The Viability of Topic Modeling to Identify Participant Motivations for Enrolling in Online Professional Development

Heather Barker
Elon University

Hollylynne Lee
North Carolina State University

Shaun Kellogg
North Carolina State University

Robin Anderson
North Carolina State University

Abstract
Identifying motivation for enrollment in MOOCs has been an important way to predict participant success rates. In this study, qualitatively coding discussion forums was combined with topic modeling to identify participants' motivation for enrolling in two successive statistics education professional development online courses. Computational text mining, such as topic modeling, has proven effective in analyzing large volumes of text to automatically identify topics or themes. This contrasts with traditional qualitative approaches, in which researchers manually apply labels to parts of text to identify common themes. Combining topic modeling and qualitative research may prove useful to education researchers and practitioners in better understanding and improving online learning contexts that feature asynchronous discussion. Three topic modeling approaches were used in this study, including both unsupervised and semi-supervised modeling techniques. The topic modeling approaches were validated and compared to determine which participants were assigned motivation themes that most closely aligned to their posts made in an introductory discussion forum. Though the three techniques have varying success rates in identifying motivation for enrolling in the MOOCs, they do all identify similar themes for motivation that are specific to statistics education.

Keywords: MOOCs, topic modeling, online professional development, discussion forums, motivation

Barker, H., Lee, H., Kellogg, S. & Anderson, R. (2024). The viability of topic modeling to identify participant motivations for enrolling in online professional development. *Online Learning, 28*(1), 175-195. DOI: https://doi.org/10.24059/olj.v28i1.3571

Introduction

Massive Open Online Courses (MOOCs) are a form of online professional development (OPD) that can be useful for learners to communicate and provide professional development (PD) that is on-demand and timely. Rising enrollment in MOOCs has led many researchers to explore participant motivation for enrolling (Boroujeni & Dillenbourg, 2019; Douglas et al., 2016; Frankowsky et al., 2015; Kellogg et al., 2014). Despite the advantages MOOCs offer, MOOCs have high dropout rates, and literature suggests this may be tied to participant motivation (Badali et al., 2022).

Motivation can have a positive impact on the participant performance in a MOOC. Researchers have linked motivational goals for taking a course to engagement levels (Littlejohn et al., 2016, Milligan et al., 2013) and retention rates (Xiong et al., 2015). Identifying motivation for enrolling in MOOCs is often achieved by analyzing enrollment surveys (Creager et al., 2018; Hollebrands & Lee, 2020; Moore & Wang, 2021; Wilkowski et al., 2014). Using enrollment surveys may limit the motivations that can be identified to just closed-ended choices. A richer data source to identify motivation may be introductory discussion forums.

Topic modeling is an unsupervised learning method that can be used for classifying large groups of texts into discrete word groups (Silge & Robinson, 2019). Since reading and identifying themes for motivation on a large volume of discussion forum posts can be a daunting task, using topic modeling may be an appropriate alternative to traditional qualitative methods of identifying themes.

This paper aims to chronicle how three different methods for topic modeling were used to identify participant motivation from discussion forum posts and why these methods may prove useful for other educational researchers. The following sections include a literature review of prior research, a methodology section detailing the three topic modeling methods, the results of each method, and finally a discussion highlighting the importance of this work. We hope to provide a different way to categorize themes for motivation specific to the courses in this study, rather than themes generic to any MOOC.

Literature Review

This literature review will provide an overview of the methodological approaches that researchers have used to identify motivations for individuals to enroll in MOOCs. We will first discuss how enrollment surveys have been used to identify motivation, the most common approach. We will then introduce how traditional qualitative methods, such as hand coding qualitative data, have been used to identify motivation. Finally, we will explore the potential advantages that topic modeling may have over traditional qualitative methods.

Enrollment Surveys

MOOCs provide participants autonomy in engagement in courses, making it important for researchers and practitioners to understand what motivates participants to enroll. In a systematic review of 50 studies, Badali et al. (2022) identified the role motivation plays in retention rates in MOOCs. Of the 50 studies, 64% used quantitative methods, 16% used qualitative methods, and 20% used mixed research methods. Badali et al. classified the motivational factors into two broad themes: need-based (academic, course, and professional) and interest-based (social, personal, and technological). The data collected for these studies were interviews (18%), surveys (70%), and a mix of interviews and surveys (12%).

Identifying motivation for enrolling in a MOOC is often achieved by asking for responses to closed-ended questions on enrollment surveys (Brooker et al., 2018; Creager et al., 2018; Hollebrands & Lee, 2020; Moore & Wang, 2020; Wilkowski et al., 2014). Wilkowski et al. (2014) sought to identify groups based on their motivations for enrolling in a MOOC hosted by Google. Participants were asked about their motivation on the enrollment survey. Possible answers included learning about aspects of the course or earning certain certificates. Moore and Wang (2020) examined the responses of an enrollment survey for a MOOC from Harvard University to identify underlying profiles for students' motivations to learn. Using Latent Profile Analysis, students were grouped as intrinsic or extrinsic learners. Moore and Wang found that those who were grouped as intrinsically motivated tended to have higher rates of course completion.

Closed-ended questions on enrollment surveys limit the types of motivations that can be identified to only those listed on the survey. Providing a space for participants to express their motivation outside of a closed-ended survey may provide a richer overview of what brings participants to a MOOC.

Qualitative Approaches

Another alternative space for identifying motivation, outside of enrollment surveys, could be in online discussion forums. Online discussion forums are spaces where participants can interact and express their individuality. Tang et al. (2018) used responses to introductory discussion forums in one MOOC to identify learners as extrinsically or intrinsically motivated. Two researchers read a total of 444 responses and used the constant comparative method to qualitatively code each introductory discussion forum post. No other studies were found that used introductory discussion forum data as an identifying source for motivation.

Qualitative approaches to analyzing discussion forum data have been done by many other researchers. Despite large numbers of discussion forum posts, researchers have tackled analyzing the data using qualitative methods by reading each post to identify themes. Nandi et al. (2012) used a grounded theory approach through open coding to identify the quality of interactions between participants and instructors in two courses that had 1,352 participants. Wang et al. (2015) used a discourse framework to hand code 7,990 discussion forum posts for a psychology MOOC. Hollebrands & Lee (2020) used open coding (of 977 posts) to identify what triggers may have caused a shift in participants' beliefs during their participation in an OPD for statistics teachers.

