OPEN ACCESS

CONTEMPORARY EDUCATIONAL TECHNOLOGY

ISSN: 1309-517X (Online) 2020, 12(2), ep275, https://doi.org/10.30935/cedtech/8341

Research Article

Identification of Cognitive Learning Complexity of Assessment Questions Using Multi-class Text Classification

Syaamantak Das
Centre for Educational Technology, Indian Institute of Technology Kharagpur, India
ORCID: 0000-0001-9896-3312

Shyamal Kumar Das Mandal
Centre for Educational Technology, Indian Institute of Technology Kharagpur, India
ORCID: 0000-0002-4088-3173

Anupam Basu
Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India
National Institute of Technology Durgapur, India
ORCID: 0000-0002-1960-9225

Received: 6 April 2020 Accepted: 21 May 2020

Abstract

Cognitive learning complexity identification of assessment questions is an essential task in the domain of education, as it helps both the teacher and the learner to discover the thinking process required to answer a given question. Bloom’s Taxonomy cognitive levels are considered as a benchmark standard for the classification of cognitive thinking (learning complexity) in an educational environment. However, it was observed that some of the action verbs of Bloom’s Taxonomy are overlapping in multiple levels of the hierarchy, causing ambiguity about the real sense of cognition required. The paper describes two methodologies to automatically identify the cognitive learning complexity of given questions. The first methodology uses labelled Latent Dirichlet Allocation (LDA) as a machine learning approach. The second methodology uses the BERT framework for multi-class text classification for deep learning. The experiments were performed on an ensemble of 3000+ educational questions, which were based on previously published datasets along with the TREC question corpus and AI2 Biology How/Why question corpus datasets. The labelled LDA reached an accuracy of 83% while BERT based approach reached 89% accuracy. An analysis of both the results is shown, evaluating the significant factors responsible for determining cognitive knowledge.

Keywords: multi-class text classification, labelled LDA, pre-trained BERT (Bidirectional Encoder Representations from Transformers), question classification

INTRODUCTION

In the field of education, it is essential to construct a cognitively well-balanced question paper. One of the standard approaches to solve this issue is the usage of Bloom’s Taxonomy (Bloom et al., 1956), created by Benjamin Bloom in the 1950s. Bloom’s Taxonomy classifies educational objectives and learning outcomes into multiple cognitive levels based on the complexity of the thinking behavior, which is required for successful completion of learning. The six levels are classified based on previous knowledge and skills such as (i) knowledge / remembering (ii) comprehension / understanding, (iii) applying, (iv) analyzing, (v) evaluating / synthesis, and (vi) creating. Each of these levels consists of several action verbs that depict the thinking required. E.g. Define, Analyze, Compare etc. However, it was observed in the work of (Stanny, 2016) that these words (Bloom's Taxonomy action verbs - (BTAV)) often co-occur in multiple levels, causing an ambiguity about the true sense of cognition. To overcome this problem, the proposed paper utilizes the effectiveness of multi-class text classification algorithms in machine learning and deep learning-based approaches as a methodology. The proposed paper uses the above methodologies to classify a given question to it's most appropriate Bloom's Taxonomy cognitive level.

Assessment questions consist of text sentences that are not cognitively structured. They also vary in multiple types. For example - (i) a question can be based on WH words such as What, Why, How etc. E.g., What factors could lead to the rise of a new species?; (ii) - a question can consist of only Bloom's Taxonomy action verbs. E.g., Explain some of the important ideas of the above section in your own words.; (iii) - a question can contain both WH and cognitive level action verbs. E.g., Why do you think average income is an important criterion for development? Explain.; (iv) - a question may contain neither cognitive level action verbs nor WH words. E.g., Will universal basic income be beneficial for the society?

The task of classifying and analysing assessment questions can fall into the category of unstructured short text classification. One of the assumptions for classifying the cognitive level of a given question is that it should belong to only one particular class. Thus, with multiple cognitive levels, it becomes a problem of multi-class classification. Considering each Bloom's Taxonomy cognitive level as a topic, the action verbs can be considered as terms/words belonging to that topic. Therefore, this makes it a topic modelling task. For topic modelling, Latent Dirichlet Allocation (LDA) is one of the standard algorithms for unstructured short text topic modelling (Massey, 2011; Uys, Du Preez, & Uys, 2008). Unlike Latent Semantic Indexing, which uses singular value decomposition and the bag-of-words representation of text documents, LDA represents texts as random mixtures over latent topics, where each topic is characterized by a distribution of words in the corpus (Blei, Ng, & Jordan, 2003). As the cognitive levels are known previously for the training data, the computation methodology will use supervised learning for the same.

The dataset used for this proposed work is an ensemble of educational assessment questions obtained from four different existing works – (i) Microsoft Search Lab - NCERT dataset (Agrawal, Gollapudi, Kannan, & Kenthapadi, 2014), (ii) the paper of Yahya et al. (Yahya, Toukal, & Osman, 2012), (iii) the paper of Jain et al. (Jain, Beniwal, Ghosh, Grover, & Tyagi, 2019) and (iv) TREC question classification dataset (Li & Roth, 2002). Apart from the mentioned datasets, two other datasets - AI2 WHY question dataset, and AI2 HOW question dataset (Jansen, Surdeanu, & Clark, 2014) were also used for testing the Why and How questions. For the computational purpose, Amazon AWS service (Amazon Comprehend) (Bhatia, Celikkaya, Khalilia, & Senthivel, 2019; Zarei & Nik-Bakht, 2019) has been used for running the labelled LDA methodology. For Deep learning, the BERT framework from Google (Devlin, Chang, Lee, & Toutanova, 2018) has been used in the Google Colab environment using GPU.

The paper describes the methodology for identifying the cognitive learning complexity of assessment questions using multi-class text Classification along with a brief overview of Bloom's Taxonomy and review of the related work. This is followed by data preparation and experimental setup. Later the results and analysis are shown, followed by future work and conclusion.

