ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2022.1.008

Identifying Optimal Baseline Variant of Unsupervised Term Weighting in Question Classification Based on Bloom Taxonomy

Anbuselvan Sangodiah✉, Tham Jee San, Yong Tien Fui, Lim Ean Heng, Ramesh Kumar Ayyasamy, Norazira Binti A Jalil

Department of Information System, Universiti Tunku Abdul Rahman, Kampar, Malaysia

anbuselvan@utar.edu.my✉, jeesan0614@1utar.my, yongtf@utar.edu.my, ehlim@utar.edu.my, rameshkumar@utar.edu.my, noraziraj@utar.edu.my

Abstract

Examination is one of the common ways to evaluate the students' cognitive levels in higher education institutions. Exam questions are labeled manually by educators in accordance to Bloom's taxonomy cognitive domain. To ease the burden of the educators, several past research works have proposed the automated question classification based on Bloom's taxonomy using the machine learning technique. Feature selection, feature extraction and term weighting are common ways to improve the accuracy of question classification. Commonly used term weighting method in the past work is unsupervised namely TF and TF-IDF. There are several variants of TF and TFIDF and the most optimal variant has yet to be identified in the context of question classification based on BT. Therefore, this paper aims to study the TF, TF-IDF and normalized TF-IDF variants and to identify the optimal variants that can be used as baseline term weighting scheme. To investigate the variants, two different classifiers were used, which are Support Vector Machine (SVM) and Naïve Bayes. The average accuracies achieved by TF-IDF and normalized TF-IDF variants using SVM classifier were 63.7% and 71.7% respectively, while using Naïve Bayes classifier the average accuracies for TF-IDF and normalized TF-IDF were 62.4% and 63.4% respectively. Generally, the normalized TF-IDF variants outperformed TF and TF-IDF variants in both accuracy and F1-measure respectively. Further statistical analysis using t-test shows that the differences in accuracy between normalized TF-IDF and TF, TF-IDF are significant. According to the results of this study, the Normalized TF-IDF2 variant had the greatest accuracy of 73.3% among normalized TF-IDF variants, whereas the TF-IDF3 variant had the highest accuracy of 70.8% among unnormalized TF-IDF variants. As a result, the normalized TF-IDF2 and unnormalized TF-IDF3 variations are useful for benchmarking and comparing with other term weighting techniques in question classification based on BT in future research.

Keywords: *Baseline Term Weighting, Question Classification, Bloom Taxonomy, Support Vector Machine, Naïve Bayes.*

Received: 21 December 2021
Accepted: 12 April 2022
Online: 22 April 2022
Published: 30 June 2022

1 Introduction

In the education field, the written examination is an assessment method that is commonly used by academicians to evaluate the student's achievement of learning [11]. When lecturers design the exam questions, they should ensure that there is a match between the course learning outcomes and assessment [18]. Therefore, it is crucial to use a suitable way to classify the exam questions into their correct category or class to measure the student's cognitive level [18]. In fact, many lecturers follow Bloom's Taxonomy (BT) as a guideline to produce a high-quality assessment [21]. This BT involves six levels: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. In Fig. 1, the levels are arranged accordingly from the lowest level of the cognitive domain (Knowledge) to the highest level (Evaluation). The description for each level is presented in Table 1.

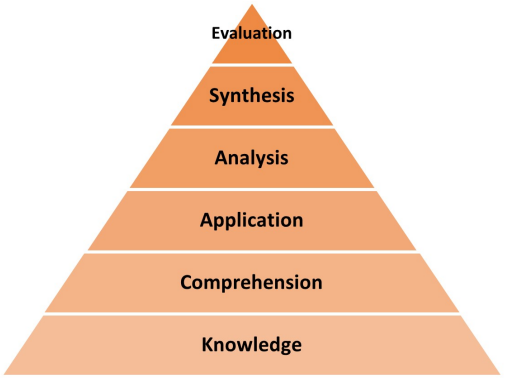

Figure 1: Bloom's Taxonomy Cognitive Domain.

Table 1: Explanation of Bloom's Taxonomy Cognitive Domain.

