AUTOMATICALLY LINKING DIGITAL SIGNAL PROCESSING ASSESSMENT QUESTIONS TO KEY ENGINEERING LEARNING OUTCOMES

S. Supraja[⋆], *Sivanagaraja Tatinati*[⋆†], *Kevin Hartman*[††], *and Andy W. H. Khong*[⋆]

[⋆] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[†] Delta-NTU Corporate Lab, Nanyang Technological University, Singapore
[††] CRADLE, Nanyang Technological University, Singapore

ABSTRACT

To deliver on the potential outcome-based teaching and learning holds for engineering education, it is important for engineering courses to provide students with different types of deliberate practice opportunities that align to the program's learning outcomes. Working from these requirements, we increased the design and measurement intentionality of a digital signal processing (DSP) course. To align the course's learning outcomes more constructively with its assessment measures, we automated the process of classifying DSP questions according to learning outcomes by introducing a model that integrates topic modeling and machine learning. In this work, we explored the effect of pre-processing procedures in terms of stopword selection and word co-occurrence redundancy issue in question classification inferences. In this work, we proposed a customized variant of the Word Network Topic Model, q-WNTM, which is able to use its pre-classified DSP questions to reliably classify new questions according to the course's learning outcomes.

Index Terms— Learning outcomes, assessment, topic modeling, extreme learning machine

1. INTRODUCTION

To comply with the Accreditation Board for Engineering and Technology's (ABET) accreditation criteria and prepare students for the workforce, all engineering programs must ensure students complete courses that collectively develop eleven categories of learning outcomes [1]. To help students demonstrate these outcomes by the end of a program, courses implement learning activities that rely on remembering concepts, applying existing knowledge to tackle problems, and generating tailored solutions to real-life scenarios [2]. As a required course in many electrical engineering programs, the design of digital signal processing (DSP) courses is crucial for achieving compliance with ABET's educational standards. The best practices of outcome-based teaching and learning suggest that course designs should identify the learning outcomes and the assessments that measure those learning outcomes before designing the course's learning activities [3]. For DSP courses, learning activities can include building circuits [4], Matlab programming [5] and laboratory experiments [6] which can be implemented to fulfill a set of learning outcomes. For a host of historical, structural, and policy reasons, the design of DSP courses often deviates from the best practices. Many courses were originally designed decades ago and incrementally updated with new content and assessment items. This slow evolution often translates to the measurement of student outcomes being grafted onto a course that was originally designed for its coverage of content [7].

While aligning deliberate practice opportunities to assessments is crucial, not all assessments and learning measures are equal [8]. Bloom's Taxonomy serves as a popular representation of how learning can be divided into categories that rely on the same domain information but different cognitive processes [9, 10, 11]. Using the convolution operation as an example, students can be asked to explain, calculate or create something as part of a deeper analysis. Integrating Bloom's Taxonomy into ABET's accreditation criteria creates a space that maps the assessment items to the learning outcomes. As a thought experiment, in this work, we focused on the space that deals with knowledge facts (*"K"*), applying a learned concept (*"A"*), and transferring the learnt concept to another domain (*"T"*). This space is obtained by merging the categories in Bloom's Taxonomy, for which the detailed explanation can be found in Section 2.2. Mapping every assessment item into our hypothetical learning outcome space would be difficult in terms of reliability and time invested. In our work, we focused on *"Is it possible to reliably classify if an assessment item aligns to a particular learning outcome?"* Hereafter, we use the word *"question"* as a proxy for assessment item.

Rule-based approaches have been adopted for question classification by combining parts-of-speech tagging, identifying verbs associated with Bloom's Taxonomy and recognizing the presence of particular punctuation marks to create features as inputs to machine learning algorithms [13, 14]. However, for a new or updated set of questions, it is observed that some questions fail to activate any of these rules [15]. To address the data dependency issue, term frequency-inverse document frequency (TF-IDF) is employed, which identifies the relevance of the words themselves by assigning weightages to each word in a question based on a representative corpus [16]. However, identical weightages assigned to different words may cause misinterpretation errors during classification according to learning outcomes. As a result, TF-IDF achieves high classification for a handpicked set of questions [12] and its performance is limited for imbalanced set of questions from a variety of sources.