These studies showcase the range of questions that can be answered by using qualitative data analysis approaches, but these approaches are time-consuming. Perhaps text mining techniques may be a more efficient way to analyze a large corpus of data, such as discussion forum data in MOOCs.

Topic Modeling

Text mining is a computational approach to analyze large collections of text to try to make meaning of data (Hearst, 2003). Topic modeling, which consists of both unsupervised and supervised machine learning methods for text mining, is used for classification of large groups of texts into discrete groups of words, or “topics” (Silge & Robinson, 2019).

Unsupervised topic modeling groups words, based on certain statistical criteria, that become the topics for a large corpus of data (Silge & Robinson, 2019). It is up to researchers to interpret these topics as they apply to the data. Ezen-can et al. (2015) used an unsupervised topic modeling technique to create seven clusters from 550 discussion posts that were part of a MOOC for educators on digital learning. Latent Dirichlet Allocation (LDA) was then used on the posts in each cluster to identify the textual themes.

Reich et al. (2016) used topic modeling to analyze themes of 350 posts in an educational policy MOOC. The topics found described patterns of discussion in the forums on the use of school vouchers and feelings about instituting the Common Core. Vytasek et al. (2017) applied four unsupervised topic modeling approaches to a set of 813 posts in a medical statistics MOOC. They found that the best way to make sense of the topics was to nest the topics as subtopics that are part of more general topics.

Seeded topic modeling is a semi-supervised learning method that identifies topics using a predetermined seeded dictionary of terms (Watanabe & Xuan-Hieu, 2020). Ramesh et al. (2014) used a semi-supervised learning method of fitting a LDA model by inputting a seeded dictionary of terms to identify topics that they assumed should be common to the context of MOOC discussion forums. Wong et al. (2019) were able to show that using a seeded LDA method was effective for tracing forum posts back to topics specific to a MOOC.

Nelson et al. (2021) recognized the gap that may exist between hand coding text and using computational methods to identify themes in socially constructed content. Nelson et al. used three common computer text mining approaches, dictionary, supervised, and unsupervised machine learning, to compare the results of the computerized text mining to previously hand coded textual data. Newspaper articles had already been coded based on themes on income inequality. Nelson et al. found that the unsupervised machine learning method worked best and 91% of the articles were coded with the same theme as the hand coding method.

Building on the motivation research, prior qualitative approaches to analyzing discussion forum data, and topic modeling approaches, has led to the research question: how can topic modeling be an effective tool for classifying the motivations of participants who enroll in online professional development courses for statistics educators?

Methods

Context and Participants

The data in this study is a large collection of posts from discussion forums in two online professional development (OPD) courses designed for statistics educators, primarily those teaching in middle schools (age 11) through introductory college courses. The context of this study is critical in understanding the outcomes of the topic modeling approaches used.

Statistics has made an impact in the mathematics curriculum, which has led to challenges in preparing teachers to teach statistics. Professional development (PD) opportunities for teachers of statistics have been implemented to foster the knowledge, skills, and dispositions necessary to effectively teach the subject. The American Statistical Association (ASA) endorsed the Statistical Education of Teachers (SET) report to guide pre- and in-service teacher preparation for teaching statistics (Franklin et al., 2015). The SET report stresses the need for professional development at the local or state level to aid mathematics teachers to teach statistics, while also recognizing the limitations of providing such professional development (Franklin et al., 2015). OPD) can be a way to provide this PD for those who need it (Lee & Stangl, 2015). The second author and her team created two OPD courses (Course 1 and 2) for statistics educators. Each course is meant to enhance teachers' understanding of statistics and teaching strategies in middle school through introductory level college courses.

The two online courses analyzed were created to provide high quality OPD for statistics educators. The "overarching goal of Course 1 is to engage participants in thinking about statistics teaching and learning in ways that are likely different from their current practices in middle school through college-level introductory statistics" (Hollebrands & Lee, p. 4). Course 2 was meant to be an extension of Course 1 while emphasizing inferential reasoning.

Course 1 was offered seven times, with the first offering in fall 2015 and the last in fall 2018. A total of 3,115 people enrolled in Course 1. Course 2 was offered three times, fall 2017, spring 2018, and spring 2019, with a total of 700 people enrolled. The courses were asynchronous. Each course had an orientation unit and five units of learning material.

Of the 3,815 total people enrolled in either course, 1,592 accessed at least Unit 1 of a course; those are the participants included in this study. Researchers have found a high drop-off rate of participants after the first unit of MOOCs (Hollebrands & Lee, 2020; Erikkson et al., 2017; Onah et al., 2014), which likely indicates participants visited and found they were no longer interested or no longer had the time to participate.

There are participants who enrolled in more than one course or enrolled in another section of the same course. Since motivation can change over time, it was decided to treat each time a person took a course as a separate participant. Participants are identified using their numeric user identification and course identification numbers (userid_bycourse). Of the 1,592 unique participants, 357 registered for more than one course, resulting in 1,949 participants for analysis purposes.

Discussion Forum Data

The data for this study is from the first discussion forum, in the orientation unit of each course titled *Meet Your Colleagues*. The prompt asked participants to introduce themselves and share why they enrolled in the course. Participants can either create a new thread or respond to other participants. Initial posts and replies were included in the data for this study.

Discussion Forum Data Preparation

In unsupervised topic modeling there are often topics that are found that do not always make sense to the user (Hu et al., 2014). To avoid the general topics that naturally arise, the data was prepared prior to modeling so that the discussion, or noise, that is not centered on motivation was reduced as much as possible. An exploratory topic modeling approaches was used to isolate relevant data.

Identifying parts of posts that may prove useful for identifying themes for motivation for taking these specific courses may not be obvious to anyone able to perform topic modeling. Thus, it was critical that the researchers were familiar with the data. The authors' expertise includes OPD for statistics teachers, so they are familiar with what motivates people to enroll in courses like these. Additionally, the authors have worked with discussion forum data from these course offerings in the past, offering a unique perspective to the best ways to prepare this specific set of data for topic modeling.