	Level	Definition	Verbs Example
1	Knowledge	Remembering, memorizing of previously learned material	Name, define, describe, list
2	Comprehension	Understanding the meaning of learned material by interpreting, translating, and comparing	Illustrate, identify, discuss, classify
3	Application	Applied learned knowledge in concrete and new situations	Apply, demonstrate, calculate, develop
4	Analysis	Break down material into components to classify, distinguish or identify relationship between them	Analyze, compare, contrast, differentiate
5	Synthesis	Integrating ideas or elements together to form a new solution	Synthesize, establish, create, prepare
6	Evaluation	Judge or criticize the value of material based on definite criteria	Evaluate, propose, argue, judge

To produce a high-quality assessment that matches course learning objectives, many educators applied the exam question classification based on BT. Unfortunately, most of them faced some problems throughout the manual classification process, for example, the problem stated in [14]. The educators need to spend a long time to conduct the classification process if there are lots of question items, through the identifying of BT keyword exist within the question. For example, the educator classified the below question: "Define E-commerce business." into knowledge level. Therefore, the automatic classification of exam questions based on BT is highly required to solve their difficulty. Some researchers proposed their approach to classify questions automatically by using machine learning algorithms in their study [14, 28, 29, 33, 45]. Exam question classification is more challenging than text classification although both classification processes are similar, the presence of words in a question is limited and less when the question item is being classified. The purpose of exam question classification is to identify the difficulty level of given question and assign it into pre-defined categories. Using machine learning technique, the level of difficulty of exam question can be determined automatically in accordance with BT cognitive level.

Past research work in question classification focused on feature extractions, feature selections, and term weighting [1, 4, 27, 43, 46]. Lately, some studies in exam question classification and text classification have shown that the term weighting method can improve the performance of the classifier in classifying exam questions and text effectively [4, 12, 13, 21]. Term weighting is a process that can indicate the presence of each term in a document and assign weight to the term accordingly. Generally, the term weighting scheme can be divided into two types, which are unsupervised and supervised. The unsupervised term weightings that have been used widely in text classification include Binary, Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) [16]. Besides these commonly used methods, other unsupervised term weighting methods such as TF Probabilistic Inverse Document Frequency (TF-PIDF), Modified TF (mTF), Modified IDF (mIDF) are proposed in some past work and discussed in [3]. As for supervised term weightings, the study of some commonly used supervised term weightings is conducted also in [3], which consists of Term Frequency Information Gain (TF-IG), Term Frequency-Relevance Frequency (TF-RF), Term Frequency Chi-Square (TF-x^2), Term Frequency Bi-normal Separation (TF-BNS) and others.

Despite the term weightings used in exam question classification adopted from text classification, not all the recent unsupervised and supervised term weighting aforementioned in the text classification can be directly used in the exam question classification based on BT. So far in the context of exam question classification, the unsupervised term weightings used are TF, Binary, TF-IDF, E-TFIDF and TFPOS-IDF. Unlike in text classification, the variants of TF and TF-IDF have not been explained and studied well hence optimal variant of TF-IDF has not been identified. Identifying the appropriate variant particularly TF-IDF in increasing classification accuracy is crucial. This is because most of the researchers in this area whose work involves comparison of question classification accuracy in terms of term weighting, feature selection, feature extraction or combination of classifiers may not use the optimal TF-IDF variant to compare with other improved term weighting schemes or advanced classification technique such as deep neural network. In view of this, this paper aims to evaluate the unsupervised term weighting schemes by using classification algorithms. The most optimal variant of TF-IDF was identified in this paper. Several classifiers such as Support Vector Machine (SVM) and Naïve Bayes were used to compare the effectiveness of variants in classification accuracy. The results indicated how the usage of these unsupervised term weighting variants could affect the accuracies of classifying exam questions based on Bloom's Taxonomy.

This paper is divided into four main sections. Sec-

Table 2: Previous Research Work in Question Classification.

