How we can use Natural Language Processing to analyse free responses

Lucia Chen, Junru Lin and Huiyu Sun

Pharmacology and Toxicology Department & Philosophy Department, University of Toronto.
Computer Science Department, University of Toronto.
Computer Science Department, University of Toronto.

Contributing authors: lucia.chen@mail.utoronto.ca; junru.lin@mail.utoronto.ca; huiyu.sun@mail.utoronto.ca;

Abstract

Our research investigates the potentials and limits of natural language processing (NLP) and its application towards efficiently and effectively generating a summary of a large textual dataset such as the free-response survey results used in Nguyen et al., 2021. Our results demonstrate that vocabulary statistics for preprocessed responses and clustering topics with TF-IDF and k-means can provide an overview of the data; topic modeling algorithm, LDA, can be applied to generate key terms and representative texts which can then be analyzed to draw hypothesized conclusions and summaries; and no significant bias concerning online learning from the dataset responses have carried over to the GloVe word embeddings. The findings of our research can offer insights into methods of analyzing free responses, and offer an approach to identifying bias in algorithms.

Keywords: NLP, TF-IDF, K-Means, LDA, GloVe, bias test, RIPA

1 Introduction

In many processes such as research, feedback is important. Precise feedback helps researchers to improve current designs, make predictions, and learn. Feedback requires data processing, which has many trade-offs. When working with a large survey dataset such as the one in Nguyen et al., 2021., four researchers first read hundreds of responses, and a list of themes was developed inductively from the survey data. An initial codebook was also revised collaboratively based on feedback from researchers whom each coded 20-80 responses. Researchers aligned in identifying the same major themes from the data. (Nguyen et al., 2021) The process is time-consuming and requires much manual labor. Different methods of collecting data all have advantages and disadvantages, but among all the forms of data collection, free-response answers are both the most rewarding yet hardest to analyze. While free response allows for more flexible and personalized answers as researchers give up control conditions, this allows other variables to be introduced such as stylistic choices and bias from language, as well as potential incoherence and personal biases. It is also a tedious and long process to go over each response manually.

Many methods have been developed to help analyze survey data with free responses. The development in the field of machine learning and natural language processing over recent years allows surveys with both qualitative and quantitative data to be efficiently summarized and analyzed.

For our project, we thought about how machine learning could be used in free response processing. Natural Language Processing (NLP) is a growing field. For example, O'Cathain and Thomas (2004) put forward four steps for analysis:

- Read a subset of the free responses
- Devises a coding frame to describe the thematic content of the free responses

- Assigns the codes to all the comments responses
- The codes can be entered into a statistical package alongside the data from the closed questions and treated as variables in a quantitative analysis

Instead of the manual work of labeling and classifying, NLP techniques such as topic modeling and clustering can be explored to generate a hypothesized description on qualitative data such as free responses in a much shorter time frame. Similar research in the field of text summarization with NLP has been done in mainly two ways, extractive and abstractive. Josef Steinberger et al., described a generic text summarization method that used the latent semantic analysis technique (LSA) to identify semantically important sentences and suggested two new evaluation methods based on LSA, which measure content resemblance between an original document and its summary. Jen-Yuan Yeh et al., used a trainable summarizer for summarization. A trainable summarizer considers several features such as position, positive keyword, negative keyword, centrality, and the resemblance to the title, to generate summaries. In our research, we are applying some similar techniques to generate computed results from free responses in survey data instead of pure texts. For analysis, conclusions are drawn by subjective hypotheses, which are not identified facts. Combining machine-assisted thematic findings with human analysis, NLP introduces a new procedure of survey analysis which is more efficient, possibly more effective, and could involve less human subjectivity and biases.

We also wanted to measure the limits of NLP against variables such as bias. Bias can become a problem in our society if it turns to discrimination. Ingrained in cultures, bias can spread rapidly through language. (Ferrer et al., 2019) We are careful to detect, mitigate, and rethink our biased views in spoken language but the same cautions are not always taken in the languages of code.

Computers understand very little of human language, and so vector-space models have been developed, e.g. by Pennington et al., 2014 to give meaning to our language. They turn written language into a binary that can easily be processed by machine learning. Vector-space models require less labor and are good at filtering information, however, contain limits regarding semantics. (Turney & Pantel, 2010) Even with the same sentence structure and context, the same sentence can have different meanings depending on social cues often omitted in code. Current research has found AI systems can inherit unconscious human biases by processing and learning human languages, which can be a big problem with machine learning and the expanded use of AI. (Zhao et al., 2019) (Caliskan et al., 2017) (Papakyriakopoulos et al., 2020) (Swinger et al., 2019) (Ferrer et al., 2019) Discrimination could be introduced in new fields in ways that might even have serious repercussions in physical life. If AI such as NLP were to be more broadly used, it would be important to know if inputted data could have introduced or magnified unconscious biases.

Past research into bias embedded into word embeddings have focused on:

- Detecting biases in word embeddings through bias tests, such as the development of WEAT (Word Embedding Association Test) by Caliskan et al. (Caliskan et al., 2017)
- Proposing methods to reduce and/or remove bias from word embeddings, such as the work of Bolukbasi et al. (Bolukbasi et al., 2016)

This research combines NLP with techniques and findings from social science and psychology to target the interdisciplinary problem of bias transmitted across word embeddings.

However, the use and applications of NLP are still in their infancy. In this paper, we want to develop a simple and useful framework of NLP to analyze free responses in a survey. Our research will provide new insights into methods and considerations of data processing, innovating and accelerating the process of data processing and analysis. Compared to many published papers also on topic modeling and unsupervised learning, our goal is more tailored to the given dataset which combines the Likert scale, multiple-choice questions, and free responses. When analyzing we consider all types of data to generate a summary.

2 Data and Method

Our research uses the data collected from Nguyen et al., 2021. In this paper, a survey was distributed through the British study tips influencer Instagram account @unjadedjade to collect the perceptions of student online learning experience at the height of the COVID-19 pandemic during spring 2020. Data was collected in the forms of free-response answers, Likert scale, multiple-choice, and dropdown checkboxes. Questions were asked to compare the differences and preferences for online and in-person learning. Overall conclusions drawn from this data found that students valued social interactions and active learning in remote learning. (Nguyen et al., 2021)

From the CSV file, we preprocessed the data and clustered it with Latent Dirichlet Allocation (LDA) (M.Blei 2003) and TF-IDF (Qaiser, 2018) and k-means (Likas, 2003). The data was also put into Global Vectors for Word Representation (GloVe) word embeddings (Pennington et al., 2014)and were run through the Relational Inner Product Association (RIPA). (Ethayarajh et al., 2019)

Preprocessing with free responses includes several steps. The text was modified to lowercase only, punctuation and stop words were removed to avoid words without meaning impacting the quality of the embedding. Words were then stemmed using the nltk.PorterStemmer package. Blank rows were labeled as ‘non-response’, and were filtered from the dataset.

Extracting information from free text is time-consuming. For example, to analyze the free responses in a survey, some researchers could only focus on subsets of the data rather than attempting in-depth qualitative analyses of all comments received, and extract particular topics of comment for detailed analysis using multiple keyword searching (Garcia et al., 2004).

We chose six survey questions (two fixed-answer questions and four free-response questions) to explore how to use NLP to analyze three questions:

1. What is the preference among different modes and why
   - What is your preferred mode of teaching for online courses? (fixed answers)
   - Why is this your preferred mode? (free responses)
2. What is the preference between online and in-person courses and why
   - If you could change one thing about the way your online classes are designed, what would you change? Why? (free responses)
   - If you could change one thing about the way your in-person classes are designed, what would you change? Why? (free responses)
3. If students could change one thing about online courses VS one thing about in-person courses, what would it be?
   - Do you prefer online or in person courses? (fixed answers)
   - Why do you prefer online or in-person courses? (free responses)

2.1 Preprocessed corpus, TF-IDF and K-means

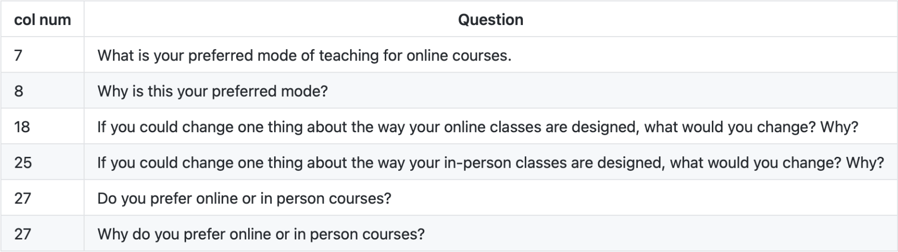
We used two different methods for the three questions, one is relatively simple and the other is more complex. We used the three questions as examples of how they could be conducted. Although the first method is simpler, this does not mean it was less useful. We can see from the result that this method can give us a quicker view of the dataset.

For questions a and b, we performed statistics of word frequency on each subset of data. We wanted to show how simple functions could provide great insight into the dataset, which would be helpful for those without an NLP background but who want to draw conclusions from a large dataset quickly. By showing the method is easy but useful, we also wanted to encourage people to ask open-end questions in a survey, which would contribute to the amount of data gathered.

For question c, we performed TF-IDF and K-means on the preprocessed data. The combination of TF-IDF (Qaiser, 2018) and k-means (Likas, 2003), is a popular and beginner-friendly clustering method. TF-IDF is used to evaluate the importance of a word in a corpus, and K-means is a clustering algorithm (see Appendix A for the details of the two algorithms). Clustering is an unsupervised machine learning method that classifies data based on the vector representations of the data.

We firstly did data reduction by instance selection. We selected responses from undergraduates, which became a subset of the original data set. One reason for doing this is because the same stage of education has more similarities and students are more likely to use similar words (for example, the word ”professor” is more common in the university than in high school). By simplifying the structure of the data set, the NLP might be less noisy. The selected data set was also used by Nguyen et al., 2021 in their research. There is overlap in our work, and thus we can compare some of the results. To make the comparison fair, we used the data set used by this paper directly to make sure we are using the same data.

For different research questions, since we would use different columns of the dataset, we did data reduction by features selection. For research question a, the data set was classified by the answers in column 7 (see Fig. 1 for column-question reference) of the CSV file (see https://github.com/JunruL/ROP299/blob/master/coded_data.csv) and then the responses in column 8 of each class will be analyzed. For research question b, column 18 was used for the analysis of ”online change” and column 25 was used for ”in-person change”. Since there are lots of blank answers in column 25, we did extra data reduction for column 25 to filter out these answers. For research question c, we classified the data set into online and in-person ones based on column 26 and did the analysis on the responses in column 27.



col num	Question
7	What is your preferred mode of teaching for online courses.
8	Why is this your preferred mode?
18	If you could change one thing about the way your online classes are designed, what would you change? Why?
25	If you could change one thing about the way your in-person classes are designed, what would you change? Why?
27	Do you prefer online or in person courses?
27	Why do you prefer online or in person courses?

Fig. 1 Column-Question reference

2.2 LDA

LDA is a probabilistic topic model that is used to learn a set of latent topics for a corpus, and predict the probabilities of each word in each document belonging to each topic (see Appendix B for an explanation on how LDA works). (Blei et al., 2003)

LDA infers documents to topic and topic to word distributions from the co-occurrence of words in the corpus. It is selected for topic modeling on free responses of the original unlabeled Nguyen et al., 2021 dataset because it performs especially well with large sets of data. Unlike the TFIDF and K-means clustering approach mentioned above, if the input corpus is small, LDA-learnt topics will be noisy due to insufficient data on word co-occurrence. (Nguyen, 2015) It is also fast and efficient in computation when working with large sets of textual data. The results LDA generates provide more details such as representative texts for each topic which is helpful when conducting qualitative analysis to the generated result.

LDA was applied to the survey questions listed above to generate results. All of the free responses were pre-processed in the same steps as for the GloVe word embedding: words were all turned to lowercase with punctuation and stop words removed, then each is stemmed. Stemming means removing the ending of a word so words with the same meaning but different derivations all share the same form. (e.x. easy, easily, easiness are all stemmed to easi) Empty rows were also filtered and each response was tokenized and stored as an element of a list.

This dataset combines multiple-choice questions with free responses. Corresponding to column 8, question 2 of the survey, “What is your preferred mode of teaching for online courses', column 9 asks ”Why is this your preferred mode?” Free responses of column 9 are first categorized based on answers of column 8, then for each category of free response data, LDA was applied to generate one topic of key terms. These key terms will be qualitatively analyzed to extract a hypothesized summary of why students prefer each mode. Similarly, for question 10, 'Do you prefer online or in-person courses', the following column asks 'why do you prefer online or in-person courses?' Free responses were categorized based on choice selected for in-person or online and LDA was used to generate their key terms. For columns 18 and 25 that ask students what they'd like to change for online or in-person courses, responses were classified into 5 topics and analyzed.