All posts from the *Meet your Colleagues* forum in the orientation unit were collected from each of the course offerings. This resulted in 1,639 posts. (Note that not all participants posted in this forum.) These posts were blinded by removing all mentions of names or locations. Many entries included introductory information about the participant, such as what they teach, where they are from, etc. For instance, in the following post, the first part is introductory information about where they teach. The second sentence was retained for analysis. "Hi, all. I have taught an Elementary Statistics course at ---- Community College for 14 years.; I have a few classroom activities that I use regularly, and I would like to get additional ideas for activities to keep my students engaged."

After reading all posts, 1,099 were considered to pertain to motivation. The 1,099 posts include multiple posts that may have been made by the same participant. The posts could have been initial posts creating a new discussion thread or replies to posts. Replies were kept for analysis purposes as well as initial posts since there were often clues to their motivation for taking the course within reply threads. Since we are interested in what motivates each participant to take the course, any posts that were made by the same participant in a specific course were merged so that when performing topic modeling the corpus of posts from each user would be read as a single document, rather than multiple documents from each user. This eliminated the possibility that more than one topic could be applied to any participant. In all, there were 946 unique participants with usable discussion posts. Thus, there were 946 documents used for topic modeling. These documents are the unit of analysis.

Identifying Text Terms in Document

To perform topic modeling, posts must be broken down, or tokenized, into strings of individual words (Silge & Robinson, 2019). These individual words form what are called a document term matrix (DTM). In the DTM, each row represents one participant's document and each column represents one word. The count of each word is recorded for each participant in the corresponding cells. The DTM was created in R using the CreateDTM function which is part of the textmineR package (v.3.0.4; Joanes & Doane, 2019). Stop words were removed from DTM prior to performing topic modeling to ensure that common English words such as *a*, *the*, *and*, etc. did not become grouped into topics (Silge & Robinson, 2019).

Stemming can be used in topic modeling to group words with the same stem (Wu et al., 2017). For instance, *learn*, *learning*, and *learned* all have the same connotation. The Porter stemmer method (Porter, 1980) was used to stem words in the posts using the stemmer function in the SnowballC package (v.0.7.0;Bouchet-Valat, 2020). There are those who caution that the use of stemming can degrade the topic modeling process (Schofield & Mimno, 2012). The decision to use stemming was made after an exploratory topic modeling approach was done without stemming words. This exploratory approach had words such as *statistics*, *statistical*, *statistic* or *learned*, *learning*, and *learn* appear so often in the topics that other words that may be helpful in identifying topics did not appear as top words. After the stemming approach was used, which combined *statistic*, *statistical*, and *statistics* to just the stem *statist*. This made room for other meaningful words to appear such as *engage* or *science*.

The DTM can be made using one word grouping or any n-sized groups of words. For this analysis, the DTM was made of one- and two-word groups, unigrams and bigrams, respectively. Wang et al. (2007) developed a topic modeling approach using groups of words to identify relevant two-word groupings such as “white house” as well as unigram and other n-gram phrases. Similarly, we used two-word groups were used to capture terms such as *build confidence* or *statistical thinking*. Including these terms would help to distinguish between words such as *learn_statistics* and *teach_statistics*. If we did not use bigrams, *statistics* would just be counted once, but we know that the motivation to learn statistics is much different than being motivated to teach statistics. Any two successive terms were considered bigrams. It is possible to look at n-gram groupings higher than n = 2 to capture more phrases. Researchers have found interpreting topics with these higher order phrases is possible, but requires programming methods specific to phrases, instead of words, which were not used in this study (Das et al., 2016; Huang, 2018; Schmiedel et al., 2019).

After the DTM was created, topic modeling was performed on a random set of 100 posts to test if any words outside of common stop words appeared more often that may not have meaning when identifying motivation. This topic model had the following terms appear most often: *ways*, *wait*, *looking forward*, *hope*, and *take*. These terms were removed from posts before creating the DTM.

The following illustrates how the steps identified above were used to clean the posts that were used to create the DTM. Below are the combined posts for participant 4451_9.

> My Name is xxx, I teach at xxx in xxx. I teach AP Statistics and am hoping to get some ideas of how I can encourage my colleagues to incorporate more statistics and data collection into their courses so that a course such as mine isn't the first time that students are exposed to Stats. It seems that most of high school courses lead students to Calculus, but I think that Statistics is much more interesting and applicable to more students.

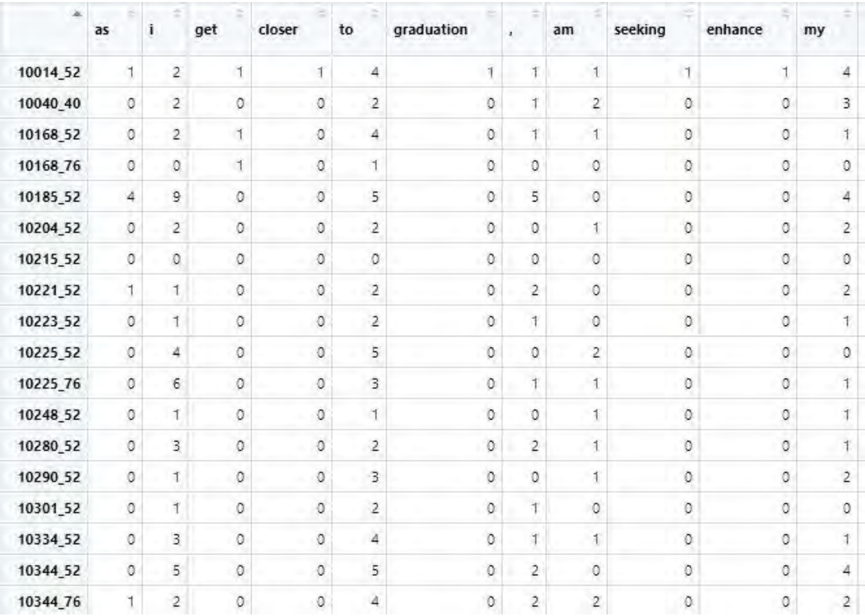
After going through the steps described above, the following words were included in the DTM for participant 4451_9. Cleaned documents, like the one below, were used to create the DTM.