No.	Author (Year)	Reference	Term Weighting	Feature Selection	Feature Extraction	Machine Learning
1	Abduljabbar and Omar (2015)	[1]		✓		
2	Osman and Yahya (2016)	[29]		✓	✓	
3	Sangodiah et al. (2017)	[33]	✓		✓	
4	Mohammed and Omar (2018)	[21]	✓		✓	
5	Aninditya et al. (2019)	[4]			✓	
6	Mohammed and Omar (2020)	[22]	✓		✓	
7	Waheed et al. (2021)	[42]				✓
8	Shaikh et al. (2021)	[35]				✓
9	Sangodiah et al. (2021)	[34]	✓			

tion 1 is the introduction to the term weighting schemes in exam question classifications. Section 2 reviews the existing research associated with the aforementioned term weightings. Section 3 demonstrates the methodology that entails the question classification model and the variants of term weightings that will be used in the study. Section 4 discusses the results and discussion. Section 5 presents the conclusion of the research.

2 Literature Review

The exam question classification is a procedure that determined the difficulty level of an exam question and assigned it to pre-determined categories based on BT. To improve the classification performance, term weighting is one of the useful solutions despite feature selection or feature extraction methods. Term weightings can be divided into two types, which are unsupervised term weighting and supervised term weighting. As the word that existed in a question is limited, the unsupervised term weighting methods that focused on the contribution of each word accordingly by calculating the weight value on each term that exists in the document [15], is more suitable to be implemented in exam question classification compared with the supervised term weighting. Therefore, to review the usage of unsupervised term weighting, feature extraction, or feature selection methods, those previous work that is related to text and exam question classification are discussed. Since this study focused on the comparison of term weighting variants, some existing comparison work for text classification is also being presented.

2.1 Related Work in Text Classification

For text classification, some researchers performed a comparative study on term weighting schemes. They compared the effectiveness of different term weighting schemes in improving the text classification result. Since the unsupervised term weighting methods are used extensively in exam question classification, therefore only the result by using unsupervised term weighting methods will be analyzed. In [19], the authors conduct their comparative study by using different unsupervised term weighting methods. The unsupervised term weighting methods used in this study are TF and TF-IDF. The highest average f-score result of 87.22 was obtained when using TF-IDF variant, $1 + \log(f_{t,d}) \cdot \log \frac{|D|}{n}$ indicated that TF-IDF performed better than TF in text classification. Another work by [23], the researchers evaluated and compared the text classification result obtained by using various unsupervised term weighting methods, such as Binary TF, TF, LogTF, TF-IDF, LogTFIDF and BM25. Based on the classification result on 20Newsgroups dataset obtained by using Random Forests (RF) classifier, TF-IDF generated the highest F1-measure value of 0.592 in classifying this dataset. The result indicated that TF-IDF variant, $f_{td} \cdot \log \frac{|D|}{n}$ can work effectively in text classification. In [6], the unsupervised term weighting methods used are TF, TF-IDF and TF-IDF-ICSDF. By using SVM classifier in classifying Reuters-21578 dataset with 3000 features, the micro F1-measure result obtained for TF-IDF variant is the highest, which is a value of 0.966. Besides that, the micro F1-measure result of 0.8893 get when using TF-IDF variant, $f_{td} \cdot \log \frac{|D|}{n}$ for the same dataset is considered as the highest and most satisfied result. By reviewed these past related works, it concluded that identifying the most optimal variant in text classification is crucial as it can increase the classification accuracy.

2.2 Related Work in Exam Question Classification

Besides text classification, some researchers have focused their studies on exam question classification. Some researchers implemented the exam question classification by applying different methods, such as term weighting, feature selection and feature extraction methods in their study to increase the classification accuracy. Table 2 summarized some existing research works that applied term weighting, feature selection or feature extraction methods in exam question classification.

In (1), the authors proposed a voting algorithm that integrated the strength of three machine learning classifiers, which are SVM, Naïve Bayes (NB), and k-Nearest Neighbour (k-NN). Besides that, they applied three Feature Selection (FS) methods, Mutual Information (MI), Chi-Square statistic, and Odd Ratio (OR) to simplify the classification process. The

researchers used a voting

algorithm that combined the outputs of three classifier approaches with each FS method. By comparing the macro F1-measure results obtained from three base-level classifiers separately with the combination approach by using MI method in a weighted feature size equal to 250, the combination approach gave the highest value of 92.28, indicating that the combination approaches able to determine the cognitive level for programming questions effectively.

