of analysis. The themes/concepts may be subdivided based on the number of similar responses and extrapolated from responses the respondent would provide in response to a query.

There exists a large quantity of qualitative research in construction, with nearly half of the papers published by the International Journal of Construction Education and Research using these methods (Collins et al. 2024). The fact that construction educators and researchers regularly use qualitative methods, illustrates the need for a reliable method which minimizes the time spent analyzing the data. While NLP can support coding, as in the use of a Word Cloud, there is not a current method to providing a thematic analysis of long-answer questions with software. There is a need for a system which will produce repeatable results, the use of NLP systems was incorporated. A case study using existing data is provided to validate and illustrate the usage of this new system which provides qualitative analysis.

Background

Qualitative Comparative Analysis (QCA) is not new to construction, and it has been used as part of Artificial Intelligence (AI) and Boolean logic to make determinations from interview and questionnaire data (Ragin 1987, Guo et al. 2022, and Ma and Fu 2020). Another qualitative analysis application is through Systematic Literature Review (SLR), which includes data mining and bibliometric network visualization (Karimi and Iordanova 2021, Collins et al. 2024). Bibliometric network visualization aids the analysis of networks by visualizing existing patterns in authorship, citations, and keywords (Karimi and Iordanova 2021).

In order to validate the proposed method a case study was identified. The case study focuses on an existing survey on changes in construction educator perceptions about learning during the largest externally imposed experiment in online education. The survey was distributed via Qualtrics, using an online survey format which provides easy access and increased participation (Kılınç and Fırat 2017). The construction education online delivery method survey included both open-ended and closed questions or multiple choice for demographics, etc. It was the intent that open-ended questions would allow respondents to provide insights that the survey author might not have considered.

The survey research objectives were A) Employing open-ended question to obtain insights into the opinions of construction educators regarding concerns with content delivery in face-to-face setting during a pandemic (collecting comprehensive and qualitative data) and B) Determine and explain the main issues and factors that construction educators have post-pandemic. The analysis objectives were A) Analyze and process the unstructured qualitative data gathered from educators by using NPL techniques and B) Examine the suggested Artificial Intelligence (AI) approach's efficacy and reliability in qualitative data analysis in correlation to construction education.

While existing research methods have been automated, they still can be time-consuming. However, with current software packages, the time can be reduced. As part of the process of automating the work, an approach to illustrate the process was developed. This paper provides an Artificial Intelligence (AI) approach for automated qualitative analysis, which is validated through an existing survey and responses to open-ended questions. The output of the analysis of one of those survey questions is presented here.

Methodology

A case study was determined to be the most appropriate method to validate a proposed NLP method for an NLP approach to qualitative analysis. A survey was issued in the Summer of 2020 asking Construction educators to reflect on the effects moving to an all-online platform. Over 1,800 educators were surveyed. The survey collected responses to over 70 closed and open-ended questions. While the closed questions focused on demographics, including gender, race, location, tenure status, and years of service, open-ended questions allowed educators to share their concerns and perceptions of the novel transition. The open-ended survey question used for validation asked respondents to provide the top three concerns with content delivery in face-to-face setting during a pandemic.

With an open-ended question allowing multiple lines of response and requesting three unique responses, there is the opportunity for the respondent to leave a null or invalid response, provide the same response multiple times, and provide more than three unique responses. For the purpose of this study, the responses are all given equal weight. Rather if a respondent provided the same response three times, that was equivalent to three respondents providing a similar response. Respondents who provided the same response three times to the question were adding their own weighting or emphasis to their response. In this case, it was included to support the respondents intention.

The response data was collected via a Qualtrics survey and exported into Excel. There are a total of 161 respondents. For the question used in the case study, there were 52 participant responses. The question requested respondents provide the top three concerns, which provides an opportunity for 156 unique responses, or 52 respondents with three responses each. AI was used to automate the process to determine the most frequent response or top concern in the most time-effective way. An AI approach was implemented (Figure 1), using five distinct steps (*tokenization*, *stop word removal*, *stemming*, *lemmatization, and part of speech tag*), with a validation performed using the question on the top three concerns with return to traditional in person learning.