Stop words and stemming still appear in this step, those were not filtered out until the creation of the DTM.

> I teach AP Statistics and am get some ideas of how I can encourage my colleagues to incorporate more statistics and data collection into their courses so that a course such as mine isn't the first time that students are exposed to Stats.

The resulting DTM was a matrix with 946 rows (representing the participants) and 8,933 columns (representing the 1- or 2-word groups). The cells of the matrix are the number of times each word(s) occurred for that participant. Figure 1 shows that participant 10014_52 used the word *as* 1 time, whereas participant 10185_52 used the word *as* 4 times. The complete row for each participant has all possible unigrams and bigrams, 8,933 columns.

Figure 1
Truncated View of the DTM



	as	i	get	closer	to	graduation	,	am	seeking	enhance	my
10014_52	1	2	1	1	4	1	1	1	1	1	4
10040_40	0	2	0	0	2	0	1	2	0	0	3
10168_52	0	2	1	0	4	0	1	1	0	0	1
10168_76	0	0	1	0	1	0	0	0	0	0	0
10185_52	4	9	0	0	5	0	5	0	0	0	4
10204_52	0	2	0	0	2	0	0	1	0	0	2
10215_52	0	0	0	0	0	0	0	0	0	0	0
10221_52	1	1	0	0	2	0	2	0	0	0	2
10223_52	0	1	0	0	2	0	1	0	0	0	1
10225_52	0	4	0	0	5	0	0	2	0	0	0
10225_76	0	6	0	0	3	0	1	1	0	0	1
10248_52	0	1	0	0	1	0	0	1	0	0	1
10280_52	0	3	0	0	2	0	2	1	0	0	1
10290_52	0	1	0	0	3	0	0	1	0	0	2
10301_52	0	1	0	0	2	0	1	0	0	0	0
10334_52	0	3	0	0	4	0	1	1	0	0	1
10344_52	0	5	0	0	5	0	2	0	0	0	4
10344_76	1	2	0	0	4	0	2	2	0	0	2

Topic Modeling Analysis

The purpose of this study is to determine the ways in which topic modeling could be an effective tool to identify themes without traditional qualitative coding. There were three topic modeling approaches used in this study, referred to as Method 1, 2, and 3.

Method 1

Method 1 used an unsupervised learning method, using a computer algorithm to determine a list of unknown topics without input from the researcher. Though the number of topics must be predetermined, which topics are chosen is entirely determined by the topic modeling algorithm. A LDA model was used to assigned topics to the DTM described above

(Silge & Robinson, 2019). LDA considers every document as a mixture of topics and every topic a mixture of words (Silge & Robinson, 2019). This means that for any document the LDA model may deduce that the terms in document A are 60% from topic 1 and 40% from topic 2, not assigning each document only one topic. Each topic is made up of a mixture of words, which can also overlap. Topic 1 may have the words *bell, ring, and chime* and Topic 2 could have *married, ring, and partner* LDA is a mathematical model that determines the likelihood of a document relating to each topic while simultaneously determining the likelihood that a word belongs to a topic (Silge & Robinson, 2019). The *LDA* function uses probabilistic functions to determine a beta value and gamma value using a predetermined number of topics. The beta value is the probability that a single word belongs to a topic. The gamma value is an estimated proportion of words from each document that belong to a topic (Hornik & Grün, 2011).

Identifying topics using method 1. Method 1 used the *LDA* function in the *topicmodels* (Grun et al., 2021) package in R to assign a mixture of words to each topic as well as assign a topic to each document (Silge & Robinson, 2019). The *LDA* function requires a DTM and a user assigned number of topics, *k*. For this function, the DTM constructed for the discussion posts was an input as well the number of topics, k = 6, which was based on the results of the *FindNumberTopics* function that is part the of the *ldatuning* package (v1.0.2; Nikita & Chaney, 2020).

Themes for the six topics were determined using their respective "bag of words"—the top 20 words with the highest beta value. Each bag of words was analyzed to identify a motivating theme. These were distributed to other mathematics and statistics education researchers to ask for their interpretation. Six volunteer researchers read the 20 words associated with each topic and completed the prompt "This group of participants is motivated to take this course because…" Using these responses, as well as knowledge of the goals of the course, and reading many of the discussion posts prior to analysis, we decided on themes for each topic. The themes identified will be discussed in the results section.

The *LDA* function also assigns each document and its assigned topic a gamma value. The gamma value is the proportion of words from each document that are generated from the assigned topic (Silge & Robinson, 2019). The higher the gamma value, the higher the probability that the document aligns to a given topic. The *LDA* function can assign a document to more than one topic. For instance, a gamma value of 0.55 for topic 1 and 0.45 for topic 2 would show that about 55% of the words in the document are generated from topic 1 and 45% from topic 2.

It was decided to include topic assignments that had a gamma value greater than 0.5 for each participant. For the purposes of this research, we were interested in the one topic most likely associated with a participant. The topic for each participant was recorded. Of the 946 documents (collection of posts), all but six had a gamma value greater than 0.5. Thus 940 documents were assigned topics. Table 1 shows the topics generated using Method 1. For brevity, the 10 top words for each topic are shown. The title and theme of each topic were determined after trying to make sense of the bag of words applied to each topic.

Method 1 validation. After each document was assigned a topic, then each document was read to determine if the topic assigned was appropriate. Of the 940 documents assigned to a topic, 573 participants' posts, or 61%, were assigned to a topic that seemed appropriate for that collection of posts.

Table **2** shows the percentage of times it was determined each document was assigned correctly to a topic. Topic 6 was assigned "Yes" the lowest percentage of the time, with about 50% of the documents applying to the topic assigned. Topic 1 had the highest percentage of agreement, with 70% of the documents assigned appropriately.

