Sentiment AND *Topic* MODELING

for Analysis of
Open-Ended Survey Results
for EvaluATE

TEAM AND ACKNOWLEDGEMENTS

Authors

Nolen Ackerman
Data Scientist, First Analytics

FIRST ANALYTICS®

Larry Mallak
Academic Director of Systems Engineering
Worcester Polytechnic Institute, Worcester, MA

WPI

Lyssa Wilson Becho
Principal Investigator, EvaluATE

Megan López
Co-Principal Investigator, EvaluATE

EvaluATE

Western Michigan University
The Evaluation Center

Report design and data visualization

Maureen Green

Acknowledgements

The research team would like to thank Emma Binder for her contributions to this work as well as Carolyn Williams-Noren for her copyediting. We would also like to thank all survey respondents who provided meaningful feedback data to EvaluATE about their event participation experience.

Suggested Citation: Ackerman, N, Mallak, L, Becho, L.W., López, M. (2024). Sentiment & Topic Modeling for Analysis of Open-Ended Survey Results for EvaluATE. Kalamazoo, MI: The Evaluation Center, Western Michigan University. Retrieved from https://evalu-ate.org/research

Table of Contents

continued on next page

INTRODUCTION

EvaluATE is the evaluation hub for the National Science Foundation's Advanced Technological Education (ATE) program. EvaluATE educates the ATE community – including evaluators, project leaders and staff, grant specialists, and college administrators – about evaluation. To this end, EvaluATE has provided webinars, workshops, newsletters, a blog, and other resources.

EvaluATE has amassed a wealth of data about its educational activities, which include 42 webinars and 7 workshops conducted since 2008. In this study, we (Larry and Nolen) used big data analytic techniques (sentiment analysis and topic modeling) to mine these data. This project's contribution concerns the analysis of textual responses to open-ended questions in surveys administered by EvaluATE. Traditional statistical methods were also used to analyze the data and provide comparisons with textual content. By combining big data analysis and statistical analysis, we explore new methods and insights that can be further used in EvaluATE's and similar settings.

Converting unstructured data, such as open-ended responses, into coded categories usually involves tedious and error-prone manual coding. But, using new machine learning algorithms, these processes can be automated and scaled to large data sets. This study deployed sentiment analysis to automatically code the emotional content of each response as either positive or negative and compute an associated sentiment score. In addition, this study used computer algorithms to cluster responses into similar categories or groups based on common phrases and words, a process known as topic modeling.

This material is based upon work supported by the National Science Foundation under Grant No. 1841783. The content reflects the views of the authors and not necessarily those of NSF.

Table of Contents

Importance of This Study

This study explored innovative methods, namely Sentiment Analysis and Topic Modeling, to make better use of evaluation data for program design, monitoring, and improvement. This study's approach provides a replicable framework for other researchers to further study *unstructured feedback data*. As an additional benefit to the EvaluATE project, this study produced a *harmonized database* of responses from the many evaluation surveys EvaluATE has conducted since 2008.

EvaluATE intends to use these findings in the following ways:

1. To gain ongoing formative feedback that can be used to assess the quality, strengths, and opportunities for improvement of EvaluATE webinars and workshops;
2. To inform the development of a standardized post-event feedback survey that can be used after all webinars and other training opportunities;
3. To inform the development of webinar resources that can be used to document and standardize internal practices and can serve as a model for external groups to increase the quality of their webinars;
4. To inform the evaluation of EvaluATE and increase understanding of users' satisfaction with webinars and workshops.

Brief Definitions

Harmonized database: a collection of data that has been standardized and made consistent for easier analysis and comparison

Unstructured feedback data: Information provided by participants or users that doesn't have a predefined structure or organization

Research Aim and Questions

Research Aim: to deepen EvaluATE's understanding of users' feedback and to support improvements to training activities using computer models and advanced data visualizations.

Secondary Goal: to illuminate how computer algorithms can be applied to unstructured text for evaluative purposes.

Research Questions:

1. How do participants in EvaluATE's training events perceive the value and impact of training?
2. What relationship exists between open-ended responses' emotional content and close-ended quantitative responses?
3. What types of survey questions can most sensitively and accurately gauge participants' satisfaction and self-assessments of learning?
4. What terms and topics in open-ended responses are most highly related with close-ended quantitative responses?

Sentiment Analysis Defined

Sentiment analysis is used to detect the level of positive or negative *sentiment* in unstructured data, such as written text. It uses a *standard or domain-specific library of terms* and *inversion factors* to identify positive and negative words or phrases. The words or phrases' strength and frequency are identified and tallied, resulting in a number known as a sentiment value for each analyzed text. The *sentiment value* is an indicator for the level of positive or negative emotion expressed by the respondent.