Bloom's Taxonomy Cognitive levels

This sub-section discusses Bloom's Taxonomy cognitive levels. Bloom's Taxonomy is a standard concept developed by Benjamin Bloom to classify thinking behaviors (Bloom et al., 1956). It is a set of three hierarchical models that are used to classify the educational learning objectives and outcomes based on the level of thinking complexity. The three domains are - Cognitive, Affective, and Psychomotor domain. The Cognitive domain is divided into six levels of learning which are - (i) knowledge / remembering, (ii) comprehension / understanding, (iii) applying, (iv) analysing, (v) evaluating, and (vi) creating. The revised Bloom's Taxonomy (Krathwohl, 2002) is a modified and upgraded version of the original Bloom's Taxonomy (Bloom et al., 1956), where the noun form of original taxonomy was changed into verb forms (Krathwohl & Anderson, 2010). Thus, in existing research works, the nomenclature of revised taxonomy – "Remembering, Understanding, Applying, Analysing, Evaluating, Creating" is often interchanged with the nomenclature of original taxonomy – “Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation”. There is no strict convention about which taxonomy to be followed, and thus it was observed that both taxonomies were equally used today. The following explains each of these levels:

A. Knowledge / Remembering - This level involves memorization, recognition, recalling and retrieving of relevant knowledge and information from long-term memory.

E.g. - State any three merits of roadways.

B. Comprehension / Understanding - This level involves determining the meaning of instructional messages along with interpretation of the information.

E.g. - Explain the importance of fossils in deciding evolutionary relationships.

C. Application/Applying - This level involves carrying out a procedure by the use of particulars and principles.

E.g. – “use the second-derivative test to determine whether critical points where f0(x) = 0 yield relative maxima or relative minima.”

D. Analysis/Analyzing - This level involves breaking down the subject into it’s constituent parts and determining how related they are. Also, it identifies the overall purpose or the structure.

E.g. - Using examples from your area compare and contrast that activities and functions of private and public sectors.

E. Evaluate/Evaluating - This level involves making judgements based on a given set of criterion.

E.g. - Are antibiotics better than traditional medicine?

F. Synthesis/Creating - This level involves combining ideas to form a novel, coherent and original product.

E.g. - Design a cost-effective strategy to generate reliable data.

The list of Bloom’s taxonomy action verbs is shown in **Table 1**.

Table 1. Our list of Bloom's Taxonomy action verbs

Knowledge / Remembering	Comprehension / Understanding	Application /Applying	Analysis /Analyzing	Evaluation /Evaluating	Synthesis /Creating
cite	convert	act	analyze	argue	arrange
define	discuss	apply	categorize	assess	assemble
label	explain	calculate	contrast	conclude	combine
list	express	compute	diagram	critique	compose
match	extend	demonstrate	differentiate	evaluate	create
memorize	generalize	dramatize	discriminate	judge	design
name	paraphrase	employ	divide	manage	develop
recall	predict	illustrate	examine	rearrange	devise
recite	report	manipulate	point out	reconcile	formulate
record	restate	operate	question	set up	generate
repeat	review	practice	separate	synthesize	invent
reproduce	rewrite	schedule	subdivide		organize
state	summarize	show	test		plan
	translate	sketch			rate
		solve			revise
		use			write
duplicate	associate	change	breakdown	decide	compile
quote	characterize	complete	correlate	grade	facilitate
read	give examples	backup	deduce	weigh	hypothesize
tabulate	indicate	implement	dissect	counsel	integrate
copy	represent	interview	prioritize	mediate	originate
draw	clarify	paint	survey	probe	propose
underline	extrapolate	utilize	break	release	role-play
	give	adapt	detect	supervise	improve
	interpolate	simulate	diagnose	verify	make
	articulate		figure	attach	specify
	observe		inspect	core	tell / tell why
			inventory	determine	collect
			investigate	value	reconstruct
			debate		reorganize
			group		

Note: This list shows only the unique (non overlapping) Bloom's Taxonomy action verbs (BTAV)

LITERATURE REVIEW

Literature on Bloom's Taxonomy Cognitive Level Distribution on Assessment Questions

In the work of Swart and Daneti (2019), the results show that the two lower cognitive levels of Blooms Taxonomy (Knowledge and Comprehension) constitutes approximately 58% of the total learning outcomes. The next cognitive level application, is about 27% of the learning outcomes. The remaining two levels, (Synthesis (Creating), and Evaluation), constitutes 15%. This observation was based on the learning outcomes of an electronics course. The paper by Jones (Jones, Harland, Reid, & Bartlett, 2009) indicates that academics are using more lower cognitive order than higher cognitive order questions in creating examination papers. Each question was evaluated and categorized multiple cognitive levels based on the verb list provided by Dalton and Smith (1989). Lee et al. (2017) stated that despite the practicality and simplicity of the model, Bloom's Taxonomy is criticized for generalized and uni-dimensional domains of knowledge and skills that could not clearly explain the levels. More-over, the levels of cognitive demands in the analysis of instructional objectives for students' learning and assessment plans also remain ambiguous. Also, although the revised Bloom's Taxonomy tries to overcome the generalization of cognitive dimensions, "there is still the challenge of identifying the level of thinking".