Table 1
Topics for Motivation to Enroll Identified Using Method 1

Topic	Bag of Words	Theme for Motivation Described by Researchers	Description This group of participants...
1	Statist, teach_learn, student, teach_statist, excited, interested, class, educate, love, mooc	Teach and understand statistics using data	... is excited and interested in learning to teach and understand statistics with data.
2	statist, teach, learn, student, stat, teach_statist, class, data, teacher, curriculum	Preparing to teach new curriculum that uses statistics	...is interested in learning and teaching statistics (using technology and data) especially as it pertains to new curriculum.
3	learn, student, statist, data, teach, class, engag, experi, understand, teacher	Teach with data, make class engaging, and interact with others	... is interested in learning how to teach students using data and make the class more engaging and interesting.
4	statist, teach, learn, student, math, understand, mooc, knowledge, al, improv	Improve teaching/knowledge of statistics by incorporating data and technology	….is interested in learning how to teach statistics, to improve their knowledge of teaching statistics and excited to incorporate interesting data and technologies.
5	statist, teach, student, year, learn, school, high, high_school, interest, ap	Preparing to teach high school students, particularly AP students	... is preparing to teach high school students this year, particularly AP students, and wants to learn ideas to engage and interest students.
6	teach, statist, student, learn, teach_statist, class, year, stat, time, idea	Looking to get new ideas and resources to prepare for the upcoming year	... is excited to learn to teach statistics for this upcoming year and gain new ideas and resources.

Table 2
Method 1 Topics and Validity Count and Percentages

Topic	Researcher Determined Theme for Motivation	Total Documents Assigned Topic	Topic Correct by Qualitative Coding?			
			Yes		No	
			n	%	n	%
1	Teach and understand statistics using data	182	128	70.3%	54	29.7%
2	Preparing to teach new curriculum that uses statistics	155	93	60.0%	62	40.0%
3	Teach with data, make class engaging, and interact with others	168	107	63.7%	61	36.3%
4	Improve teaching/knowledge of statistics by incorporating data and technology	141	87	61.7%	54	38.3%
5	Preparing to teach high school students, particularly AP students	143	83	58.0%	60	42.0%
6	Looking to get new ideas and resources to prepare for the upcoming year	151	75	49.7%	76	50.3%
Total		940	573	61.0%	367	39.0%

The following illustrates an example of the validation process for a participant's collection of posts. Participant 13262_58 posted the following:

> I hold a Masters in Curriculum and Instruction and am completing this course because I despise numbers, despite the fact that I'm quite good with them. I tend to face my fears, lol.

Method 1 assigned this document Topic 4, "improve teaching/knowledge of statistics by incorporating data and technology." This post does not mention anything about this individual wanting to improve their teaching or knowledge of statistics. Though it may be implied that this person is trying to improve their knowledge of statistics, every effort was given to apply validation on what was written, not what was implied. This topic assignment was not correct.

Method 2

Method 2 used a semi-supervised learning method, seeded topic modeling, to determine the topics specific to motivation. Qualitative methods were used to create a list of topics for motivation based on a sample of randomly chosen posts to create a seeded dictionary of topics.

Determining topics for the seeded dictionary. For Method 2 a seeded word dictionary was needed to create predefined topics to run a supervised topic model. The dictionary was created using a priori coding and in vivo coding (Creswell, 2013) to identify themes for motivation based on 10% of the discussion forum posts. Since there were 1,639 original posts from the introductory discussion forums, 164 random posts were chosen to code to identify motivation themes. The 164 posts were a stratified random sample of the 10 courses based on the percentage of posts to this forum in each course.

A priori codes were based on questions that were asked in the enrollment survey. Additional codes were created as the posts were read. The codes were combined to identify the themes for motivation (topics) as well as words to seed each topic. This resulted in 14 topics for the seeded dictionary (Table 3). For example, one topic was titled “confidence.” Those participants were motivated by an opportunity to increase their own confidence to teach statistics. The seed words for this topic were *confidence*, *confident*, and *build** (the * indicates the stem is used as the seed word).

Instead of defining the DTM as it was for Method 1, a data frame was made with a list of rows of two variables, userid and post. This data frame was then converted to a document feature matrix (DFM), which is the acceptable input for the *seededlda* function. Like the DTM from Method 1, the DFM has the rows of the matrix as the participants and the columns are all the words that appear in the corpus of posts. The dictionary and the DFM were fed into an LDA function that is part of the *seededlda* package to assign each participant a topic.

Assigning and validating topics per participants. To assign topics, the *textmodel_seededlda* function in the *seededlda* package was used (Watanabe & Xuan-Hieu, 2020). The function assigns each topic (see
Table) to a user based on the frequency of times the words appear in the DFM. The *seededlda* function returns a list of words that define each topic for each user. This will include the seed words from the dictionary as well as other words that fit into the theme based on the likelihood that each topic produces each term (Watanabe & Xuan-Hieu, 2020).

Like Method 1, the posts for each user were read and determined whether the topic assigned was appropriate. The results of this validation process were recorded for the overall posts as well as for each topic.
Table shows the number of documents assigned to each of the 14 topics for Method 2 and the number of valid assignments for each topic. Of the 946 documents, 463, or 48.9%, were

considered to have an appropriate topic assigned to them by Method 2. Additionally, 483, or 51.1%, were not considered to be assigned correctly.

Table 3
Method 2 Topics and Validity Count and Percentages

Topic	Title	Total Documents Assigned Topic	Topic Correct by Qualitative Coding?			
			Yes		No	
			n	%	n	%
1	library of resources	99	68	68.7%	31	31.3%
2	collaborate	88	63	71.6%	25	28.4%
3	repeater	88	25	28.4%	63	71.6%
4	students reasoning	78	25	32.0%	53	68.0%
5	learn statistics	73	37	50.7%	36	49.3%
6	confidence	71	29	40.9%	42	59.1%
7	engaging class	67	27	40.3%	40	59.7%
8	requirement	61	13	21.3%	48	78.7%
9	pedagogy	60	38	63.3%	22	36.7%
10	technology	59	24	40.7%	35	59.3%
11	professional practice	58	35	60.3%	23	39.7%
12	preparing	54	39	72.2%	15	27.8%
13	real data	48	29	60.4%	19	39.6%
14	stats investigation	42	11	26.2%	31	73.8%
	Total	946	463	48.9%	483	51.1%

The following example illustrates how validity was determined. Participant 11997_58 posted "Statistics is not a strength of mine. I do not want my students to struggle because their teacher struggles with concept."

