Final Report

Author: Qing Liu (tdn4tv@virginia.edu) DS 5001 Spring 2023
Date: 05/02/2023

Purpose: This project implements a series of text analytic methods (e.g., PCA, LDA, word2vec, sentiment analysis) to analyze the transcript decouments from a teacher coaching intervention

Introduction

Delivering educational interventions in field settings may cause program deviations, especially compared to labotary settings . Evaluators can learn how a program functions in practice and determine whether it has been implemented according to design by using implementation research. Traditional methods of fidelity assessment involve valid metrics for each intervention and session ratings by observers, which makes the process time-consuming, expensive, and frequently impractical.

Anglin et al. (2021) (https://journals.sagepub.com/doi/full/10.1177/23328584211028615) suggest a novel and scalable way for evaluating implementation structures using semantic similarity in natural language processing. Researchers can assess whether an intervention was carried out according to a systematic protocol and how well it was repeated across sessions, sites, and studies by applying semantic similarity to the transcripts of the intervention sessions. The authors use semantic similarity approaches to quantify variations in intervention transcripts and measure intervention adherence.

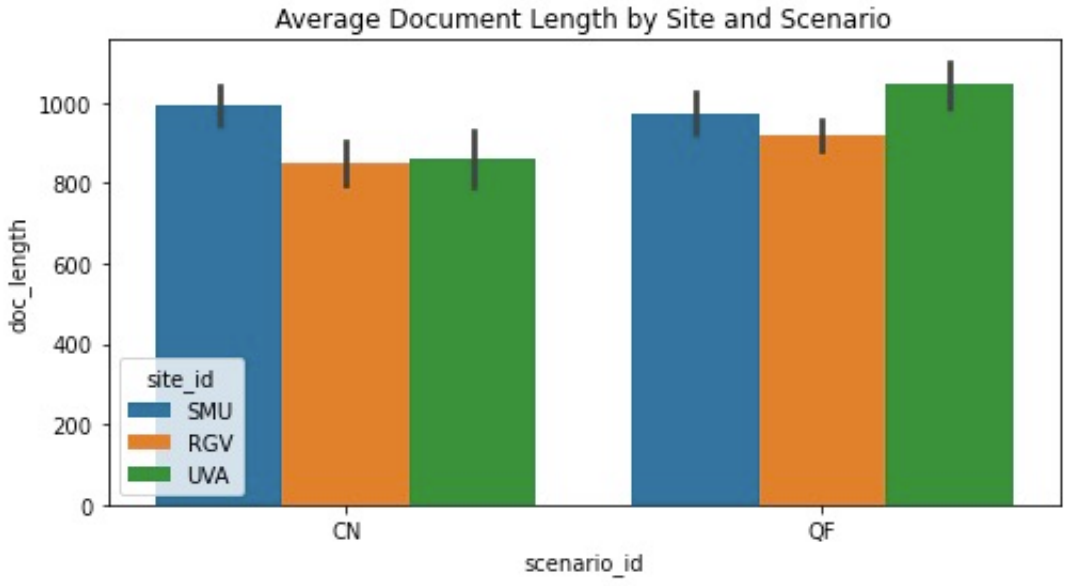
Motivated by the authors' previous research on measuring treatment adherence and consistency using natural language processing techniques, the current study will apply a series of text analytic methods to analyze transcripts from a teacher coaching intervention program. The dataset consists of transcript documents collected from the same program but across three different study sites and from different data collection cycle. The corpus includes all conversations between preservice teachers and coaches who provide suggestions to enhance the teacher candidates' pedagogical skills.

Source Data

Cohen et al. (2021) (https://journals.sagepub.com/doi/abs/10.3102/0162373720906217?journalCode=epaa) evaluated whether offering coaching to teacher candidates in their teacher education courses will improve their pedagogical skills on managing students' off-task behaviors while establishing norms and supporting text-focused instructions. This project makes use of the intervention transcripts from this teacher coaching program.

