

Article

Semantic Similarity Analysis for Examination Questions Classification Using WordNet

Thing Thing Goh ^{1,2}, Nor Azliana Akmal Jamaludin ^{2,*}, Hassan Mohamed ², Mohd Nazri Ismail ² and Huangshen Chua ¹ 

¹ School of Engineering, Department of Electrical and Electronics, UOW Malaysia KDU University College, Shah Alam 40150, Malaysia; ttgoh@uow.edu.my (T.T.G.); hs.chua@uow.edu.my (H.C.)

² Faculty of Defense Science and Technology, National Defence University of Malaysia (NDUM), Sungai Besi 57000, Malaysia; hassan@upnm.edu.my (H.M.); m.nazri@upnm.edu.my (M.N.I.)

* Correspondence: azliana@upnm.edu.my

Abstract: Question classification based on Bloom's Taxonomy (BT) has been widely accepted and used as a guideline in designing examination questions in many institutions of higher learning. The misclassification of questions may happen when the classification task is conducted manually due to a discrepancy in the understanding of BT by academics. Hence, several automated examination question classification systems have been proposed by researchers to perform question classification accurately. Most of this research has focused on specific subject areas only or single-sentence type questions. There has been a lack of research on question classification for multi-sentence type and multi-subject questions. This paper proposes a question classification system (QCS) to perform the examination of question classification using a semantic and synthetic approach. The questions were taken from various subjects of an engineering diploma course, and the questions were either single- or multiple-sentence types. The QCS was developed using a natural language toolkit (NLTK), Stanford POS tagger (SPOS), Stanford parser's universal dependencies (UD), and WordNet similarity approaches. The QCS used the NLTK to process the questions into sentences and then word tokens, such as SPOS, to tag the word tokens and UD to identify the important word tokens, which were the verbs of the examination questions. The identified verbs were then compared with the BT's verbs list in terms of word sense using the WordNet similarity approach before finally classifying the questions according to BT. The developed QCS achieved an overall 83% accuracy in the classification of a set of 200 examination questions, according to BT.

Keywords: question classification; bloom taxonomy; similarity; POS tagger; universal dependencies



Citation: Goh, T.T.; Jamaludin, N.A.A.; Mohamed, H.; Ismail, M.N.; Chua, H. Semantic Similarity Analysis for Examination Questions Classification Using WordNet. *Appl. Sci.* **2023**, *13*, 8323. <https://doi.org/10.3390/app13148323>

Academic Editors: Valentino Santucci and Yudong Zhang

Received: 20 April 2023

Revised: 19 May 2023

Accepted: 13 June 2023

Published: 19 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Assessments via written examinations are one of the most common methods used to evaluate students' achievement in most educational institutions. They evaluate students' achievements referring to the course learning outcomes (CLO) that are aligned with the level of learning and understanding based on Bloom's Taxonomy (BT). Therefore, it is crucial to prepare a high-quality examination question that can align with the intended CLO. The correctness of examination question classification is checked by the accreditation bodies and quality organizations [1]. However, there are always discrepancies in examination question classifications into BT due to the inefficiency of manual classification. Some BT verbs can be found in different BT levels that cause inconsistent labeling when examination questions are being classified. Although discrepancies in examination question classifications can be reduced with many rounds of manual moderation, this is time-consuming, and a good deal of effort must be put into ensuring the understanding of moderators in question classification.

In recent years, various studies have been conducted on the application of natural language processing (NLP) in automation question classification (AQC). Humans can

classify examination questions easily based on their understanding of the meaning of the questions and BT. However, it is not easy to teach the machine to understand the meaning of these questions. Questions have to be processed in order to be understood by machines. NLP has been implemented in many fields, such as text classification, document summarization, question answering, etc. It enables the machine to read the text, interpret it and determine the important keywords. NLP has the computer ability to process information from a large database of text with very little human intervention [2]. Generally, question classification systems using NLP consist of three main modules: the pre-processing module, the keyword extraction module, and the classification module.

This paper proposes a question classification system (QCS) that performs examination question classification based on the revised BT using semantic similarity techniques. This study started with an overview of commonly used tools and techniques in a pre-processing module and was then followed by semantic similarity analysis to improve the accuracy of the classification of the questions. The test results were analyzed, the best-performing methods or approaches adopted, and the possible enhancement of QCS was discussed. Other than NLP, multiple-sentence examination questions were included in this study. Many studies of examination question classifications are performed on single-sentence examination questions only and in specific areas only. Even in question-answering systems, this classification also involves mostly single sentences and not multiple sentences. However, in reality, examination questions consist of single sentences and multiple sentences, especially for examination questions with a higher thinking order that consists of problem statements or a case study with many sentences. Consequently, this QCS could benefit institutions of higher learning (IHL) in the development and implementation of question classification systems.

Furthermore, this research studied examination question classifications from different areas, including Engineering Schools such as physics, programming, electronics, mechanical, management, etc. According to the research [3–6], with rule-based and machine learning for question classification, accuracy can be improved with a huge set of rules implemented or a large training set to train the machine. In this research, 83% accuracy in 200 examination question classifications from different areas involving single and multiple sentences was achieved with only three rules in NLP text processing and optimization in semantic similarity analysis.

2. Related Work

Various studies were carried out in the automation of examination question classifications based on BT using NLP. BT has been widely accepted and implemented as an important tool to guide academics, develop a holistic assessment, and promote a higher thinking order to design examination questions [1]. The revised BT consists of six cognitive levels—Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating, as shown in Table 1 [7]. The verbs listed in BT are used as an indicator for question classification. However, some of the listed verbs exist in multiple BT levels [8], which often leads to inconsistencies when categorizing the examination questions by academics due to their different understanding of BT. Thus, an automated system is proposed to classify these questions using NLP.