Brief Definitions

Inversion factor: linguistic element or contextual clue that reverses the polarity of a statement and allows for an accurate interpretation of its underlying sentiment

Sentiment: the emotional content within written text

Sentiment value: a numerical value representing the level of positive or negative terms used in a text response

Standard or domain-specific library of terms: a collection of terms coded specifically for a domain of interest and study

The diagrams below provides three sentences that highlight how sentiment analysis is applied. The computer algorithm identifies positive (**green bold**) and negative (red underlined) terms. Mathematical formulas are then used to calculate the total positive or negative value of the sentence. In the examples provided below, a higher score indicates a more positive sentiment. A lower sentiment score indicates a more negative sentiment. Using these generated scores across large data sets, trends of positivity and negativity can be detected.

Examples of Sentiment Analysis Calculations

Statement	Word Count	Extent Positive	Extent Negative	Overall Sentiment Score
"The **simple** but striking visuals and text combined with the **excellent** speaker and moderators were **clear**, **consistent**, and also **engaging**!"	20	34.57	1	**Highly Positive with a 2.43 Sentiment Score**
"Way too many people in the chat box, it was hard to read and follow."	15	1	22.1	**Highly Negative with a .46 Sentiment Score**

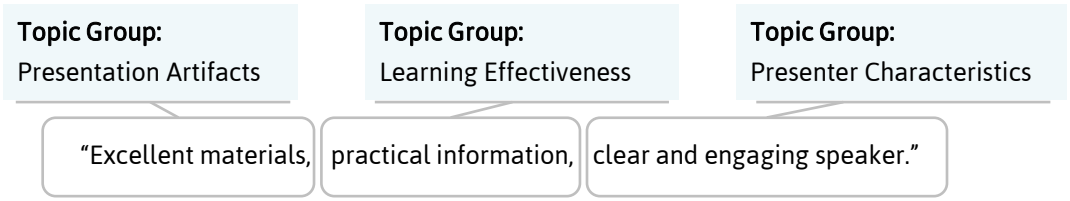
The diagram below provides an example of mixed sentiment, where a sentence contains positive or negative linguistic elements to express the opposite sentiment. In these cases, the algorithm has difficulty accurately identifying the overall sentiment. Mixed sentiment is represented in linguistic constructs where descriptions of items are provided in contrast to opposite items. Traditional sentiment analysis techniques have difficulty with these sentence structures, and automated processing is limited. Using mixed sentiment descriptions, however, is atypical. The example below demonstrates how a respondent uses a contrast to a negative item (boring seminar) to praise the good item (most recent webinar).

Example of Mixed Sentiment Analysis Calculation

Statement	Word Count	Extent Positive	Extent Negative	Overall Sentiment Score
"I sit through a LOT of boring webinars where I may find one or two nuggets, not this one."	19	1	7.98	**Negative with a .6 Sentiment Score**

Topic Modeling Defined

Topic modeling is a statistical modeling technique for unstructured data. Its purpose is to detect the topics or key concepts expressed in text. Its main objective is to identify themes (an *exploratory* technique) or to properly cluster relevant topics (a *confirmatory* technique). Topic modeling uses a standard or domain-specific text matrix to classify *opinion phrases*. The image below is an example of how responses would be broken down into different topic groups.



Brief Definitions

Confirmatory: seeks to confirm an already existing theory or hypothesis

Exploratory: seeks to understand data without preexisting theories or hypotheses

Opinion phrases: segments of text, often the lowest level of complete communication in the response

METHODS

In this study, our tasks followed four major phases:

Data preparation: To conduct an analysis on data collected over more than a decade, a significant amount of time was required to *harvest and harmonize data*. This step involved building a consolidated data model and manually coding a subset of responses to train the sentiment analysis and topic models.

Model development: Using these coded responses as a training set, the sentiment analysis and topic models were developed iteratively by creating domain-specific term matrices.

Model validation: Using a subset of coded responses that had been held back to use ahttps://evaluate.org/research/sentiment-analysis/videos/s a *validation data source*, the models were tested for the intended use.

Analysis and visualization: The models were deployed to code the entire response set, and traditional statistical techniques were used to complete the cycle of analysis. An interactive data visualization was created to facilitate the use and interpretation of the findings.