Functions were coded to extract information from the LDA model into data frames and outputted as CSV files, such as the most probable topic each word belongs to, distribution of topics over the whole corpus, and the most representative texts for each topic (the top 10 texts with the highest probability belonging to each topic). Visualizations are generating using pyLDAvis from scikit.learn package, an open-source python library that helps in analyzing and creating interactive visualization of the clusters generated from the corpus.

2.3 GloVe and RIPA

We chose GloVe over other embedding models for two reasons. The first, being that GloVe consistently outperforms other models such as word2vec. (Pennington et al., 2014) We more decisively chose GloVe because of our first proposal plan, in which the WEAT bias test (Caliskan et al., 2017) would be used instead of the RIPA bias test. As WEAT relies and builds upon the GloVe model of embeddings, we began our work towards this goal. When our research plan changed, we decided to keep the research and work put into creating our Glove embeddings (see Appendix C for more information about the GloVe word embedding model). We downloaded the GloVe package (https://github.com/stanfordnlp/GloVe) to upload our pre-processed data into GloVe.

Past research has found evidence for a real-world bias concerning views on online learning. Traditional learning has been face-to face, but online learning rapidly advances with the advent of technology. However, many people are skeptical as to how the two compare, especially as online learning requires less hands-on experience (Kizilcec et al., 2019). With stigmas as being less qualified and employable than traditional learning, online learning has previously been less favored until the recent pandemic forced a large majority of the population online. (Kizilcec et al., 2019) Past research also shows that moving online for learning can cause students to become less successful. (Glazier et al., 2019) As our data collected comparisons between online and in-person learning, we explored whether the collected data had fragments of this bias. It should also be noted that previous research has been focused on a complete online or in-person format, as opposed to a hybrid style which could have been included in some of the data.

Among the many word embedding bias tests published, we chose RIPA for our research. We had originally planned to use the more extensively published and widely used WEAT, but this plan changed as we conducted further research into RIPA, which claimed many misgivings to WEAT. WEAT compares distances through cosine similarity between a target vector and attribute word in a GloVe embedding model building on similarities to the implicit association test. (Caliskan et al., 2017) Their research implies that AI can perpetuate cultural stereotypes. The RIPA paper, however, concludes WEAT to have flaws to systematically overestimate bias. As it requires two sets of attribute words to occur with equal frequency in the training corpus, the test can be easily manipulated to show false bias. The notion of statistical significance of WEAT is also disingenuous as word embeddings alone are not enough to give statistical significance to association. (Ethayarajh et al., 2019) We only ran the results of undergraduate students, column 8 which had been used to solve the survey question a, through RIPA as a test for bias (see Appendix D for more information about the RIPA bias test).

All code was uploaded onto the github https://github.com/JunruL/ROP299.

3 Analysis

3.1 Preprocessed corpus, TF-IDF and K-means

NLP can turn qualitative data into quantitative data. Based on word frequency, we can have an overview of the whole dataset and know what students most cared about. With the clustering results, we can easily summarize some potential topics of the free responses. To understand the dataset, we don't have to read the whole dataset. Instead, we can just read a subset of each cluster and then combine the information.

3.1.1 What is the preference among different modes and why

We found that over half of the students chose live classes (ie: Zoom, google meet, etc.) as their preferred mode. Among these students, most talked about interaction. Besides, the opportunity to ask questions in live classes was also valued. The second most popular model was recorded Lectures/videos, which were chosen by more than one third of the students. The primary factor for it was the time flexibility of recorded lectures. Less than 10% of students chose Uploaded or emailed Materials or Discussion forums/chats. As more than 90% of students preferred live classes (ie: Zoom, google meet, etc.) or recorded lectures/videos, we may infer that recorded live classes can meet most students' needs. (see Table 1)

3.1.2 What is the preference between online and in-person courses and why

Based on the responses to column 26, which asked whether students preferred online or in-person courses, most students (86.02%) preferred in-person courses and less than 14% (13.98%) preferred online courses.

Table 1 Word Frequency of Responses for preference among different modes

	Live[1]		Recorded[2]		Materials[3]		Discussion[4]	
Rank	Word	F[5]	Word	F[5]	Word	F[5]	Word	F[5]
1	interact	765	time	583	time	93	discuss	37
2	question	655	lectur	477	work	71	question	33
3	ask	596	paus	275	class	54	get	29
4	lectur	532	watch	263	lectur	51	time	29
5	class	493	class	230	materi	51	ask	25
6	engag	393	pace	217	pace	51	interact	23
7	live	345	go	207	like	35	lectur	19
8	like	328	live	206	live	33	class	18
9	feel	326	record	202	easier	29	learn	18
10	get	312	back	176	allow	25	abl	17
11	time	279	understand	176	learn	25	student	17
12	it'	217	work	172	prefer	24	like	16
13	record	202	take	164	onlin	23	live	16
14	actual	200	note	163	get	22	better	15
15	learn	198	also	159	inform	22	engag	15
16	student	190	like	149	also	21	feel	15
17	discuss	185	want	148	go	21	allow	14
18	easier	183	learn	141	find	20	peopl	14
19	real	168	abl	138	note	20	understand	14
20	allow	163	video	135	studi	20	also	12
21	also	163	allow	129	zoom	20	way	12
22	abl	158	need	122	feel	19	zoom	12
23	make	155	whenev	116	video	19	chat	11
24	professor	143	easier	113	abl	18	make	11
25	teacher	142	materi	113	access	18	easier	10
26	face	129	make	98	it'	18	help	10
27	motiv	125	zoom	91	take	18	less	10
28	work	119	someth	88	want	18	materi	10
29	understand	116	get	86	need	17	particip	10
30	give	108	it'	83	read	17	record	10

[1] Students preferring Live classes (ie: Zoom, google meet etc.): 2622/4807 = 54.55%

[2] Students preferring Recorded Lectures/Videos: 1726/4807 = 35.91%

[3] Students preferring Uploaded or emailed Materials: 298/4807 = 6.20%

[4] Students preferring Discussion forums/chats: 161/4807 = 3.35%

[5] F stands for Frequency

In the responses for in-person choice, ”online” appeared as many times as ”person”, which means the students mentioned the disadvantages of online classes when they gave the reasons for the preference for in-person classes. We may infer that some students preferred in-person courses because online courses were not a good experience. It’s interesting that ”feel” is the 3rd most frequent word, which means students cared about their subjective feelings and experience and in-person courses made them feel better. Besides, ”learn”, ”motivation” and ”engagement” also appeared many times and these were the dominant advantages of in-person courses. ”Distracted” appeared in the top 30 list. Since it’s a negative word, we can infer that those who preferred in-person courses thought it easy to be distracted during online courses.

As for those who preferred online courses, they valued the convenience of time and space, since ”time” and ”home” appeared many times in their responses. Compared with the responses of in-person preference, ”feel” was not a high-frequency word as in the responses for in-person courses, which means some students preferred online courses because of more objective reasons. However, some students did have a better feeling in online courses. They mentioned they would feel more comfortable. (see Table 2)

3.1.3 If students could change one thing about online courses VS one thing about in-person courses, what would it be

For the change about online courses, making online courses live was a distinct topic among students’ expectations for improvements of online courses. Based on the clustering results, about 10% of students were unsatisfied with the interaction and engagement of online courses. About 6.3% of students mentioned that they hoped teachers could provide recording courses. About 5% of students answering the question did not know what to change. (see Table 3)

As for changes about in-person courses, we notice that there were lots of blank answers to these questions. There were 4780 responses for ”change one thing about online courses”, while there were only 3263 for ”change one thing about in-person courses’ ’. The reason why the first question has a much higher response rate is unknown. Among the 3263 students who provided answers, about 24% of students answering the question wanted a smaller class or group.

Table 2 Word Frequency of Responses for Preference in Online and In-person Courses

	In-person Courses[1]		Online Courses[2]	
Rank	Word	Frequency	Word	Frequency
1	person	1218	time	302
2	onlin	1209	onlin	279
3	feel	1152	class	185
4	class	1115	work	151
5	cours	1081	lectur	142
6	learn	1054	cours	129
7	motiv	961	like	116
8	engag	956	feel	115
9	easier	760	prefer	101
10	like	744	learn	100
11	inperson	715	get	92
12	get	705	person	90
13	lectur	665	home	84
14	interact	630	pace	79
15	also	577	less	77
16	work	554	studi	77
17	much	498	also	73
18	peopl	484	better	71
19	better	468	much	71
20	ask	426	go	68
21	question	420	take	63
22	student	409	inperson	58
23	make	403	easier	57
24	it'	401	comfort	56
25	less	394	abl	54
26	prefer	382	dont	53
27	distract	375	uni	53
28	friend	369	would	53
29	see	364	day	52
30	studi	355	don't	52

[1] Students preferring In-person Courses: 4135 / 4807 = 86.02%

[2] Students preferring Online Courses672 / 4807 = 13.98%

About 11% of students wanted more interaction (with teachers and other students), engagement, and the opportunity to ask questions. In the clustering results, about 5% of students wanted more discussion. In addition, from the results, about 8.5% of 3263 students answering the question thought there was nothing to change. (see Table 4)

From the above analysis, more interaction and engagement were expected in both online and in-person courses. However, for in-person courses, students

Table 3 Clustering Statistics of Responses to Changing One Thing about Online Courses

cluster	topic	percentage
0	interaction	3.93%
1	groups	4.44%
2	live lectures	8.05%
3	nothing	1.57%
4	record	6.36%
5	don't know	3.33%
6	live	4.46%
7	???[1]	14.67%
8	engagement	5.06%
9	???[1]	47.99%

Note: For a more detailed clustering result with several example responses in each cluster, see Appendix E

[1] Could not summarize a topic for the cluster since the data in this cluster is heterogeneous.

Table 4 Clustering Statistics of Responses to Changing One Thing about In-person Courses

cluster	topic	percentage
0	???[1]	36.32%
1	class size and class length	14.22%
2	engagement, ask questions	7.85%
3	more discussion	5.00%
4	smaller groups	7.07%
5	smaller class size	2.67%
6	nothing	5.76%
7	???[1]	15.84%
8	no change	2.76%
9	interaction	3.52%

Note: For a more detailed clustering result with several example responses in each cluster, see Appendix F

[1] Could not summarize a topic for the cluster since the data in this cluster is heterogeneous.

not only emphasized the interaction with teachers but also with their classmates. The response rate of the "in-person" questions was only 68.6%. Among the students who answered the questions, the "in-person" question was more likely to receive the "no change" answers (about 8.5% vs 5%). Thus, we may infer that students were more satisfied with in-person courses.

3.2 LDA

3.2.1 What is the preference among different modes and why

There are four choices for column 8 ‘What is your preferred mode for teaching online courses?’: ‘Live classes (ie: Zoom, google meet, etc.)’, ‘Discussion forums / Chats’, ‘Recorded Lectures/Videos’ and ‘Uploaded or emailed Materials’.