NLP enables machines to identify keywords from the questions and then classify them according to BT. Both Haris and Omar [9] and Biawas et al. [10] proposed an automated analysis of examination questions to determine the appropriate BT levels using a rule-based approach to the syntactic structure of the questions. Text pre-processing techniques such as stop-words removal, stemming, lemmatization and part-of-speech (POS) tagging have also been implemented. The Haris and Omar [9] approach achieved a satisfactory result with a macro F1 of 0.77. However, it required more rules and greater training questions to improve the accuracy of the system. On the other hand, Biawas et al. [10] proposed a reduction in the number of question categories to three categories in order to achieve a

more satisfactory result. Reducing these question categories is not a good resolution to be implemented in examination question classifications with BT.

Table 1. Adapted from Anderson’s revisions on Bloom’s taxonomy verbs [7]. 2015, Jayakodi, K., Bandara, M. & Perera, I.

Categories	Cognitive Verb List of Anderson Taxonomy	
	Description	Verb List
Creating	Builds a structure or pattern from diverse elements	categorizes, combines, compiles, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, reconstructs
Evaluating	Makes judgments about the value of ideas or materials	appraises, compares, concludes, contrasts, criticizes, critiques, defends, describes, discriminates, evaluates, explains, interprets
Analyzing	Separates material or concepts into component parts	analyses, breaks down, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies, illustrates
Applying	Uses a concept in a new situation or unprompted use of an abstraction	applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts
Understanding	Comprehends the meaning, and interpretation, of instructions and problems	comprehends, converts, defends, distinguishes, estimates, explains, extends, generalizes, gives an example, infers, interprets, paraphrases
Remembering	Recalls or retrieves previously learned information	defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognizes, reproduces, selects, states

Subsequently, the machine learning (ML) approach was explored to overcome the defect of rule-based classifications. The questions tagged with associated target responses were supplied to the machine prior to predicting the correct response. Based on the training dataset, the machine could classify this question automatically without many rules set when given a new set of data. Kusuma et al. [11] suggested an automatic system to classify the Indonesian language based on the support vector machine (SVM) algorithm and achieved promising classification results with 88.6% accuracy.

Osman and Yahya [1] and Sulaiman et al. [12] also studied examination question classification with a SVM. The researchers compared different machine learning methods such as Naïve Bayes, SVM, KHH logistic regression and decision trees in their study. However, this research only focused on computer science examination questions or single-sentence questions. Osman and Yahya’s [1] research provided the best classification result with 0.7667 while Sulaiman’s [12] research obtained an accuracy of 75.2%. In our research, we aimed to include examination questions from different areas so that these could be implemented in different faculties later. Further, we would expect to have an accuracy greater than 80% in the real implementation of examination question classification.

Mohamed et al. [5] classified examination questions with semantic and syntactic methods. POS tagging was used to determine the verbs and nouns from the questions while WordNet was used to identify the required sense of the words and remove irrelevant information. Machine learning SVM was then used to classify the examination questions with an accuracy of 82%. However, this research only focused on a specific area which was computer programming with single-sentence questions only as the training dataset. Typically, ML was more flexible in question classifications than rule-based classifications because questions could be trained on a new taxonomy, while a good deal of time was spent maintaining the rules. However, it was challenging to obtain enough data for the experiment, especially collecting examination questions in order to improve the accuracy of examination classification. More training sets resulted in more accurate results. Further, many researchers have proposed further investigations into semantic knowledge in question classifications to improve the accuracy of classification.

Jayakodi et al. [3] proposed rule-based examination question classifications with WordNet and the cosine similarity algorithm. Verbs were extracted from the questions and similarities were identified by comparing them with BT verbs. Only a 71% accuracy was obtained. However, this research focused on one specific area only. Mohammed and Omar [13] proposed the classification of questions according to BT using the term frequency-inverse document frequency (TF-IDF) and Word2Vec. Modified TF-IDF was used to identify the most important word in the questions. Word2Vec was used to denote the identified verbs and BT verbs with a semantic similarity. Finally, the classifier provided an average accuracy of 89.7% [6]. However, this research only focused on single-sentence questions. Contreras et al. [14] studied the integration of statistical and machine learning in essay question classifications. TF-IDF and SVM were used in the question classification. However, this system could only process the questions with BT keywords. They obtained 82.60% of the F1 measurement.

Table 2 summarizes the approaches and techniques used in examination question classifications and their performances. Overall, rule-based approaches were commonly used in examination question classifications. However, these questions were restricted to certain syntactic patterns, and more rules were set to achieve better accuracy. It was not easy to maintain a huge set of rules. Thus, ML was proposed by researchers to achieve better results in examination question classifications. The SVM has been the popular approach used in examination classifications. However, this accuracy can only improve with a large training set of examination questions.

Table 2. Comparison of different studies on question classification.

Studies	Processing and Classification Method	Performance
Osman and Yahya [1]	Machine Learning SVM with unigram features	Accuracy 76.67%
Jayakodi et al. [3]	Rule-based (POS tag + Wordnet + Cosine sim)	Accuracy 71%
Diab and Sartawi [4]	Rule-based (POS tag + Wordnet + Cosine sim)	Accuracy 83%
Sangodiah et al. [15]	Rule-based + machine learning (SVM)	Accuracy 82%
Mohamed et al. [5]	WordNet + machine learning (SVM)	Accuracy 82%
Sulaiman et al. [12]	Statistical TF-IDF + machine learning (SVM)	Accuracy 75%
Mohammed and Omar [6]	TF-IDF + Word2Vec (semantic similarity) on single sentence questions	F1 89.7%
Contreras et al. [14]	TF-IDF + machine learning (SVM)	F1 82.6%

In general, examination question classifications have been studied by many researchers in recent years. However, there has been no implementation carried out from the research and no public dataset exists that contains real examination questions tested in a question classification system. Much research has focused on text pre-processing and classification using rule-based or machine learning with single sentences or one subject area. However, in terms of implementation, questions from different subject areas consist of multiple sentences. Some of these examination questions are more than three sentences long. In order to develop a system that can classify questions in actual implementation, the system developed should be able to classify questions from different areas, including multiple sentences, with fewer rules set and achieve at least 80% accuracy. Another issue to be considered is the availability of examination questions as a training set if a machine-learning approach is chosen. According to Osman and Yahya [1], Minaee and Liu [16], and Mohammed and Omar [6], significant results can be obtained with a huge amount of data to train the machine. However, it is not easy to collect huge amounts of the dataset from examination questions. Therefore, a rule-based semantic approach with a minimum rule set is proposed in this study.