Literature on Question Classification

Teachers ask a variety of cognitive questions for different purposes, and these questions tend to form a cognitive taxonomy. Lower order cognitive questions are used extensively by instructors to check the knowledge level of students. These questions require Remembering from memory or Understanding the explanations given in the text content. Higher-order cognitive questions targets assessing higher cognitive skills such as Analysis, Synthesis, and Evaluation. As stated in the work of (Redfield & Rousseau, 1981), the higher-order questions have led to higher student learning. There are two major types of questions - (i) in-chapter or adjunct questions, and (ii) end-chapter questions. The adjunct question is located in multiple positions (inserted before or after paragraphs). The work of (Peverly & Wood, 2001) gives a comprehensive overview of adjunct questions. The paper gave the example of Rothkopf (1970), who stated that adjunct questions would help to learn by "encouraging readers to attend to relevant portions of text". The paper also cited examples of Andre (1979) and R. J. Hamilton (1985) who extended Rothkopf's theory by showing that various types of adjunct questions allows different levels of information processing. Factual questions lead to lower cognitive processing (e.g., Remembering) and higher-level questions (e.g., Evaluation) lead to more in-depth processing. The paper's results indicated that inserted adjunct questions were more effective than chapter end questions. Finally, a series of research (R. J. Hamilton, 1985; R. Hamilton, 1992) on types of questions have indicated that higher-order questions (inferences) lead to better learning performance than lower-order (fact-based) questions.

Literature on Unstructured Short Text Classification using Labelled LDA and BERT

Rationale for algorithm selection: Word co-occurrence models (Bicalho, Pita, Pedrosa, Lacerda, & Pappa, 2017), topic modelling (Zhang & Zhong, 2016), and word embedding clustering (Wang et al., 2016), are all examples of standard short text analysis methods. However, these models are useful when there is a sufficiently large training set. Transfer learning (Pan & Yang, 2009) was developed as an alternative method to reduce the need for training data. Transfer learning can be an effective method for short-text classification and requires little domain-specific training data (Long, Chen, Zhu, & Zhang, 2012; Phan, Nguyen, & Horiguchi, 2008), however, it demands to create a new model for every new classification task which is one of the major drawbacks.

Two algorithms have been chosen for this research work. First is Labelled LDA for machine learning approach and the second is Pre-trained BERT model for deep-learning approach.

LDA: The traditional LDA model (Blei et al., 2003) uses a multinomial mixture distribution θ (d) over all K topics, for each document d, from a Dirichlet prior α. For this research using Labelled LDA (Ramage, Hall, Nallapati, & Manning, 2009), the topics (Bloom's Taxonomy cognitive levels) are already known. Therefore, θ (d) is restricted to be defined only over the topics that matches to its labels Λ (d). Since the word-topic assignments z_i (**Table 1**) are taken from this distribution, this restriction ensures that all the topic assignments are limited to the document's labels. This essentially makes the algorithm to learn a bag of words model for each label, but with a shared prior in the form of η. As the document has only a single label, its topic assignment is limited to the corresponding topic, and all its words are generated from the same multinomial distribution. This is because Λ (d) will ensure that only one value of θ (d) is nonzero. β refers to the topic multinomials. The model of labelled LDA is shown in **Figure 1**.

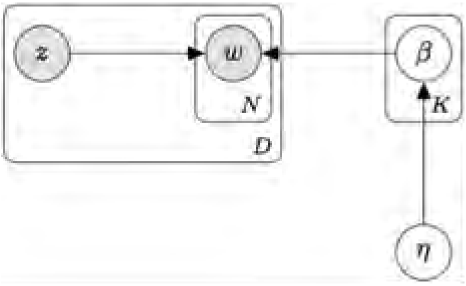

Figure 1. Labelled LDA model using only one label per document

Note: Based on the original Labelled LDA model by (Ramage et al., 2009). The multinomial topic distributions over vocabulary for each topic K, from a Dirichlet prior, where N is the document D length. w represents a list of word indices, and z represents word-topic assignments.

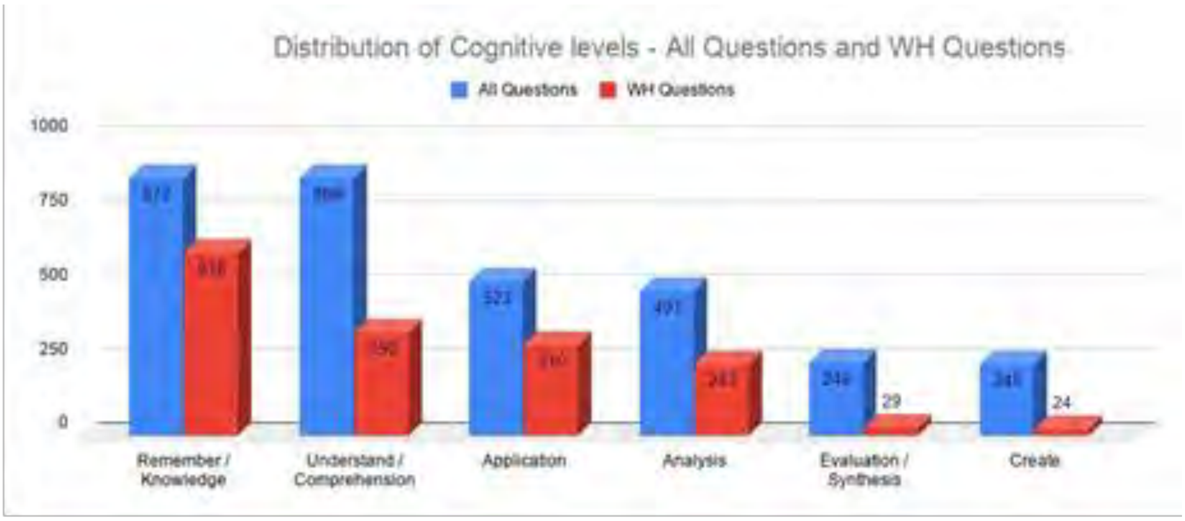

Figure 2. Distribution of Cognitive levels - All Questions and WH Questions

BERT: BERT stands for Bidirectional Encoder Representations from Transformers. The algorithm uses pre-train deep bidirectional representations from the unlabelled text by jointly conditioning on both the left and right contexts. Therefore, the pre-trained BERT model can be optimized with just one additional output layer to use in a wide range of NLP tasks. With the continuous growth of unlabelled text data, the development of pre-trained language models like BERT can give better results (Howard & Ruder, 2018; Radford et al., 2019). Even for tasks such as short text classification, which is difficult to model statistically, due to fewer features and training data, using a pre-trained language model can be useful (Luo & Wang, 2019). The method takes advantage of general language understanding to comprehend contextually relevant new words, without the requirement of additional domain data, where data volume is limited.

