Elia Kostyrka

IST 736 – Text Mining

October 2019

Final Report

**Text Mining Training and Evaluation Data**

1. **Introduction**

Evaluation is critical for measuring the effectiveness of training. Following the Kirkpatrick Model for analyzing and evaluating the result of training, an organization collects data after each training event which measures “level 1 reaction” to the training (Kirkpatrick, 2019). Level 1 evaluation measures the degree that participants found the training favorable, engaging, and relevant to their job position/daily role (see Figure 1 for all levels). Responses to two open-text questions from a post-training survey were analyzed for this report to identify areas for improvement. The two questions are as follows:

* Q1. Why would you recommend- or not recommend- this training?
  + Precursor question on Likert scale: Would you recommend this training to your peers?
  + Q1a. Why would you recommend this training?
  + Q1b. Why would you not recommend this training?
* Q2. How would you improve this training or its delivery?

*What benefits has the organization experienced as a result of the training?*

*Have participants applied what they learned from the training?*

*How much did participants learn from the training and have their skills improved?*

*How did participants respond to the training?*

Figure 1. Kirkpatrick Evaluation Model

**Problem:** The organization requires an efficient way to analyze the survey phrases from over 11,000 participants trained each month to identify sentiment and training themes to help improve the effectiveness of training.

1. **Method** 
   1. **Data collection**

Survey responses were collected after each training event and consolidated over a one-year period (fiscal year). The data was organized by participant and training event in a .csv file. The file contained both structured and structured data, although only the unstructured text data was analyzed for this report. The unstructured data contained raw responses from participants for the two open-text questions listed above. Responses varied in text size from one-to-two words to a paragraph long. The dataset contained dirty data, including misspellings, acronyms, excessive punctuations, informal grammar, unnormalized case, organization lingo, and in some cases bad language. The dataset was preprocessed to normalize the data before analysis.

* 1. **Data cleaning**

Various experiments were needed while preprocessing the data. Data preprocessing was conducted for all of the analysis methods and included removing punctuation and digits, making all text lowercase, removing English stopwords and extended stopwords that could be associated with the question itself (e.g., words like “training”, “improvement”, and “would”). Additional words that showed nonresponsive phrases were also removed (e.g., words like “none,” “nothing”, or “n/a”). Some word transformations were also made for acronyms that came up as part of the most common words (e.g., “thx” was transformed to “thanks”). Lemmatization was conducted after tokenization to transform words to their base form to improve the summarization of word frequency. Correct spelling was attempted with little success, so this method was not used (e.g., the word “audio” was correct to “audit”). The Python libraries pandas, nltk, sklearn, and Textblob were used to help preprocess the dataset.

In order to analyze the question “Why would you recommend- or not recommend- this training?” the data was split into two separate datasets. Using the precursor question Likert scale question “Would you recommend this training to your peers?” as a guide, the data was split to represent phrases for those that would recommend training (Likert scale responses 4 and 5) and those that would NOT recommend training (Likert scale responses 1 and 2). Phrases from the neutral response (Likert scale response 3) were not used. There were 2700+ phrases associated with recommending training and 4300+ phrases associated with not recommending training.

For the question “How would you improve this training or its delivery?” all phrases were maintained as one dataset. There were 31,000+ phrases for analyzing. However, rare words were removed as there were various recommendations in the responses. In addition, extended stopwords were not removed for TF-IDF analysis.

* 1. **Models**
     1. **Unsupervised learning**

Three models were used for unsupervised learning, N-gram vectorization (word frequency), term frequency–inverse document frequency (TF-IDF) weighting, and the Latent Dirichlet Allocation (LDA) topic model algorithm were implemented using Python’s nltk, Texblob, sklearn, and genism libraries. Figure 2 shows the parameters and observation for each model. Each method required that the corpus be vectorized for processing.

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameters** | **Observation** |
| N-gram vectorization | - unigram, bigrams, and trigrams  - top 60 most common words then top 10 most common words | bigrams and a top-10 cutoff provided good insight for the question on whether a participant would recommend training, but unigrams and a top-10 cutoff provided better insight for the question on how a participant would improve training |
| TF-IDF weighting | - weighting of features  - word frequency multiplied by the log of the number of documents divided by the number of documents in dataset that contain the term  - extended stopwords not removed to maintain the identify of comments to search as an example | highest weight provided insight into specific issues |
| Topic modeling | - number of topics selection was 10  - number of words selected was 4 | visualization of distance map and most salient terms helpful |

Figure 2. Unsupervised learning model parameters and observation

* + 1. **Supervised learning**

One model was used for supervised learning, Multinomial Naïve Bayes on a labeled dataset for the question on improving training was implemented using Python’s sklearn library. Figure 3 shows the parameters and observation for the model.