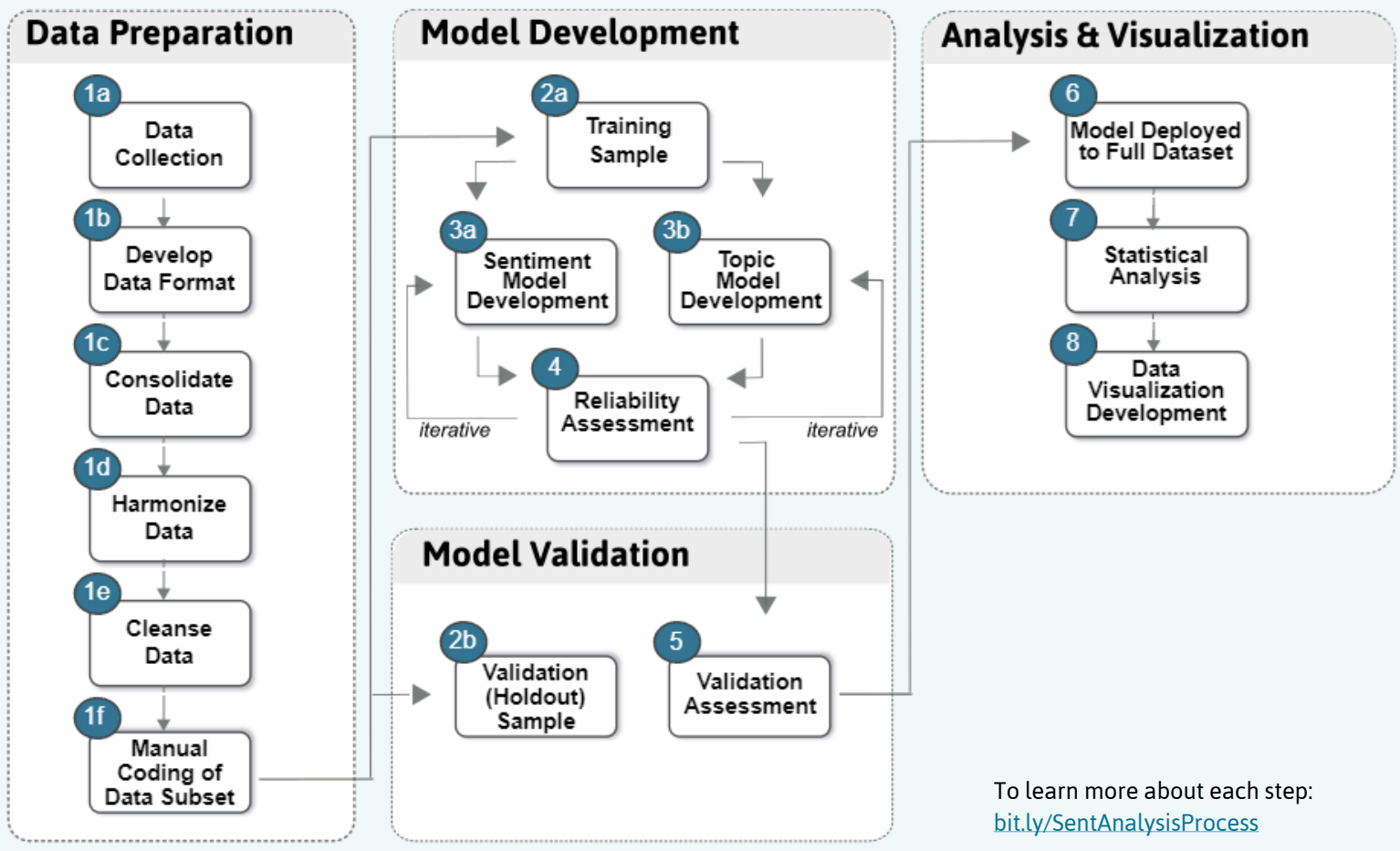
Brief Definitions

Data harvesting: the collection of all data and metadata from the surveys. Metadata relates to data elements such as date of survey, webinar name, and presenter.

Data harmonization: the standardization of data elements to facilitate statistical analysis and collation. As an example, a date format from one survey may exist as 1/2/2020 and another may exist as January 2nd 2020. Harmonization sets the standard format and converts all dates to that format.

Validation data source: a validation data source is a subset of data elements excluded from the model building process. It is used as a testing data source to ensure the model is responsive to unseen data and is not over-fit from the training data source.

The diagram below provides additional details on individual process steps. A short video has been produced to provide an overview of each step.



Data Preparation

In this study, we looked at attendees' responses to post-event feedback surveys about EvaluATE webinars and workshops over the past 12 years. Surveys asked attendees to report their overall satisfaction with training events through a number of open- and close-ended questions. Most relevant to this study were the following questions:

1. Open-ended question: "What aspect of this session was especially good?"
2. Open-ended question: "What aspect of this session needs the most improvement?"
3. Close-ended question: "Please rate the quality of the session in which you participated."
4. Close-ended question: "What is your overall opinion of the quality of this session?"

EvaluATE provided 3,320 completed surveys across 49 different post-event feedback surveys. The surveys were conducted after 42 webinars and 7 workshops between 2008 and 2020. It is worth noting that small changes in the wording of the above-mentioned questions were made between some event surveys throughout the 12- year period. However, the researchers determined that data could safely be collapsed across surveys as the wording differences would not reasonably lead to differences in interpretation or responses.