DATA PREPARATION AND EXPERIMENTAL SETUP

Dataset preparation: Four different datasets were used for this research. The NCERT dataset of Microsoft Search Lab [1] was collected from NCERT text-books. Five subject experts manually annotated them. Apart from the NCERT dataset, the question dataset of existing papers Yahya et al. (2012), Jain et al. (2019), and Li and Roth (2002) were also used. The dataset of Jain et al. (2019) and Yahya et al. (2012) were already annotated, while only DESCRIPTION class of (Li & Roth, 2002) were manually annotated for the research. Two types were considered for the experiment - (a) questions of all types and (b) questions with WH words. This is because, questions can have action verbs only, WH words only, both of them and neither. The first one covers all and the second one specific to WH questions. The Cognitive distribution of both type of questions is shown in **Figure 2**. It was observed that both Why and How had a similar distribution when not paired with BTAV. Emphasis was given on WH questions as two types of questions Why and How showed almost equal distribution across Cognitive levels as shown in **Figure 3**.

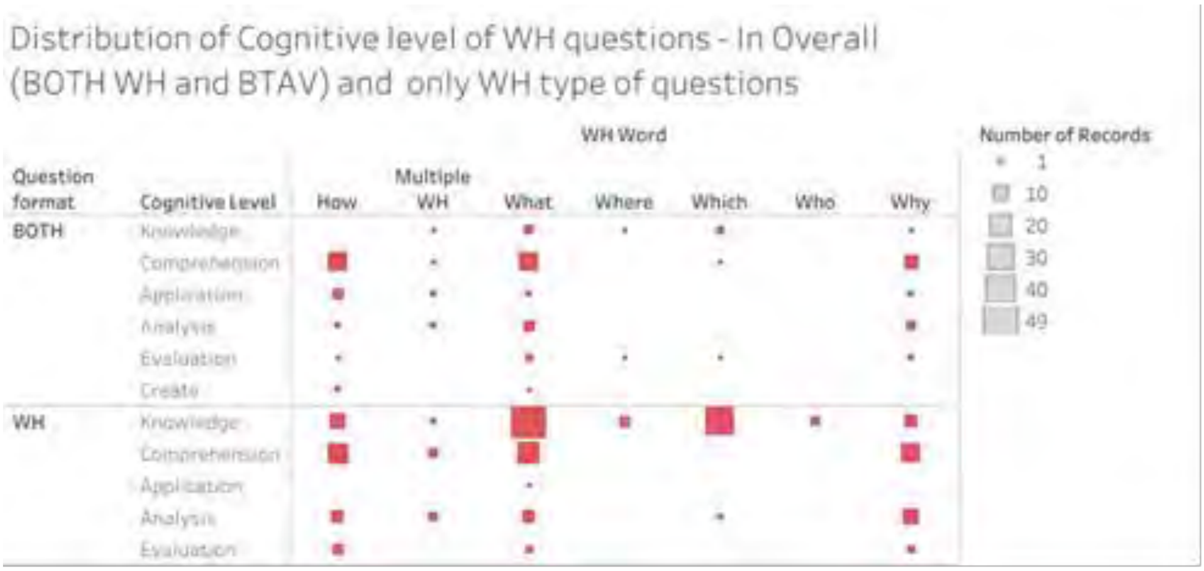

Figure 3. Distribution of WH words over cognitive levels

Training Data: For this research, the following approach was considered for identifying the cognitive level of an assessment question. First, a set of 434 questions from the NCERT dataset (Agrawal et al., 2014) was manually annotated by five subject experts with the substantial agreement of 0.65 Fleiss kappa value (Landis & Koch, 1977) as inter-annotation agreement. Apart from the NCERT dataset, questions from (Yahya et al., 2012) and (Jain et al., 2019) were also used as questions from these datasets also were manually tagged by human experts, making them the gold standard data. Furthermore, the DESCRIPTION (DESC:) class of TREC Question dataset (Li & Roth, 2002) was also used. It was observed that the subclasses of DESCRIPTION class - (i) definition, description, manner, and reason can be mapped to Remember/Knowledge, Understanding / Comprehension, Applying/Application, and Analysing/Analysis of Bloom's Taxonomy Cognitive level respectively.

Computation Environment: The first experiment was performed using Labelled LDA methodology using the Amazon AWS Comprehend service. Amazon AWS comprehend service uses standard LDA as a text classification algorithm making it an unsupervised learning approach. For this research, a custom classifier was used where it was trained on labelled data making it a supervised learning approach. For the second experiment, a BERT based multi-class classification model was developed using the Google Colab GPU environment. The BERT base model (uncased) was used, which has 12-layer, 768-hidden, 12-heads, 110M parameters. Tokenization was performed, and the softmax function was used. While computing probability, to understand cross-entropy loss, a score called logit, which is a raw unscaled value associated with a class, is used. For neural network architectures, a logit is an output of a dense (fully-connected) layer. Five epochs were used, along with a learning rate of 3e-5, and maximum sequence length was taken 128, which covered the length size of 99% questions. For both experiments, ~90% data was used for training.

Testing data[1]: The TREC testing dataset was used for testing the result. As observed in the manual annotation of NCERT dataset, that for questions with WH words - Why and How showed an almost even distribution over all Bloom's Taxonomy cognitive levels as shown in **Figure 3**. Thus additionally, the AI2 Why and How question datasets were used for testing purposes. The overview of the are shown in **Table 2**. The distribution of the Cognitive level of the training dataset is shown in **Table 3**.