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameters** | **Observation** |
| Multinomial Naïve Bayes classifier | - unigram count vectorizer  - stopwords not removed to allow for better classification  - 70% of dataset for question on improving training was labeled with one of 17 categories  - both hold-out method and cross-validation used  - 60%/40% train/test split for hold-out method used | the confusion matrix was difficult to interpret given the large number of categories, but the accuracy rate on the test dataset was satisfactory |

Figure 3. Supervised learning model parameters and observation

1. **Results by model**
   1. **Sentiment Analysis**

Sentiment analysis was conducted using Python’s Textblob library. The data set was analyzed separately for the first survey question about whether a participant would recommend training. Results in Figures 4a. through 4d. show that there are many neutral responses, indicating that a Likert scale survey item would likely provide better results that using text mining to determine the sentiment of responses.

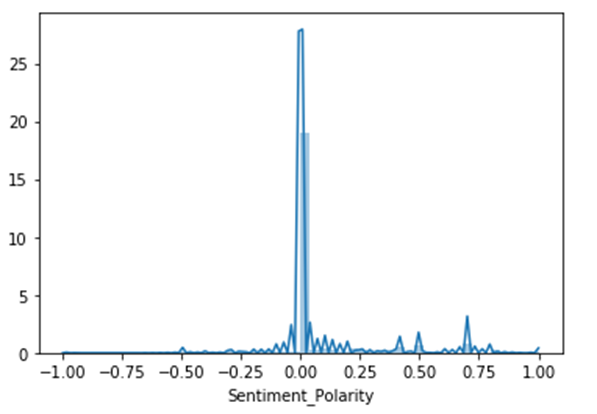
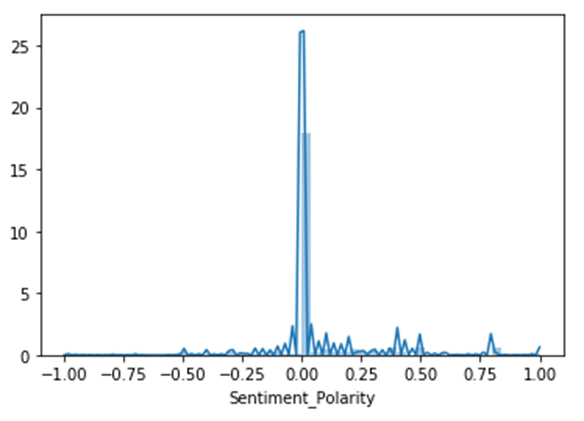


Figure 4a. All Recommend training sentiment polarity Figure 4b. Improve training sentiment polarity

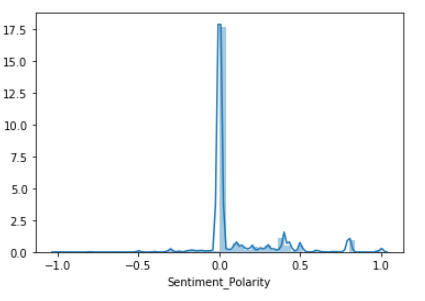
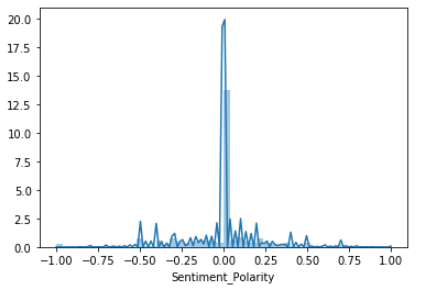
 

Figure 4c. Yes, Recommend training sentiment polarity Figure 4d. Not Recommend training sentiment polarity

* 1. **N-gram vectorization**

N-gram vectorization showed that participants that would recommend training to others felt that overall training provided a good refresher of the various topics covered over the fiscal year and provided good information that was needed on the job (see Figure 5). Participants that would NOT recommend training felt that there were issues with the assessment questions and the mock case file provided as reference material (see Figure 6). Participants largely recommended improving the questions in the assessment (see Figure 7).