Model Development

Sentiment Model

To conduct a correlation of sentiment scores to overall satisfaction, responses to both the overall satisfaction question (Question 4 listed above) and the open-ended questions (Questions 1 and 2 listed above) were required. From the initial 3,320 surveys provided, this meant that answers to close- and open-ended questions from 1,058 unique respondents were analyzed. From these, over 7,600 opinion phrases were coded for sentiment and topic classification. Therefore, opinion phrases were retained only when these conditions were met (n = 1,547 for Question 1; n = 1,186 for Question 2).

Summary statistics for sentiment scores are presented below.

Survey Question	Opinion Phrases (n)	Mean Sentiment Score	Sentiment Standard Deviation	Minimum Sentiment Score	25%	50%	75%	Maximum Sentiment Score
"What aspect of this session was especially good?"	1,547	1.2192	0.4268	0.2974	1.0000	1.0605	1.6432	2.4248
"What aspect of this session needs the most improvement?"	1,186	0.9810	0.4125	0.1522	0.6135	1.0000	1.1172	2.3023

Topic Model

In order to carry out the topic modeling, a word or phrase matrix was created to classify opinion phrases based on six topic groups that were identified through a literature review and modified to encompass the specificities of the data set.

Seventy percent (70%) of opinion phrases were classified into one of these six groups.

Thirty percent (30%) of opinion phrases were considered "low information" (e.g., "Great Job"; "Learned about from email") and could not be placed into topic groups.

Individual opinion phrases sometimes pertained to more than one topic group, resulting in multiple classifications for some opinion phrases. This resulted in 4,439 codes across the 2,733 opinion. The table below summarizes the topic groups and identifies the percentage of opinion phrases that pertained to each group.

Topic Group Assignment	Number of Opinion Phrase Codes (n)	Percentage of Coded Opinion Phrases
Physical Characteristics (Technology): Aligned to the "physical learning characteristics" identified by Fu (2010), this grouping represents learning prerequisites that enable information transfer. As most of the training events in this data set took place virtually, these comments relate mostly to the technology required to hear, see, interact with, and download information provided in the session. In a physical learning environment, this topic may need to be expanded to encompass classroom design and layout.	325	4.40%
Presentation Design Characteristics: Aligned to the "course structure" and "content of course" categories identified by Peltier (2007), this grouping represents the design of the structure of the workshop / webinar. This extends to PowerPoint slides, amount of information presented, overall time allocated, logical design structure, and balance of material.	1,335	18.20%
Presenter / Delivery Characteristics: Aligned to the "teacher characteristics" identified by Fu (2010), this grouping represents elements associated with the presenters, speakers, facilitators, and moderators in the workshops and webinars, including their speed, volume, and perceived professionalism in presenting. This topic group also contains elements related to how "engaging" speakers are with learners.	994	13.50%

Topic Group Assignment	Number of Opinion Phrase Codes (n)	Percentage of Coded Opinion Phrases
Presentation Artifacts / Assets Characteristics: This is somewhat aligned to the "content of course" category identified by Peltier (2007), but in this context it specifically focuses on artifacts provided for workshop or webinar participants to use in the future. These artifacts include transcripts, samples, checklists, guides, videos, templates, and assorted one-pagers and flowcharts. This seemed to be of particular interest in the data set responses and, as such, indicated an establishment of a separate topic group.	769	10.50%
Perceived Utility of Learning Characteristics: Aligned to "usefulness" as identified by Cheok (2015), this grouping represents the training's perceived usefulness to and reception by participants. This topic group represents "feelings" of helpfulness, inspiration, new perspectives, and "mind-shift" among participants. This does not gauge true learning transfer (as verified by pre- and post-test analysis), but is related to the subjective feeling of having gained knowledge. The negative elements of this topic group are related to redundancy, lack of new knowledge, and stale content.	1,016	13.80%
Interaction / Interactivity Characteristics: Fu's dramaturgical model (2010) provides two interaction groups (1.teacher to student, and 2. student to student). This topic group combines all interactions during the webinar, including Q&A sessions, chat features, role-play activities, and other interactive elements. Subgrouping by keyword search can provide additional information on individual sub groups.	1,068	14.50%
Unassigned: This category includes unassociated low-information phrases, as described above.	1,842	25.10%

Model Validation

To assess the model's fitness for the intended use, the sentiment analysis and topic model algorithms were tested with two data sets. The first data set, referred to as the training sample, (bit.ly/SATMN-Training) was used to train the models and to verify the accuracy of the algorithms. A second, hold-out data set (bit.ly/SATMN-HoldOut) was used to validate the models and ensure the models were not overfit to the training data. Using the hold-out data set—previously unseen by the model—allowed for the assessment of generalizability to the larger data set. The results of the verification and validation activities for both models are summarized in the corresponding tables below:

Sentiment Analysis Model Testing	
Verification (Training Sample)	Validation (Hold-out Sample)
Accuracy: 467 / 581 = 80.55%	Accuracy: 96 / 130 = 78.46%
.70 kappa	.66 kappa
.76 weighted kappa	.70 weighted kappa

Topic Model Testing	
Verification (Training Sample)	Validation (Hold-out Sample)
Accuracy: 143 / 172 = 85.6% for "low-information items"	Accuracy: 149 / 182 = 81.9% accuracy for "low-information" items
Accuracy: 319 / 379 = 84.2% for items assigned to the six topic groups	Accuracy: 292 / 380 = 76.8% accuracy for items assigned to six topic groups

Since the survey questions examined in this study oppose one another (one question asked for elements that drove satisfaction and one question asked for elements required for improvement), it was hypothesized that the mean sentiment scores for the responses to the two opposing questions would be statistically different. The results of sentiment analysis model testing were used for face validity for the sentiment scoring model.