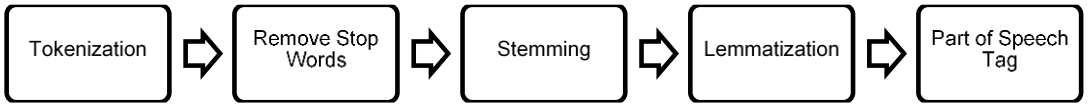

Figure 1. Steps for analyzing qualitative data using AI (Sulbaran et al. 2024)

To have a computer perform Natural Language Processing (NLP), language must still be broken down into its base parts, such as parsing. Tokenization allows words to have tokens represent them, i.e., a

variable or number that is easily identified by a computer (Tay et al. 2021). As some languages do not use articles like “a,” “an,” or “the,” the use of articles in language is not necessary for understanding meaning. Prepositions and similar types of words are also removed for the same reason. These “stop words” are removed. However, the English language and the United States vernacular uses articles as a part of the language. NLP can also be performed by hand as part of thematic or concept analysis. The following steps are performed by individuals through an iterative process or now utilizing a computer or AI. In this case, the qualitative analysis automation was performed using Python for coding and Jupyter Notebook as the interface.

Further, base words may have prefixes and suffixes, which in verb usage only change the tense, from past to present and then to future tense. In order to consider the base meaning, “stemming” must also be performed to match all of one verb idea, i.e., the words pair, paired, and pairing are all converted to the base word pair (Jivani 2016). Lemmatization is also performed. Whereas stemming reduces a word to its stem, it does not recognize the part of speech associated with the word (Jivani 2016), i.e., a pair can be used as a noun (pair of pants) or a verb (pairing a matching sets of socks). While lemmatization includes morphology or the study or words as part of computational linguistics (Jivani 2016) and must be further considered, it is outside the realm of this discussion. While there are many method options for Part of Speech (POS) Tags, it is simply assigning a value, or POS tag, to a word to allow the computer to determine how that word is used (Tian et al. 2009). For the United States English dialect, basic POS definitions include noun, verb, or adjective.

Results

After using the approach described above, a word cloud can be generated with the results. While there are many online word cloud generators, the steps utilized in the above framework, remove stop words, stemming, lemmatization, tokenization, etc., make this output more useful.

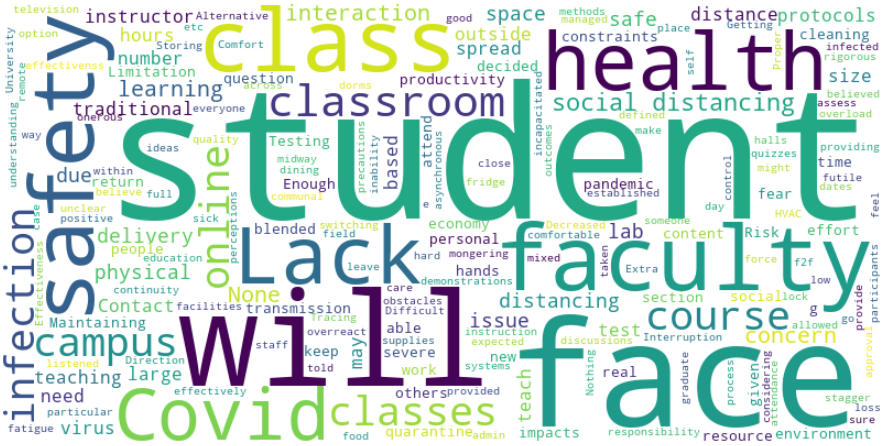

Figure 2. Top Three Concerns Word Cloud

The word cloud output (Figure 2) is based on responses to the survey statement: “If your University decides to offer courses in a traditional setting during fall 2020, please list the top three concerns that you might have with the delivery of content in this setting.” It is evident that educators felt strongly about “students,” “face,” “will,” and “faculty.” It is also evident that there is context missing. The verb “will” does not add value to the word cloud. Further, where does “lack” fit into the discussion? Similarly, is “class,” “classes,” and “classroom” the same theme or are they pointing to different

concepts? What makes theses themes or concepts important? While the initial steps of the AI word cloud approach did provide output, we are missing the context of how these words fit together. A direct quote from the participant responses is "lack of personal interaction with students," which is difficult to ascertain from the work cloud. The work cloud prioritizes the term "face" over "lack," but does little to provide context as to the why.