The source dataset comprises transcripts (https://virginia.box.com/s/zw88b5qqt93wg0pozpn1g3clc1uhv0ou) in plain text format, with each subfolder name indicating the site (e.g., UVA, SMU, RGV) or teaching task scenario (e.g., Classroom Norm (CN) vs. Quality of Feedback (QF)). The cleaned corpus (https://virginia.box.com/s/w89qtu05rruo02rwag8lnbzc9yl072z4)file has a hierarchical structure defined by the OHCO as document_id, paragraph_id, sentence_id, and token_id. It should be noted that the paragraphs are parsed based on each speaker's (e.g., preservice teacher vs. coach) talk within each transcript document. In most analyses, instead of embedding the scenario and site as additional layers into the OHCO, these two important pieces of information are saved as labels in the LIB (https://virginia.box.com/s/q1kgg1va4s11q0ov2gfuy68cra8fr6x8) dataset.

After tokenization, removal of stopwords and insignificant interjections, the corpus contains 200,702 rows distributed among 214 documents, with an average document length of 938. Analysis of figure 1 reveals that the average document length of the text-based discussion scenario is greater than that of the classroom norm scenario. Additionally, the UVA site shows the largest average document length for the QF scenario, while the SMU site has the largest average document length in the CN scenario.



Data Model

I. Data Processing and Tokenization

During the data processing and tokenization, four main tables were generated. The first table, named LIB (https://virginia.box.com/s/q1kgg1va4s11q0ov2gfuy68cra8fr6x8), contains the path and file name of each document, along with the type of site and scenario. The second table, called TOKENS (https://virginia.box.com/s/w89qtu05rruo02rwag8lnbzc9yl072z4), provides information on the corpus after tokenization, including the OHCO structure represented by document_id, paragraph_id, sentence_id, and token_id. Other columns include the speaker's position, name, and a flag indicating whether they are a coach or a preservice teacher. Flags are also used to identify stop words and insignificant words such as "yeah" or "okay". The third table, known as VOCAB (https://virginia.box.com/s/lyg6rmi6gu9ztn44x8cxdgxvjevdq60h), includes statistics for each term in the corpus, such as its frequency "n", probability of occurrence "p", information "i", the most probable part of speech "max_pos", tfidf, dfidf, and stem. Finally, the BOW (https://virginia.box.com/s/ivipft5fmesk1o80aa73fkrg182qrj90) table is generated from the corpus by collapsing at the document and paragraph level. It provides the document-term count matrix to calculate measures related to TFIDF.

II. Principle Component Analysis

Before the PCA process, I reduced the feature space of the TFIDF by keeping the top 1000 significant nouns, verbs and adjectives. There are some major analytic tables generated after applied the PCA on the normalized the TFIDF table. LOADINGS (https://virginia.box.com/s/j82o047jprfvvn4n1bzzvpi0cdx0s370) table shows the contribution of each term to the top five components. For instance, it shows the that top terms contributes the most positively to the 1st component that explains the most variance, including "think", "lisa", "text". The COMPONENT_WORD (https://virginia.box.com/s/z5xijs5wsjc5fdvhotqtknkiw0j1agbr) table includes the 5 top components that explained the most variance of the transcript data, and the top terms that contribute thhe most to each components. For instance, as table 1 shows, The 1st component explained the most variance of the data. It has a evident pattern that the 1st principle component is positively correlated with words that are from text_beased dicussion scenario, such as preservice teachers provides feedback to the avatar "Lisa" in the simulated classroom, while the 1st component is negatively correlated with behavior mangament scenario. For example, the negative columns include words of Ethan, talking, classroom, which refer to the Avatar Ethan perform those off-task behaviors (singing, humming, etch.), and perservices teachers are trying to redirect his behavior and build the classroom norms. So it seems the PCA is able to distinguish these two scenarios.

On the other hand, the second component seems to play a role to differentiate the conversations between speaker 1 (coach) and speaker 2 (preservice teacher). The positive column includes words such as "feel", "sorry", "thank", "um", which are most likely from preservice teacher. And the negative column includes words such as "want", "ethan","stop", "students", "ask", which appears to refer to when coaches provide the training to those preservice teacher.

Table 1: Components and Top Words

1	pos	neg
comp_id		
PC0	think lisa text nervous paragraph evidence cal...	stop please ethan talking classroom behavior n...
PC1	went think alright feel good simulation teache...	text evidence going want ethan stop student as...
PC2	think went good text evidence students um want...	feel lie detector brings results likely pismo ...
PC3	think stop lisa please ethan calm detector lie...	go thank text evidence simulation questions gr...
PC4	paragraph says relaxed nervous heart smiled ch...	going simulation feel went alright evidence te...