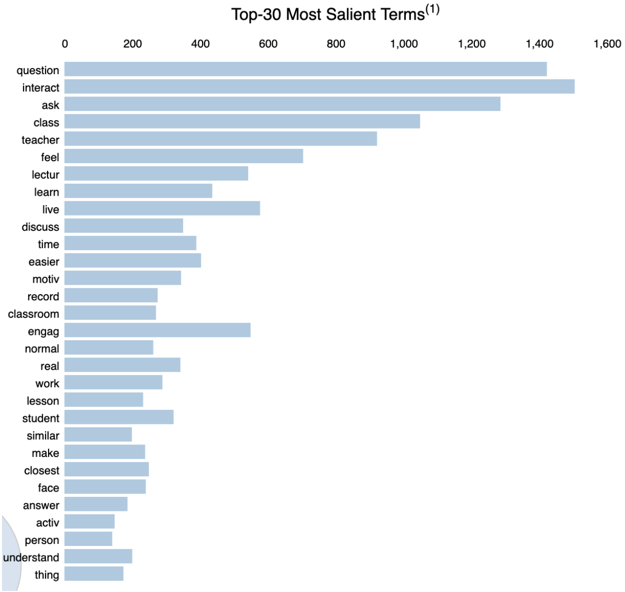


Fig. 2 Key terms and their frequencies generated for ‘Live classes (ie: Zoom, google meet etc.)’ responses

54% (5020/9351) of students chose ‘Live classes (ie: Zoom, google meet etc.)’ (see Fig. 2). Some of the key terms generated from their responses are ‘interact’ (0.048), ‘ask’(0.040), ‘question’(0.044), ‘motiv’, ‘engag’(0.018), ‘feel’(0.022), and ‘live’(0.017). Numbers in brackets indicate the weighting of the term for the topic. The words ‘interact’, ‘feel’, ‘live’ imply students seem to appreciate how synchronous online courses offer opportunities for interaction with the instructor and other students. ‘Engag’ as a highly frequent key term suggests students find this approach engaging. Students may feel more

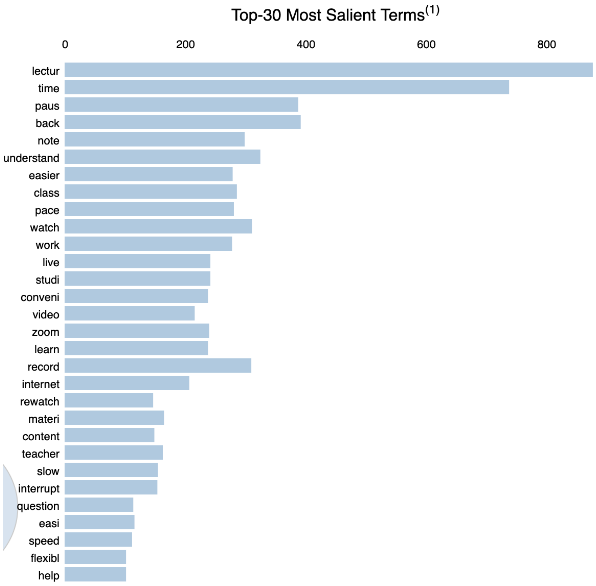


Fig. 3 Key terms and their frequencies for 'Recorded Lectures/Videos' responses

inclined to ask questions, possibly because the instructor or other students can respond to it immediately with no waiting time. The word 'motiv' implies students are motivated to keep up with the pace of the course, I (author H.S.) hypothesize this may be from knowing other students are also attending weekly synchronous live meetings.

The second most popular choice (30%: 2781/9531) is 'Recorded Lectures / Videos' (see Fig. 3) . Key terms and most frequently appearing words include 'time', 'paus', "back', 'note', 'understand', 'pace', 'flexib' and 'conveni'. These terms possibly imply students feel they have more control over their time and find recorded materials more convenient for them to learn at their own pace. Recorded materials can be paused for taking notes and played back when confusion arises which is convenient. This approach requires more independent learning which may work better for some students.

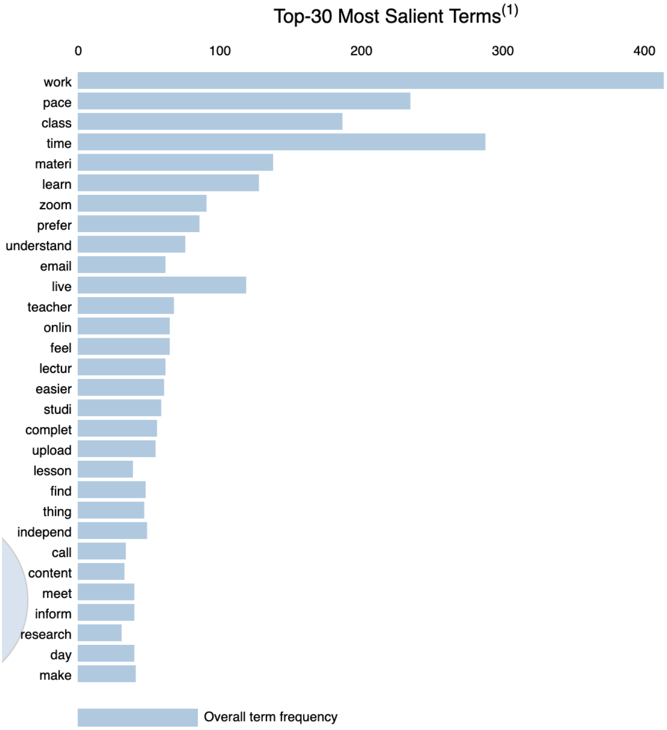


Fig. 4 Key terms and their frequencies for 'Uploaded or emailed Materials' responses

12% (1187/9531) of students choose 'Uploaded or emailed Materials'(see Fig. 4). The key terms are very similar to ones for the ones for 'Recorded materials', many actually overlap (as seen in the figures: 'time', 'lectur', 'work', 'class', 'pace', etc.) Key terms 'independ' suggests students may prefer to work at their own pace and manage their own time. Around 4% (363/9531) of students choose 'Discussion forums/chats' (see Fig. 5), some key terms are 'interact', 'discuss', 'question', 'time', 'peopl' and 'peer'. I (author H. S.) propose these students find forums a great way to associate with peers and tackle a question collaboratively.

I (author H.S.) think some students prefer independent studying over attending classes in a group setting. But for another group of students, it could

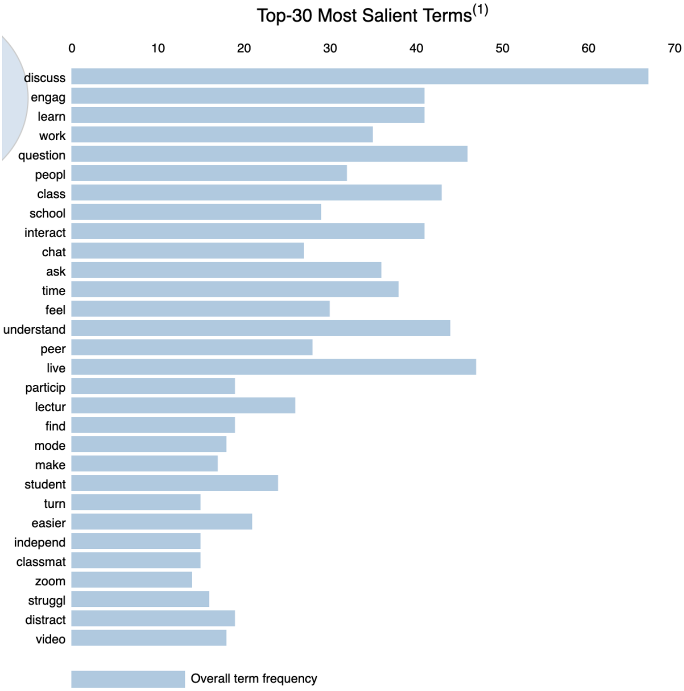


Fig. 5 Key terms and their frequencies for 'Discussion forums/chats' responses

be the opposite. Online learning has definitely made socialization and bonding between peers more difficult which creates the needs for ways students can interact online. A possible conclusion is more than half of people selected synchronous meetings because a fixed schedule creates a sense of order and routine, and more importantly, a sense of community.

3.2.2 What is the preference between online and in-person courses and why

Around 85% (7935/9351) of students answered question '10. Do you prefer online or in-person courses?'chose in-person (see Fig. 6) and 15% (1416/9531) chose online (see Fig. 7). At first, the number of topics is set to 1 for both categories of response to generate their overall key terms. The key terms turned out to be quite general. For in-person responses, key terms include 'pace', 'question', 'teacher', 'live', 'easier', 'engag', 'interact', 'discuss', 'motiv' and 'real'. They seem to imply the possible benefits of going to classes offline: people can meet each other face to face, interact more with the instructor and ask questions, and classes are live and engaging.

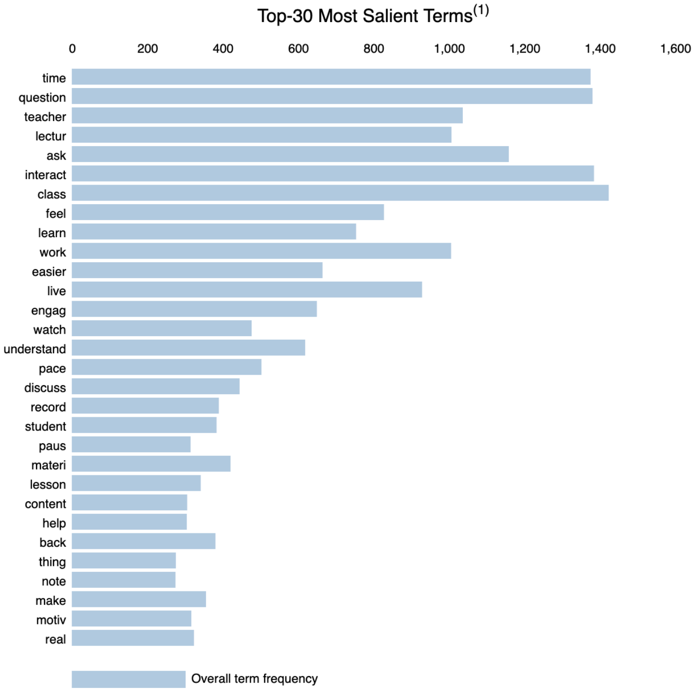


Fig. 6 Key terms and their frequencies for prefer in person responses

But the representative texts are interesting. Out of the 10 representative texts generated by the LDA model, half of them referred to online learning and commented why they don't enjoy it. A few are even negative, e.g. 'Our online exam workshops didn't give us any answers to our questions, and we're pretty much pointless', 'Zoom is somehow mentally draining', 'online classes are a bit more chaotic' (see Appendix G).

After the number of topics for in-person responses is set to 5 to experiment, I(author H.S.) see that there exists one topic that is composed of more

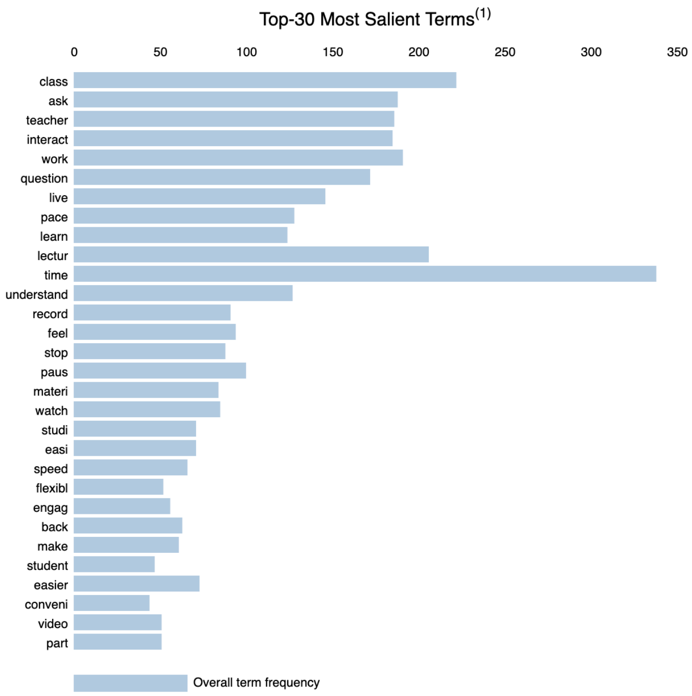


Fig. 7 Key terms and their frequencies for prefer online responses

negative terms such as 'difficult', 'hard', and 'procrastin', 'prefer', 'forc', 'interrupt', and 'distract'. 'Onlin', 'zoom' and 'prefer' are also included.(see Fig. 8) I(H.S.) think this could be LDA picking up responses that are negative about online learning and responses that are directly comparing learning online v.s. in person. A possible conclusion that I (author H.S.) make from these terms is that students find online learning more difficult and more distracted than in person.

For online responses, key terms include 'easi', 'pace', 'time', 'record', 'stop', 'speed', 'flexibl', and 'conveni'. They may imply this group of students find the method of delivering online materials such as recordings convenient since it allows pauses and replays. I make the conjecture that students who prefer online learning possibly enjoy independent studying, deciding their own pace of learning, and appreciate more freedom with time.

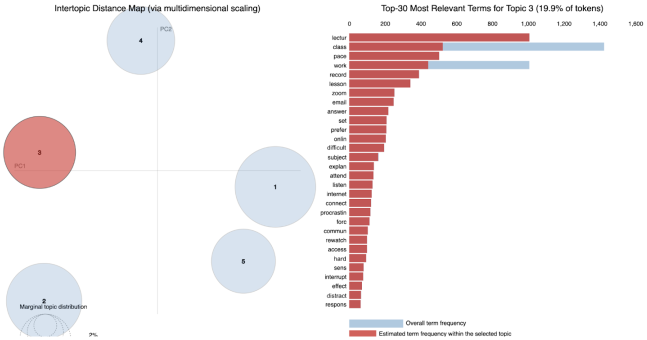


Fig. 8 visualization of key terms and their frequencies for Topic 3 of prefer in person responses generated using pyLDAvis

3.2.3 If students could change one thing about online courses VS one thing about in-person courses, what would it be

For in-person courses, 67.5% (7131/10562) of students answered the question without leaving it blank, compared to 98.4% (10391/10562) of students leaving feedback for online courses. It seems that students have more opinions about online courses and are more satisfied with in-person ones.