By using the method suggested by Sulbaran et al. (2024) and illustrated in Mosier et al. (2024), the Python output is the series of words, based on word count (times used) and the use percentage. While the table provides the same information as the Word Cloud, it lacks the graphical presentation.

Table 1. Word Used Percentage from ALL Group Snowball Stemmer: Total words (n)= 146

	Word	Word Count	Use Percentage
1	student	37	25.34%
2	face	22	15.06%
3	class	17	11.64%
4	distanc	13	8.90%
5	faculti	12	8.22%
6	safeti	11	7.53%
7	health	9	6.16%
8	lack	9	6.16%
9	social	8	5.48%
10	covid	8	5.48%

Once parsed, the word cloud method only provides codes, or individual words, but not concepts or themes. Similarly, Table 1 also only provides the individual, parsed words, but not themes. However, a bar chart can be used to describe the output in a more graphical style than the table (Figure 3). The results remain the same as with the word cloud, while concept codes or theme words are identified, there is no connection between the words. A deduction could be made that the respondents who are educators are most concerned with students, based on the respondent group and question analyzed.

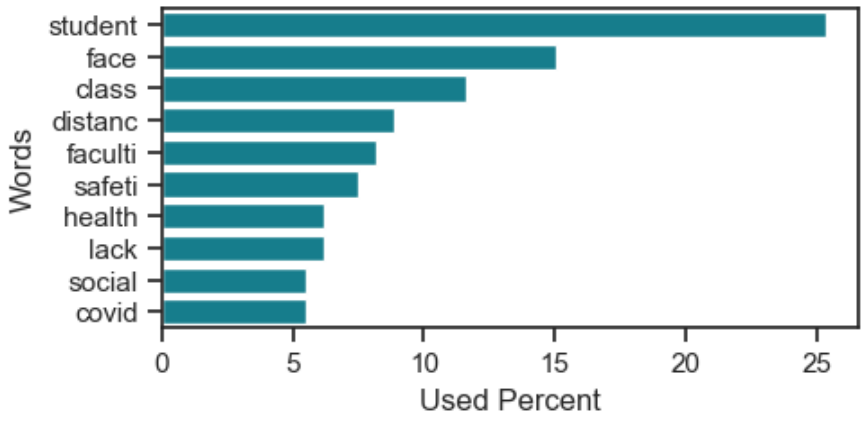

Figure 3. Word Used Percentage Bar Graph

By utilizing a qualitative analysis framework method (Sulbaran et al. 2024), a concept map or bigram (Figure 4) has been added to the word cloud, word use table, and bar graph. The bi-gram provides the

qualitative solution that is missing from a traditional AI word cloud approach. For the bi-gram, through the above framework steps, stemming, lemmatization, and tokenization, parts of speech articles like "a," "an," and "the" have been removed. Therefore, the relationships between the codes become evident in the bi-gram which are typically made up of two-word phrases.