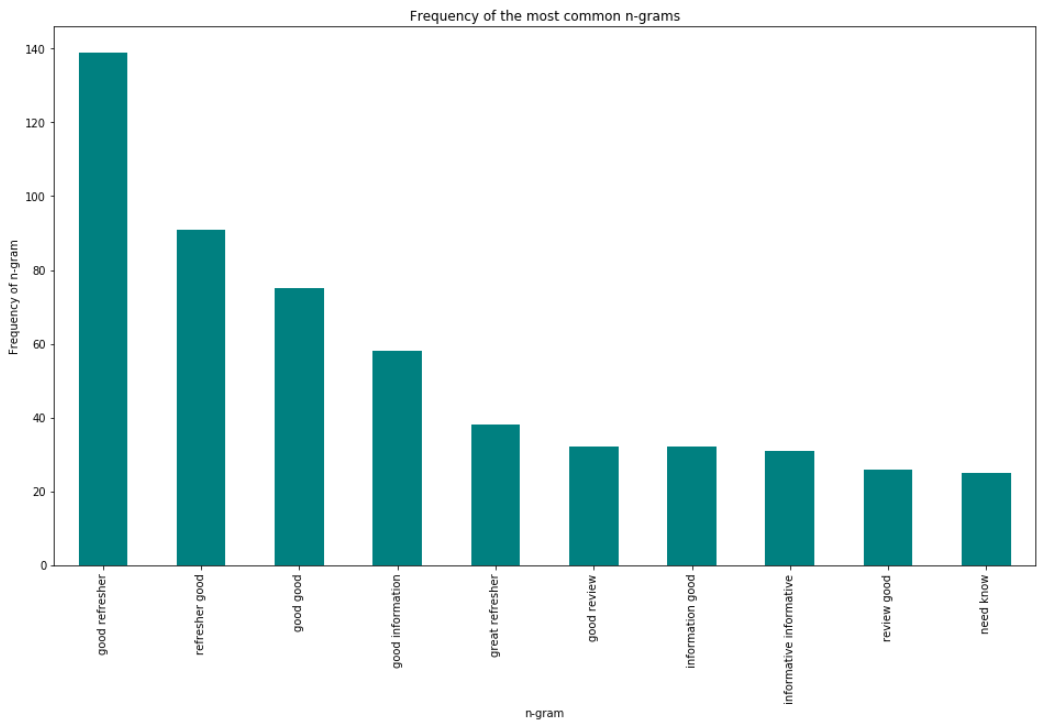


Figure 5. Yes, Recommend (bigram) Top 10 terms

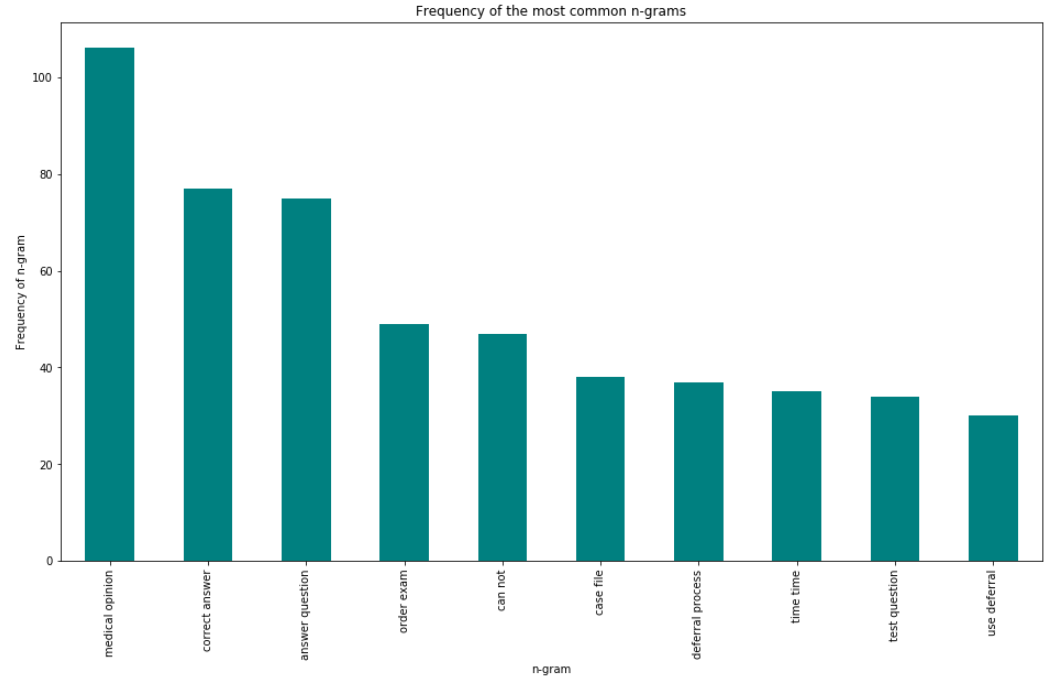


Figure 6. Not recommend (bigram) Top 10 terms

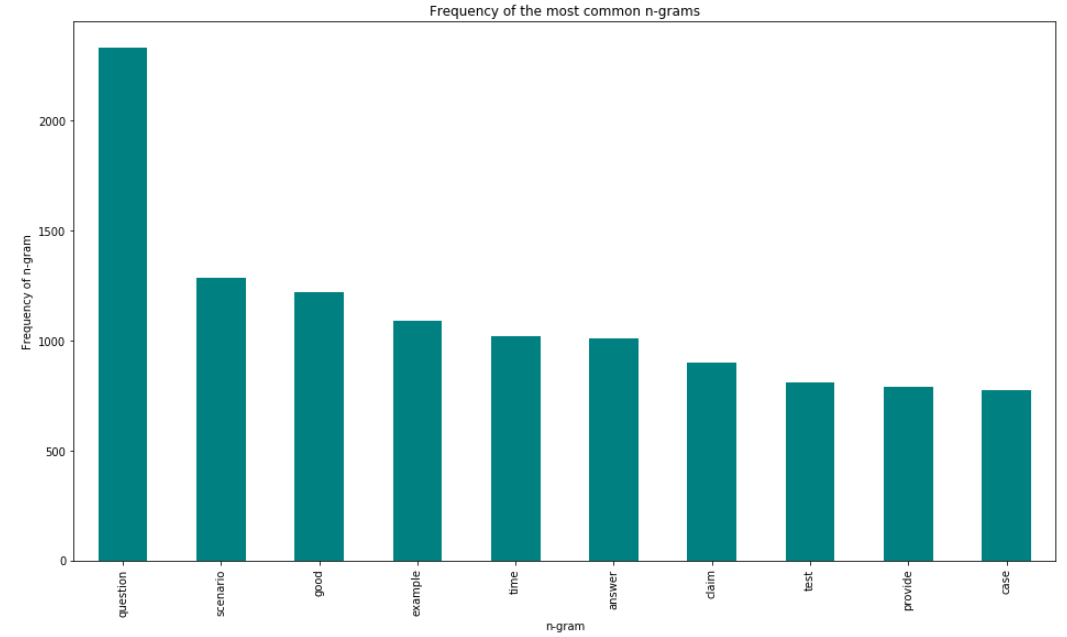


Figure 7. Improve training (unigram) Top 10 terms

* 1. **TF-IDF weighting**

TF-IDF measured how important a word was in the document (dataset). TF-IDF analysis showed that participants that would recommend training to others felt that the exam in the mock case file was helpful (see Figure 8a). Participants that would NOT recommend training felt that there were issues with the assessment scenario and its questions and answer responses (see Figure 8b). Participants recommended improving training by adding examples and reviewing the response answers to assessment questions (see Figure 8c). TF-IDF analysis provided a more refined list of specific issues with training.