5 topics are generated separately for each column of responses. I use 5 because by experimenting with different numbers, 5 gives a visual representation of topics that are not small or overlapping, meaning topics are sufficiently distinct; too many topics will give better complexity scores but are harder for analysis since smaller topics can be trivial.

Refer to Appendix H for the table of generated topics and representative texts of in-person responses

For in-person courses, Topic 0 has representative texts on students' suggestions of where to spend more or less class time, such as leaving more time for answering questions, shorter or longer school/lecture hours. (E.x. 'More face to face time for questions', 'Less time wasting - waiting for people to arrive etc', 'Change the amount of lessons so we have more time for revision')

I (author H.S.) think this sort of information is helpful for instructors constructing lesson plans. Using this set of rough results, teachers can even ask students to fill out surveys on how they would like to divide up lecture hours to further enhance students' experience. Topic 1's texts focus more on interactions between students and teachers, which are also valuable feedback. Topic 2 has representative texts that convey students' message of hoping to incorporate more online media activities and make use of more online platforms and services. Putting textbooks materials and recordings online seems to be encouraged by students. The term 'record' I think is from students wanting lectures to be recorded for later revision. Topic 4's representative texts are mainly about smaller class sizes and students wanting more interactions and seminars.

Refer to Appendix I for the table of generated topics and representative texts of online responses

For online courses, representative texts of Topic 0 are a mix of different themes, each document is about different themes. It is worth noting 2 texts are written in foreign languages and LDA is picking up on that, putting them in the same group. Topic 1 has representative texts on students mostly wanting more live and engaging lectures. Topic 2 has over half of the texts on students finding the current workload too much and proposing less work. Topic 3's representative responses are almost all about wanting more live classes and referring to benefits such as live classes are more engaging, have more interactions, etc. Topic 4 has texts about communication relevant suggestions. Some students find professors hard to reach and don't have access to immediate help.

3.3 Bias Analysis with RIPA

Through bias testing, we wanted to see if bias between online and in-person learning was introduced or magnified by the code through the word embedding algorithm. For RIPA we chose the terms ‘person’ for x, and ‘onlin’ for y, so we would be working on the difference vector between these terms in the vector-space model (see Equation J1 in Appendix J for the difference vector) concerning the third term, w. In an unbiased space, w should have no association between either x and y, demonstrated through taking the inner product (dot-product) between the vector w and the difference vector between x and y. (see Equation J2 in Appendix J for the RIPA test for association within word embeddings) For w we chose the terms ‘easi’,’interact’, ‘engag’, ‘work’, ‘easier’, ‘understand’, ‘help’, ‘explain’, and ‘hard’, all of which should neither be associated with online nor in-person learning. These terms are neutral in the quality that they do not objectively describe either the x or y term.

The terms were chosen in consideration of the vocabulary of the space, column 8 of the free-response answers. As a result, many words did not have a co-occurrence, in which the denominator or numerator would equate to 0 and the equation would be null or had to be stemmed. Stemming words did not affect the equation as one of the benefits of RIPA is that similar words to describe the same context should show the same result. (Ethayarajh et al., 2019)

The same terms were then put through the RIPA equation for corpus bias (see Equation J3 and Equation J4 in Appendix J for the RIPA equation for corpus bias).

Refer to Appendix J for more logistics about the math behind RIPA.

Using the same w term, the absolute difference was taken between the bias from the word embeddings and the corpus. This makes the results of

RIPA interpretable, as RIPA is intended for embedding models that factorize a matrix containing a co-occurrence statistic. (Ethayarajh et al., 2019) A standard of 0.001 was set from the Ethayarajh et al. paper, and associations less than this standard would be considered statistically significant. The results are shown in Table 5.

Table 5 Bias Test Results

w term	RIPA	embed	stat
Easi	-0.2512	0.15198	0.40315883
interact	0.06604	-0.0037	0.06978304
engag	0.19957	0.0733	0.12626851
work	-0.2953	-0.2965	0.00125913
easier	0.07685	0.16428	0.08770115
understand	0.00833	-0.6984	0.70677148
help	-0.7202	-0.0179	0.70228041
explain	0.41962	-0.4155	0.8351725
hard	-0.2512	-0.5394	0.28827091

Note: This table compares the absolute difference (under the 'stat' column) between bias in the word embeddings ('RIPA' column') and the corpus ('embed') for the same w term

In my research I (L.C) found none of the w associations to have an absolute difference of less than 0.001, and therefore to be statistically significant. This suggests that no significant bias concerning online bias from this dataset was introduced into the GloVe embedding model.

There are many possible reasons for this. Firstly, our research was limited to a very small pool of data compared to previous applications of RIPA. (Ethayarajh et al., 2019) As such our vocabulary size was greatly diminished which decreased the number of words we could have measured association for under the w variable. The survey results were also very subjective, asking for students' own experiences and perceptions. There is also a slight chance results may have been affected by the informal language used, as slang and incomplete sentences change sentence structure, and therefore context differently. This could have impacted the vector space embeddings, co-occurrences, as well as vocabulary. These conditions might affect data as we know GloVe to work optimally on large corpora, with small and asymmetric context windows, with syntactic info drawn from intermediate context and word order; these results may not demonstrate GloVe at its best function. However, this point should be generalized regardless as AI will without a doubt encounter a variety of writing styles to be analyzed.

While this data could be very hopeful towards other applications of GloVe, it is still important to consider past research which finds that bias affects different algorithms differently. (Lauscher & Glavaš, 2019) As we only used GloVe, these results do not apply to all models, and research similar to this should be conducted across different models, such as word2vec (Mikolov et al., 2013) to help developers better understand and choose their algorithms. It is theorized that pre-processing affects the bias intake, which is a point that would also need to be further researched. (Lauscher & Glavaš, 2019)

4 Discussion

4.1 Word Frequency

Word frequency was used in Section 3.1.1 to analyze Question 1 and in Section 3.1.2 to analyze Question 2.

4.1.1 Comparison with Nguyen et al.'s Analysis

In Nguyen et al.'s (2021) paper, they also analyzed students' preference between online and in-person courses. They counted the number of keywords such as social interaction, engagement, and motivation. From their analysis of the responses for preference in in-person courses, social interaction appeared 693 times, engagement appeared 639 times, and motivation appeared 440 times. In our preprocessed corpus of reasons for preference of in-person courses (see Table 2), “interact” appeared 630 times, “engag” appeared 956 times, and “motiv” appeared 961 times. (See the comparison between results of the two papers in Table 6)

Table 6 Comparison of Word Frequency between Nguyen et al.’s Research and Our Research

Nguyen et al.’s Research		Our Research	
Word	Frequency	Word	Frequency
social interaction	693	interact	630
engagement	639	engag	956
motivation	440	motiv	961

When students talked about interaction, engagement, or motivation, they might use all kinds of words, including nouns, verbs, and adjectives. For example, responses “Having the human interactions. In online class you miss out on the chats and banter that make class a community”, “I prefer in-person courses because (in my case) it provides a more interactive learning” and “Because I enjoy being in the physical learning environment and I really miss interacting with my peers.” are all about interactions. But if we only count the mention of “interaction” in these three responses, it would be only 1 instead of 3. This means when we restrict word frequency to the occurrence of nouns or verbs, we are at risk of losing information. Besides, we can see that “motiv” appeared 961 times, more than twice of “ motivation”, which appeared 440 times. So if we use the preprocessed data for analysis, we might be closer to the real data and the analysis will be more accurate and reliable. However, ”social interaction” was counted 693 times in Nguyen et al.’s research, while ”interact” was counted only 630 times in our research. The reason might be human selections of ”social interaction” in Nguyen et al.’s research, which means that judging the appearance of ”social interaction” is not just based on the appearance of "interact". This implies adding human work can sometimes make the analysis more accurate.

In addition to the topic count accuracy, the preprocessed corpus also makes it easier to have an overview of the potential topics (see Table 1 and Table 2).

4.1.2 Limitations

Although this method is easy to conduct, it can only give us an overview of the dataset. We could only draw a limited conclusion from the result of word frequency lists. If we want to have a better understanding of the dataset, word frequency lists will not be enough. For example, with the word frequency of "interact", we are not sure whether most students are talking about the interaction with teachers or classmates. Even if we also know the word frequency of "teacher" and "classmate", we are still unable to answer the question, since the words are independent in the word frequency statistics.

4.2 TF-IDF and K-means clustering

The combination of TF-IDF and K-means was used in Section 3.1.3 to analyze Question 3.

4.2.1 Limitations of K-means

K-means has trouble clustering data where clusters are of varying sizes and densities. In both online and in-person clustering results, over 50% of responses are in two clusters. In online clusters, 14.67% of responses are in cluster 7 and 47.99% of responses are in cluster 9 (see Table 3). In in-person clusters, 36.32% of responses are in cluster 0 and 15.84% of responses are in cluster 7 (see Table 4). Those clusters themselves seem to be meaningless since they're hard to summarize. For example, in cluster 9 of online clusters, there are responses such as "It would be better if professor can have top students in the live asking important questions, we would learn more about the subjects.", "The way lessons are performed because it's not sufficient", and "I would love if all the teachers used live call/ video techniques rather then just emailing the materials, considering most of my teachers don't use live, there some subjects I find hard and can not reach the teacher for further learning and understanding.". The first response is about having top students ask questions; the second one is about changing the way lessons are performed; the third one is about having live classes. These three responses are talking about different topics, but they are put in one cluster by the algorithm. Therefore, it's hard to summarize such clusters since it contains responses with several topics.

Although some responses may be long and thus contain lots of information, each of them can only be in one cluster. For example, in online clusters, the response "I would like to have more interaction with my professors and I wish that our lectures were not just PDFs uploaded online. Live Q&As would be great" is in cluster 2 that has the topic of "live lectures". However, these responses also talked about interaction, so it can also be put into cluster 0 that has the topic of interaction. In the clustering results, there are 3.93% of responses in cluster 0 (see Table 3). With the limitation of one response in one cluster, this number could not be accurate anymore.

4.2.2 Limitations of the Combination of TF-IDF and K-means

Since TF-IDF and k-means make no use of semantic similarities between words, there are two clusters with the same topic. For example, when clustering responses to changing one thing about online courses, "I wouldn't change anything" and "Nothing" should be put into one cluster, but in fact, they are in different clusters in the result. Since "wouldn't change anything" appears many times, the responses with similar structures form a cluster. And "nothing" also appears many times and forms another cluster. This is because we use TF-IDF, which ignores the semantic meaning in language.

Besides, some responses are in the wrong clusters. For example, when clustering responses to changing one thing about online courses, "I wouldn't change anything per say, But maybe having more workshops would be good" was put in cluster 8 since it contains "I wouldn't change anything". But if we do the clustering manually, we would not consider it the same as "I wouldn't change anything".

4.3 LDA

Our research question is exploring the limits and potentials of NLP techniques and specifically its application to survey data such as Stein et al. By applying LDA and drawing hypothesized conclusions from the results, we developed an approach that quickly summarizes free responses. This approach, which is also our research question applied in action, will be analyzed by its algorithmic limitations. Its results on the data will be compared to manual coding results in Stein et al paper. At last, we will make connections with previously published papers on similar research questions and propose what we have done well on or what can be explored further.

4.3.1 Comparison with Nguyen et al.'s Analysis

Compared to the hypothesized qualitative analysis conducted from the outcomes of LDA, the conclusions drawn in Stein et al paper are much more detailed and structured. Both conclude that students prefer synchronous classes. The LDA analysis hypothesizes students feel motivated and engaged from the key terms which line up with conclusions from the paper. The paper additionally mentions ‘students whose synchronous classes include active-learning techniques (which are inherently more social) report significantly higher levels of engagement, motivation, enjoyment, and satisfaction with instruction’ which LDA did not address since the key terms are generated using separate categories of responses hence comparative conclusions can not be made. We can only hypothesize one approach ‘engag’ (key term from LDA) students but can not conclude one approach is possibly more engaging than another.

In the paper, 86.1% of students prefer in-person courses and 13.9% prefer online, which is slightly different from the 85% and 15% statistics that are generated from LDA. ‘Students who prefer in-person courses most often mention the importance of social interaction (693 mentions), engagement (639 mentions), and motivation (440 mentions)’, similar to NLP techniques, the researchers also identified key themes by keeping track of frequencies. However, topic modeling algorithms work by reverse engineering, using frequencies and co-occurrence of key terms to generate themes whereas researchers can directly identify topics.

‘Students’ suggestions for improvements in online learning focus primarily on increasing interaction and engagement, with 845 mentions of live classes, 685 mentions of interaction, 126 calls for increased participation and calls for changes related to these topics such as, “Smaller teaching groups for live sessions so that everyone is encouraged to talk as some people don’t say anything and don’t participate in group work’, these themes are also identified from LDA topics (see Appendix I).