3. Methodology

The automated examination question classification system proposed in this paper starts with question preprocessing (QP), followed by verb extraction (VE) and ends with a question classification module (QC), as shown in Figure 1.

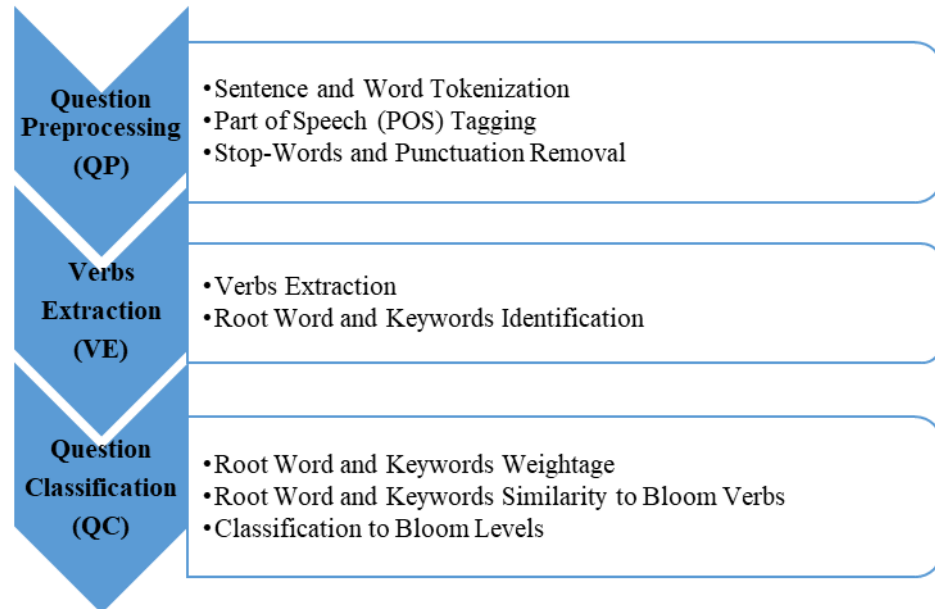


Figure 1. General flow of question classification system.

3.1. Question Collection

A total of 200 final examination questions in the English medium were collected from the School of Engineering (SoE), UOW Malaysia KDU University College over three years, from 2017 to 2019. The examination questions were collected randomly from various subjects such as Programming, Electrical, Electronics, Mechanical, Automation, Mechanical and Management of Engineering, as shown in Table 3. These examination questions were classified manually by SoE lecturers and moderated by three lecturers who had more than 10 years of experience in lecturing. In total, 77.5% of these examination questions were single-sentence questions, while 22.5% were multi-sentence questions.

Table 3. Statistics of the 200-examination question dataset.

(a) Statistics of the 200-Examination Question Dataset by Bloom Levels								
Bloom Level	Subject Domains							Total
	Automation	Electronics	Management	Mechanical	Physics	Power	Program	
1	0.5%	2.0%	2.0%	3.0%	1.0%	3.0%	4.5%	16.0%
2		2.5%	2.0%	5.0%	0.5%	4.5%	5.0%	19.5%
3		2.5%		7.0%	5.0%	7.5%	6.5%	28.5%
4	2.5%	3.5%	2.5%	5.0%	0.5%	1.0%	4.5%	19.5%
5	1.0%		1.5%	0.5%	0.5%	4.5%		8.0%
6	2.5%	1.0%			0.5%		4.5%	8.5%

Table 3. Cont.

(b) Statistics of the 200-examination question dataset by subject domains and question types								
Question Type	Subject Domains							Total
	Automation	Electronics	Management	Mechanical	Physics	Power	Program	
Single-sentence	3.5%	10.5%	4.0%	17.0%	6.5%	16.0%	20.0%	77.5%
Multi-sentence	3.0%	1.0%	4.0%	3.5%	1.5%	4.5%	5.0%	22.5%
Total	6.5%	11.5%	8.0%	20.5%	8.0%	20.5%	25.0%	100.0%

3.2. Question Preprocessing

The question preprocessing (QP) module is an important step in NLP. It is executed to ensure that words in the sentence are tagged with verbs, nouns, adjectives, etc., for the subsequent VE module. Examination questions with multiple sentences are tokenized into a single sentence; questions with a single sentence can then be tokenized into word tokens. Generally, the QP module uses the default tokenizer provided by the NLTK package, which is a PunktSentenceTokenizer. Part-of-speech (POS) tagging is the subsequent step after tokenization. It labels each word in the sentence with nouns, verbs, adjectives, etc. This is the most critical process in QP because an error in POS tagging may cascade into the VE module and then further into the final QC module. There are many POS taggers, such as the TnT POS tagger, ClassifiedBasedPOS tagger, Perceptron POS tagger, and Stanford POS tagger, etc. The Stanford POS tagger was adopted in this module because it generally outperformed other taggers, according to Goh et al. [17], Tian and Lo [18] and Go and Nocon [19]. The “English-left3wrods-distsim” and “English-bidirectional-distsim” are two commonly used models for the Stanford POS tagger. These two models were evaluated in the QP module, and the best-performing model was the one adopted in the QP module. The POS tagging process was followed by stop words and punctuation removal. Stop words are common, uninformative words such as “a,” “an,” and “the” that can be removed to reduce the size of the dataset and focus on more meaningful words.