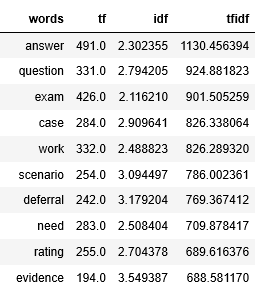
 

Figure 8a. Yes, Recommend Top 10 TF-IDF terms Figure 8b. Not Recommend Top 10 TF-IDF terms

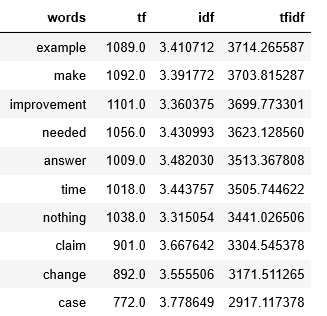
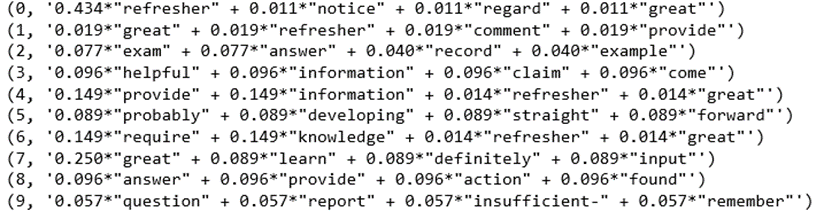


Figure 8c. Improve training Top 10 TF-IDF terms

* 1. **LDA topic modeling**

LDA topic modeling analysis showed that participants responses were clustered into three different quadrants for whether a participant would recommend training or not. Participants noted the exam and answer as one reason, training serving as helpful and a refresher as a second reason, and finally issues surrounding the questions as a third reason for recommending training (see Figure 9). Its likely participants would recommend training but then noted confusion with questions as an area for improvement.