[1] Note: Since no previous papers provided BTAV list, experiments with BTAVs could not be performed, as the dataset of BTAVs may vary.

Table 2. Overview of the datasets

Dataset Name	Data Quantity
Aggrawal et.al (NCERT Dataset)	434
Jain et. al Dataset	1053
Yahya et.al Dataset	600
TREC Training (Description Class) Dataset	1162
TREC Testing (Description Class) Dataset	138
AI2 Biology HOW questions Dataset	185
AI2 Biology WHY questions Dataset	193

Table 3. Distribution of Cognitive levels in the training dataset

Cognitive Levels	Knowledge	Comprehension	Application	Analysis	Evaluation	Create
Count	872	869	523	491	249	245

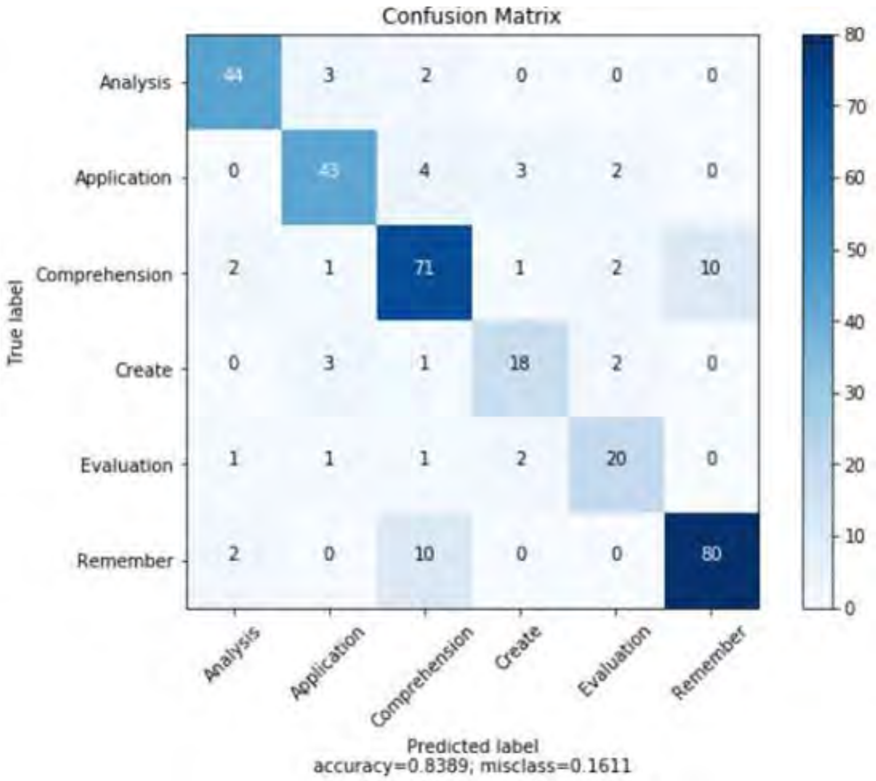

Figure 4. Confusion Matrix of Labelled LDA experiment - OVERALL

RESULT AND ANALYSIS

Labelled LDA

For overall questions, a set of (ensuring all classes are being covered) 2967 in-stances of training data were used to train the customized classifier service of Amazon Comprehend, and 329 instances was used for testing purpose. The evaluation metrics are as follows: “Accuracy”: 0.8389, “Precision”: 0.8245, “Recall”: 0.8268, “F1Score”: 0.8255. The confusion matrix is shown in **Figure 4**. The accuracy (0.83) obtained is more than previous literature (Jain et al., 2019)’s result.

For WH questions, a set of 1417 WH questions were used as training and 157 questions were used for testing. The evaluation metrics are as follows: “Accuracy”: 0.7898, “Precision”: 0.6188, “Recall”: 0.5875, “F1Score”: 0.5982.