Method 2 assigned this document to Topic 4, "students' reasoning." Since the post includes the word "students" it makes sense why the assignment was made, but this participant is clearly struggling with their own confidence, not with how students are reasoning with statistics. This post was not assigned a valid topic.

Method 3

Method 3 generated topics based on a condensed version of topics from Method 2. The number of overall topics in Method 2 was higher than Method 1 (14 topics versus 6), which could have led its lower validity rate.

Topics and validation. The topics from Method 2 (**Error! Reference source not found.**) were collapsed into four topics (
). The first motivation topic in Method 3 grouped six topics from Method 2 (real data, technology, stats investigation, students reasoning, library of resources, collaborate) together that were all related to course objectives. When enrolling, participants could see the course objectives. We assumed that some people were motivated by these visible goals.

Table 4
Method 3 Topics and Validity Count and Percentages

Topic	Title	Total Documents Assigned Topic	Topic Correct by Qualitative Coding?			
			Yes		No	
			n	%	n	%
1	Course specific goals	414	367	88.6%	47	11.4%
2	Continuing professional learning	207	113	54.6%	94	45.4%
3	Learn statistics/increase confidence	144	81	56.3%	63	43.7%
4	Pedagogical goals	181	156	86.2%	25	13.8%
	Total	946	717	75.8%	229	24.2%

The second topic combines goals specific to continuing professional learning (professional practice, requirement, repeater). Since these courses give people the opportunity to earn continuing education credits, some people may enroll for professional learning goals outside of specific course goals or take a course again. Another common theme in discussion forums is for participants to want to become better teachers, hence the third topic of pedagogical goals (engaging class, preparing, pedagogy). The fourth topics centers on those who want to learn statistics/increase confidence (learn statistics, confidence). These could be separate topics, but when reading posts, it was found they often occur together. We recognize that other groupings may have been appropriate. The 14 topics were collapsed based on the researchers' prior experience with the data and course context.

The topic a participant was assigned in Method 2 carried over to Method 3. Then each topic was renamed to the appropriate topic in

. If the topic applied to a participant in Method 2 was valid, it remained valid in Method 3. The participant documents that were not considered valid in Method 2 were then reassessed for the validity based on the new topics. Of the 946 documents, 75.8% were determined to be assigned to an appropriate topic and 24.2% were not.

Discussion

If the goal of this study was to determine which method was better at assigning topics, then we could say that Method 3 is “best” since it has the highest validity score. It is of interest that Nelson et al. (2021) found that the dictionary-based method they used was not the most aligned to hand-coded data, suggesting “dictionary methods will struggle with the identification of broader concepts but can play a role when specific phrases are of interest” (p. 228).

Instead, the goal of this study was to investigate how topic modeling can be used for analyzing qualitative data, particularly analyzing the motivation of participants to enroll in OPD for statistics teachers using discussion forum data. Badaldi et al. (2022) identified 50 articles that sought to identify motivation, none of which used discussion forum posts as a source of data. None of these articles used topic modeling as a means to identify motivation. By comparing the different topic modeling methods to qualitative analysis results, this study suggests that topic modeling can be a useful tool for qualitative researchers in their analysis process. Analyzing qualitative data is a “process of bringing order, structure, and meaning to a mass of collected data” (Marshall & Rossman, 1990, p. 111). Though qualitative data analysis produces rich and informative results, the process can be tedious, time-consuming, and messy (Creswell, 2013; Hilal & Alibri, 2013). This is even truer now in our world of large data.

The unsupervised technique in Method 1 was used to identify themes for motivation using a computer algorithm, rather than researchers reading every post to identify themes. Though the validity score for Method 1 was not very high (61%), the themes identified gave good insight into why people enrolled in these courses, outside of the choices they indicated on an enrollment survey. For instance, the two top assigned topics were “Teach and understand statistics using data” and “Preparing to teach new curriculum that uses statistics.” From this, we can encourage those that are designing OPD for statistics educators to create content that centers around data.

We also believe that anyone with access to large amounts of discussion forum data could use these techniques to lessen the work of traditional coding methods. Methods 2 and 3 used a semi-supervised learning method to assign predetermined themes for motivation to individuals. Methods 2 and 3 required the input of a seeded dictionary. This is not unlike traditional qualitative methods, where a group of researchers may code a subset of the data, identify, and define themes, then create a “codebook”, and apply those codes to the rest of a dataset (Roberts et al., 2019). We created a codebook with the seeded dictionary, then let the computer algorithm code the remainder of the data. This semi-supervised learning method is particularly useful when there is a lot of data, but not a lot of research capacity (i.e., people hours) to apply the codes to a large dataset. Though the validity scores for Method 2 were very low (~49%), the validity for Method 3 rose to approximately 76%.

The seeded dictionary for the approach used in Methods 2 and 3 was created using the expert knowledge of the first two authors. Others have attempted to create seeded dictionaries from more diverse knowledge sources. For instance, Resnik et al. (2015) qualitatively analyzed 6,459 stream of consciousness essays written by college students about depression. Using this prior research, they then created a seeded dictionary describing general themes about depression. They used this dictionary to analyze a series of random Tweets about depression and determined the topics identified by college students were useful in identifying themes for the general public. Perhaps other researchers may find it useful to created seeded dictionaries from large datasets that have already been qualitatively analyzed to then attempt to model other datasets, rather than letting a random sample of the data inform the dictionary as was done in this study.

Limitations

Several limitations to this study should be taken into consideration. This study used the *topicmodels* and *seededlda* packages in R, there are many more methods available not only in R but also in other statistical software tools. The discussion forum posts that were used as the dataset included only text. There were posts that included pictures, hyperlinks, or other html inputs such as emojis, that were not part of the data analyzed.