Another key analytic table is the Document Component (https://virginia.box.com/s/6o0efonzke9kn9vboqutkovdxga6dokg) table which projects each document onto those five components.

III. Topic Modeling (LDA)

The parameters in the LDA include: n_gram_range = [1,2], n_terms = 1000, n_topics = 10, max_iter = 20 and n_top_terms = 9. Key analytic tables include THETA (https://virginia.box.com/s/aico38paikef160r2zozo8yxb0ren4la), which contains the probability of each topic in the corresponding document. On the other hand, the table PHI (https://virginia.box.com/s/zj0t6bt6qmee5gwowjp24n6ssjg4fzl6) contains the probability of each term occuring in the topic. The table TOPIC (https://virginia.box.com/s/qmsclw4ors2zapzq61r0kqs5vq7v3ou1) includes top ten topics from transcript corpus and nine top terms in each topic. For example, the T00 topic includes words like "detector", "lie", "results", feel". For more interpretation , please see the detailed exploration section.

IV. Word Embeddings

The parameters in the word2vec include: windown = 5, vector_size = 200, min_count = 30, and workers = 4. The key analytic table is GENSIM_DOCS (https://virginia.box.com/s/yuu2t4ofh1xnavt62wltwmmdyl9912lz) which includes a collection of documents that are represented as a list of words. Those words are staying together in each 'document' because they tend to occur in the similar contexts simultaneously. For example, the 7th row is a short document only has two words "paragraph" and "four", these two words often occur in the transcripts from the text-based discussion, indicating some instruction on the quality of text-based feedback.

V. Sentiment Analysis

This project applies the NRC lexicon to explore documents similarity across study sites and different scenarios regarding sentiment scores from NRC lexicon and Vader. The analytic tables DOCUMENT VADER (https://virginia.box.com/s/9743rwn3u6zijb5v8lktbp3hae6yhqoz) includes the NRC sentiment scores for each transcript document, such as anger, anticipation, disgust, fear, joy, negative, postive, sadness, surprise, trust, and sentitiment score. But it also includes metrics such as pos, neg, neu and compound from Vader method.

Detailed Exploration

I. PCA

Figure 1 projects each document onto the first and second components derived from principal component analysis. It is clear that the first component effectively distinguishes documents into two clusters, which align with each document's scenario (e.g., quality of feedback, classroom norms). From the plot, we observe that the quality of feedback (QF) documents are dispersed on the right side, whereas the classroom norms (CN) documents are dispersed on the left side. This further confirms that the first component is positively correlated with the text-based discussion scenario and negatively correlated with the classroom norm scenario. Regarding the second component depicted on the y-axis, it is evident that the majority of documents are concentrated towards the bottom, with fewer documents situated at the top. This observation is consistent with the fact that each transcript contains a higher proportion of coach discussions due to the intervention's nature. Furthermore, it is apparent that the quality of feedback (QF) scenario exhibits greater dispersion along the second component compared to the classroom norms (CN) scenario. This phenomenon also aligns with the expectation that text-based discussions would involve a broader range of conversations between coaches and preservice teachers.

Figure 2 shows that the text-based discussion conversations from RGV and SMU sites exhibit greater similarity in terms of distribution along the 1st component, while those from the UVA site demonstrate a higher degree of variation. Interestingly, when it comes to the classroom norm conversations, RGV and UVA display a more comparable distribution, while SMU exhibits less variation.