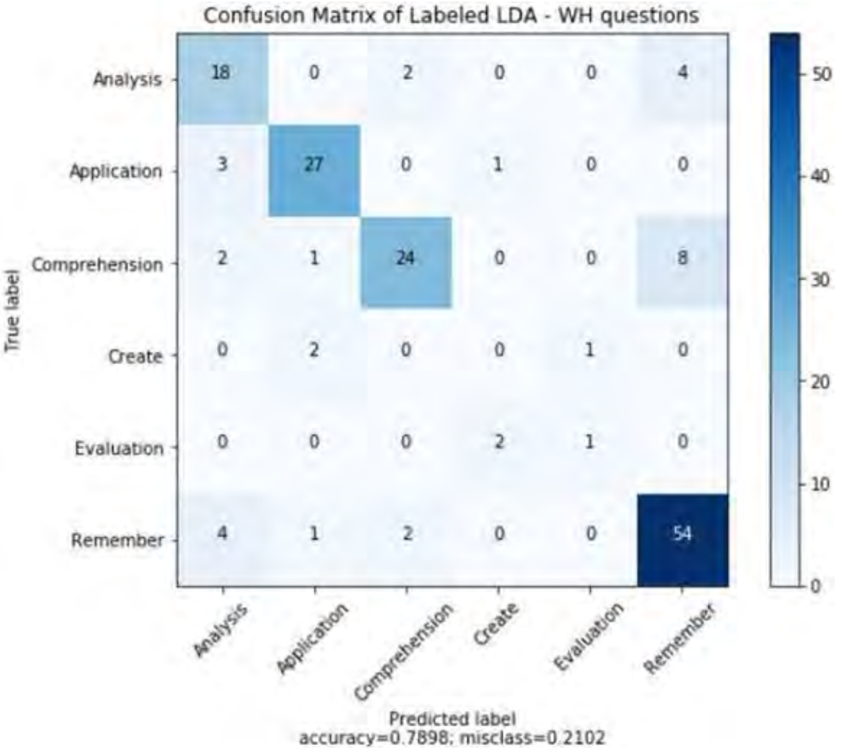

Figure 5. Confusion Matrix of Labelled LDA experiment - WH questions

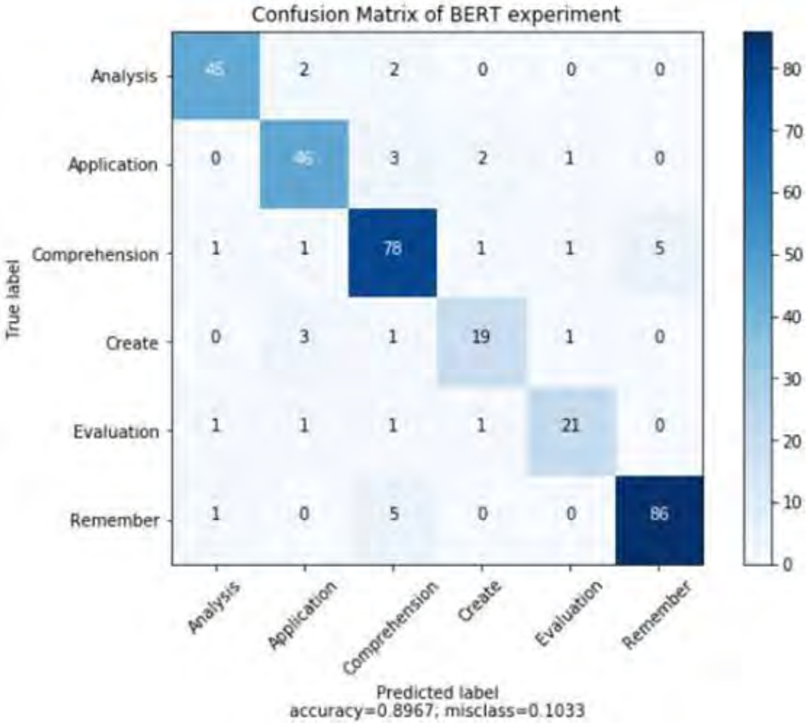

Figure 6. Confusion Matrix of BERT experiment - OVERALL

BERT

For deep-learning model, the training loss at the end of five epochs was 0.113230795. The confusion matrix of overall test data is shown in **Figure 6**. The accuracy obtained was 0.8967. The results showed significant improvement from LDA methodology. Furthermore, the AI2 How and Why data sets (**Figure 8a** and **Figure 8b**) were tested to see if the prediction patterns are matching or not. The accuracy obtained for Overall is 89.67% and for WH questions is 88.68%.