Brief Definitions

Accuracy: a measurement used to quantify the difference between an estimated or predicted value and the accepted "true" value

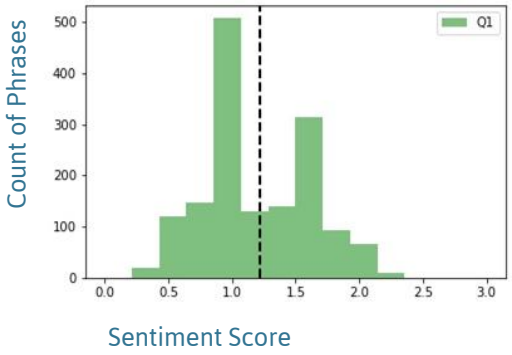
Kappa: a metric that reduces the accuracy calculation by eliminating agreement of ratings simply due to chance

Validate: within data science, the testing of a generated model against data that has been withheld from the training step used for model development

Verify: within data science, the testing of a generated model against training data

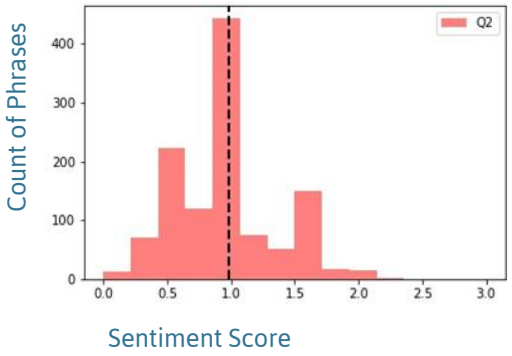
The figures below illustrate the mean sentiment scores for two different questions. The mean sentiment score related to 'good' aspects of the sessions was higher (1.2) than elements that needed improvement (.97). This difference was tested using an independent groups analysis, assuming unequal variances. This analysis revealed a statistically significant difference in mean sentiment scores between responses to the positive and negative questions. This finding provides face validity for the sentiment model regarding fitness for use in differentiating negative and positive feedback.

Mean Sentiment Scores for Questions 1 and 2



Question 1 (POS):

"What aspect of this session was especially good?"



Question 2 (NEG):

"What aspect of this session needs the most improvement?"

Brief Definitions

Independent group analysis: the comparison of two distinct groups to determine if there is a statistically significant difference between them

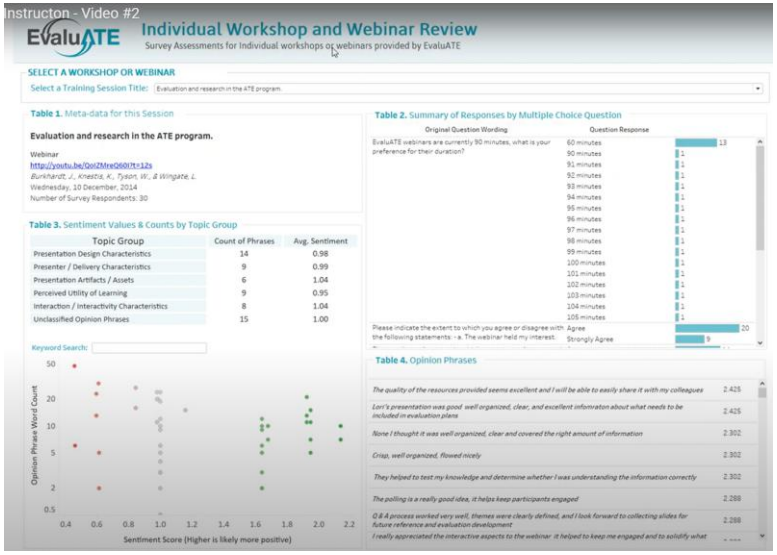
Unequal variances: a situation where the standard deviations between two data sets are different

Data Visualization

To facilitate distribution of the research findings and allow EvaluATE to gain additional insight from the data, two interactive data visualizations were created. These data visualizations were generated by connecting directly to the large database of survey results as well as the responses coded according to sentiment and topic group during the research effort. The following two instructional videos describe the functionality of the data visualizations.