The paper concludes one detail ‘anxiety was mentioned 12 times in the much larger group that prefers in-person learning’ which could not be captured by LDA because the frequency of such words is too low. However, this is vitally important information because researchers/instructors need to know even one case exists. This is where manual coding outperforms the NLP approach.

4.3.2 Limitations

Though LDA can perform topic modeling and generate key terms for free responses, I have noticed lots of the key terms for different responses are the repeated same ones so they lose significance in terms of analysis. For example, terms such as 'time', 'class', 'questions', 'interact', 'motiv' are included in topics of every category of free response data, as can be seen in figures above displaying key terms for different categories of responses. This indicates key terms on their own are not enough information. Representative texts do provide more perspective but sometimes there still exists ambiguity between each topic. One example is the representative texts generated shown in Appendix H and Appendix I: each topic's representative texts may not necessarily be about the same theme and topics are not distinct enough to tell each other apart.

Due to the nature of LDA, each topic distribution contains every word but assigns a different probability to each of the words. (Seyd et al, 2017) A word is not fixed to one specific topic hence the key terms for topics can overlap. As mentioned, the frequently appearing terms are key terms for every topic and this decreases the difference between topics. These terms may have different weightings for each topic, but they are very commonly used such that their weights are high for each topic (which is exactly why they always become the key terms). How to prevent topics from having overlapping key terms is something that can be considered for further analysis. Possible solutions that can be explored include trying clustering instead of topic modeling algorithms or removing these terms as stop words.

Two scores are used to measure how good an LDA model is: complexity score and coherence score. Coherence scores measure the degree of semantic similarity between words within a topic. A low coherence score results in too few or very broad topics, whereas a high coherence score results in uninterpretable topics or topics that ideally should have been merged. (Seyd et al, 2017) I notice the coherence scores for responses are not very high and most don't even reach 0.5 (see Table 7), indicating the model is only moderately good or even bad. A lot more information from free responses is not included in generated outcomes of key terms. This can be improved by adjusting parameters of the LDA model such as alpha, chunk size, and the number of topics. More topics will surely include more information from a smaller cluster of responses but as mentioned before, sometimes the distinction between topics is not enough in application to determine if some topics are too small and meaningless.

Table 7 Coherence Scores for Each Category of Responses

Question	Category of Responses	Coherence Score
1	Live classes (ie: Zoom, google meet etc.)	0.20818362424143583
1	Discussion forums / Chats	0.23444281984893398
1	Recorded Lectures/Videos	0.30863480313159475
1	Uploaded or emailed Materials	0.36944281817061825
2	Prefer in-person courses	0.23931662689453198
2	Prefer online courses	0.2106252332198232
3	Change one thing about online courses	0.3346182003512511
3	Change one thing about in person course	0.31277377929541506

Words that appear frequently together should be considered as phrases instead of separate words. Especially in the case of negation, such as 'not easi', 'dont enjoy', etc. Key terms which are phrases can provide more semantic meaning. If a word appears many times as a key term, there is a difference if it appears solely by itself many times or it is attached to another word. LDA does allow n-grams, K. Daniel (2017) even concludes 'LDA/n-gram model did show a significant improvement over the baseline once it had enough data that it was able to generalize well.' In our example, the n-grams model is not used but this is a direction that can be further explored to enhance LDA models' performance in such applications.

4.3.3 Previous Research

Pietsch et al.(2018) applied LDA to analyze open-response survey data provided by a market research company. This paper uses topic coherence and document classification to evaluate generated models quantitatively and qualitatively to appraise whether topic models can replace human coding (Pietsch et al., 2018) which is different but related to our research question. This research has several places where our research can learn from: Firstly, their LDA model is used with GloVe word embedding trained on an enormous external corpus which generates improved results Secondly, for preprocessing, the length of documents is restricted for consistency in using specifically shorter length documents for experimentation

We are interested in how they qualitatively evaluate the quality of results: 'The qualitative evaluation, on the other hand, requires the involvement of market research experts to judge extracted topics and compare the outputs of different short text topic models to one another.' (Pietsch et al., 2018) Experts look at the key terms and comment on which they think are representable and which are confusing. Unexpectedly, expert opinions 'are again inconsistent with the coherence scores.' Higher coherence scores don't necessarily imply the better quality of topics. Labels assigned from key terms and pre-manually coded labels are also compared.

This paper mentions a similar limitation of LDA that we have found: 'Many words appear in every method (e.g., "easy" for topic A) but only few words are unique to one method. Further, the unique words are rather positioned at the end of the lists, meaning that the topics are even more similar when focusing only on the top words.'

5 Conclusion

As some researchers might be in a dilemma of whether or not to analyze free responses when conducting surveys (O'Cathain, 2004), NLP offers ways to process large amounts of text data such as free responses of surveys. In this paper, we summarized several possible ways to process and analyze free responses, using the online learning survey data (Nguyen et al., 2021) as an example. Many researchers have also put forward frameworks to process such data. Rohrer et al. (2017) summarized six possible steps that can help understand a text dataset: analysis of selection effects, preprocessing of textual data, word-level analyses, topic modeling, topic-level analyses, and visual representation of results. The way we followed in this research was very similar, but we do not have the step of analysis of selection effects (which means identifying variables that determine whether an open question is answered) since we are focused on extracting information from non-blank responses.

Preprocessing documents and listing the number of occurrences of each preprocessed word can give a quick view of the data, which is simple and convenient. TF-IDF and k-means can give a good result for clustering in a way — some clusters have distinct topics. For most of the clusters, we can manually summarize the topic quickly. Therefore, with the clustering results, we can easily find the potential topics and the spread(weight) of each topic. However, in the clustering result of questions c about changing one thing about online courses and in-person courses, lots of documents (over 50%, see in Table 3 and Table 4) could not be well classified and they will be put into one or two clusters, which is meaningless. In short, NLP can help us have a basic understanding of a large dataset, but it is still not accurate enough.

There are several possible improvements and directions for future work. For K-means, the number of clusters is needed before we run the algorithm, which is known as “K” in “K-means”. The choice of K is important and it can influence the performance of the algorithm. Kodinariya and Makwana (2013) summarized six approaches to determine the value of K. Following these steps, it’s possible to have a better clustering result, which can give us a more accurate analysis. In addition, visualization of data can also be helpful to give us a better understanding of data. Rohrer et al. (2017) used word clouds to visualize the corpus in their research, which is more direct to understand the word frequency spread than reading numerical statistics.

The topic modeling algorithm LDA generates topics of key terms which provides clues for summaries of responses. It works fast and efficiently, providing a general overview of topics in the corpus of given free responses. However, the topics generated are not distinct enough to classify documents or present significant differences in semantic meaning. Frequently appearing terms that are in key terms of every topic lose semantic significance and offer fewer insights since they are also repetitive and overlap between topics. Representative texts for topics are thought to offer more information and perspective, but they have also shown the classification did not work that well. A topic can combine texts on several themes, see Appendix K and L tables.

Previous work such as Pietsch et al.(2018) concludes that NLP techniques can be used to summarize data to a degree but the results are still very limited and incomparable with human coding. Similar conclusions can be drawn from our analysis. This approach can be used to gather a quick overview of data but is limited in terms of validity and details hidden. There are a lot of previously published papers on testing the validity of NLP techniques and how good they are in terms of summarizing and classifying documents. More work can be done in the future addressing improving the accuracy of such models, for example, implementing the model with word embedding (Nguyen et al, 2015).

In addition, no significant bias concerning online learning has been carried over to a GloVe word embedding based on our RIPA analysis using column 8 of the Nguyen et al. dataset. This suggests that this GloVe embedding model of NLP could be a reliable method of processing and carrying information found in this dataset, without any bias between online and in-person learning. While these conclusions can only apply to this dataset for this bias, the same methods could be repeated to find other biases from this dataset, or to look for the same biases in other datasets. Current findings also show that bias is found differently across vector-space embedding models and hypothesizes bias effects in hyper-dimensional space to greatly depend on the preprocessing stages of data. (Lauscher & Glavaš, 2019) As we have only conducted research using the GloVe embedding model, the same conclusions can not be drawn for other embedding models, including the results of Ngyuen et al., who qualitatively coded their data manually. Additionally, our research was very limited by our choice of data and the size of our vocabulary. The next step in the further analysis should be to replicate our results with an expanded vocabulary and different datasets, such as Wikipedia articles or other free-response survey results, to see if they can generalize.

If our results did show bias, another next step is to address what should be done with this bias. Methods of de-biasing algorithms have been suggested such as in the Bolukbasi et al. paper but carry much criticism. The RIPA paper questions some of the conceptual concepts of this method, and research such as in the Ferrer et al. paper question their practicality. (Ethayarajh et al., 2019) (Ferrer et al., 2019) De-biasing is thought to be helpful as introduced bias into AI can cause real-world altercations and through removing bias in one of our increasingly more common methods of communication, real-world bias could potentially be reduced. (Bolukbasi et al., 2016) However, this process alters the AI model of the world instead of how AI acts on this perception. This causes 'fairness through blindness', which is comparable to other ethical problems such as 'race blindness' to combat racism. Ferrer et al. instead suggest we focus on developing fairness through awareness to classify individuals and prevent discrimination against membership of the group. (Ferrer et al., 2019) (Dwork et al., 2012) Bender and Friedman propose using data statements as a design solution for NLP, to mitigate emergent bias and diagnose pre-existing bias without 'solving' it. (Bender & Friedman, 2018) These approaches are very different, and could seem more preferable depending on the situation; what bias has been detected, what will the word embeddings be applied to, and what is the purpose of the research. In this paper, we have not modified to bias in any way, as firstly the purpose of our research was to investigate if bias had been introduced into the word embeddings, and secondly as our word embeddings are not being put towards an application.

Availability of data. The data used in this research can be found through https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2TGOPH

Code availability. All code used in this research is availability on github: https://github.com/JunruL/ROP299

Authors' contributions. Chen: bias analysis and contextual relations; Lin: word frequency, TF-IDF and K-means applications; Sun: focused on NLP data processing and LDA topic modelling with free responses.

Conflict of interest. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments. A special thanks to Kawin Ethayarajh and Kritika Sharma for helpful discussions, faculty advisor Carolyn Sealfon, and the UofT Research Opportunity Program for this opportunity.

Appendix A TF-IDF and K-means

TF-IDF can be used to evaluate the importance of a term in a document. It extracts keywords and calculates document similarity. Here is a detailed description of TF-IDF:

- TF stands for term frequency. Term frequency means the frequency of a word in a document and is the count of a term in a document divided by the number of words in this document.
- DF stands for document frequency. Document frequency measures the importance of a document in the whole set of corpus and is the occurrence of a term in documents divided by the total number of documents.
- IDF stands for inverse document frequency. Inverse document frequency measures the informativeness of a term and we take logarithm from document frequency (log(N/(df+1)), where N is the count of corpus).
- Finally, TF-IDF score is the product of TF score and IDF score.

K-means is a clustering algorithm. Followings are how it works:

- Set the number of clusters, which is k.
- Select k points randomly and choose the initial centroids.
- For the remaining points, assign each to the closest cluster. The distance calculated here only involves a point and the centroid of a cluster. In this step, there will be k times to calculate the distance and then do a comparison to choose the smallest one.

- After every point gets assigned to a cluster, reset the centroids for the clusters by taking the average of all data points that belong to each cluster.
- Reassign the points with the same method until there is no change to the centroid or the number of iterations reaches a certain value.

To combine TF-IDF and K-means for clustering:

- Obtain the vector representations for each piece of data
  - Set the length of the vector to the number of words in the whole corpus. Each entry in the vector represents the weight of the corresponding word.
  - For words appearing in a response, calculate its weight using TF-IDF and put the values in the corresponding position in this response's vector.
  - For words not appearing in a response, put zero instead.
- Run K-means to cluster responses using the vectors obtained following the above steps.