In (2), a comparative study of various machine learning methods and linguistically motivated features used in classifying exam questions based on BT cognitive levels automatically is presented. Through the experiment conducted, the average accuracy result of above 0.6 obtained by four classifiers which are SVM, Logistic regression, decision trees and NB using the unigram feature concluded that using machine learning models in question classification can achieve a high level of accuracy. Besides that, the researchers reported that the Logistic regression model using a combination of Unigrams and Bigrams features generated a higher accuracy result of 0.7683 compared to the accuracy result of 0.7667 obtained by SVM model and unigram feature. The result indicated that the implementation between machine learning models with the combination of linguistically motivated features, which features can perform syntactic analysis of text deeply, able to increase the classification accuracy. Lastly, the authors also concluded that it is important to focus on machine learning models and linguistically motivated features, such as the combination of Unigrams and Bigrams features that can increase the classification accuracy result simultaneously when performing exam question classification.

In another research (3), an exam classification framework is proposed by using different feature types. Besides some general feature types such as Bag-of-Words (BOW) and POS Tagging, a new feature type that has strong dependence with BT cognitive levels called taxonomy based is proposed by the authors to classify exam questions from various areas. The performance of question classification by using different feature types such as BOW, the combination of BOW and POS (BWP), the combination of BOW and general taxonomy (BWG), and the combination of BOW with general specific taxonomy (BWGS) are evaluated and compared based on the accuracy results obtained. Through the experiment, the accuracy result obtained from the application of general feature types is lower than taxonomy-based feature types. The highest accuracy result of 0.729 was obtained when the experiment is conducted with SVM classifier using BWGS feature, one of the taxonomy-based features. It can be concluded that the proposed taxonomy-based feature such as BWGS feature can improve the accuracy result in exam question classification.

Enhanced TF-IDF is introduced in the research (4) to improve the effectiveness of exam question classification based on BT cognitive domain. The part-of-speech tagger is applied to assign impact factor for each word that exists in the exam question. After that, the classification performance of several classifiers such as SVM, Naïve Bayes and K-Nearest Neighbour is evaluated. The highest average F1-measure result of 86% obtained from SVM classifier indicated that the usage of enhanced feature E-TFIDF works more effectively in increasing the classification accuracy compared to TF-IDF. It is because a higher value of impact factor for a related word in the document is reached when using E-TFIDF, but a lower value is obtained when using traditional TF-IDF. In summary, the proposed E-TFIDF can enhance the effectiveness of SVM classifier in classifying exam questions based on BT cognitive domain.

Besides that, the authors for research (5) proposed an approach using TF-IDF and Naïve Bayes classifier to conduct the exam question classification in accordance with BT cognitive level. The researchers examined various indexing terms for instance Words, Characters and N-gram to choose the best approach that can classify exam questions accurately. The approach of using Naïve Bayes classifier, TF-IDF with N-gram indexing terms reported a superior performance with the accuracy precision result of 85%, which meant that the proposed approach could classify exam questions accurately.

In (6), a classification model is proposed to classify exam questions from multiple domains based on Bloom's taxonomy. In this study, the authors introduced a new feature type, W2V-TFPOSIDF based on the combination of two feature extraction methods: TFPOS-IDF, which is a modified TF-IDF with Part-of-Speech (POS) and pre-trained word2vec to produce question vectors with high-quality representation. There are two datasets used in this study, the first dataset containing 141 questions and the second dataset containing 600 questions. The satisfactory result obtained by using W2V-TFPOSIDF and different machine learning classifiers indicated that this feature could perform more effectively compared to TF-IDF and TFPOS-IDF in classifying exam question. Among these classifiers, SVM classifier generated the highest F1-measure result for both datasets, which are 0.837 and 0.897 respectively. This F1-measure result showed that the proposed feature type can classify the exam question from multiple domains accurately.

In another research (7), the authors proposed BloomNet, a transformer-based model to reduce the effort of institution administrators in mapping the course learning outcomes (CLOs) and exam questions to BT levels manually. In this study, two datasets are tested for a different purpose. For the first dataset, they applied various baseline models and compared the IID (independent and identically distributed) performance of these models with BloomNet throughout the text classification process. As a result, BloomNet outperforms other baseline models and obtained the

highest accuracy of 87.5%. Whereas for the second dataset, the same experimental setup is conducted to evaluate the OOD (out-of-distribution) performance. Same with the expectation, BloomNet achieved the highest accuracy result of 70.4% among others baseline models. However, BloomNet is difficult to deploy in production since it consists of three language encoders, and these encoders make it became memory heavy. In summary, the work focused on NLP models, word embedding to achieve better results but there is no evidence that the BloomNet performs better than past research work focusing on enhanced unsupervised term weighting schemes [21, 22].