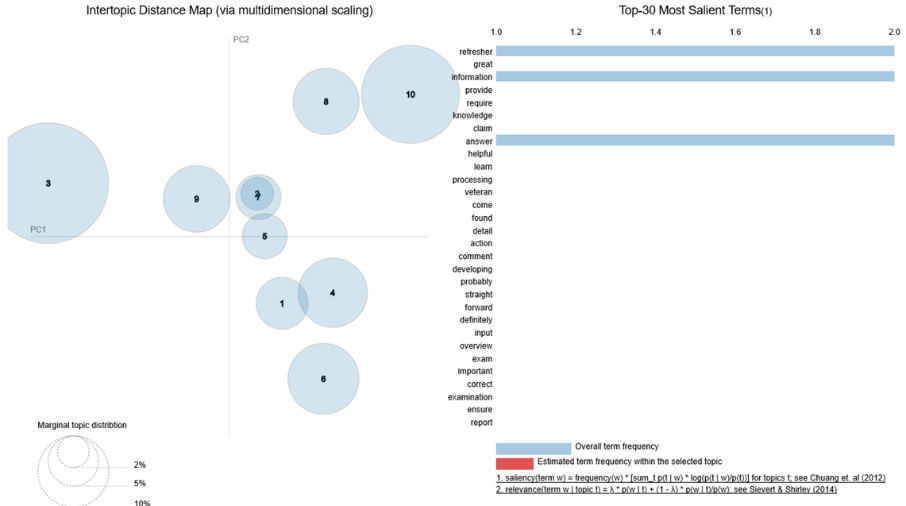
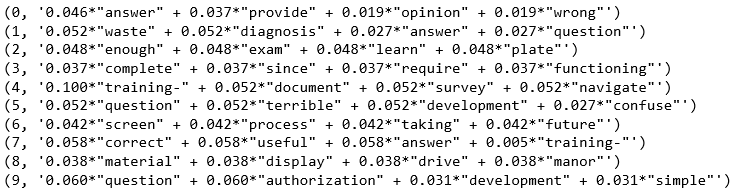


Figure 9. Yes Recommend 10 Topics

Participants noted the (correct) answer as one reason, the question as a second reason, and finally issues surrounding the content as a third reason for NOT recommending training (see Figure 10).



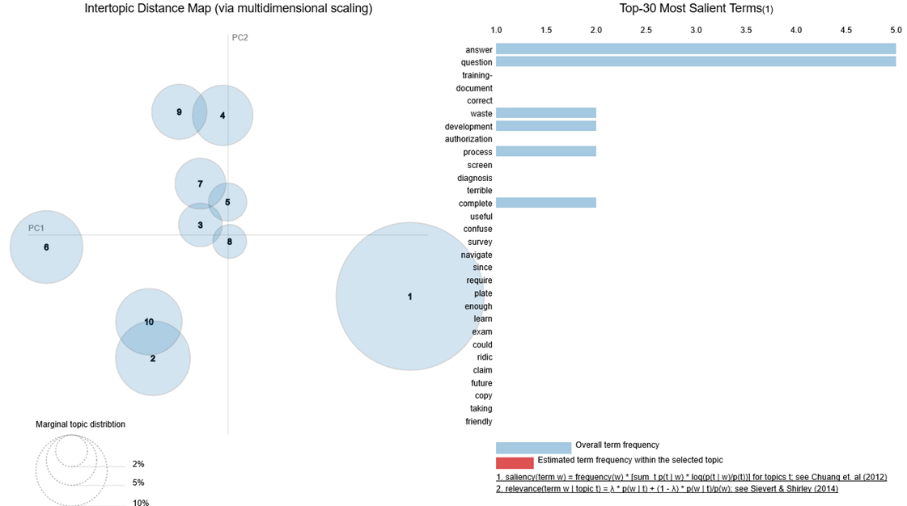
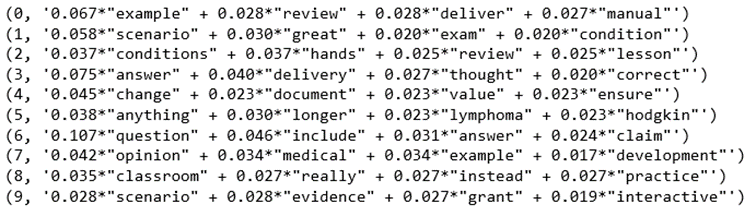