Appendix B LDA Topic Model

LDA is a generative statistical model that discovers hidden latent topics/themes among the input corpus. A topic LDA generates is a set of terms related by cooccurrence. A term may be ambiguous belonging to more than one topic, but its neighboring terms which belong to only one topic will clarify its ambiguity in the context. (Nguyen, 2015) A document is perceived as a composite of parts and a part is a word or a phrase that belongs to distribution of topics. LDA has two major parameters, Φ is the 'words versus topics' matrix and Θ is the 'document versus topics' matrix. What these two matrices represent is that each word and each document belongs to different topics with different probabilities. It is a probabilistic model since some topics are more likely to appear in a document than others, there exists a probability distribution. What distinguishes LDA from clustering algorithms is that it is multinomial, the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics. (M. Blei, 2003)

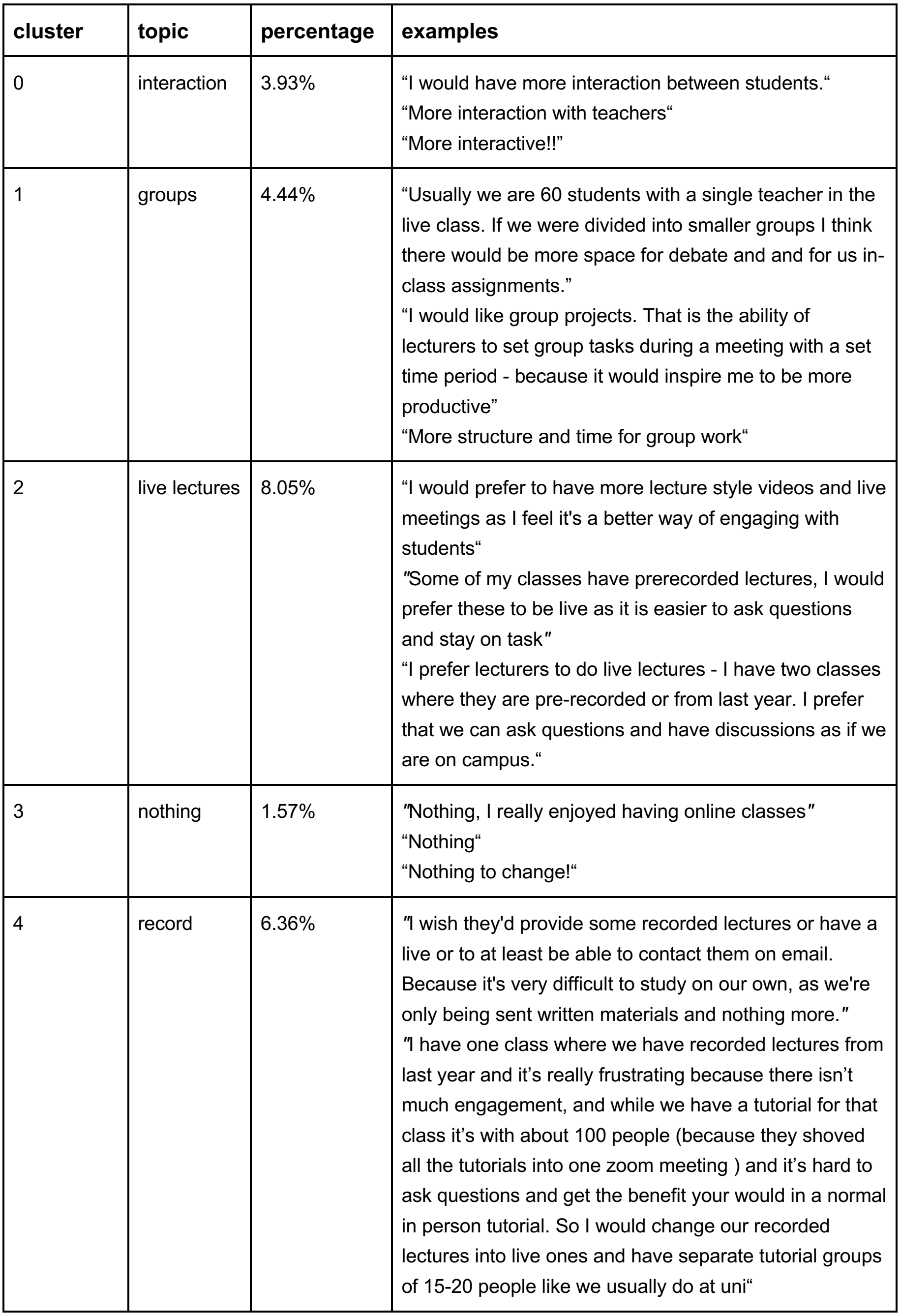
Appendix C the GloVe Word Embedding Model

GloVe is a log bilinear regression model which combines global matrix factorization and local context methods. It takes the relationship of words by examining co-occurrence probabilities with various probe words. In model analysis, they found semantic information is better captured on large corpora, with small and asymmetric context windows. Syntactic information is drawn from intermediate context and word order. (Pennington et al., 2014)

Appendix D the RIPA Bias Test

RIPA uses the relational inner product between a word and a difference vector to measure association. A relation vector is created through scalar projection onto a one-dimensional subspace defined by unit vector b as opposed to a higher-dimensional subspace to make results more interpretable. A positive result indicates an association towards the first term, and a negative, the second.

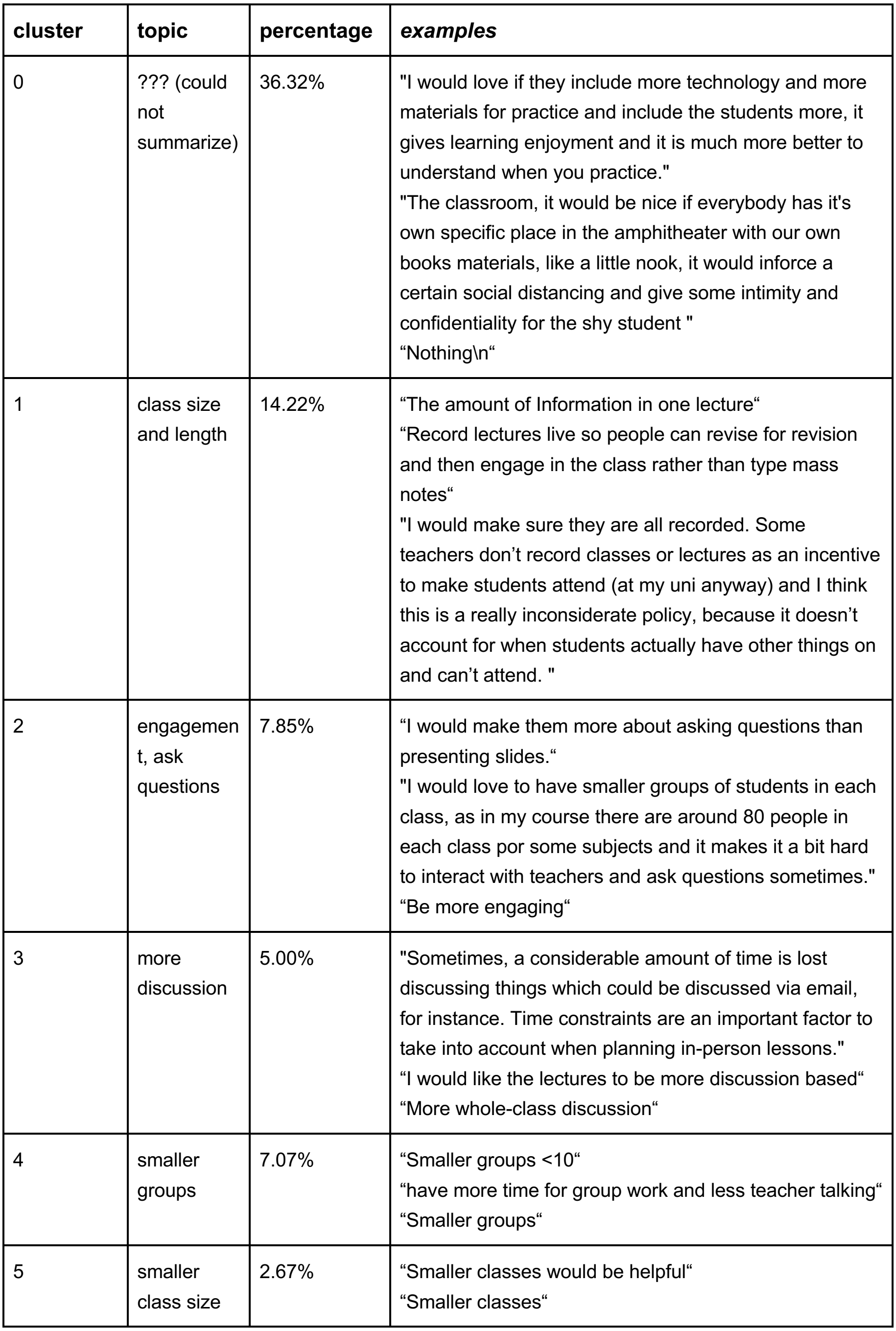
Appendix E Clustering Result for Changing One Thing about Online Courses



cluster	topic	percentage	examples
0	interaction	3.93%	“I would have more interaction between students.“ “More interaction with teachers“ “More interactive!!”
1	groups	4.44%	“Usually we are 60 students with a single teacher in the live class. If we were divided into smaller groups I think there would be more space for debate and and for us in-class assignments.” “I would like group projects. That is the ability of lecturers to set group tasks during a meeting with a set time period - because it would inspire me to be more productive” “More structure and time for group work“
2	live lectures	8.05%	“I would prefer to have more lecture style videos and live meetings as I feel it's a better way of engaging with students“ ″Some of my classes have prerecorded lectures, I would prefer these to be live as it is easier to ask questions and stay on task″ “I prefer lecturers to do live lectures - I have two classes where they are pre-recorded or from last year. I prefer that we can ask questions and have discussions as if we are on campus.“
3	nothing	1.57%	″Nothing, I really enjoyed having online classes″ “Nothing“ “Nothing to change!“
4	record	6.36%	″I wish they'd provide some recorded lectures or have a live or to at least be able to contact them on email. Because it's very difficult to study on our own, as we're only being sent written materials and nothing more.″ ″I have one class where we have recorded lectures from last year and it’s really frustrating because there isn’t much engagement, and while we have a tutorial for that class it’s with about 100 people (because they shoved all the tutorials into one zoom meeting) and it’s hard to ask questions and get the benefit your would in a normal in person tutorial. So I would change our recorded lectures into live ones and have separate tutorial groups of 15-20 people like we usually do at uni“

			“Be able to speed up the video if watching a recorded session on black board collaborate.“
5	don't know	3.33%	“I don't know“ “I would like to be examined weekly. In that way I would know I'm doing everything right.“ ″I don’t think I would change something, obviously some professors give more energy to online classes than other, but overall my experience have been great“
6	live	4.46%	“more live classes“ “Have live discussions about assessment changes“ “I would like more online live classes to be conducted“
7	??? (could not summarize)	14.67%	″I would like the instructors be more engaged in their work and nit just sending us the handouts via the university website, but instead do live classes on Zoom or record a vocal memo and send it. It would be more beneficial and engaging″ ″I wish it was more organized. I know they did what they could with the sudden situation, but I wish there wasn't so much uncertainty. And I wish classes were recorded so I don't have to miss class if I don't have wifi″ ″the way that professors are not aware about the fact that lots of hours of classes, with them talking nonstop it's really unhelpful to the learning progress″
8	engagement	5.06%	″I would love to have video chats such as zoom, as it would make the course much more engaging and motivating.″ ″I would made them live lectures, because I think it’s more interactive and engaging″ ″I would make everyone turn on their video camera, it makes class far more engaging″
9	??? (could not summarize)	47.99%	″It would be better if professor can have top students in the live asking important questions, we would learn more about the subjects.“ “The way lessons are performed because it's not sufficient“ ″I would love if all the teachers used live call/ video techniques rather then just emailing the materials, considering most of my teachers don't use live, there some subjects I find hard and can not reach the teacher for further learning and understanding.″