3.3. Verb Extraction

Verb extraction (VE) identifies and extracts verbs in the word tokens from the QP module. These word tokens are tagged with the Penn Treebank tag by POS taggers of the QP module. By referring to each verb form’s description, the “VB” tag has the largest likelihood of active actions, indicating the thinking level enabling its use in classification according to the BT level. Generally, a sentence should have a verb, and this verb is tagged with “VB”, but there is a possibility that no verb is tagged in the question and another possibility that more than one verb is tagged. When there is no “VB” tagged verb found in the word tokens, “VBP” and “VBZ” must be included in the consideration of VE, as shown in Table 4. When there is more than one verb identified, one of the verbs should be used in classification, while the rest of the verbs play supporting roles when there is ambiguous classification. The main verb can be referred to as the root word (RW), while the supporting verbs are referred to as keywords (KW). The Stanford parser universal dependencies (UD) approach can be used to identify the RW and KW, as shown in Table 5. The UD approach identifies the verbs based on the priority from the RW, then the KW level 1, etc.

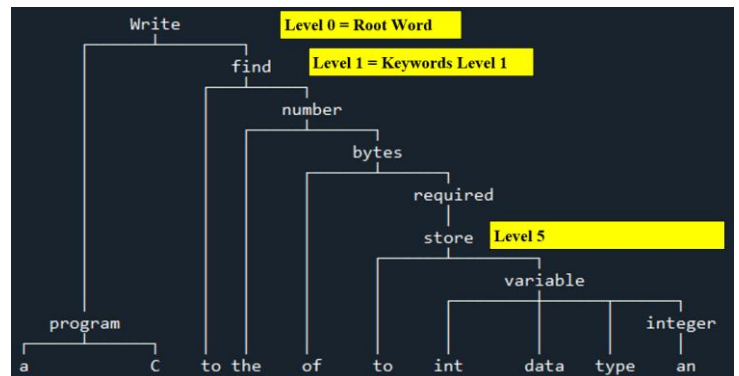
Table 4. Scenario of no verb form with “VB” tag but with other verb forms such as “VBP” in a single-sentence question.

Input Sentence	For the Wye-Wye Circuit in Figure 2, Calculate the Line Currents.
After POS Tagging, Stop Words and Punctuation Removal	[('for', 'IN'), ('wye-wye', 'JJ'), ('circuit', 'NN'), ('figure', 'NN'), ('2', 'CD'), ('calculate', 'VBP'), ('line', 'NN'), ('currents', 'NNS')] - No 'VB' tagged verb. - There is a 'VBP' tagged verb—'calculate'.

Table 5. Scenario of more than one verb tagged with the “VB” tag in a single-sentence question.

Input Sentence	Write a C Program to Find the Number of Bytes Required to Store an Integer (Int) Data Type Variable.
After POS Tagging, Stop Words and Punctuation Removal	[('write', 'VB'), ('c', 'NN'), ('program', 'NN'), ('find', 'VB'), ('number', 'NN'), ('bytes', 'NNS'), ('required', 'VBN'), ('store', 'VB'), ('integer', 'NN'), ('int', 'NN'), ('data', 'NNS'), ('type', 'NN'), ('variable', 'NN')] 3 verbs tagged with 'VB'—write, find and store.

Stanford Parser's UD Tree



Root word = ['write']

Keywords Level 1 = ['find']

Keywords Level 5 = ['store']

3.4. Question Classification

The question classification (QC) module aims to classify questions based on the RW and KW extracted from the VE module, using WordNet similarity. The RW and KW are compared with the verbs listed from BT. Each verb has a set of synsets. The WordNet similarity identifies how similar the two sets of synsets—a set from the RW and KW—are before utilizing another set from the verbs listed from BT. The similarity score is in the range of 0 to 1, where the higher similarity score obtained means that the synsets are closer to each other in terms of word sense. Generally, the algorithm for the QC module based on WordNet similarity can be grouped into three possible scenarios. The first scenario is the exact matching between the RW and one of the BT verbs (BV). In this scenario, the question can be directly classified according to the BV. The second scenario is when the RW matches with exactly one of the BVs, and the BV exists in multiple BT levels. The QC requires WordNet similarity between the KW and BV for classification. The third scenario is where no exact matching occurs, and the WordNet similarity score between the RW, KW and BV is used in classification. Each RW, KW and BV from BT level 1 to level 6 has its own synset, so a huge number of synset similarity scores can be generated. A three-dimensional matrix was designed to summarize the scores and tabulate them into each Bloom level. The method used to summarize the matrix was evaluated, with the best result implemented in the QC module. QC identifies the BT level with the highest similarity score as the BT level of the questions.

As for the multi-sentence type of questions, the questions were tokenized into single sentences such as S1, S2, etc. Every single sentence was processed in the same manner. QC identified the BT level with the highest similarity score as the BT level of every single sentence. The highest BT level among the sentences was then identified as the BT level of the multi-sentence type of question. For example, S1 was classified as BT Level 2, while S2 was classified as BT Level 3. The multi-sentence type of question was then classified to the highest BT level: BT Level 3. The highest BT level selection in the multi-sentence questions was consistent with the manual question classification practice.

4. Result and Discussion

This section aims to discuss the evaluation results of the approaches implemented in the QP, VE, and QC module in the QCS.