Figure 1. Scatter plot of document on the aixs of 1st and 2nd components

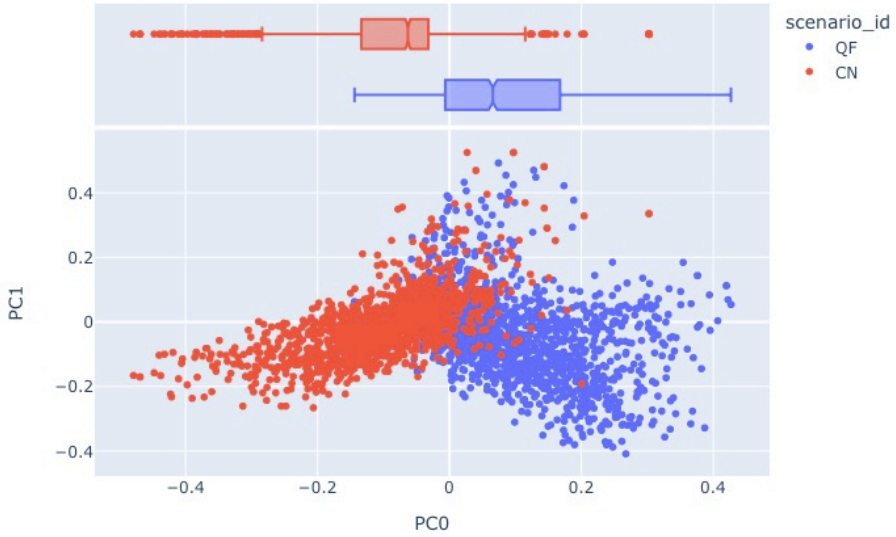
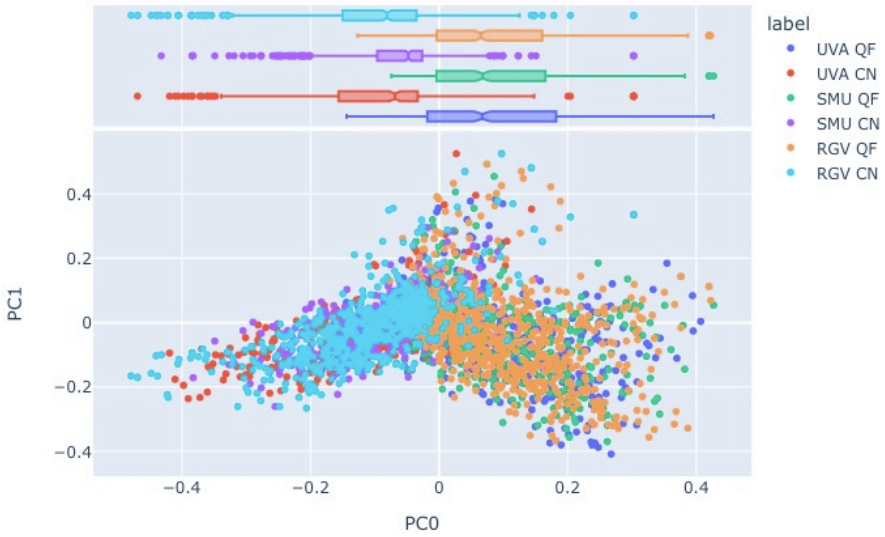


Figure 2. Documents by study site and scenario

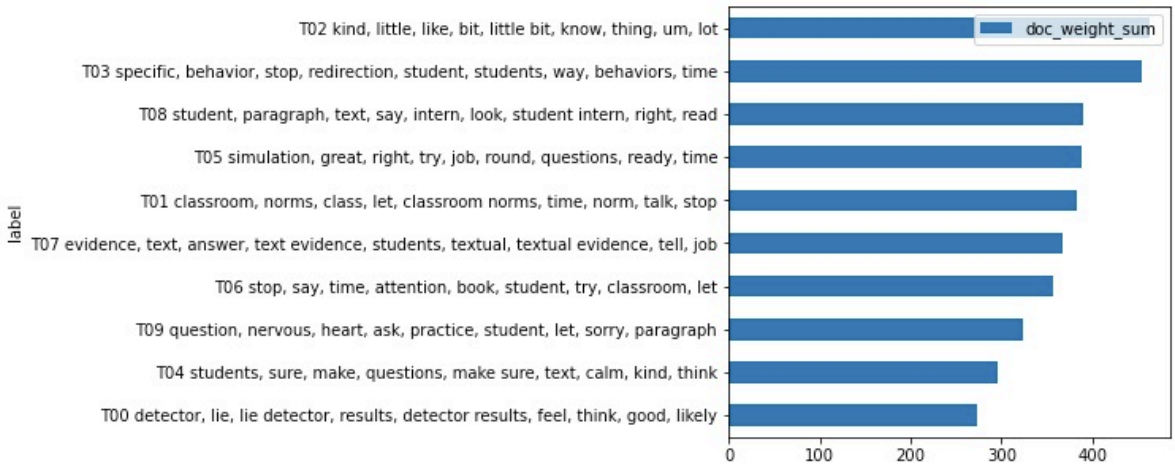


II. LDA

The document weight sum provides an overall distribution of topics within the entire corpus. Aftering sorting topics by their document weight sum, according to figure 3, we observe that T02 topic, which has the largest document weight and indicates its dominancy in the corpus, includes words such as "kind, little, like, bit, little bit, know, thing, um, lot". Within the context of conversations between coaches and preservice teachers, this topic appears to center around expressions frequently employed by preservice teachers to temper their statements, conveying a sense of uncertainty and hesitation.