Main Data Visualization – Explorer: bit.ly/SATMN-1

Specific Session Data Visualization - Explorer: bit.ly/SATMN-2



FINDINGS

Related to initial research questions, this study demonstrated that using sentiment analysis and topic modeling in large survey data sets can produce insights otherwise not easily discoverable with manual review.

Research Question 1:

How do participants in EvaluATE's training events perceive the value and impact of training?

Insights: Overall satisfaction scores were very high across the entire data set collected over the ten-year period; this lack of variation seemed to indicate a high level of overall satisfaction with the usefulness of webinar and workshop content. The table below provides mean sentiment scores and overall satisfaction ratings from the collected surveys. Additional analysis for each individual webinar and workshop is provided in the data visualization for in-depth analysis of overall satisfaction scores.

		2009	2015	2016	2017	2018	2019	2020	Total
Please rate the quality of the session in which you participated (or) What is your overall opinion of the quality of this session?[1]	**Mean Rating**	4.4	3.6	4.6	4.6	4.9	4.9	4.9	4.7
	Standard Deviation	0.6	0.5	1.2	1.2	1.1	1.1	1.2	1.2
	Survey Responses (n)	54	94	441	390	416	367	213	1,975

[1] *The response options for the survey items were displayed on a fully anchored 5-point Likert scale: poor (1), fair (2), good (3), very good (4), and excellent (5). The standard deviation indicates how much the responses varied. A higher standard deviation shows greater variety in responses, while a lower standard deviation indicates more similar responses.*

Research Question 2:

What level of relationship exists between emotional content in open-ended responses and close-ended quantitative responses?

Insights: There was substantially more variation and information richness in the open-ended survey responses than in the Likert scale ratings of overall satisfaction. This increased variation and richness, when coded in a structured manner using topic modeling and sentiment analysis, allows for identification of potential interventions and modifications to webinars and workshops to address drivers of dissatisfaction or enhance the elements most highly correlated with satisfaction.

The sentiment scores of responses related to "aspects of the sessions that were especially good," had no statistically significant correlation with the quantitative responses. Respondents who gave a session very positive ratings in response to close-ended questions were not more likely to provide highly positive open-ended feedback.

Survey Question	Opinion Phrases (n)	Mean Sentiment Score	Sentiment Standard Deviation	Minimum Sentiment Score	25%	50%	75%	Maximum Sentiment Score
"What aspect of this session was especially good?"	1,547	1.2192	0.4268	0.2974	1.0000	1.0605	1.6432	2.4248

When comparing the sentiment scores of responses related to "aspects of the sessions that need the most improvement" there was no statistically significant correlation. Respondents who gave very negative Likert scale ratings to a session overall were slightly more likely to provide highly negative open-ended feedback.

Survey Question	Opinion Phrases (n)	Mean Sentiment Score	Sentiment Standard Deviation	Minimum Sentiment Score	25%	50%	75%	Maximum Sentiment Score
"What aspect of this session needs the most improvement?"	1,186	0.981	0.4125	0.1522	0.6135	1	1.1172	2.3023

Research Question 3:

What types of survey questions can most sensitively and accurately gauge participants' satisfaction and self-assessments of learning?

Insights: As identified in the first research question, overall session assessments on a 1–5 scale were high across collected surveys. However, the analysis of responses to open-ended questions showed more variation in sentiment (see table below). In addition, the open-ended questions were included in all surveys during the 10-year period, whereas the surveys from 2010 through 2014 did not include the close-ended overall assessment questions. For those years, the sentiment scores provide some additional information on suggested improvements and drivers of satisfaction.

Additionally, although overall satisfaction as gauged by the overall assessment question was nearly perfect (4.9 / 5.0) in the last three years' surveys (2019-2021), during that same period negative feedback was collected in the open-ended questions. Viewing only the overall 4.9 / 5.0 score without assessment of the open-ended responses would indicate little room for improvement. Analysis of the topic groups within the open-ended responses, however, provides specific opportunities for continuous improvement. Therefore, in this study, the open-ended responses provide more sensitivity than close-ended questions in gauging satisfaction and self-assessment of learning.

Overall Satisfaction Rating Close-Ended Survey Question	
Mean Rating	4.7
Standard Deviation	1.2
Coefficient of Variation	0.255

Sentiment Scores Open-Ended Survey questions	
Mean Rating	1.28
Standard Deviation	0.48
Coefficient of Variation	0.375

Research Question 4:

What terms and topics in open-ended responses are most highly related with close-ended quantitative responses?

Insights: "Presenter characteristics" (related to the presenter's knowledge and quality of delivery) was the topic group most highly correlated to overall session satisfaction or dissatisfaction. Comments in the "technology-related characteristics" topic group were the least correlated to overall sentiment scores, suggesting that these characteristics are baseline expectations and are less likely than other characteristics to drive satisfaction or positive sentiment.