4.1. Question Preprocessing

The question preprocessing (QP) module uses the default tokenizer provided by the NLTK package, which is the PunktSentenceTokenizer. The PunktSentenceTokenizer performs the sentence and word tokenization with a minimum error which shows 99% and 92% accuracy, respectively. Subsequently, the words are tagged with “NN”, “VB”, etc., using the POS tagger. Evaluation of POS tagging using the Perceptron model provided by the NLTK package, the “English-left3wrods-distsim” model and the “English-bidirectional-distsim” model provided by the Stanford parser package was conducted. The outcome of the tagged “VB” was compared with the expected verbs. The default Perceptron POS tagger showed a 55.5% accuracy, while the Stanford POS tagger with the “English-bidirectional-distsim” model and the “English-left3wrods-distsim” model showed 72.5% and 80.5%, respectively. Therefore, the Stanford POS tagger with the “English-left3wrods-distsim” was implemented in the QP module. The POS taggers were evaluated further by extending the POS tagging result—the word tokens tagged with VB, VBD, VBG, VBN, VBP, and VBZ—against the expected verbs. This yielded a better result with an accuracy of 60.5%, 82.0% and 83.5%, respectively, using Perceptron, the “English-left3wrods-distsim” model, and the “English-bidirectional-distsim” tagger.

4.2. Verb Extraction by Verb Form Analysis and Rules Applied

Verb extraction (VE) is a module that identifies the type of word to be extracted from a question, such as “VB”, “VBD”, “VBG”, “VBN”, “VBP”, and “VBZ”. These verb forms are generally a verb but with additional conditions to differentiate themselves from each other, as shown in Table 6. Experiments were conducted to determine the accuracy of “VB” extracted from the questions with the verb tagged manually by an expert. Based on the result from the experiment conducted on 200 examination questions, an 80.5% accuracy was obtained.

Table 6. Verb forms description and its example.

Verb Form	Verb Form Description	Verb Example
VB	Verb, base form	Take
VBD	Verb, past tense	Took
VBG	Verb, present participle	Taking
VBN	Verb, past participle	Taken
VBP	Verb, non-third person singular present	Take
VBZ	Verb, third person singular present	Takes

There were 8% of errors due to a lack of verbs tagged with “VB” in the questions. The errors always occurred in questions that had a verb as the first word of the sentence, but this was tagged as a noun by the POS tagger such as “state”, “design”, “name”, “list”,

“point”, and “estimate”. This could be rectified by setting Rule 1, converting the first-word noun into a verb and referring to Table 7. Another 3% of errors was due to no “VB” being tagged in the examination question, though “VBD”, “VBG”, “VBN”, “VBP”, and “VBZ” were tagged. For example, “calculate” was tagged as a “VBP” instead of “VB”. Rule 2 was applied to extract the verb with “VBP” and “VBZ”. After setting these two rules, the accuracy of the correct verb identified was optimized from 80.5% to 91.5%.

Table 7. Effectiveness of verb extraction process after applying rules.

Process		Grade	The 200 Single-Sentence Question Dataset
Verb extraction process to extract the word tokens tagged with “VB” only by the Stanford POS tagger with the “left3words” model.		Accurate	80.5%
		ERR1	8.0%
		ERR2	11.5%
Rule Applied	RULE #1: First Word is a VerbIf meeting following conditions: 1. There are many word tokens tagged with “VB”; 2. The first word is tagged as a noun with “NN” or “NNP”; 3. The first word is one of the words in the defined list of possible incorrectly tagged verbs—‘state’, ‘design’, ‘name’, ‘estimate’, ‘point’ and ‘list’.		8.0%
	RULE #2: “VBP” and “VBZ”If meeting following conditions: 1. There are many word tokens tagged with “VB”; 2. There are word tokens tagged with “VBP” or “VBZ”.		3.0%
Verb extraction process to extract the word tokens tagged with “VB” only by the Stanford POS tagger with the “left3words” model, was then applied to RULE #1 and #2		Accurate	91.5%

4.3. Similarity Matrix Scores Analysis and Sorting in Question Classification Module

The question classification (QC) module uses the WordNet Similarity approach to obtain similarity scores between the verbs extracted—including RWs and KWs—and the verbs from the BT list—BVs. Every verb, either a RW, KW or BV, has several synsets that can range from a few to as many as 75 synsets. WordNet generates a similarity score between each synset from the RWs or KWs and each synset from the BVs. Huge similarity scores are generated, and a two-dimensional matrix is utilized to store the similarity scores, where the matrix’s X axis is the BVs’ synset and the Y axis is the RWs’ or KWs’ synset.

Figure 2 is an example of the two-dimensional similarity scores matrix of RW—“write”—and the BVs of Bloom Level 1. The RW—“write”—consisted of ten synsets, while the Bloom Level 1 consisted of 14 BVs with a total of 77 synsets. As a result, a total of 770 similarity scores was generated and stored in the two-dimensional matrix as shown in Figure 2. The similarity scores in the matrix were then sorted from a high to low score value, as shown in Figure 3. Because there were six Bloom levels, the two-dimensional matrix extended to the three-dimensional matrix where the Z axis referred to BT levels, as in Figure 4.

Subsequently, the matrix needed to be summarized into a simple score representation for classification, which was a single similarity score for each Bloom level. Three types of statistical approaches were identified—the mean, median and maximum. Table 8 shows an example of the RW where “write” is summarized using the three statistical approaches. Both mean and maximum approaches resulted in a conclusive highest score at the single Bloom level, but the mean approach had the highest similarity score at Bloom Level 1, and the maximum approach was at Bloom Level 6. An experiment was conducted to evaluate and identify the best statistical approach to be used and summarize the similarity matrix into a single-score representation. Experiment results have shown the mean approach to be the best approach considering how all test questions can be successfully classified to a single Bloom level with the highest accuracy, as shown in Table 9.