Meanwhile, the T03 predominantly concerns student off-task behaviors and preservice teachers' redirection efforts. In contrast, the T04 appears to be more closely related to conversations within the realm of text-based discussions.

Figure 3. Document weight sum by each topic



Interestingly, according to figure 4 the T02 topic (conveying uncertainty and hesitation) and T03 topic (re-directing off-task behaviors) seems to be more prevalent in SMU site, compared to RGV and UVA. Overall, T08 and T06 topics are more prevalent in RGV site, where T08 seems to suggest that teachers may use various techniques, such as asking students to stop their off-task behaviors, redirecting their attention, and encouraging them to try different activities or focus on their books. Additionally, the words "classroom" and "student" emphasize that these conversations are centered on managing the classroom environment and addressing individual student behaviors. T07 seems to be the most prevalent in UVA site, which suggests that teachers are emphasizing the importance of finding evidence from texts to support students' answers and arguments. Additionally, words such as "tell" and "job" imply that teachers may provide feedback or guidance to students, helping them improve their skills in extracting and utilizing textual evidence effectively. Overall, this topic reflects the instructional focus on promoting students' comprehension and critical thinking abilities by engaging them with text-based tasks and providing high-quality feedback.

As shown in figure 5, in the context of teacher coaching intervention, upon examining the topic weights disaggregated by classroom norms and quality of feedback scenarios, it becomes apparent that topics T01, T03, and T06 are associated with conversations centered around redirecting students' off-task behaviors. Meanwhile, topics T07, T08, and T12 are more closely related to providing feedback or guidance to students during the teaching process.

Figure 4. Document weight on topics across sites

term_str	RGV	SMU	UVA	label
topic_id				
T00	0.071752	0.063316	0.089357	T00 detector, lie, lie detector, results, detector results, feel, think, good, likely
T01	0.095437	0.119613	0.096982	T01 classroom, norms, class, let, classroom norms, time, norm, talk, stop
T02	0.105413	0.154922	0.118960	T02 kind, little, like, bit, little bit, know, thing, um, lot
T03	0.102738	0.144769	0.125280	T03 specific, behavior, stop, redirection, student, students, way, behaviors, time
T04	0.092210	0.068314	0.076623	T04 students, sure, make, questions, make sure, text, calm, kind, think
T05	0.104231	0.106084	0.105374	T05 simulation, great, right, try, job, round, questions, ready, time
T06	0.119412	0.070967	0.094503	T06 stop, say, time, attention, book, student, try, classroom, let
T07	0.098586	0.083255	0.118117	T07 evidence, text, answer, text evidence, students, textual, textual evidence, tell, job
T08	0.123649	0.097295	0.090342	T08 student, paragraph, text, say, intern, look, student intern, right, read
T09	0.086571	0.091465	0.084461	T09 question, nervous, heart, ask, practice, student, let, sorry, paragraph

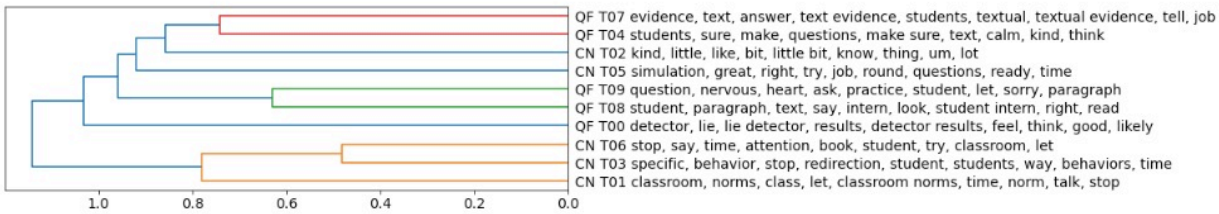
Figure 5. Document weight on topics across scenarios