In response to the limitations of these approaches, our previous work on automatic question classification employed text-based techniques (TF-IDF and latent Dirichlet allocation (LDA) [17]) to analyze the presence of words and features in a question followed by machine learning algorithms such as extreme learning machine (ELM) [18] to classify them. We hypothesize that the reasons for LDA to achieve limited performance are: 1) stopwords were not removed, hence topics were comprised of high-frequency words which

This work was conducted within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

led to uncertainty in identifying the accurate label, and 2) the sparsity of the document-level to word co-occurrences resulting from the short questions interfered with LDA [12]. The occurrences of words in short texts do not contribute as much to detecting word relationships when compared with long documents [19].

The inability of LDA to adequately classify short texts such as messages and tweets [20] prompted the development of the word network topic model (WNTM). WNTM models the distribution over topics for each word rather than directly learning the topics for each document as in LDA [21]. The semantic density of data is enriched and the global contextual information is made available through the word-word space [21]. Questions belong to the category of short texts, similar to tweets which were limited to 140 characters long, privileging WNTM over LDA for question classification [20].

WNTM classifies short texts by sieving out the content words in a document. Thus, WNTM cannot be directly applied to question classification for two reasons: 1) pre-processing stopword removal is different, and 2) redundancy of content-agnostic versus technical word combinations is not considered. With respect to the former issue, in traditional document classification, stopwords (question words, prepositions, articles, conjunctions, action verbs etc. [22]) are generally defined as high-frequency words which do not contribute to a document's subject matter. However, when classifying a question according to its learning outcome, common stopwords become key to determining the proper category. Thus, in this work, a list of stopwords which functionally did not contribute to classifying the questions is generated. With respect to the latter issue, to identify the learning outcome of a question, if we only consider the content-agnostic words such as *"what", "explain"*, we may not be able to identify the learning outcome without knowing the context reflected by the technical words and other content-agnostic words that surround it. Since we are not performing document classification based on content, but rather question classification according to learning outcomes, we sought to observe the relationship among content-agnostic words, as well as, between content-agnostic and technical words that correspond to the categories of learning outcomes. These two issues prompted us to further expand WNTM by customizing the pre-processing procedures to better fit our use case. In line with the above, we propose *"q-WNTM"* algorithm, a tailor-made WNTM for question classification.

By addressing limitations observed in prior work which include the lack of consideration of appropriate stopwords and word co-occurrence redundancy, we sought to provide a more reliable tool for DSP instructors to use when linking questions to their courses' learning outcomes. We picture the intended use case as follows: Say that an actual *"A"* or *"T"*-type question is misclassified as *"K"*-type. This misestimates a course's prioritization of memorization. The converse implies that students who merely memorized material demonstrated more outcomes than they should have. By focusing on word context, our model avoids such misinterpretations of student abilities in terms of course evaluation by minimizing misclassification. The tool-assisted course improvement process would lead to a more reliable matching between questions and learning outcomes. Such a system also lends itself to the generation of custom question sets students could use as deliberate practice opportunities. With this vision in mind, we have prioritized our model to minimize the number of falsely identified categories compared to existing techniques.

2. METHODS FOR QUESTION CLASSIFICATION

In this section, we describe our question dataset, the taxonomy we have adopted to categorize the questions, and our proposed technique for classifying questions according to learning outcomes.

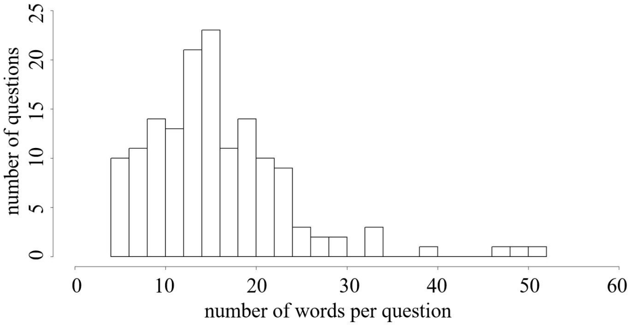

Fig. 1. Frequency distribution of length of DSP questions.

2.1. Dataset of questions

The corpus of 150 DSP questions underlaying this work aggregates questions published in well-known textbooks [23, 24, 25], obtained from online question banks and generated by an instructor of an undergraduate DSP course. The mean length of our questions was 16.2 words (SD = 8.01) or 88.54 characters (SD = 44.20). The frequency distribution of question length in terms of the number of words is depicted in Fig. 1. The prevalence of short questions confirms the need to forgo LDA and instead apply WNTM.