RW - 'WRITE"	write.v.01	0.2222	0.2857	0.1333	0.2222	0.2857	0.2857	X.XXXX	X.XXXX	0.2500	0.2000	0.2000											
	write.v.02	0.2000	0.2500	0.1250	0.2000	0.2500	0.2500	X.XXXX	X.XXXX	0.2222	0.1818	0.1818											
	publish.v.03	0.2222	0.2857	0.1333	0.2222	0.2857	0.2857	X.XXXX	X.XXXX	0.2500	0.2000	0.2000											
	write.v.04	0.2500	0.3333	0.1429	0.2500	0.3333	0.3333	X.XXXX	X.XXXX	0.2857	0.2222	0.2222											
	write.v.05	0.2000	0.2500	0.1250	0.2000	0.2500	0.2500	X.XXXX	X.XXXX	0.2222	0.1818	0.1818											
	compose.v.02	0.2500	0.3333	0.1429	0.2500	0.3333	0.3333	X.XXXX	X.XXXX	0.2857	0.2222	0.2222											
	write.v.07	0.2000	0.2500	0.1250	0.2000	0.2500	0.2500	X.XXXX	X.XXXX	0.2222	0.1818	0.1818											
	write.v.08	0.1667	0.2000	0.1111	0.1667	0.2000	0.2000	X.XXXX	X.XXXX	0.1818	0.1538	0.1538											
	spell.v.03	0.2857	0.4000	0.1538	0.2857	0.4000	0.4000	X.XXXX	X.XXXX	0.3333	0.2500	0.2500											
	write.v.10	0.2222	0.2857	0.1333	0.2222	0.2857	0.2857	X.XXXX	X.XXXX	0.2500	0.2000	0.2000											
Bloom Level 1 Synsets		match.v.02		reproduce.v.02		recall.v.04		label.v.05		label.v.01		define.v.04		*****		*****		match.v.03		submit.v.02		list.v.04	
	1	2	3	4	5	6			75	76	77												

Figure 2. A two-dimensional similarity score matrix.

RW - 'WRITE'	write.v.01	0.8571	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	X.XXXX	X.XXXX	0.1429	0.1333	0.1333
	write.v.02	0.6667	0.4615	0.4615	0.4286	0.2857	0.2857	0.2857	X.XXXX	X.XXXX	0.1667	0.1538	0.1250
	publish.v.03	0.5714	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	X.XXXX	X.XXXX	0.1429	0.1333	0.1333
	write.v.04	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000	X.XXXX	X.XXXX	0.1538	0.1429	0.1429
	write.v.05	0.6667	0.4615	0.4615	0.4286	0.2857	0.2857	0.2857	X.XXXX	X.XXXX	0.1667	0.1538	0.1250
	compose.v.02	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000	0.4000	X.XXXX	X.XXXX	0.1538	0.1429	0.1429
	write.v.07	0.8571	0.2857	0.2857	0.2857	0.2857	0.2857	0.2857	X.XXXX	X.XXXX	0.1333	0.1250	0.1250
	write.v.08	0.6667	0.2222	0.2222	0.2222	0.2222	0.2222	0.2222	X.XXXX	X.XXXX	0.1176	0.1111	0.1111
	spell.v.03	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	X.XXXX	X.XXXX	0.1667	0.1538	0.1538
	write.v.10	0.5714	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	X.XXXX	X.XXXX	0.1429	0.1333	0.1333
Bloom Level 1 Synsets													
Top 1 Score synset													
Top 2 Score synset													
Top 3 Score synset													
Top 4 Score synset													
Top 5 Score synset													
Top 6 Score synset													
.....													
.....													
Last 3 Score synset													
Last 2 Score synset													
Last 1 Score synset													
	1	2	3	4	5	6			75	76	77		

Figure 3. A sorted two-dimensional similarity score matrix in descending order.

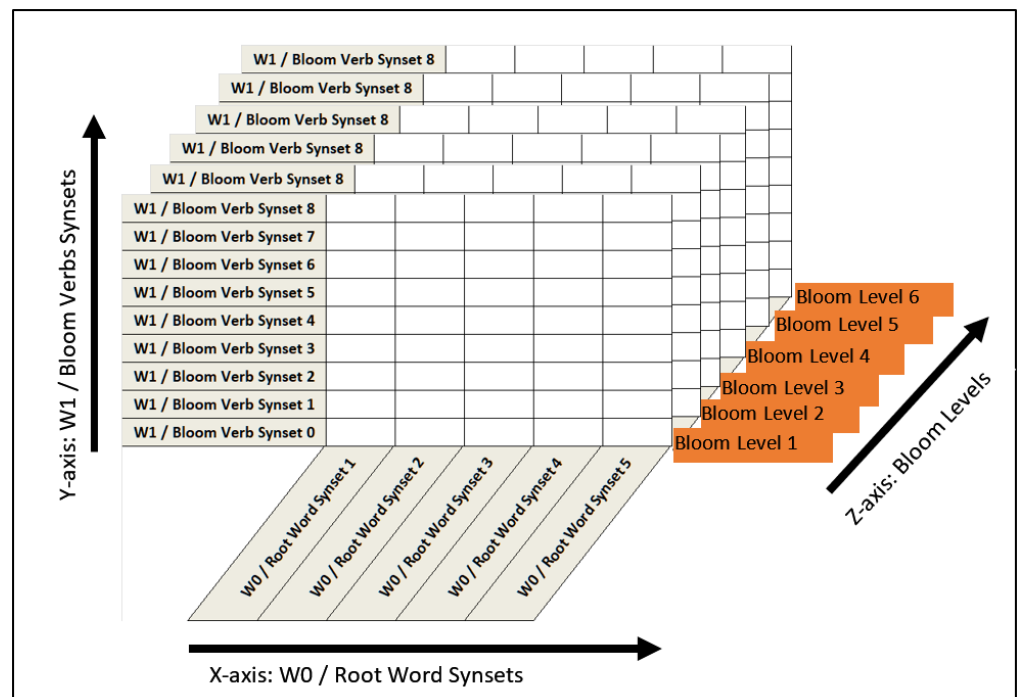


Figure 4. A three-dimensional similarity score matrix.

Table 8. Similarity score matrix summarized into single score for each Bloom level using statistical approaches—mean, maximum and median.