It was hypothesized that each topic group's mean sentiment score for a given session would be correlated to respondents' overall numerical ratings of the webinar or workshop. The responses to the question "What aspect of this session was especially good?" were correlated with the sentiment score of the overall webinar quality. There were no statistically significant correlations among various topics and overall opinion ratings, with the exception of the group called "presenter / delivery characteristics." There is a small but statistically significant correlation between that topic group and the overall opinion rating.

Spearman Rank-Order Correlation Analysis

The study correlated the sentiment score of the question "What aspect of this session was especially good?" with the numerical ratings shared on the Likert-style evaluation item concerning the overall webinar quality.

Topic	Total_(n)	Mean Sentiment	SD	Spear_(n)	p	r_s
Physical Characteristics (Technology)	61	0.883	.4125	47	.26937	.1678
Presentation Design Characteristics	448	1.114	.5165	338	.57455	.03063
Presenter / Delivery Characteristics	345	1.059	.4811	285	.002*	.1879
Presentation Artifacts / Assets Characteristics	470	1.047	.4600	343	.2054	.06854
Perceived Utility of Learning Characteristics	331	1.122	.4895	239	.2318	.07763
Interaction / Interactivity Characteristics	520	1.055	.4668	375	.32008	.05148
Unassigned: Unassociated Low-Info Phrases	513	1.325	.3656	330	.15905	.07769

**statistically significant at p<=.05*

Research Question 4:

What terms and topics in open-ended responses are most highly related with close-ended quantitative responses? (*continued*)

For the responses to the question, "What aspect of this session needs the most improvement?", a correlation analysis was conducted regarding the sentiment to the Likert rating of the overall webinar quality. There was no significantly significant correlation between any topic and the overall opinion rating, with the exception of the group called "presenter / delivery characteristics." There is a small but statistically significant correlation between that topic group's sentiment score and the overall opinion rating. The "unassigned" topic group also has a small but statistically significant correlation with the overall opinion rating.

Spearman Rank-Order Correlation Analysis

The study correlated the sentiment score of the question "What aspect of this session needs the most improvement?" with the numerical ratings shared on the Likert-style evaluation item concerning the overall webinar quality.

Topic	Total (n)	Mean Sentiment	SD	Spear (n)	p	r_s
Physical Characteristics (Technology)	151	0.703	.2974	101	.794	.0262
Presentation Design Characteristics	371	0.751	.4544	251	.0727	.113
Presenter / Delivery Characteristics	127	0.709	.5032	81	.042*	.2264
Presentation Artifacts / Assets Characteristics	179	0.803	.4361	119	.8803	-.01394
Perceived Utility of Learning Characteristics	142	0.784	.4736	106	.9371	.00774
Interaction / Interactivity Characteristics	243	0.734	.4385	163	.1995	.1010
Unassigned: Unassociated Low-Info Phrases	628	1.140	.3683	423	.0039*	.13994

**statistically significant at p<=.05*

IMPLICATIONS FOR USING SENTIMENT ANALYSIS AND TOPIC MODELING

- Over 75% of the phrases used in the open-ended survey responses related to one or more topic groups identified in previous literature on satisfaction and dissatisfaction with learning sessions. This provides additional support for previous research in this area through a new approach and analysis technique.
- Data visualization tools and portrayals allow for data queries that go beyond the initial study objectives and provide tools for EvaluATE to continue to investigate data sets on their own.
- The establishment of reliability between the manual coding of the hold-out sample and the sentiment analysis model results was analogous to establishing inter-rater reliability and provides support for the use of sentiment analysis and topic modeling in these and similar settings.
- When the data dictionary that is used to analyze text is defined, once prohibitive or expensive analyses can be replaced with machine learning techniques such as sentiment analysis and topic modeling to provide deep insights at a relatively economical cost.
- Going beyond "black box" machine learning techniques via custom-coded projects allows for contextually -nuanced phrasing to be more accurately coded. Examples of this custom-coded advantage include:

  - Differentiating between uses "professional" (to reference a presenter's quality or to refer to an occupation).
  - Differentiating between uses of "application" (to mean a way of putting learning to use, or to mean a document used to apply for a grant).
  - Distinguishing between technical issues related to "sound quality" and the terminology "sounds like" when describing the understanding of presentation content.

RECOMMENDATIONS TO EVALUATE

Survey Design & Data Management:

Based on the data review of questions collected from post-event feedback surveys over the ten-year history, there are several suggestions regarding data management in the survey collection process:

1. Ensure questions are worded consistently over time (remove "webinar" and/or "workshop" and use "session").
2. Ensure numbering system in Qualtrics is consistent over time for the same questions.
3. Potentially break down the "overall satisfaction" question into six categories aligned to the identified topic groups.
4. Evaluate establishing a "data governance" committee that reviews and approves any changes to survey questions in order to ensure consistency in wording and data structure over time.
5. Establish a centralized advanced database (MS SQL, MySQL, S3, Redshift, etc.) that allows for easier data retrieval, storage, and analysis.