A subject matter expert manually classified all of the training questions that were passed into the extreme learning machine (ELM). To obtain a ground truth measure when evaluating the classification performance of the testing set, the same instructor manually classified each of the test questions. Although the questions covered a range of DSP topics such as discrete-time signals, discrete-time Fourier transform and z-transform, classification was done without any analysis of the content. The classification was conducted based solely on the learning outcome the instructor intended to measure with each question.

2.2. Customized set of learning outcomes

As a proof of concept, we concerned ourselves with ABET's content-based learning outcomes. We collapsed all of these outcomes into a single dimension. We then used a reduced version of Bloom's Revised Taxonomy to stratify the content related outcomes. The taxonomy starts with the recollection of information at the lowest level, ascends to the application of knowledge, and peaks with creative outcomes [26]. The reduction to Bloom's Revised Taxonomy reflects the philosophy of the DSP course instructor who viewed the different levels of reasoning about course content as pertaining to knowledge, application and transfer. The framework of cognitive levels we adopted is depicted in Fig. 2. We combined the lowest two levels in Bloom's Revised Taxonomy into *"Knowledge"* (*"K"*), retained the subsequent level as *"Application"* (*"A"*) and combined the top three levels into *"Transfer"* (*"T"*). These three categories form the basis of our analysis and are consistent with ABET's engineering education accreditation criteria.

To illustrate the classification process with examples from our DSP question dataset, Fig. 2 shows an example question for each of the learning outcomes. Generally, *"K"*-type questions require students to recall and understand DSP facts and information. *"A"*-type questions require students to apply their DSP knowledge to solve a closely related problem. Lastly, *"T"*-type questions require students

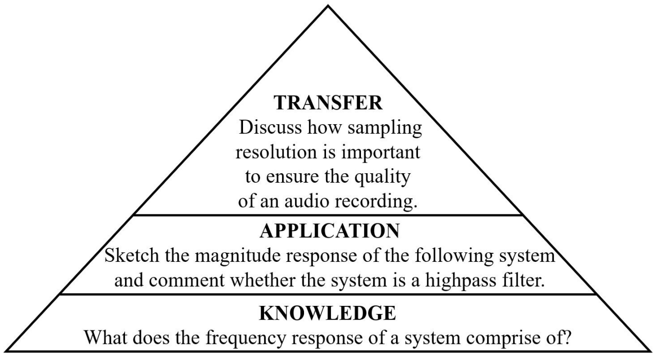
TRANSFER
Discuss how sampling resolution is important to ensure the quality of an audio recording.
APPLICATION
Sketch the magnitude response of the following system and comment whether the system is a highpass filter.
KNOWLEDGE
What does the frequency response of a system comprise of?

Fig. 2. Our custom learning outcomes framework.

to transfer their understanding of DSP principles to analyze, evaluate, and generate real-life situations not presented in the learning materials.

2.3. Proposed q-WNTM model

Our proposed model includes a three-step pre-processing approach and a question classification technique.

2.3.1. Text pre-processing

We began with basic data cleaning. Since we only focused on the questions' text, we removed symbols, diagrams, equations, numbers and punctuation marks. We ensured all remaining characters were in lower case before removing stopwords. In line with our use-case to preserve the essence of a question, we removed only four words *"the", "and", "a", "an"*. These words do not affect a question's context. Apart from these four words, we created a separate list of content-agnostic words that included all other general stopwords such as *"how", "state", "why"* etc. The corpus's remaining words such as *"DFT", "filter", "FIR"* are considered technical words. With the above procedure, there were a total of 109 unique content-agnostic words and 437 unique technical words in our dataset. Segregating content-agnostic and technical words from each other ensures that our proposed question classification technique ignores the relationships between the presence of technical words and provides each question's contextual information.

2.3.2. Question classification technique

Our proposed model employs WNTM [21] to generate the topic probabilities for each question. After generating the weighted word co-occurrence network, the adjacency lists for every unique word in the corpus are constructed. Unlike the conventional WNTM algorithm, for every technical word present in the corpus, the other co-occurring technical words in the corresponding adjacency lists are removed. This removal ensures that there will only be combinations of 1) content-agnostic with content-agnostic and 2) content-agnostic with technical words (and vice-versa). These resulting lists allow us to perform context-based instead of content-based classification.