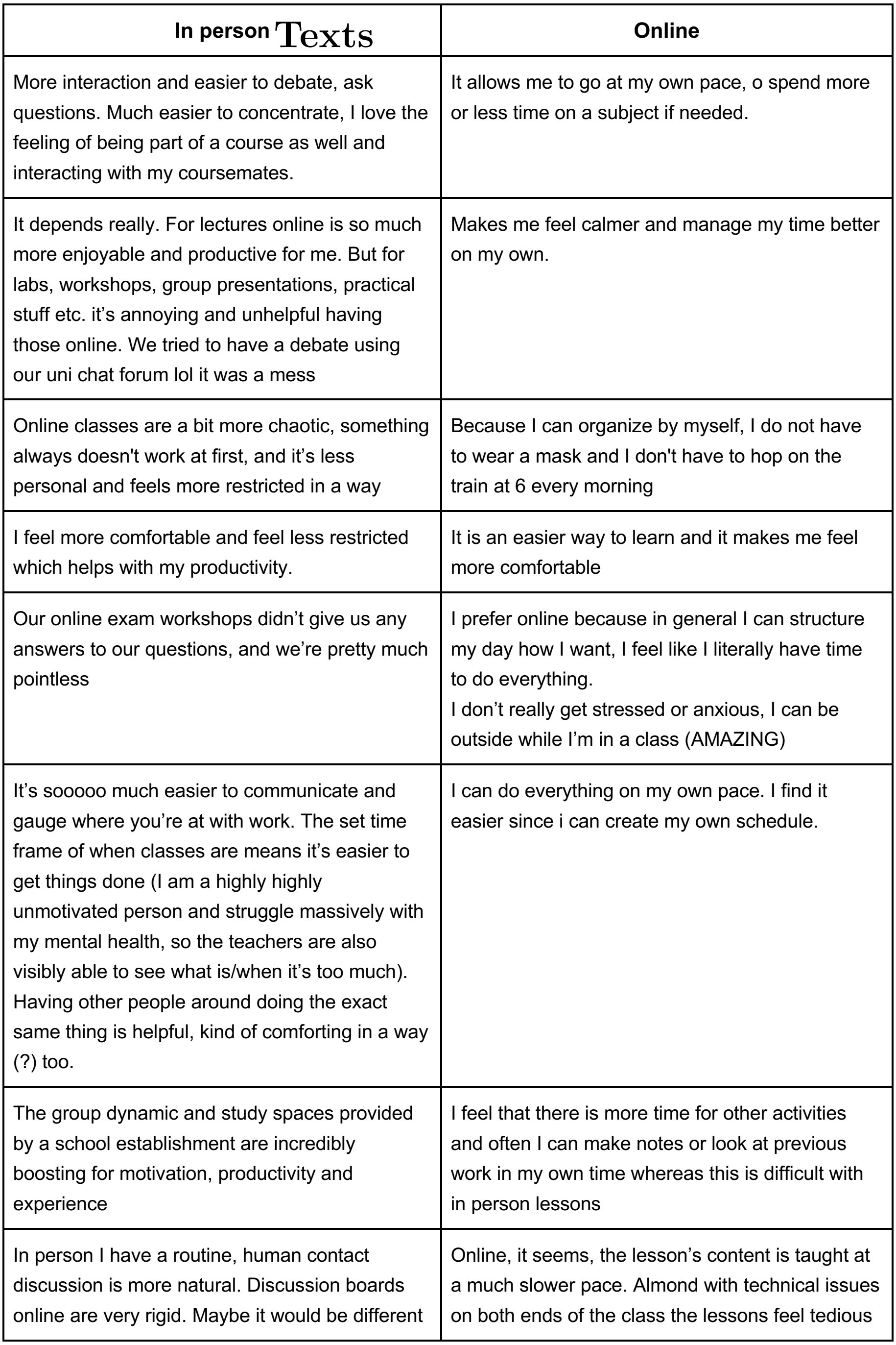
Appendix F Clustering Result for Changing One Thing about In-person Courses



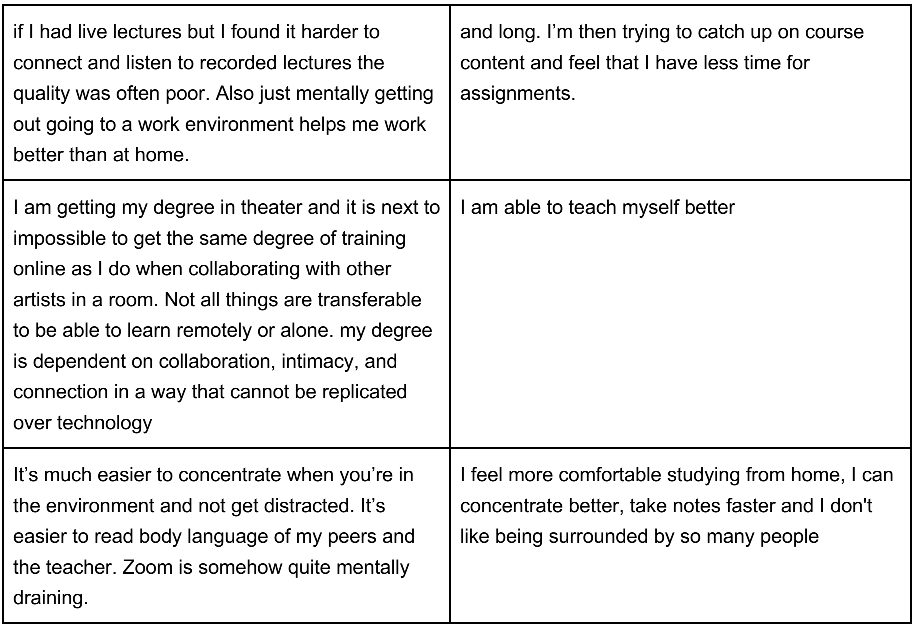
cluster	topic	percentage	***examples***
0	??? (could not summarize)	36.32%	"I would love if they include more technology and more materials for practice and include the students more, it gives learning enjoyment and it is much more better to understand when you practice." "The classroom, it would be nice if everybody has it's own specific place in the amphitheater with our own books materials, like a little nook, it would inforce a certain social distancing and give some intimity and confidentiality for the shy student " “Nothing\n“
1	class size and length	14.22%	“The amount of Information in one lecture“ “Record lectures live so people can revise for revision and then engage in the class rather than type mass notes“ "I would make sure they are all recorded. Some teachers don’t record classes or lectures as an incentive to make students attend (at my uni anyway) and I think this is a really inconsiderate policy, because it doesn’t account for when students actually have other things on and can’t attend. "
2	engagement, ask questions	7.85%	“I would make them more about asking questions than presenting slides.“ "I would love to have smaller groups of students in each class, as in my course there are around 80 people in each class por some subjects and it makes it a bit hard to interact with teachers and ask questions sometimes." “Be more engaging“
3	more discussion	5.00%	"Sometimes, a considerable amount of time is lost discussing things which could be discussed via email, for instance. Time constraints are an important factor to take into account when planning in-person lessons." “I would like the lectures to be more discussion based“ “More whole-class discussion“
4	smaller groups	7.07%	“Smaller groups <10“ “have more time for group work and less teacher talking“ “Smaller groups“
5	smaller class size	2.67%	“Smaller classes would be helpful“ “Smaller classes“

			“Class sizes“
6	nothing	5.76%	“Nothing “ "Nothing, I enjoy them" "Nothing, I love them"
7	??? (could not summarize)	15.84%	"It easier to participate in class when it's online. I guess it's an individual thing, not so much something that my classes should change. I do feel like I would change the examination process. My uni has very strict exams and I get REALLY nervous, and even if I know the subject I do terrible in exams because of it. It takes away the joy of learning" “I’d rather have more debates during classes instead of being just the professor’s development of the topic“ “Some classes could definitely be online without trouble. It'd save us time and money.“
8	no change	2.76%	“I wouldn't change anything“ “i wouldn't change anything“ “Wouldn’t change anything“
9	interaction	3.52%	“My interaction activies in tutorials“ “More interactive“ “Smaller lectures which will allow more student to Professor interaction“

Appendix G LDA-Q2: Top 10 Representative Texts



In person	Online
More interaction and easier to debate, ask questions. Much easier to concentrate, I love the feeling of being part of a course as well and interacting with my coursemates.	It allows me to go at my own pace, o spend more or less time on a subject if needed.
It depends really. For lectures online is so much more enjoyable and productive for me. But for labs, workshops, group presentations, practical stuff etc. it’s annoying and unhelpful having those online. We tried to have a debate using our uni chat forum lol it was a mess	Makes me feel calmer and manage my time better on my own.
Online classes are a bit more chaotic, something always doesn't work at first, and it’s less personal and feels more restricted in a way	Because I can organize by myself, I do not have to wear a mask and I don't have to hop on the train at 6 every morning
I feel more comfortable and feel less restricted which helps with my productivity.	It is an easier way to learn and it makes me feel more comfortable
Our online exam workshops didn’t give us any answers to our questions, and we’re pretty much pointless	I prefer online because in general I can structure my day how I want, I feel like I literally have time to do everything. I don’t really get stressed or anxious, I can be outside while I’m in a class (AMAZING)
It’s sooooo much easier to communicate and gauge where you’re at with work. The set time frame of when classes are means it’s easier to get things done (I am a highly highly unmotivated person and struggle massively with my mental health, so the teachers are also visibly able to see what is/when it’s too much). Having other people around doing the exact same thing is helpful, kind of comforting in a way (?) too.	I can do everything on my own pace. I find it easier since i can create my own schedule.
The group dynamic and study spaces provided by a school establishment are incredibly boosting for motivation, productivity and experience	I feel that there is more time for other activities and often I can make notes or look at previous work in my own time whereas this is difficult with in person lessons
In person I have a routine, human contact discussion is more natural. Discussion boards online are very rigid. Maybe it would be different	Online, it seems, the lesson’s content is taught at a much slower pace. Almond with technical issues on both ends of the class the lessons feel tedious



if I had live lectures but I found it harder to connect and listen to recorded lectures the quality was often poor. Also just mentally getting out going to a work environment helps me work better than at home.	and long. I’m then trying to catch up on course content and feel that I have less time for assignments.
I am getting my degree in theater and it is next to impossible to get the same degree of training online as I do when collaborating with other artists in a room. Not all things are transferable to be able to learn remotely or alone. my degree is dependent on collaboration, intimacy, and connection in a way that cannot be replicated over technology	I am able to teach myself better
It’s much easier to concentrate when you’re in the environment and not get distracted. It’s easier to read body language of my peers and the teacher. Zoom is somehow quite mentally draining.	I feel more comfortable studying from home, I can concentrate better, take notes faster and I don't like being surrounded by so many people

Appendix H LDA-Q3: In-person Topics

Topic number (number + hypothesized theme)	Top 10 Key words (topic number, weighting * key term)	Top 10 Representative Texts	Distribution (number of doc, overall percentage)
0 Class time distribution (where to spend more/less)	(0, '0.085*"time" + 0.035*"chang" + 0.030*"lesson" + 0.024*"read" + ' '0.017*"school" + 0.016*"studi" + 0.016*"give" + 0.015*"distract" + ' '0.015*"answer" + 0.014*"assign"')	0,″SHORTER DAYS. Waking up at 6:30am and getting home at 6:30pm just from school (no extra-curriculars or time to do homework), with only a 30minute lunch break is way too tiring.″ 1,More productive lessons- I realise now how much time is wasted- I have got much more done at home 2,I would change it so that we could choose the professors/teachers for the course instead of randomly being assigned one 3,Less time wasting - waiting for people to arrive etc 4,″Reduce the time I have to be there, not have to hear again an explanation of something I already understood″ 5,Change the amount of lessons so we have more time for revision 6,I wish they were timed better in the day because I can get super tired because I haven't had the right breaks 7,More face to face time for questions 8,Less pressure to answer questions in front of everyone 9,I would change the deadlines too many assignments and no time	1400.0, 0.1963
1 Instructor-student dynamic	(1, '0.073*"teacher" + 0.065*"student" + 0.039*"onlin" + 0.032*"person" + ' '0.026*"materi" + 0.014*"classroom" + 0.014*"inperson" + 0.013*"debat" + ' '0.012*"particip" + 0.012*"motiv"')	10,″The student teacher dynamic was very off, teachers had 0 respect for students and were just downright mean and horrible quite often″ 11,I'd allow the integration of online resources and tech into the classroom 12,Less competitiveness and favouritism if highest achieving students 13,also have materials online in case i don't go (bit ironic but oh well)	1261.0, 0.1768

		14,Stop teachers from treating students like crap 15,To BAN the no hands up policy it gives me anxiety. 16,″They're quite diverse, each having strengths and weaknesses″ 17,I would want teachers to encourage students to participate more. 18,Teachers to stop picking on students 19,Use of laptops or online resources	
2 Duration and form of lectures	(2, '0.064*"lectur" + 0.051*"engag" + 0.038*"question" + 0.034*"feel" + ' '0.026*"ask" + 0.020*"inform" + 0.018*"talk" + 0.018*"professor" + ' '0.018*"note" + 0.016*"lot"')	20,″More lectures per week for each module, currently it is only one lecture a week for each module during third year. ″ 21,Sometimes with large lectures it's difficult to ask questions and become more engaged 22,Less overly long lectures. They can feel tedious and rambly. 23,start them after 10am because I can't focus so early in the morning 24,I wish lecturers would slow down and not be dismissive of questions and actually explain what they're doing. 25,″The use of PBL is sometimes very useless Imo, e.g. pre-discussions″ 26,More Engagement; sometimes my Profs are bored by their own lectures 27,Not so early in the morning so my brain is awake!! 28,Include more relevant information rather than than blabbering 29,less copying from textbook/board. it doesn't help absorb information	1604.0, 0.2249
3 More involvement with media	(3, '0.053*"make" + 0.046*"learn" + 0.037*"lectur" + 0.036*"interact" + ' '0.033*"work" + 0.022*"understand" + 0.022*"teach" + 0.022*"content" + '	30,″More small group teaching rather than lectures, make lectures optional as they are recorded anyway ″ 31,More small group interactive sessions and continuous mini assessments. 32,″Same, I would make them more graphic and video base dinstead of presentations″ 33,More assessed work with individual feedback rather than generic comments	1502.0, 0.2106