The Porter stemmer method was used to stem words in the creation of the DTM in Method 1 and DFM in Methods 2 and 3. The Porter stemmer method is susceptible to over-stemming words or causing faulty conflation of words (Farrar & Hayes, 2019; Krovetz, 1993), meaning that words seem to appear more often because they were shortened so much. There are other methods that could have been used, such as the Krovetz method which attempts to help this over-stemming process but is also known to under-stem words (Farrar & Hayes, 2019). This study acknowledges the limitations of the Porter method in the data cleaning process of the discussion forum data.

Conclusions

Isoaho et al. (2021) states that many studies that employ topic modeling interpret results of the topics in isolation from the documents used to produce the topics. This study did not make that mistake. The topics produced and validity of the model were interpreted and evaluated with the context of the data always present. The researchers' knowledge of the course and experience in the context of statistics teaching was critical in making decisions related to all aspects of the process. We assert the context of the textual data used to produce topic models must always be the biggest consideration in every step of the process, especially when interpreting and sharing results.

Ability to replicate a study is often hard to do for any qualitative analysis, since researchers do not often share the steps of how codebooks are made or how thematic coding is applied (Roberts et al., 2019). It is the hope that enough detail is provided in this article so that the topic modeling methods can be replicated and built upon so that topic modeling can become a useful tool in analyzing online discussion forum data.

Ethics Board Approval

The data used in this study was approved for research for use through the institutional review board of the institution of the second, third, and fourth author. All data was blinded prior to analysis.

References

Lee, H., & Stangl, D. (2015). Taking a chance in the classroom: Professional development MOOCs for teachers of statistics in K–12. *CHANCE*, *28*(3), 56–63. https://doi.org/10.1080/09332480.2015.1099368

Hollebrands, K.F., & Lee, H. S. (2020). Effective design of massive open online courses for mathematics teachers to support their professional learning. *ZDM*, 52, Feb. 2020, 1–17. https://doi.org/10.1007/s11858-020-01142-0

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha, M. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran, and Vikram Pudi (Eds.), *Advances in knowledge discovery and data mining* (pp. 391–402). Springer. http://doi.org/10.1007/978-3-642-13657-3_43

Badali, M., Hatami, J., Banihashem, S. K., Rahimi, E., Noroozi, O., & Eslami, Z. (2022). The role of motivation in MOOCs retention rates: A systematic literature review. *Research and Practice in Technology Enhanced Learning*, *17*(5). https://doi.org/10.1186/s41039-022-00181-3

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3*(30), 774. doi: 10.21105/joss.00774, https://quanteda.io.

Bouchet-Valat, M. (2020). *Package 'SnowballC'. R package version 0.7.0.* https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf

Boroujeni, M. S., & Dillenbourg, P. (2019). Discovery and temporal analysis of MOOC study patterns. *Journal of Learning Analytics*, *6*(1), 16–33. http://dx.doi.org/10.18608/jla.2019.61.2

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change, 36*, 89–100. https://doi.org/10.1016/j.gloenvcha.2015.12.001

Brooker, A., Corrin, L. de Barba, P., Lodge, J., & Kennedy, G. (2018). A tale of two MOOCS: How student motivation and participation predict learning outcomes in different MOOCS. *Australasian Journal of Educational Technology, 34*(1), 1–15. https://doi.org/10.14742/ajet.3237

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing. *16th European Symposium on Artificial Neural Networks 2008*, *72*(7–9): 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

Creager, J. H., Wiebe, E. N., & Kellogg, S. B. (April, 2018). Time to shine: Extending certificate deadlines to support open online teacher professional development. Presented at *AERA Annual Meeting*, New York, NY.

Creswell, J. (2013). Qualitative inquiry and research design: Choosing among five approaches. Sage.

Das, A., Shrivastava, M., & Chinnakotla, M. (2016). Mirror on the wall: Finding similar questions with deep structured topic modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 454–465). Springer. https://doi.org/10.1007/978-3-319-31750-2_36

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course" reconceptualizing educational variables for massive open online courses. *Educational Researcher*, *43*(2), 74–84.

Douglas, K. A., Bermel, P., Alam, M. M., & Madhavan, K. (2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192, http://dx.doi.org/10.18608/jla.2016.33.9.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of Psychology*, *53*(1), 109–132.

Eriksson, T., Adawi, T., & Stöhr, C. (2017). "Time is the bottleneck": A qualitative study exploring why learners drop out of MOOCs. *Journal of Computing in Higher Education*, *29*(1), 133–146. doi: 10.1007/s12528-016-9127-8

Ezen-Can, A., Boyer, S.E, Kellogg, S., & Booth, S. (2015, March). Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 146–150). https://doi.org/10.1145/2723576.2723589

Farrar, D., & Hayes, J. H. (2019). A comparison of stemming techniques in tracing. In *2019 IEEE/ACM 10th International Symposium on Software and Systems Traceability (SST*). doi: 10.1109/SST.2019.00017

Franklin, C., Bargagliotti, A., Case, C., Kader, G., Scheaffer, R., & Spangler, D. (2015). *The statistical education of teachers.* American Statistical Association.

Frankowsky, M. H., Wiebe, E., Thompson, I., & Behrend, T. (2015). *Data analytics for modeling user behavior within MOOCs: A comparison of clustering techniques*. AERA 2015 Annual Meeting, Chicago, IL, United States. https://www.aera.net/Events-Meetings/Annual-Meeting/Previous-Annual-Meetings/2015-Annual-Meeting/2015-Annual-Meeting-Details

Gao, F., Wang, C., & Sun, Y. (2009). A new model of productive online discussion and its implications for research and instruction. *Journal of Educational Technology Development and Exchange*, *2*(1), 65–78. https://scholarworks.bgsu.edu/vcte_pub/25

Gao, F., Zhang, T., & Franklin, T. (2013). Designing asynchronous online discussion environments: Recent progress and possible future directions. *British Journal of Educational Technology*, *44*(3), 469–483. https://doi.org/10.1111/j.1467-8535.2012.01330.x

Garrison, D. R., Anderson, T., Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *The American Journal of Distance Education*, *15*(1), 7–23. https://doi.org/10.1080/08923640109527071