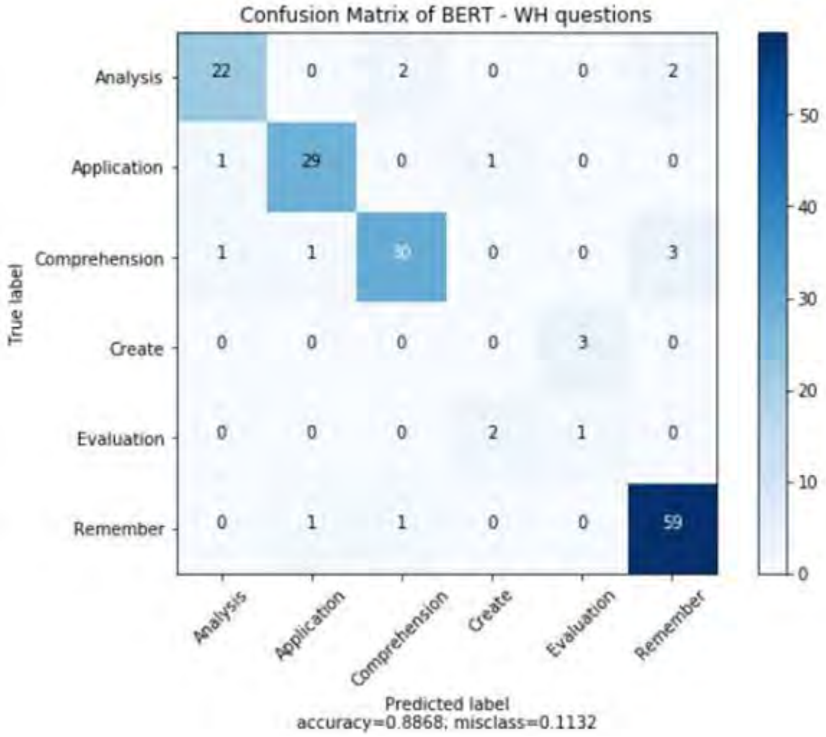

Figure 7. Confusion Matrix of BERT experiment - WH questions

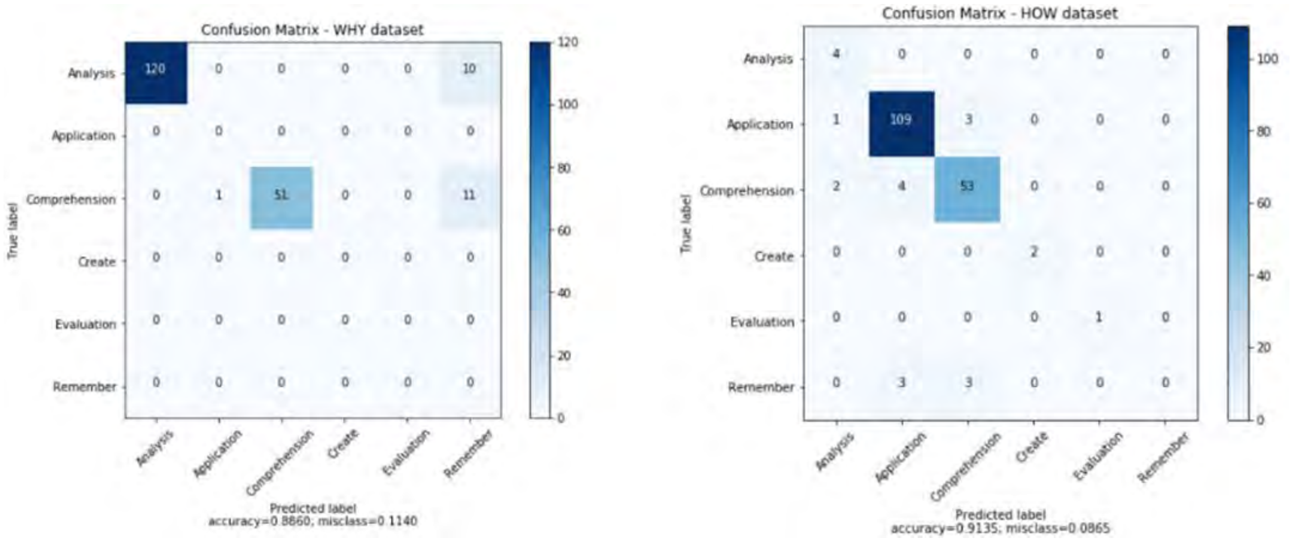

Figure 8. (a) WHY dataset confusion matrix (b) HOW dataset confusion matrix

Observation and Discussion

A. General observation: The deep-learning-based approach performed better than the machine learning approach, as language models were better in classifying the cognitive levels when compared to using the bag-of-words based models. However, the error occurred in cases where the text structure of the question stem was similar. E.g., What does < subject > mean ? was tagged as Remember / Knowledge level and What does < subject > do ? was tagged as Comprehension / Understanding level, as the former has a more specific answer. Both the proposed methodologies yielded a significantly better result when compared to existing works (Yahya et al., 2012; Jain et al., 2019). The comparison of results is shown in **Table 4**. However, the use of deep-learning-based pre-trained language model methodology yielded an exceptionally better result when compared to all existing methods. The errors that occurred may be reduced with a better training dataset.

Table 4. Comparison of multiple algorithm's accuracies

	Algorithm	Accuracy
1.	K-Nearest Neighbors (KNN)	75%
2.	Random Forest	82.2%
3.	Decision Trees	82.2%
4.	SVM	82.2%
5.	Neural Network	82.2%
6.	Linear Discriminant Analysis	83.3%
7.	Logistic Regression	83.3%
8.	Labelled LDA - Overall	83.89%
9.	Labelled LDA - WH questions	78.98%
10.	Deep Learning (BERT) - Overall	89.67%
11.	Deep Learning (BERT) - WH questions	88.68%

B. Bloom's Taxonomy action verbs: From the results, it was observed that a set of Bloom's Taxonomy action verbs were truly ambiguous in nature as it is challenging to identify the required Cognitive level unless the context is known. Examples of such Bloom's Taxonomy action verbs are - Choose, Describe, Design, Explain, Show, and Use. These words have distributions across multiple Cognitive levels.

C. WH questions: It can be concluded from these observations that when multiple WH words are used in the question stem, it leads to a higher cognitive requirement, mainly if Why is used. The word Why, when occurring individually, has the highest frequency in Comprehension level. However, when occurring with another WH word (e.g., What, Which having individual higher occurrence frequencies at lower cognitive level), which acts as a context; the Why acts as a higher cognitive level (e.g., Analysis) identifier. The ambiguity of both How and Why can be controlled based on the context.

D. Other questions: For questions that do not contain either action verbs or WH words, it was observed that the cognitive level distribution was almost uniform throughout.

E. Better performance of Deep learning models: This is because BERT uses bidirectional training of Transformer, an attention-based model, for language modelling. Most machine learning models, trains themselves on the text input sequentially, while the Transformer encoder reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surrounding words. Classification tasks are done by adding a classification layer on top of the Transformer output.

F. Limitations: First, the limitations of the proposed methodology are that unless there is a lot of training data, the accuracy cannot be improved. Second each training questions need to be carefully annotated by the annotators so that the training data is correct. Third, the algorithm being essentially dependent on training data, do have it's limitation on open ended philosophical questions (e.g. Who is god?)

FUTURE WORK AND CONCLUSION

In this proposed work, a feasible solution to a significant challenge in the educational domain was provided. The proposed paper tries to solve the problem of the cognitive learning complexity of school textbook assessment questions. This is a problem that is dependent on human experts for solutions which is subject to biasness and ambiguity about the true sense of cognitive level. The previous research work uses Bloom's taxonomy as a methodology to approach the said problem. However, due to the overlapping of Bloom's taxonomy cognitive action verbs across multiple cognitive levels, the existing methods were not efficient to solve the ambiguity. The proposed research work uses computational approaches to solve the problem. Using machine learning and deep learning models trained across multiple existing assessment question datasets, it was observed that the proposed methodologies provide a significant improvement from existing approaches in terms of accuracy. Also, questions without Bloom's Taxonomy action verbs, too, were assigned a cognitive level correctly by the algorithms due to the usage of bidirectional approaches in understanding the context of a word. The paper also contributes by providing cognitive levels of WH questions, which was little explored previously. This can act as a helping tool to identify the cognitive level of a question and can help instructors in setting up the learning materials of the curriculum and assessment questions for evaluation. The present limitation of the proposed methodology is the need for a massive amount of training data. While getting academic questions from a textbook is not a concern; getting them annotated is a time-consuming and challenging task. This is because annotations will vary from expert to expert based on previous knowledge of the domain. For future work, the tasks should be considering images (graphs, photos, equations) as a part of the question to identify the cognitive level of the give question. Also, the objective should be building a corpus of academic questions across multiple subjects annotated by domain experts as a standard dataset for future related works.