Statistical Approaches to Summarize Similarity Scores of RW—“Write”	Bloom Level 1	Bloom Level 2	Bloom Level 3	Bloom Level 4	Bloom Level 5	Bloom Level 6
MEAN	0.2549	0.2604	0.2702	0.2555	0.2574	0.2687
MEDIAN	0.2500	0.2500	0.2500	0.2500	0.2500	0.2500
MAXIMUM	0.8571	0.8571	0.7500	0.6000	0.8571	1.0000

Table 9. Effectiveness of statistics measures—mean, maximum and median in the 200-examination question dataset.

Statistics Measures	Grade	The 200-Examination Question Dataset
Mean	Successful Bloom Level Classification and Accurate	83.0%
	Successful Bloom Level Classification but Inaccurate	17.0%
	Inconclusive Bloom Level Classification	0.0%
Maximum	Successful Bloom Level Classification and Accurate	37.0%
	Successful Bloom Level Classification but Inaccurate	33.0%
	Inconclusive Bloom Level Classification	30.0%
Median	Successful Bloom Level Classification and Accurate	30.0%
	Successful Bloom Level Classification but Inaccurate	5.0%
	Inconclusive Bloom Level Classification	65.0%

According to Diab and Sartawi [4], they filtered and modified the data lists in their research by collecting only 75–100% of the synset verbs in order to obtain better accuracy with a similarity measurement. Table 10 shows the mean value of the top 20, 40, 80, 100 and 110 similarity scores per each Bloom level, and the Bloom level with the highest similarity score switched from Level 2 to Level 3 when the pick of the top similarity scores increased from the top 80 to the top 100. Hence, another experiment was conducted to identify the suitable amount of similarity scores to be summarized. This ensured that the most relevant synsets verbs could be considered in the QC. The experiment result showed the top 120 similarity scores and the pick of similarity scores yielding the highest Bloom level classification accuracy in the 200-examination question dataset. However, the accuracy was improved if different top similarity scores were picked depending on the condition of the RW matching exactly with any BVs. If the RW matched exactly with any of the BVs from multiple BT levels, the top 120 similarity scores picked provided the best Bloom level classification accuracy. If no exact matching between RWs with any BVs was found, the top 80 similarity scores picked provided the greatest accuracy.

Table 10. Bloom level with highest score depending on the amount of similarity scores summarized.

Top Scores Pick	Bloom Level 1	Bloom Level 2	Bloom Level 3	Bloom Level 4	Bloom Level 5	Bloom Level 6	Bloom Level with Highest Score
Top 20 Pick	0.329	0.371	0.337	0.328	0.309	0.322	Bloom Level 2
Top 40 Pick	0.297	0.330	0.299	0.303	0.268	0.289	Bloom Level 2
Top 60 Pick	0.275	0.305	0.282	0.284	0.268	0.276	Bloom Level 2
Top 80 Pick	0.265	0.287	0.282	0.267	0.268	0.276	Bloom Level 2
Top 100 Pick	0.265	0.270	0.282	0.257	0.268	0.276	Bloom Level 3
Top 110 Pick	0.265	0.260	0.282	0.257	0.268	0.276	Bloom Level 3

As a result, the top 80 similarity scores picked were used if the RW did not match exactly with any BVs; the top 120 similarity scores picked were used if the RW matched exactly with any of the BVs. This pick combination resulted in an overall 83% accuracy in the Bloom level classification of the 200-examination question datasets. Table 11 shows an

analysis of the Bloom level classification result of the 200-examination question datasets by subject domain and the multi-sentence question type. The Programming, Electronics, Power, Physics, and Mechanical domains obtained a greater than 80% accuracy in QC. The Automation and Management domains experienced a relatively low accuracy due to higher inaccurate Bloom level classifications in their domains with multi-sentence type questions. Thus, further analysis and study were needed to improve the accuracy of the QC in multiple sentences.

Table 11. Analysis of 83% accuracy obtained in Bloom level classification of the 200-question dataset.

Subject Domain	200-Question Dataset	Bloom Level Classification Accuracy	Multi-Sentences Question Type Bloom Level Classification Accuracy
Programming	25.0%	92.0%	90.0%
Electronics	11.5%	91.3%	100.0%
Power	20.5%	87.8%	77.8%
Physics	8.0%	81.3%	66.7%
Mechanical	20.5%	80.5%	57.1%
Automation	6.5%	61.5%	16.7%
Management	8.0%	56.3%	25.0%
	100%	83.0%	

4.4. Evaluation of Question Classification Approaches

The 83% accuracy result obtained in this paper is considered satisfactory if compared to the research conducted by different researchers, as shown in Table 2. This paper proposed the WordNet-Wu similarity approach in examination question classifications because the WordNet-Wu similarity approach is the best performer among other approaches based on the experimental results, as shown in Table 12. The experiment was conducted to evaluate other approaches, such as the WordNet-Path similarity approach, the Cosine Similarity approach recommended by Diab and Sartawi [4], and the Word2Vec semantic similarity approach recommended by Mohammed and Omar [6]. All approaches were tested on the same 200-examination question dataset, which consisted of 22.5% multi-sentence question types and was randomly picked from several subject domains.

Table 12. Comparison of different approaches on question classification using same 200-question dataset with 22.5% of multi-sentence type questions and questions from multi-domains.

Question Classification Approach	Performance Result
WordNet—Wu Similarity (proposed in this study)	Accuracy 83.0%
WordNet—Path Similarity	Accuracy 64.5%
Cosine Similarity	Accuracy 43.0%
Word2Vec (semantic similarity)	Accuracy 55.0%

5. Conclusions

This paper presents a continuous study of the framework proposed by Goh et al. [17] for the examination question classifications according to Bloom's taxonomy. We aimed to study the framework of question classifications in order to improve its accuracy. Question pre-processing (QP), verb extraction (VE), and question classification (QC) modules were the three important modules in the pipeline processing structure. The right keywords were identified from 200 final examination questions from various subjects covering areas such as Programming, Electrical, Electronics, Mechanical, Automation, Mechanical, and Management. The PunktSentenceTokenizer, the left3words model of the Stanford POS

tagger, and the two rules implemented improved the accuracy of keyword identification from 80.5% to 91.5%.