After forming the adjacency lists for every unique word present in the corpus, we treat the lists as a new set of documents and apply the standard LDA Gibbs sampling to iterate through the word-topic allocation counts and topic-adjacency list allocation counts [27]. Each topic *t* generated in this model is a multinomial distribution ϕ_t over the vocabulary of words, with a symmetric Dirichlet prior β. Similarly, each new adjacency list *a* generated by our model is a multinomial distribution θ_a over the topics, with a symmetric Dirichlet prior α. After obtaining the topic probabilities for each adjacency list corresponding to the global set of co-occurrence relationships for each word, the topic probabilities for each original question based on every individual word w_j are inferred as

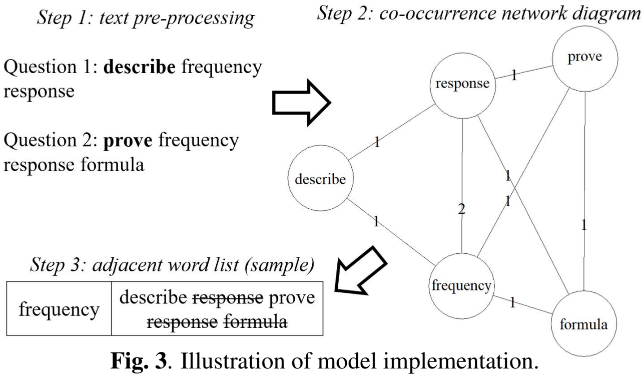
Step 1: text pre-processing
Question 1: describe frequency response
Question 2: prove frequency response formula
Step 2: co-occurrence network diagram
response
prove
describe
frequency
formula
1
1
1
1
2
1
1
1
1
Step 3: adjacent word list (sample)
frequency
describe ~~response~~ prove ~~response formula~~

Fig. 3. Illustration of model implementation.

$$P(t|q) = \sum_{w_j} P(t|w_j)P(w_j|q), \tag{1}$$

where $P(t|q)$ refers to the probability of every topic *t* in each question *q*, $P(t|w_j)$ refers to the probability of every topic *t* in each adjacency list belonging to each word w_j (θ_a), and $P(w_j|q)$ refers to the frequency of each word in the original question divided by the total number of words in that question. The vector of N topic probabilities for each question forms the input to the ELM which eventually labels these vectors to the learning outcomes space.

2.4. Implementation

To highlight the above procedure, Fig. 3 illustrates the steps taken to convert a sample set of two questions into adjacency lists. The words presented in boldface in Step 1 refer to each question's pre-defined set of content-agnostic words. After constructing the weighted network of word co-occurrences in Step 2, the adjacent words co-occurring with each word are identified. The string of adjacent words for the technical word *"frequency"* is shown in Step 3. Since we wanted to determine the difference in topic probabilities between including and excluding technical words in the adjacency list of a technical word, the other technical words in the adjacency list of *"frequency"* are removed (struck out).

To illustrate the impact of our choice of stopwords, given a question *"what does the frequency response of a system comprise of"*, the phrase *"what does"* serves as a signifier of its category *"K"*. If standard stopword removal is applied, it will appear as *"frequency response system"*, and the learning outcome becomes difficult to identify. Conversely, our choice of stopword removal transforms the question into *"what does frequency response of system comprise of"* which does not limit the ability of the algorithm to categorize it as a *"K"*-type question. Various examples from our dataset can also be used to illustrate the importance of context. With reference to content-agnostic words, *"explain what"* signifies a different category than *"explain why"*. With respect to content-agnostic and technical words, *"find DTFT"* differs from *"describe DTFT"*. However, when comparing *"phase response"* and *"magnitude response"*, the significance of co-occurring technical words is of less importance when determining the learning outcome expressed by a question.

3. RESULTS AND DISCUSSION

The objective of our experiments is to show how our model can be useful for DSP instructors to classify questions with minimal false

Table 1. Comparison of F1 scores.

Method	K	A	T	Average	s.d.
TF-IDF [16]	0.857	0.513	0.333	0.583	0.218
LDA [17]	0.444	0.941	0.737	0.707	0.204
WNTM [21]	0.545	0.800	0.848	0.744	0.133
q-WNTM	**0.700**	**0.903**	**0.923**	**0.848**	**0.101**

"s.d." refers to the standard deviation among the three categories.

classifications, and with the intention of providing the right set of practice opportunities to students. To achieve this, we explored the impact of word combination redundancy. We compared our proposed q-WNTM model with TF-IDF, LDA and WNTM.