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, suppl 1: 5228–5235. http://www.pnas.org/content/101/suppl_1/5228.full

Grün, B., Hornik, K., Blei, D.M., Lafferty, J.D., Phan, X., Matsumoto, M., Nishimura, T., & Cokus, S. (2021). *Package 'topicmodels'. R package version 0.2-12*. https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

Hammer, D., & Berland, L. K. (2014). Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences*, *23*(1), 37-46. https://doi.org/10.1080/10508406.2013.802652

Hara, N., Bonk, C., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science*, *28*, 115–152. https://doi.org/10.1023/A:1003764722829

Hilal, A. H., & Alabri, S. S. (2013). Using NVivo for data analysis in qualitative research. *International Interdisciplinary Journal of Education, 2*(2), 181–186.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software, 40*(13), 1–30. http://www.jstatsoft.org/v40/i13

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine learning*, *95*(3), 423–469. DOI 10.1007/s10994-013-5413-0

Huang, W. (2018). PhraseCTM: Correlated topic modeling on phrases within Markov random fields. In Proceedings of the *56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), p. 521-526, 56th ACL Meeting, Melbourne, Australia. https://aclanthology.org/volumes/P18-1/

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

Joanes, T., & Doane, W. (2019). *textmineR: Functions for text mining and topic modeling. R package version 3.0.4*. https://cran.r-project.org/web/packages/textmineR/textmineR.pdf

Kellogg, S., Booth, S., & Oliver, K. (2014). A social network perspective on peer supported learning in MOOCs for educators. *International Review of Research in Open and Distributed Learning*, *15*(5), 263–289.

Kop, R., Fournier, H., Sui, J., & Mak, F. (2011). A pedagogy of abundance or a pedagogy to support human beings? Participant support on massive open online courses. *The International Review of Research in Open and Distance Learning*, *12*(7), 74–93. https://doi.org/10.19173/irrodl.v12i7.1041

Krovetz, R. (1993). Viewing morphology as an inference process in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '93, p. 191 – 202. SIGIR93: 16th International ACM/SIGIR '93 Conference on Research and Development in Information Retrieval, Pittsburgh, PA. https://dl.acm.org/doi/proceedings/10.1145/160688

Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education*, *29*, 40-48.

Marshall, C., & Rossman, G. (1990). *Designing qualitative research.* Sage.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW),* 1–23. https://doi.org/10.1145/3359174

Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist MOOCs. *Journal of Online Learning and Teaching*, *9*(2), 149–159.

Moore, R. L., & Wang, C. (2021). Influence of learner motivational dispositions on MOOC completion. *Journal of Computing in Higher Education*, *33*(1), 121–134. https://doi.org/10.1007/s12528-020-09258-8

Nandi, D., Hamilton, M., & Harland, J. (2012). Evaluating the quality of interaction in asynchronous discussion forums in full online classes. *Distance Education*, *33*(1), 5–30.

https://doi.org/10.1080/01587919.2012.667957

Nelson, L. K., Burk, D., Knudsen, M. & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, *50*(1), 202–237. https://doi.org/10.1177/0049124118769114

Nikita, M., & Chaney, N. (2020). *Tuning of the Latent Dirichlet Allocation model parameters. R package version 1.0.2*. https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: Behavioural patterns. *EDULEARN14 Proceedings*, *1*, 5825–5834.

Porter, Martin F. 1980. An Algorithm for Suffix Stripping. *Program 14*(3): 130–37.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. Ninth workshop on on innovative use of NLP for building educational applications, Baltimore, Maryland. https://aclanthology.org/W14-1804.pdf

Reich, J., Stewart, B., Mavon, K., & Tingley, D. (2016). The civic mission of MOOCs: Measuring engagement across political differences in forums. Third (2016) ACM Conference on Learning@ Scale, Edinburgh, Scotland. https://doi.org/10.1145/2876034.2876045

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. The 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado. https://doi.org/10.3115/v1/W15-1212

Roberts, K., Dowell, A., & Nie, J. B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC medical research methodology*, *19*(1), 1–8. https://doi.org/10.1186/s12874-019-0707-y

Schmiedel, T., Müller, O., & vom Brocke, J. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, *22*(4), 941–968. https://doi.org/10.1177/1094428118773858

Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, *4*, 287–300.

Silge, J., & Robinson, D. (2019). *Text mining with R: A tidy approach*. O'Reilly.

Tang, H., Xing, W., & Pei, B. (2018). Exploring the temporal dimension of forum participation in MOOCs. *Distance Education*, *39*(3), 353–372. doi:10.1080/01587919.2018.1476841

Vytasek, J. M., Wise, A. F., & Woloshen, S. (2017). Topic models to support instructors in MOOC forums. Seventh international learning analytics & knowledge conference, Vancouver, Canada. https://doi.org/10.1145/3027385.3029486

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. *Seventh IEEE international conference on data mining (ICDM 2007)* Washignton, D.C. https://dl.acm.org/doi/proceedings/10.5555/1335998?id=31

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rose, C. (2015). Investigating how students' cognitive behavior in MOOC discussion forums affect learning gains. *Proceedings of the 8th International Conference on Data Mining*, 226-233.

Watanabe, K. & Xuan-Hieu, P. (2020). *Package 'seededlda'. R package version 0.5.1*.

https://cran.r-project.org/web/packages/seededlda/seededlda.pdf

Wilkowski, J., Deutsch, A., & Russell, D. M. (2014, March). Student skill and goal achievement in the mapping with google MOOC. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 3-10).

Wong, A. W., Wong, K., & Hindle, A. (2019). Tracing forum posts to MOOC content using topic analysis. arXiv preprint arXiv:1904.07307.

Wu, Z., Lei, L., Li, G., Huang, H., Zheng, C., Chen, E., & Xu, G. (2017). A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, *84*, 12–23.

Xiong, Y., Li, H., Kornhaber, M. L., Suen, H. K., Pursel, B., & Goins, D. D. (2015). Examining the relations among student motivation, engagement, and retention in a MOOC: A structural equation modeling approach. *Global Education Review, 2*(3), 23–33.