term_str	CN	QF	label
topic_id			
T00	0.045820	0.103978	T00 detector, lie, lie detector, results, detector results, feel, think, good, likely
T01	0.161372	0.042690	T01 classroom, norms, class, let, classroom norms, time, norm, talk, stop
T02	0.135887	0.114022	T02 kind, little, like, bit, little bit, know, thing, um, lot
T03	0.209762	0.030814	T03 specific, behavior, stop, redirection, student, students, way, behaviors, time
T04	0.055867	0.105567	T04 students, sure, make, questions, make sure, text, calm, kind, think
T05	0.113566	0.096270	T05 simulation, great, right, try, job, round, questions, ready, time
T06	0.151942	0.038207	T06 stop, say, time, attention, book, student, try, classroom, let
T07	0.034087	0.168139	T07 evidence, text, answer, text evidence, students, textual, textual evidence, tell, job
T08	0.036566	0.178521	T08 student, paragraph, text, say, intern, look, student intern, right, read
T09	0.055132	0.121791	T09 question, nervous, heart, ask, practice, student, let, sorry, paragraph

When clustering topics by document weight, it shows (figure6) that T03 and T06 topics are very close, we might interprete the topic cluster relate to classroom management and redirecting off-task behaviors. For instance, T03 emphasizes the importance of specificity in redirection, acknowledging and stopping problematic behaviors, and managing students' actions, and T06 highlights the role of teacher communication ("say"), time management, and maintaining students' attention. The term "book" may refer to teaching materials or resources, suggesting that the discussion might involve lesson planning and organization.

Figure 6. Topic similarity

<Figure size 432x288 with 0 Axes>



III. Word2vec

As figure 7 shows, The bottom cluster in the plot, which is centered around the word "Lisa", indicates the presence of discussions in a text-based scenario. In this particular scenario, Lisa is a character who serves as a point of reference for the student avatars, and the preservice teacher is responsible for providing constructive feedback to the students based on their perspectives of Lisa. An example of such a perspective is given in the original text, where Lisa is portrayed as being excited about her new job. The student avatars in the discussion are named Ava and Jasmine, and they are expressing their thoughts and opinions about Lisa based on the information provided in the text. In contrast, for the cluster on the right, centered around the word "Ethan", indicates off-task behaviors from Ethan, such as "Ethan is drumming", "Ethan is whistling", "Ethan plays Darth Vader". These two word clusters are further confirmed by the output from the function to get similar words as table 2a and table 2b shows.

Figure 7. Word cluster

Table 2a. Similar words for 'ethan'

	term	sim
0	making	0.998582
1	whistling	0.998301
2	talking	0.997219
3	quiet	0.996824
4	stop	0.996718
5	specific	0.996617
6	instead	0.996586
7	need	0.996495
8	darth	0.996325
9	drumming	0.996221

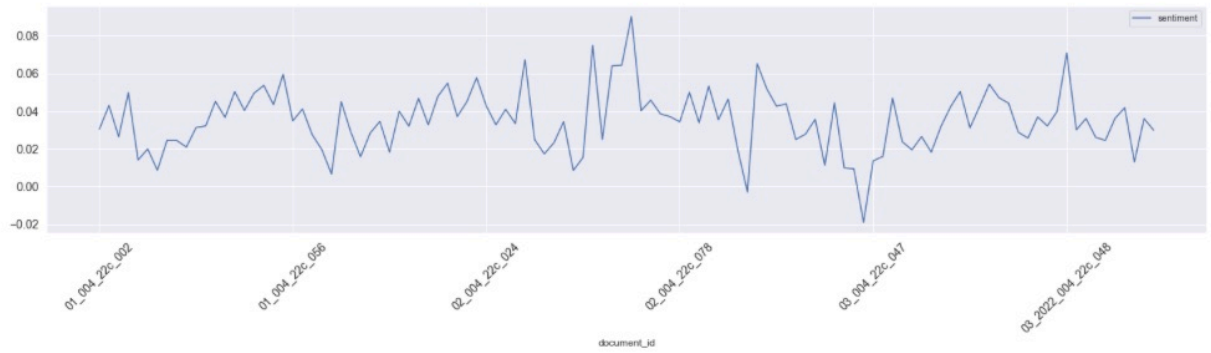
Table 2b. Similar words for 'lisa'

	term	sim
0	feeling	0.991205
1	lie	0.989978
2	pizmo	0.987431
3	pismo	0.987161
4	brings	0.986692
5	nervous	0.986241
6	says	0.985507
7	relaxed	0.985117
8	22	0.984527
9	likely	0.983045

IV. Sentiment

It is notable that I initially expected minimal variability in the sentiment across documents, as the discussions and topics in the transcript documents were limited to teacher coaching interventions. However, I observed significant fluctuations in the sentiment across various documents(shown in figure 8).