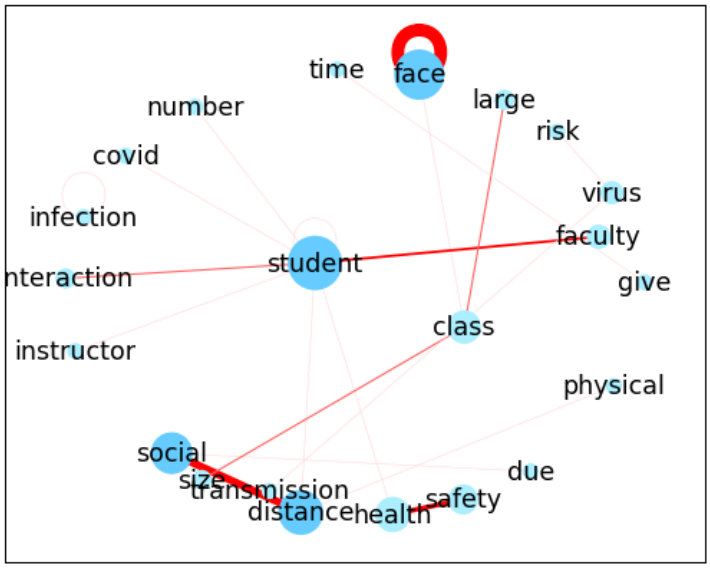

Figure 4. Bigram of top three concerns with the Traditional Delivery of Courses

The bi-gram provides a connection between code words "student" and "interaction." Similarly, a connection is evident between code words "student" and "faculty." The line thickness indicates the strength of relationship or how many times these two words were connected. In the previous two examples, student to interaction was not as strongly connected as was student to faculty, based on the line thickness. The size of the solid circle or dot indicates the number of times an individual code word was used. The open circle or line which connects to itself, shows that "face" is connected to "face", as in "face to face." The thickness of the line indicates a very strong connection.

While this method seeks to use AI as an alternative to the typically laborious coding of a thematic or content analysis, the analysis is not yet complete. It is necessary to reconstruct the concepts by replacing the missing articles, conjunctions, and prepositions that were removed during the process. A discussion, as was provided above is necessary to describe the missing pieces. While less laborious, the AI or NLP (Sulbaran et al. 2024) processes for qualitative analyses have not provided a complete result.

The initial steps shown in Table 1 and Figure 3 illustrate that the proposed method provides a similar result to the Word Cloud (Figure 2). Student occurs most frequently in the Word Cloud making it the largest and occurs at the highest rate or 25.34%. Where the Word Cloud falls short and what the proposed method provide is re-connecting the individual words. Using the bi-gram, the themes which occur most frequently are largest and easily identifiable.

Summary and Conclusions

To create a replicable and reproducible analysis, this research has sought to provide an AI NLP approach for describing qualitative data through a computer algorithm with a data mining process. The survey

collected responses from construction educators throughout the United States to determine the top three concerns with the traditional delivery. A total of 161 construction educators participated in a survey providing their concerns about the changes in delivery modality.

The response of unstructured qualitative data was successfully converted into meaningful data through a computer-generated algorithm using Python and Jupyter Notebook. The qualitative data created by open-ended responses can be challenging to analyze. Thus, this paper described an AI NLP approach to describe the data collected through tokenization, removing stop words, stemming, lemmatization, and applying POS tags. This paper was also successfully able to perform a case study to validate the proposed approach to describe the highest priority of educators. The number highest value syntactical words were 1) students, 2) face, 3) class, 4) distance, 5) faculty, 6) safety, 7) health, and 8) lack. This individual terms easily connect to concepts such as 1) student interaction, 2) large class size, 3) face to face, 4) health and safety, and 5) social distancing, etc.

Future Research and Limitation

The paper presents a qualitative data analysis process using Natural Language Processing (NLP) to describe big open-ended qualitative data sets in quantitative analysis with a specific case as an example. The case used to explain the qualitative data process aims using NPL aimed at determining construction educator concerns with education delivery in face-to-face setting during the pandemic. One of the limitations with the research is the validation of the findings and ensuring the finding generalizability especially if the data set were analyzed by humans. Future research could investigate the validity of the process by comparing the human qualitative analysis process with the one to outlined in the paper and determine if there was a convergence in the findings between the data analysis conducted by the two different entities.

Additionally, the researcher could utilize the AI approach in the next step easily translates to statistical analyses such as Anova, T-test, correlation, etc., to benefit research communities.

References

Braun, V. and Clarke, V. (2006). “Using Thematic Analysis in Psychology.” Qualitative Research in Psychology. V.3. p.77-101.

Collins, W. Salman, A., Olbina, S. and Mosier, R. (2024). “Scientometric, Thematic, and Methodological Analysis of IJCER Construction Education Focused Publications – 2004 – 2023.” International Journal of Construction Education and Research. DOI: 10.1080/15578771.2024.2404019

Fossey, E., Harvey, C., Mcdermott, F., & Davidson, L. (2002). Understanding and Evaluating Qualitative Research. Australian & New Zealand Journal of Psychiatry, 36(6), 717–732. https://doi.org/10.1046/j.1440-1614.2002.01100.x

González Canché, M. S. (2023). “Machine driven classification of open-ended responses (MDCOR): An analytic framework and no-code, free software application to classify longitudinal and cross-sectional text responses in survey and social media research.” Expert Systems With Applications, 215. https://doi.org/10.1016/j.eswa.2022.119265

Guo, W., Lu, W. and Kang, F. (2022). "Combining Transaction Characteristics and Governance Mechanisms to Suppress Opportunism in Construction Projects: Qualitative Comparative Analysis." Engineering, Construction, and Architectural Management.