Once the correct verbs were extracted from the examination questions, the WordNet similarity scores between the verbs extracted and the Bloom verb list were used to identify the examination question category. The mean score of the tabulated similarity result was computed, and the top 80 similarity scores were summarized. Finally, the framework achieved 83% accuracy in question classifications with semantic similarity and a rule-based approach.

Author Contributions: Methodology, T.T.G. and H.M.; software, T.T.G. and N.A.A.J.; formal analysis, T.T.G.; data curation, T.T.G. and M.N.I.; writing—original draft preparation, T.T.G. and H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from National Defence University of Malaysia (NDUM) and UOW Malaysia KDU University College.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank National Defence University of Malaysia (NDUM) through its Centre for Research and Innovation Management (PPPI) and UOW Malaysia KDU University College to support this paper for publication and funded the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Osman, A.; Yahya, A.A. Classifications of Examination Questions using linguistically-Motivated Features: A Case Study based on Bloom's Taxonomy. In Proceedings of the Sixth International Arab Conference on Quality Assurance in Higher Education (IACQA), Khartoum, Sudan, 9–11 February 2016; Volume 467, p. 474.
- Gupta, M.; Verma, S.K.; Jain, P. Detailed Study of Deep Learning Models for Natural Language Processing. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 18–19 December 2020; pp. 249–253. [\[CrossRef\]](#)
- Jayakodi, K.; Bandara, M.; Perera, I.; Meedeniya, D. WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy. *Int. J. Emerg. Technol. Learn.* **2016**, *11*, 142. [\[CrossRef\]](#)
- Diab, S.; Sartawi, B. Classification of Questions and Learning Outcome Statements (LOS) into Bloom's Taxonomy (BT) by Similarity Measurements Towards Extracting of Learning Outcome from Learning Material. *Int. J. Manag. Inf. Technol.* **2017**, *9*, 2. [\[CrossRef\]](#)
- Mohamed, O.J.; Zakar, N.A.; Alshaikhdeeb, B. A Combination Method of Syntactic and Semantic Approaches for Classifying Examination Questions into Bloom's Taxonomy Cognitive. *J. Eng. Sci. Technol.* **2019**, *14*, 935–950.
- Mohammed, M.; Omar, N. Question Classification based on Bloom's Taxonomy Cognitive Domain using modified TF-IDF and Word2vec. *PLoS ONE* **2020**, *15*, e0230442. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jayakodi, K.; Bandara, M.; Perera, I. An Automatic Classifier for Examination Questions in Engineering: A Process for Bloom's Taxonomy. In Proceedings of the 2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Zhuhai, China, 10–12 December 2015; pp. 195–202. [\[CrossRef\]](#)
- Goh, T.T.; Mohamed, H.; Jamaludin, N.A.A.; Ismail, M.N.; Chua, H.S. Questions Classification According to Bloom's Taxonomy using Universal Dependency and Word Net. *Test Eng. Manag.* **2020**, *82*, 4374–4385.
- Haris, S.S.; Omar, N. A Rule-based Approach in Bloom's Taxonomy Question Classification through Natural Language Processing. In Proceedings of the 7th International Conference on Computing and Convergence Technology (ICCCT), Seoul, Republic of Korea, 3–5 December 2012; pp. 410–414.
- Biswas, P.; Sharan, A.; Kumar, R. Question Classification using Syntactic and Rule-based Approach. In Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, 24–27 September 2014; pp. 1033–1038. [\[CrossRef\]](#)
- Kusuma, S.F.; Siahaan, D.; Yuhana, U.L. Automatic Indonesia's questions classification based on bloom's taxonomy using Natural Language Processing a preliminary study. In Proceedings of the 2015 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 16–19 November 2015; pp. 1–6. [\[CrossRef\]](#)
- Sulaiman, S.; Wahid, R.A.; Ariffin, A.H.; Zulkifli, C.Z. Question Classification based on Cognitive levels using Linear SVC. *Test Eng. Manag.* **2020**, *83*, 6463–6470.
- Mohammed, M.; Omar, N. Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2018**, *8*, 1679–1685. [\[CrossRef\]](#)

14. Contreras, J.O.; Hilles, S.; Bakar, Z.A. Essay Question Generator based on Bloom's Taxonomy for Assessing Automated Essay Scoring System. In Proceedings of the 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Cameron Highlands, Malaysia, 15–17 June 2021; pp. 55–62. [[CrossRef](#)]
15. Sangodiah, A.; Ahmad, R.; Wan Amand, W.F. Taxonomy Based Features in Question Classification using Support Vector Machine. *J. Theor. Appl. Inf. Technol.* **2017**, *95*, 2814–2823.
16. Minaee, S.; Liu, Z. Automatic Question-Answering using a Deep Similarity Neural Network. In Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 14–16 November 2017; pp. 923–927. [[CrossRef](#)]
17. Goh, T.T.; Mohamed, H.; Jamaludin, N.A.A.; Ismail, M.N.; Chua, H.S. A Comparative Study on Part-of-Speech Taggers' Performance on Examination Questions Classification According to Bloom's Taxonomy. *J. Phys. Conf. Ser.* **2022**, *2224*, 012001. [[CrossRef](#)]
18. Tian, Y.; Lo, D. A Comparative Study on the Effectiveness of Part-Of-Speech Tagging Techniques on Bug Reports. In Proceedings of the 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), Montreal, QC, Canada, 2–6 March 2015; pp. 570–574. [[CrossRef](#)]
19. Go, M.P.; Nocon, N. Using Stanford Part-of-Speech Tagger for the Morphologically-rich Filipino Language. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC), Cebu City, Philippines, 16–18 November 2017; pp. 81–88.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.