Figure 8. Overall sentiment score across documents in CN scenario



Based on the analysis of sentiment distributions across different sites, it is observed that transcripts from the UVA site exhibit a normal distribution of sentiment values. Conversely, the SMU site stands out as having documents with negative sentiment values, which warrants further exploration of the underlying text and factors contributing to the extreme values observed in the sentiment distribution.

When looking at sentiment value histogram for two distinct scenarios, it is noteworthy that the quality of feedback scenario exhibits a greater incidence of documents with negative sentiment, as well as an overall lower average positive sentiment, in comparison to the classroom norm scenario. One would have perhaps expected the classroom norm scenario to display a greater frequency of negative sentiment, given the nature of the student's off-task behaviors and the high levels of reported stress among preservice teachers, including reports of tears in some instances. As such, the findings of the sentiment analysis between the two scenarios appear to suggest unexpected trends that merit further investigation.

Figure 9a. Histograms of CN sentiment by Groups

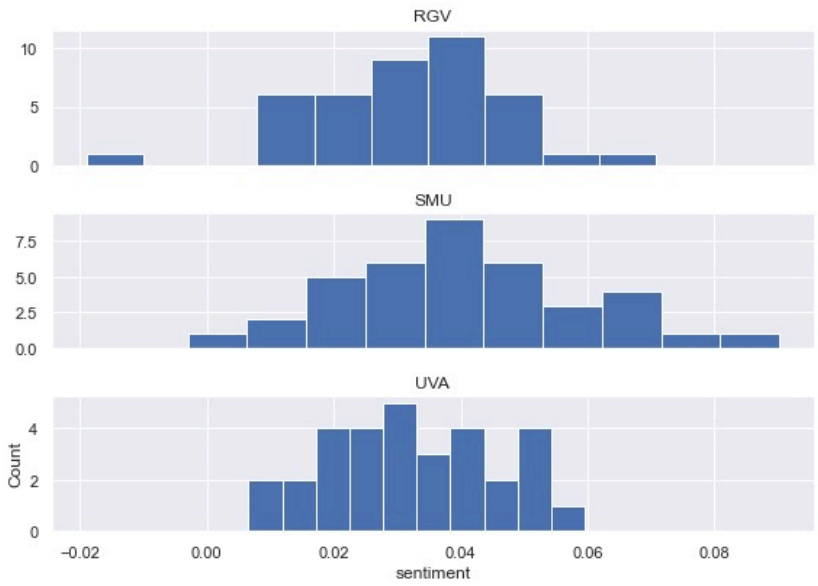
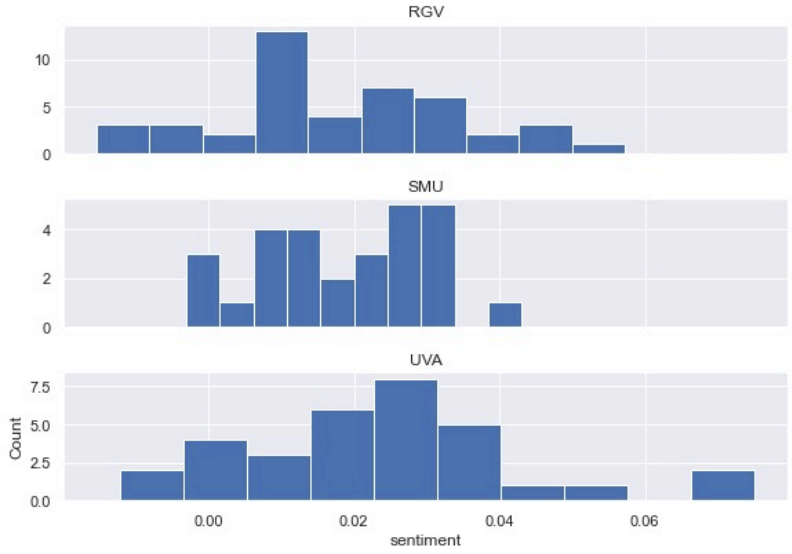