Jivani, A. G. (2016). "A comparative study of stemming algorithms." International Journal of Computer Technology Applications. V.2. N.6. pp. 1930–1938.

Karimi, S., and Iordanova, I. (2021). "Integration of BIM and GIS for Construction Automation, a Systematic Literature Review (SLR) Combining Bibliometric and Qualitative Analysis." Archives of Computational Methods in Engineering. V.28. N.7. pp. 4573–4594.

Karlsson, M., Nilsson, T., and Pichler, S. (2014). "The Impact of the 1918 Spanish Flu Epidemic on Economic Performance in Sweden: An Investigation into the Consequences of an Extraordinary Mortality Shock." Journal of Health Economics. V. 36. http://dx.doi.org/10.1016/j.jhealeco.2014.03.005

Kılınç, H., & Fırat, M. (2017). Opinions of Expert Academicians on Online Data Collection and Voluntary Participation in Social Sciences Research. Educational Sciences: Theory & Practice, 17(5), 1461–1486. https://doi.org/10.12738/estp.2017.5.0261

Lee, J.S., Shin, J.C., & Ock, C.Y. (2022). "The multi-hot representation-based language model to maintain morpheme units." Applied Sciences. V.12. N.20. 10612. doi:https://doi.org/10.3390/app122010612

Ma, L., and Fu, H. (2020). "Exploring the Influence of Project Complexity on the Mega Construction Project Success: a Qualitative Comparative Analysis (QCA) Method." Engineering, Construction, and Architectural Management. V.27. N.9. pp. 2429–2449.

Mosier, R.D., Adhikari, S., Langar, S., and Sulbaran, T. (2024). "Program Affiliation Impact on Educator Perspective of Online Learning Environments." Journal of Architectural Engineering. V.30. N.4. DOI: 10.1061/JAEIED.AEENG-1815

Ragin, C.C. (1987). "The Comparative Method: Moving beyond Qualitative and Quantitative Strategies." Univ of California Press, Berkeley, CA, pp. 19-34.

Sulbaran, T., Langar, S., Adhikari, S., and Mosier, R.D. (2024). "Framework for Analysis of Qualitative Data for Construction and Engineering Disciplines: A Case of Faculty Perspective during Covid-19." International Journal of Modern Engineering, Fall 2023. DOI: https://zenodo.org/records/10728066

Swanson, R. A. and Holton, E. F. (2009). Research in Organization: Foundations and Methods of Inquiry, Berrett-Kohler Publishers Inc., Seattle.

Sherratt, F. and Leicht, R. (2020). "Unpacking Ontological Perspectives in CEM Research: Everything is biased." J. Constr. Eng. Manage. 146 (2): 04019101. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001734.

Tay, Y.; Tran, V.Q.; Ruder, S.; Gupta, J.; Chung, H.W.; Bahri, D.; Qin, Z.; Baumgartner, S.; Yu, C.; Metzler, D. (2021). “Charformer: Fast character transformers via gradient-based subword tokenization.” arXiv. V.2106.12672.

Tian, S., Ibrahim, T., Umal, H., and Yu, L. (2009). "Statistical uyhur POS tagging with TAG predictor for unknown words," 2009 ISECS International Colloquium on Computing, Communication, Control, and Management. Sanya, China. pp. 60-62. doi: 10.1109/CCCM.2009.5267823.

Timmermans, S., and Tavory, I. (2012). “Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis.” Sociological Theory. V.30. N.3. pp. 167–186.

Wolff, B., Mahoney, F., Lohiniva, A. L., & Corkum, M. (2018). Collecting and Analyzing Qualitative Data | Epidemic Intelligence Service | CDC. https://www.cdc.gov/eis/field-epi-manual/chapters/Qualitative-Data.html