3.1. Hyperparameters selection

The hyperparameters that require optimal initialization for topic models are the number of topics (N), prior for document/adjacency list to topic probabilities (α), and prior for topic to word probabilities (β). An empirically determined N = 10, α = 0.1 and β = 0.01 were selected as the optimal parameters for our dataset. We confined the number of Gibbs sampling iterations to 2000. For TF-IDF, the traditional bag-of-words approach results in each question being represented as a vector of weightages for the entire vocabulary including irrelevant zeroes and insignificant weightages. Hence, our strategy was to sort the word weightages for each question in ascending order and choose the 10 largest weightages per question based on the mean number of significant weightages. 105 questions (70%) were randomly selected to train the extreme learning machine (ELM). The remaining 45 questions (30%) were used to test the model. A 10-fold cross validation was performed on the training dataset to initialize ELM optimally. A grid search performed for the number of hidden nodes and the activation function showed that 27 hidden nodes using the sigmoid activation function yielded the best representation for our dataset.

3.2. Comparison analysis

We compared F1 scores to evaluate the performance of question classification using the four techniques. Given a model, we calculated an individual F1 score for each category and the model's macro-average F1 score which aggregates the mean of the model's precision and recall values and thereafter calculating the harmonic mean between them. Table 1 shows the F1 scores for each algorithm. Apart from comparing the relative F1 scores, Table 2 shows the confusion matrices. The purpose of calculating F1 scores is to differentiate the extent to which each model falsely identifies the true category of a question, thereby hindering the appropriate cognitive level of practice opportunity provided to a student.

Our results suggest that our proposed q-WNTM model links assessment questions to learning outcomes more accurately than existing models. TF-IDF achieves the highest macro-average F1 score with the lowest standard deviation among the 3 categories as seen in Table 1. From Table 2, it can be seen that *"K"*-type questions are identified best, indicating that TF-IDF is biased towards length, while the other two categories are misinterpreted more often due to the lack of consideration of semantic context spaces. LDA achieves the lowest F1 score for *"K"*-type questions due to severe sparsity in terms of question length to topic allocation. It classifies the other two categories of question types more accurately, but not to a large extent, which underscores the limitations of using LDA for short texts.

Table 2. Confusion matrices for each method.

	TF-IDF				**LDA**		
	Predicted				Predicted		
True	K	A	T	True	K	A	T
K	9	0	0	K	4	0	5
A	1	10	6	A	1	16	0
T	2	12	5	T	4	1	14
	WNTM				**q-WNTM**		
	Predicted				Predicted		
True	K	A	T	True	K	A	T
K	6	3	0	K	7	0	2
A	3	14	0	A	3	14	0
T	4	1	14	T	1	0	18

WNTM returns a higher macro-average F1 score with a lower standard deviation than the previously implemented methods as seen in Table 1. This suggests the importance of using features derived from word level co-occurrences when conducting short text topic modeling. However, the results exhibit significant level of error which limits the performance of directly applying WNTM for question classification.

The solution to reducing this error lies in our approach to eliminating redundant word co-occurrences which forms a more diagnostic abstraction of the original question. The impact of excluding co-occurring technical words yields a high macro-average F1 score of 0.848 and a low standard deviation of 0.101 with q-WNTM as shown in Table 1. The relative proportion of word occurrence for a question $P(w_j|q)$ is held constant because the words in the original question are not deleted. However, the main difference is due to the computation of $P(t|w_j)$ in the adjacency list for every technical word. If the co-occurring technical words were not excluded, θ_a (topic probabilities) in the adjacency list for each of the 437 technical words would be much different based on the count of topics during the Gibbs sampling iterations. Without any such removal, the original WNTM yielded misinterpreted topic probabilities. As seen in Table 2, almost all of the questions are correctly classified. The results obtained with q-WNTM highlight that for this dataset, the proposed stopwords list and the consideration of word combination redundancy is necessary for enhancing the question classification performance.

4. CONCLUSION

To help instructors match their assessment questions to learning outcomes, we constructed a classification model and compared it with previously implemented methods. Our model augments the performance of question classification beyond the previously described work by addressing issues concerning the selection of stopwords and considering redundant edges in the network of word co-occurrences. Our work served as a tool and resource repository for constructively aligning the assessment items found in a DSP course to its learning outcomes. With more courses and more questions, our work could lead to more deliberate practice opportunities, balance and measurement consistency to assessments of learning.