Figure 9b. Histograms of QF sentiment by Groups

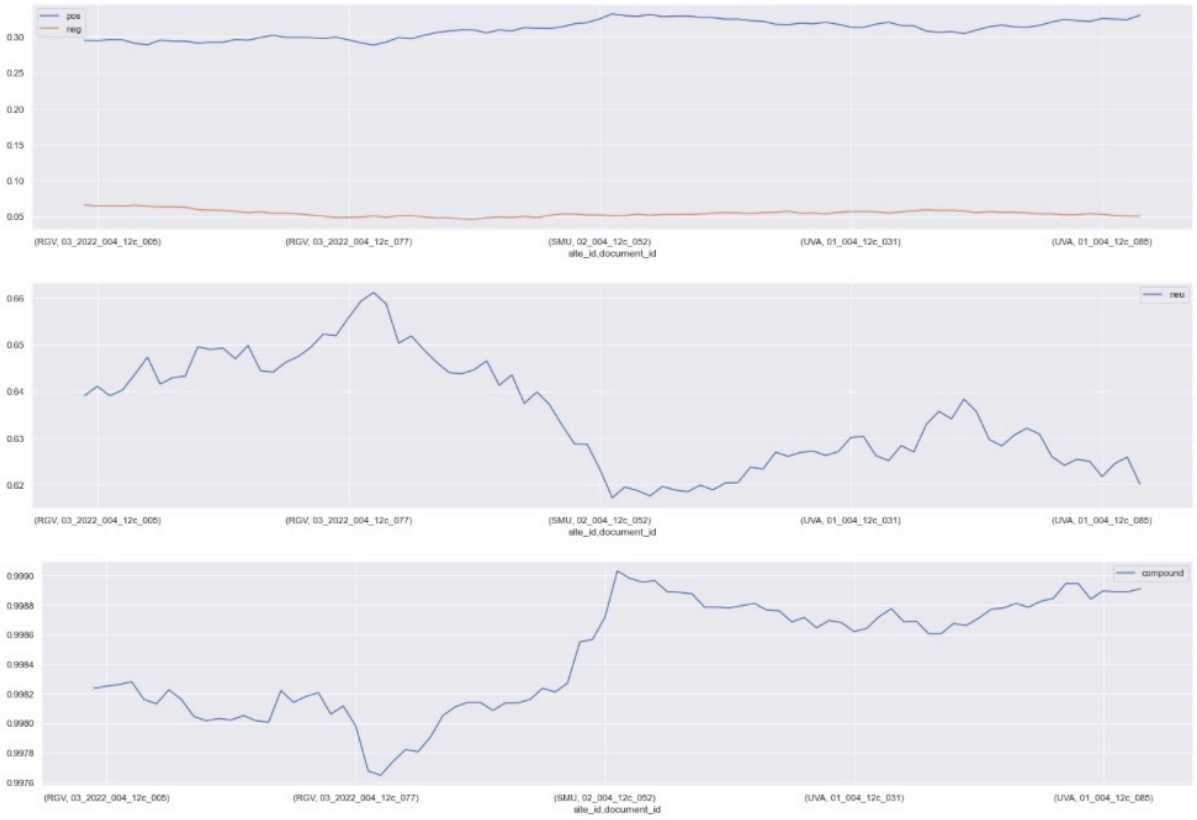


In general, the sentiment polarity scores of transcript documents across all sites, regardless of the CN or QF scenario, exhibited a remarkably high degree of positivity, with scores exceeding 0.98. Notably, documents from the UVA site displayed a marginally higher average sentiment polarity score than those from the other two sites. This observation may suggest that conversations from the UVA site were more positively oriented than those from the other sites.

Furthermore, the sentiment polarity patterns exhibited a remarkable degree of consistency across sites, regardless of scenario. For instance, the RGV and SMU sites displayed a greater prevalence of neutral words in comparison to the UVA site, across both scenarios.

Figure 10. Sentiment across documents in CN scenario

: <AxesSubplot:xlabel='site_id,document_id'>



Conclusion and Interpretation

Both PCA and LDA effectively distinguish the differences in the text between CN and QF scenarios, dividing the terms or topics into two distinct clusters. Additionally, the sentiment analysis reveals that documents from different sites, even in the same scenario, exhibit variations, requiring further investigation.