Webinar Management:

Based on the findings from the sentiment analysis and topic modeling development described in this report, the data support several considerations for webinar design:

1. Excellent execution with regard to use of technology will not overcome poorly designed presentations or subpar delivery in driving satisfaction with the webinar or workshop. The results seem to indicate technology is a baseline expectation that must work, but exceptional technology delivery does not seem to be a major driver of satisfaction.
2. Focus efforts on presenter quality. The topic group most highly correlated to overall satisfaction was presenter and delivery characteristics. Focus efforts to improve delivery quality and ensure presenters possess high levels of knowledge and expertise in the area of discussion.
3. Use the collected structured data and the data visualization tool to help answer questions regarding specific aspects of webinars and workshops. Use these insights to design interventions to improve satisfaction. Use the collected open-ended survey results to check new sentiment ratings against collected sentiment ratings after making deliberate changes. For example, if the negative sentiment consistently relates to a topic such as question-and-answer sessions, attempt an intervention in this area and post-audit the sentiment related to this area after the intervention.

LIMITATIONS

Proper application of sentiment analysis and topic modeling techniques requires an understanding of the data's context, along with programming expertise to customize the phrasing for sentiment analysis. Using machine learning techniques such as these without the benefit of project-specific coding expertise subjects the resultant analyses to the assumptions made for more general populations in "off-the-shelf" solutions. As a result, accuracy may suffer, as terms that are desirable in one context may be undesirable in other contexts.

GLOSSARY OF KEY TERMS

Term	Definition
Accuracy	a measurement used to quantify the difference between an estimated or predicted value and the accepted "true" value
Confirmatory	seeks to confirm an already existing theory or hypothesis
Data harmonization	the standardization of data elements to facilitate statistical analysis and collation. As an example, a date format from one survey may exist as 1/2/2020 and another may exist as January 2nd 2020. Harmonization sets the standard format and converts all dates to that format.
Data harvesting	the collection of all data and metadata from the surveys. Metadata relates to data elements such as date of survey, webinar name, and presenter.
Exploratory	seeks to understand data without preexisting theories or hypotheses

Term	Definition
Harmonized database	a collection of data that has been standardized and made consistent for easier analysis and comparison
Independent group analysis	the comparison of two distinct groups to determine if there is a statistically significant difference between them
Inversion factor	linguistic element or contextual clue that reverses the polarity of a statement and allows for an accurate interpretation of its underlying sentiment
Kappa	a metric that reduces the accuracy calculation by eliminating agreement of ratings simply due to chance
Opinion phrases	segments of text, often the lowest level of complete communication in the response
Sentiment	the emotional content within written text
Sentiment value	a numerical value representing the level of positive or negative terms used in a text response

GLOSSARY OF KEY TERMS (*continued*)

Term	Definition
Standard or domain-specific library of terms	a collection of terms coded specifically for a domain of interest and study
Unequal variances	a situation where the standard deviations between two data sets are different
Unstructured feedback data	Information provided by participants or users that doesn't have a predefined structure or organization
Validate	within data science, the testing of a generated model against data that has been withheld from the training step used for model development
Validation data source	a subset of data elements excluded from the model building process. It is used as a testing data source to ensure the model is responsive to unseen data and is not over-fit from the training data source.
Verify	within data science, the testing of a generated model against training data

REFERENCES

Cheok, M. L., & Wong, S. L. (2015). Predictors of e-learning satisfaction in teaching and learning for school teachers: A literature review. *International Journal of Instruction, 8*(1), 75–90.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Draus, P. J., Curran, M. J., & Trempus, M. S. (2014). The influence of instructor-generated video content on student satisfaction with and engagement in asynchronous online classes. *Journal of Online Learning and Teaching, 10*(2), 240–254.

Fu, FL. (2010). Comparison of Students' Satisfaction and Dissatisfaction Factors in Different Classroom Types in Higher Education. In: Tsang, P., Cheung, S.K.S., Lee, V.S.K., Huang, R. (eds) Hybrid Learning. ICHL 2010. Lecture Notes in **Computer Science, vol 6248**. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14657-2_38

Peltier, J. W., Schibrowsky, J. A., & Drago, W. (2007). The interdependence of the factors influencing the perceived quality of the online learning experience: A causal model. *Journal of Marketing Education, 29*(2), 140–153.

ADDITIONAL READINGS

Jurka, T.P. (2012). Tools for Sentiment Analysis Package for R (v0.2). http://cran.rproject.org/src/contrib/Archive/sentiment/sentiment_0.2.tar.gz

Mohammadi, H. (2015). Investigating users' perspectives on e-learning: An integration of TAM and IS success model. *Computers in human behavior, 45*, 359–374.

R Development Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Retrieved from http://www.R-project.org.

Riloff, E., & Wiebe, J. (2003, July). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 105–112). Association for Computational Linguistics.