REFERENCES

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2014). Study navigator: An algorithmically generated aid for learning from electronic text-books. *Journal of Educational Data Mining, 6*(1), 53-75.

Andre, T. (1979). Does answering higher-level questions while reading facilitate productive learning? *Review of Educational Research, 49*(2), 280-318. https://doi.org/10.3102/00346543049002280

Bhatia, P., Celikkaya, B., Khalilia, M., & Senthivel, S. (2019). Comprehend medical: a named entity recognition and relationship extraction web service. *arXiv preprint* arXiv: 1910.07419. https://doi.org/10.1109/ICMLA.2019.00297

Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., & Pappa, G. L. (2017). A general framework to expand short text for topic modeling. *Information Sciences, 393*, 66-81. https://doi.org/10.1016/j.ins.2017.02.007

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bloom, B. S., et al. (1956). *Taxonomy of educational objectives. vol. 1: Cognitive domain*. New York: McKay, 20-24.

Dalton, J., & Smith, D. (1989). *Extending children's special abilities: strategies for primary classrooms*. Office of Schools Administration, Ministry of Education, Victoria.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. https://doi.org/10.18653/v1/N19-1423

Hamilton, R. (1992). Application adjunct post-questions and conceptual problem solving. *Contemporary Educational Psychology, 17*(1), 89-97. https://doi.org/10.1016/0361-476X(92)90050-9

Hamilton, R. J. (1985). A framework for the evaluation of the effectiveness of adjunct questions and objectives. *Review of Educational Research, 55*(1), 47-85. https://doi.org/10.3102/00346543055001047

Howard, J., & Ruder, S. (2018). Universal language model ne-tuning for text classification. *arXiv preprint* arXiv:1801.06146. https://doi.org/10.18653/v1/P18-1031

Jain, M., Beniwal, R., Ghosh, A., Grover, T., & Tyagi, U. (2019). Classifying question papers with bloom's taxonomy using machine learning techniques. In *International conference on advances in computing and data sciences* (pp. 399-408). https://doi.org/10.1007/978-981-13-9942-8_38

Jansen, P., Surdeanu, M., & Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (volume 1: Long papers) (pp. 977-986). https://doi.org/10.3115/v1/P14-1092

Jones, K. O., Harland, J., Reid, J. M., & Bartlett, R. (2009). Relationship between examination questions and bloom's taxonomy. In *2009 39th IEEE frontiers in education conference* (pp. 1-6). https://doi.org/10.1109/FIE.2009.5350598

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin c. wittrock and the revision of bloom's taxonomy. *Educational psychologist, 45*(1), 64-65. https://doi.org/10.1080/00461520903433562

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Lee, Y.-J., Kim, M., Jin, Q., Yoon, H.-G., & Matsubara, K. (2017). Revised blooms taxonomy the swiss army knife in curriculum research. In *East-asian primary science curricula* (pp. 11-16). Springer. https://doi.org/10.1007/978-981-10-2690-4

Li, X., & Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on computational linguistics-volume 1* (pp. 1-7). https://doi.org/10.3115/1072228.1072378

Long, G., Chen, L., Zhu, X., & Zhang, C. (2012). Tcsst: transfer classification of short & sparse text using external data. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 764-772). https://doi.org/10.1145/2396761.2396859

Luo, L., & Wang, Y. (2019). Emotionx-hsu: Adopting pre-trained bert for emotion classification. *arXiv preprint* arXiv:1907.09669.

Massey, L. (2011). Autonomous and adaptive identification of topics in unstructured text. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 1-10). https://doi.org/10.1007/978-3-642-23863-5_1

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

Peverly, S. T., & Wood, R. (2001). The effects of adjunct questions and feed-back on improving the reading comprehension skills of learning-disabled adolescents. *Contemporary Educational Psychology, 26*(1), 25-43. https://doi.org/10.1006/ceps.1999.1025

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web* (pp. 91-100). https://doi.org/10.1145/1367497.1367510

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing, 1*(1), 248-256. https://doi.org/10.3115/1699510.1699543

Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of educational research, 51*(2), 237-245. https://doi.org/10.3102/00346543051002237

Rothkopf, E. Z. (1970). The concept of mathemagenic activities. *Review of educational research, 40*(3), 325-336. https://doi.org/10.3102/00346543040003325

Stanny, C. (2016). Reevaluating blooms taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences, 6*(4), 37. https://doi.org/10.3390/educsci6040037

Swart, A. J., & Daneti, M. (2019). Analyzing learning outcomes for electronic fundamentals using blooms taxonomy. In *2019 IEEE global engineering education conference* (educon) (pp. 39-44). https://doi.org/10.1109/EDUCON.2019.8725137

Uys, J., Du Preez, N., & Uys, E. (2008). Leveraging unstructured information using topic modelling. In *Picmet'08-2008 portland international conference on management of engineering & technology* (pp. 955-961). https://doi.org/10.1109/PICMET.2008.4599703

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L., & Hao, H. (2016). Semantic ex-pansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing, 174*, 806-814. https://doi.org/10.1016/j.neucom.2015.09.096

Yahya, A. A., Toukal, Z., & Osman, A. (2012). Blooms taxonomy-based classi-cation for item bank questions using support vector machines. In *Modern advances in intelligent systems and tools* (pp. 135-140). Springer. https://doi.org/10.1007/978-3-642-30732-4_17

Zarei, F., & Nik-Bakht, M. (2019). Automated detection of urban flooding from news. In *Proceedings of the 36th international symposium on automation and robotics in construction* (pp. 515-521). https://doi.org/10.22260/ISARC2019/0069

Zhang, H., & Zhong, G. (2016). Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems, 102*, 76-86. https://doi.org/10.1016/j.knosys.2016.03.027

Correspondence: Syaamantak Das, Centre for Educational Technology, Indian Institute of Technology Kharagpur, India. E-mail: syaamantak.das@gmail.com