5. REFERENCES

[1] A. Kenimer and ABET. (2017, Oct.), *Proposed revised to the criteria for accrediting engineering programs general criteria introduction, criterion 3. student outcomes*,

[2] R. M. Felder and R. Brent, "Designing and teaching courses to satisfy the ABET engineering criteria," *J. Eng. Edu.*, vol. 92, pp. 7–25, 2003.

[3] D. Boud and N. Falchikov, "Aligning assessment with long-term learning," *Assess. Eval. High. Edu.*, vol. 31, pp. 399–413, 2006.

[4] R. S.-Nom, "Interactive teaching and assessment using recycled SP concepts," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2013, pp. 4354–4358.

[5] C. H. G. Wright, T. B. Welch, and M. G. Morrow, "Signal processing concepts help teach optical engineering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2016, pp. 6275–6279.

[6] M. F. Bugallo and A. M. Kelly, "An outreach after-school program to introduce high-school students to electrical engineering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. 2015, pp. 5540–5544.

[7] M. B-Sacre, M. F. Cox, M. Borrego, K. Beddoes, and J. Zhud, "Changing engineering education: Views of U.S. faculty, chairs, and deans," *J. Eng. Edu.*, vol. 103, pp. 193–213, 2014.

[8] J. Biggs, "Enhancing teaching through constructive alignment," *High. Edu.*, vol. 32, pp. 347–364, 1996.

[9] S. Z. Qamar, A. Kamanathan, and N. Z. A.-Rawahi, "Teaching product design in line with Bloom's Taxonomy and ABET student outcomes," in *Proc. IEEE Global Eng. Edu. Conf. (EDUCON)*. 2016, pp. 1017–1022.

[10] S. Goel and N. Sharda, "What do engineers want? Examining engineering education through Bloom's Taxonomy," in *Proc. 15th Annu. Conf. Aust. Assoc. Eng. Edu.*, 2004, pp. 1–13.

[11] D. A. Abduljabbar and N. Omar, "Exam questions classification based on Bloom's Taxonomy cognitive level using classifiers combination," *J. Theoretical Appl. Inform. Technol.*, vol. 78, pp. 447–455, 2015.

[12] S. Supraja, K. Hartman, S. Tatinati, and A. W. H. Khong, "Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes," in *Proc. 10th Int. Conf. Educational Data Mining (EDM)*. 2017, pp. 56–63.

[13] S. S. Haris and N. Omar, "Bloom's Taxonomy question categorization using rules and N-gram approach," *J. Theoretical Appl. Inform. Technol.*, vol. 76, pp. 401–407, 2015.

[14] K. Jayakodi, M. Bandara, and I. Perera, "An automatic classifier for exam questions in engineering: A process for Bloom's Taxonomy," in *Proc. IEEE Int. Conf. Teaching, Assessment, Learn. Eng. (TALE)*. 2015, pp. 12–17.

[15] R. H. Creecy, B. M. Masand, S. J. Smith, and D. L. Walt, "Trading MIPS and memory for knowledge engineering," *Commun. ACM*, vol. 35, pp. 48–64, 1992.

[16] A. A. Yahya, A. Osman, A. Taleb, and A. A. Alattab, "Analyzing the cognitive level of classroom questions using machine learning techniques," in *Proc. 9th Int. Conf. Cognitive Sci.*, 2013, pp. 587–595.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[18] H. Guang-Bin, "What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle," *Cogn. Comput.*, vol. 7, pp. 263–278, 2015.

[19] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. Int. World Wide Web Conf. (WWW)*, 2013, pp. 1445–1455.

[20] L. Hong and B. Davison, "Empirical study of topic modeling in Twitter," in *Proc. 1st Workshop Social Media Analytics*. ACM, 2010, pp. 80–88.

[21] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: A simple but general solution for short and imbalanced texts," *Knowledge, Inform. Syst.*, vol. 48, pp. 379–398, 2016.

[22] C. Fox, "A stop list for general text," *ACM-SIGIR Forum*, vol. 24, pp. 19–35, 1989.

[23] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, McGraw-Hill Companies, 2005.

[24] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing: A Practical Approach*, Prentice Hall, 2001.

[25] L. Chaparro, *Signals and Systems using MATLAB*, Academic Press, 2014.

[26] D. Krathwohl, "A revision of Bloom's Taxonomy: An overview," *Theory into Practice*, vol. 41, pp. 212–218, 2002.

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci.*, vol. 101, pp. 5228–5235, 2004.