Figure 10. Not Recommend 10 Topics

Topics that participants noted for improvement included the question, the scenario and medical opinion in the mock case file, as well as, hands-on application, possibly in a classroom environment versus online training (see Figure 11).



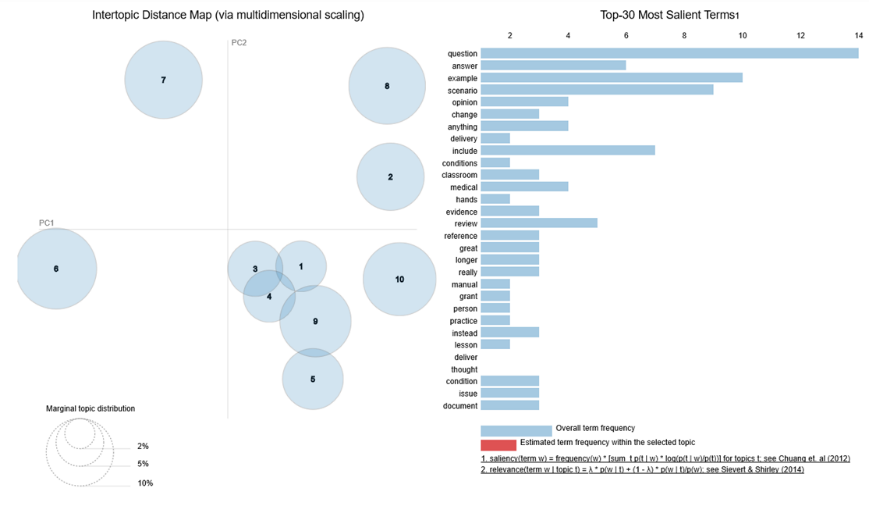


Figure 11. Improve training 10 Topics

* 1. **MNB classifier**

The MNB classifier implemented on labeled data provided a higher accuracy score on phrases classified on the test data. There was high accuracy on 6 of 15 topics (see Figure 12).

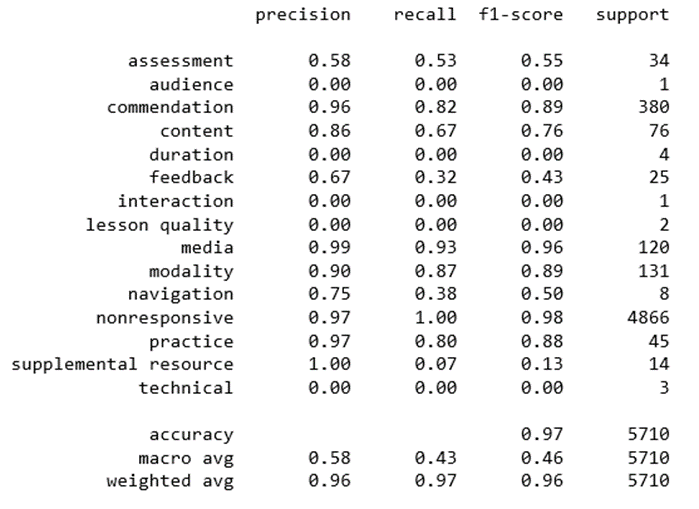
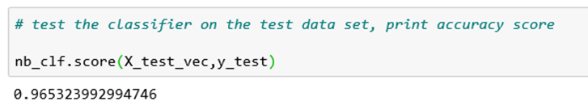


Figure 12. Improve training MNB classifier accuracy

1. **Conclusion**

Supervised learning provided the highest accuracy in predicting the category for survey comments; however, more labeled data is required for each category to assist with text classification and evolving themes over time. Unsupervised topic modeling was helpful in determining high level areas in training that required special attention and TF-IDF weighting provided specific phrase examples of importance.

**References**

Kirkpatrick Partners (2019). The Kirkpatrick Model. Retrieved from

https://www.kirkpatrickpartners.com/Our-Philosophy/The-Kirkpatrick-Model