	'0.017*"record" + 0.016*"subject"')	34,Allow more freedom in content produced (for example an essay presented in a different format) 35,More creative projects instead of strictly sticking to the course content and curriculum. 36,More realistic applications and content (real life situations) 37,″Not have them, get rid of them. Punish CHINA for this disgraceful act of destroying our education″ 38,″More variance in pedagogy, a balance between lectures and group-work/creative work. ″ 39,″More variance in pedagogy, a balance between lectures and group-work/creative work. ″	
4 Smaller class size	(4, '0.193*"class" + 0.054*"discuss" + 0.045*"smaller" + 0.032*"peopl" + ' '0.027*"group" + 0.027*"hour" + 0.016*"thing" + 0.014*"easier" + ' '0.013*"cours" + 0.013*"focus"')	40,″Definitely smaller class sizes. My classes can be up to 32 people. In the classes I have where there are more than 30 people, these are the ones that I perform the worst in.″ 41,more smaller group discussions prior to whole seminar discussions 42,I would have less people in a class because my smaller classes are the ones I'm more focused in 43,divide the big groups in smaller classes 44,More face-to face contact on sites like Zoom 45,I wish classes were smaller so it would be easier for everyone to be involved and discuss 46,″Smaller class sizes, enabling more 1-1″ 47,Smaller classes would be more ideal and also more group discussion 48,″Smaler groups, we never have class with less than 200 people″ 49,″Less catered to exams. Yes, we do have exams but not everything should be so catered to them.″	1364.0, 0.1913

Appendix I LDA-Q3: Online Topics

Topic number (number + hypothesized theme)	Top 10 Key words (topic number, weighting * key term)	Top 10 Representative Texts	Distribution (number of doc, overall percentage)
0 More virtue meetings / materials (? a mix of themes, even including foreign languages)	(0, '0.068*"learn" + 0.061*"video" + 0.049*"assign" + 0.020*"chat" + ' '0.017*"call" + 0.015*"task" + 0.014*"talk" + 0.014*"face" + 0.013*"explain" ' '+ 0.013*"cours"')	0,"Creo que solo cambiaría algunos profesores, su manera de enseñar no es entendible " 1,"Some courses could benefit from more organized information on their pages, but other than that, I have no complaints" 2,"Je ferais plus de cours sur zoom, car le lien avec les autres sont importants." 3,"More posted assignments, less mandatory video calls. " 4,Have some videos explaining as well as written text readings 5,"More large-scale discussion/debate, to gather a more comprehensive evaluation/analysis of a topic" 6,Rendre le cours plus interactif 7,The video and audio quality sucked 8,Less assignments and more learning new topics as we are mainly revising 9,More face to face virtual contact	1366.0, 0.1315
1 More live and engaging lectures	(1, '0.122*"lectur" + 0.049*"content" + 0.042*"feel" + 0.039*"engag" + ' '0.037*"live" + 0.028*"session" + 0.026*"studi" + 0.013*"method" + ' '0.012*"powerpoint" + 0.011*"problem"')	10,"more live lectures, more mack quizzes/tests" 11,Have more live lectures/seminars to engage with the content 12,I'd rather have more prerecorded lectures than live sessions 13,Live lectures so that I would feel more involved 14,I would like them to be more engaging but also to require more attention to lectures 15,How engaging the lectures are and the number of lectures 16,Move to live lectures than prerecorded 17,More live lectures as I feel I would engage more	1711.0, 0.1647

		18,I would like my lectures to be live (not from previous years/pre-recorded) 19,I would like to have live lectures as when live you are less likely to be distracted/procrastinating	
2 Less work more help	(2, '0.073*"work" + 0.058*"student" + 0.034*"materi" + 0.022*"understand" + ' '0.020*"teach" + 0.019*"professor" + 0.017*"hour" + 0.014*"set" + ' '0.014*"day" + 0.013*"school"')	20,"The amount of work should be more spread out, because (IGCSE evidence gathering aside) it is sometimes too much to handle." 21,not set such masses amount of work for a 1hr lesson 22,Colleges advise students on how to boost WiFi 23,The lack of clarity on some material because it's challenging to teach yourself 24,"More consistent outlined plans of when work must be completed, a wider understanding of the added pressure working from home may have or students. " 25,More tests(that don't count as official grades) to see if we really understand everything. 26,the amount of work i'm being issued as it seems like i'm getting so much more than I would at school 27,Nothing liverpool hope are dealing with it amazingly 28,Set a realistic amount of work 29,more materials to assist with the work we're completing	1815.0, 0.1747
3 More live classes	(3, '0.076*"class" + 0.039*"live" + 0.037*"discuss" + 0.037*"interact" + ' '0.033*"zoom" + 0.030*"group" + 0.024*"engag" + 0.022*"lesson" + ' '0.020*"meet" + 0.018*"time"')	30,Be able to break into small groups to discuss things during/after live meetings. 31,More live classes. More interactive. I would prefer that everyone was more wngaged and not passs so that we can have more interesting discussions. 32,I would have preferred more live zoom classes so it could represent the classes we had in person better. 33,More structured to a specific time so they don't take up more time than a usual lesson would	2677.0, 0.2576

		34,More regular live classes with interactive activities or discussions 35,More small quizes or something a little more interactive so that you would be more engaged during classes on zoom. 36,Time differences I am in France so 7h away (9 am classes are at 2am etc.) 37,Have more interaction within the class because I get inspired/learn more from my peers thoughts and ideas. 38,I would prefer having more visual/video lessons e.g through Zoom because they're more engaging and allow for more explanation 39,I would prefer to have live interaction with my classes over zoom.	
4 Communication relevant	(4, '0.061*"onlin" + 0.056*"teacher" + 0.055*"class" + 0.054*"make" + ' '0.050*"time" + 0.029*"question" + 0.024*"record" + 0.023*"chang" + ' '0.018*"interact" + 0.017*"peopl"')	40,"Make teacher interaction easier, at the moment it is through email and responses dont come immediately" 41,i would like that there is something that regulates who speaks (like a request to speak or smth) because it gets really messy when all people ask the teacher at the same time. 42,"I would make it in a way that is easier to ask questions during the classes, so that people wouldn't shy away from interrupting and speaking up" 43,"Not much, online classes are just bad and my teachers are trying their hardest but you can't really make online classes good." 44,I would change the rigidity of the class times - they were hard to keep up with during the initial change online 45,"A lot of my tutors are using this time as a holiday, so we don't get any emails answered, so I'd change this because I need them to answer me" 46,"Different teachers use different platforms/ to communicate with us, it would be nice if there was some sort of unity, so it's easier to find what we need to do"	2822.0, 0.2716

		47,"Have smaller classes instead of big ones. Mainly because too many people become problematic, especially if a question needs answering and there’s either no response or too many " 48,"I would make it easier to ask questions, and make it less of a ‘bother’ to teachers " 49,"I would like to be able to mute other people's microphones, because the background noise is more prominent when we're in an online class"	

Appendix J Mathematics and Logistics of the RIPA Bias Test

Ethayarajh et al. found that in practice, λ is about 1, and α to be about –1. Z measures for the learned bias, trained from the Nguyen et al. dataset, from the GloVe embeddings, and can be outputted from the algorithm. The function p measures joint word probabilities, and is normalized by N, the total amount of words, according to Church & Hanks, 1989. csPMI is the co-occurrence shifted pointwise mutual information matrix (PMI), simplified in Equation J5 from Ethayarajh, 2019. The PMI is simplified as in Equation J6, from Ethayarajh et al., 2019.

The difference vector between x and y:

$$\vec{b} = (\vec{x} - \vec{y})/\|\vec{x} - \vec{y}\| \tag{J1}$$

The RIPA test for association within word embeddings:

$$g(w;\ 'man',\ 'woman') = \frac{< \vec{w},\ m\vec{a}n - wom\vec{a}n >}{\|m\vec{a}n - wom\vec{a}n\|} \tag{J2}$$

The RIPA equation for corpus bias:

$$\beta_{GloVe} = C(\log \frac{p(x,w)}{p(y,w)} - z_x + z_y) \tag{J3}$$

where

$$C = \frac{1/\sqrt{\lambda}}{\sqrt{-csPMI(x,y) + \alpha}} \tag{J4}$$

The co-occurrence shifted PMI:

$$csPMI(x,y) = PMI(x,y) - \log p(x,y) \tag{J5}$$

PMI theorem:

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{J6}$$

Reference

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (in press). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *The 30th International Conference on Neural Information Processing Systems (NIPS'16)*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics -*. Published. https://doi.org/10.3115/981623.981633

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. Published. https://doi.org/10.1145/2090236.2090255

Ethayarajh, K. (2019a, June 21). *Word Embedding Analogies: Understanding King - Man + Woman = Queen*. Kawin Ethayarajh. https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html

Ethayarajh, K. (2019b, September 23). *Bias in Word Embeddings: What Causes It?* Kawin Ethayarajh. https://kawine.github.io/blog/nlp/2019/09/23/bias.html

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019a). Towards Understanding Linear Word Analogies. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Published. https://doi.org/10.18653/v1/p19-1315

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019b). Understanding Undesirable Word Embedding Associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Published. https://doi.org/10.18653/v1/p19-1166

Garcia, J., Evans, J., & Reshaw, M. (2004). ``Is There Anything Else You Would Like to Tell Us'' – Methodological Issues in the Use of Free-Text Comments from Postal Surveys. *Quality & Quantity*, *38*(2), 113–125. https://doi.org/10.1023/b:ququ.0000019394.78970.df

Glazier, R. A., Hamann, K., Pollock, P. H., & Wilson, B. M. (2019). What drives student success? Assessing the combined effect of transfer students and online courses. *Teaching in Higher Education*, *26*(6), 839–854. https://doi.org/10.1080/13562517.2019.1686701

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2019). Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

Kizilcec, R., Davis, D., & Wang, E. (2019). Online degree stigma and stereotypes: A new instrument and implications for diversity in higher education. *SSRN Electronic Journal*. Published. https://doi.org/10.2139/ssrn.3339768

Kodinariya, T., & Makwana, P. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*, 90–95.

Kulhanek, R. (2013). A Latent Dirichlet Allocation/N-Gram Composite Language Model. https://corescholar.libraries.wright.edu/etd_all/1143

Lauscher, A., & Glavaš, G. (2019). Are We Consistently Biased? Multidimensional Analysis of Biases in Distributional Word Vectors. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Published. https://doi.org/10.18653/v1/s19-1010

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461. https://doi.org/10.1016/s0031-3203(02)00060-2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*. Published.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, *3*, 299–313. https://doi.org/10.1162/tacl_a_00140

Nguyen, T., Netto, C. L. M., Wilkins, J. F., Bröker, P., Vargas, E. E., Sealfon, C. D., Puthipiroj, P., Li, K. S., Bowler, J. E., Hinson, H. R., Pujar, M., & Stein, G. M. (2021). Insights Into Students' Experiences and Perceptions of Remote Learning Methods: From the COVID-19 Pandemic to Best Practice for the Future. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.647986

O'Cathain, A., & Thomas, K. J. (2004). "Any other comments?" Open questions on questionnaires – a bane or a bonus to research? *BMC Medical Research Methodology*, *4*(1). https://doi.org/10.1186/1471-2288-4-25

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Published. https://doi.org/10.1145/3351095.3372843

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Published. https://doi.org/10.3115/v1/d14-1162

Pietsch, A. S., & Lessmann, S. (2018). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, *1*(2), 93–116. https://doi.org/10.1080/2573234x.2019.1590131

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1), 25–29. https://doi.org/10.5120/ijca2018917395

Rohrer, J. M., Brümmer, M., Schmukle, S. C., Goebel, J., & Wagner, G. G. (2017). “What else are you worried about?” – Integrating textual responses into quantitative social science research. *PLOS ONE*, *12*(7), e0182156. https://doi.org/10.1371/journal.pone.0182156

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019). What are the Biases in My Word Embedding? *Proceedings* *of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Published. https://doi.org/10.1145/3306618.3314270

Steinberger, J. & Jezek, K. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation.

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Published. https://doi.org/10.1109/dsaa.2017.61

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, *41*(1), 75–95. https://doi.org/10.1016/j.ipm.2004.04.003

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. W. (2019). Gender Bias in Contextualized Word Embeddings. *Proceedings of the 2019 Conference of the North*. Published. https://doi.org/10.18653/v1/n19-1064