In (8), a LSTM based deep learning model is proposed by the authors to achieve the objective stated in (9). The proposed model is expected to predict the Bloom's level for CLO and exam questions respectively. In this study, "Wiki-Word Vector", a skip-gram pre-trained embedding is used for word representation. Therefore, the proposed model can gain enough domain understanding and classify the CLOs and questions into pre-defined category by using LSTM network. LSTM network has a gated mechanism, which can control the flow of input sequences, and make a decision whether what information to keep and discard throughout the flow. As a result, the proposed model generates a satisfied accuracy result of 87% and 74% in classifying CLOs and question items. At the same time, this model outperforms a model used in the existing study [44] by improving the classification result of overall accuracy to 3%. This figure may be less than 3%, if an optimal variant of TF-IDF has been used in [44]. In summary, the work focused on comparing deep neural networks LSTM and word embedding against past research work focused on traditional machine learning techniques based on unsupervised term weighting technique which may not have used the optimal variant of TF-IDF.

In (9), the author accentuates that assigning different weights for verbs, nouns adjectives give different results on classification accuracy. This is because the presence of verbs in exam questions are important than nouns and adjectives in increasing classification accuracy. The work reaffirms that related past research work that enhanced term weightings by assigning different weights for verbs, noun and adjectives [21, 22] has the potential in increasing classification accuracies.

In short, most of the previous research used the unsupervised term weighting TF-IDF to perform exam question classification. However, the optimal variants of TF-IDF are not studied well and deeply in exam question classification. It is imperative to use the optimal variant of TF-IDF as a baseline term weighting in order to compare effectively with the improved term weighting schemes or advanced models such as word embedding and deep neural network [42, 35]. This is to ensure results are more conclusive. Therefore, this study evaluates several variants of unsupervised term weighting, TF and TF-IDF in relation to classification accuracy and identifies the optimal variants.

3 Methodology

3.1 Dataset

In this study, the data set used is a set of exam questions collected from the business domain. The data was domain-specific and labeled with BT cognitive level. The BT cognitive level of each question is identified by lecturers when they prepare the question. Throughout the question labeling process, the lecturer is moderated by expertise such as an academic lecturer to make sure each question has been labeled correctly. To ensure all data followed the BT guidelines, the collected questions were checked with the presence of at least one BT keyword when they went through a pre-processing phase. In this dataset, there are 181 open-ended questions are related to business and marketing fields [33]. Fig. 2 shows the distribution of exam questions in accordance with BT levels in the data set. Table 3 shows some sample questions in the data set for each BT level. The version of BT used is the version published by Benjamin Bloom and his collaborators in the year 1956 [8].

In this study, a question classification model, which is a simulation by using classification techniques to evaluate the unsupervised term weighting schemes is introduced. The proposed question classification model shown in Fig. 3 consists of three main phases, which are preprocessing, feature extraction, and classification phase. The preprocessing phase is Phase 1, which involves several tasks, such as tokenization, stop-word removal, and lemmatization. Phase 2 is the feature extraction phase. Once Phase 1 and Phase 2 are completed, two machine learning classifiers such as Support Vector Machine (SVM) and Naïve Bayes (NB) applied in Phase 3 to perform the exam question classification process and identify the cognitive domain of the dataset in accordance with BT. Lastly, the results generated were compared for evaluation purposes.

3.2 Question Classification Model

Phase 1:
The exam question collected initially may contain noisy data or misspelling issues. Therefore, the preprocessing is applied before proceeding to the next phase to format unstructured data. In this phase, the questions passed through several steps: lowercase conversion, tokenization, remove stop words, lemmatization and compliance with BT guidelines.

The first step was lowercase conversion. Each word that existed in the question is converted into the lowercase format. After that, a tokenization task was implemented to identify the boundaries within words in question items and split them into a list of tokens. Some unimportant words that may exist within question items such as punctuation marks, numbers, and non-letters were eliminated. Besides that, an additional stop words list that contained proper nouns, ab-

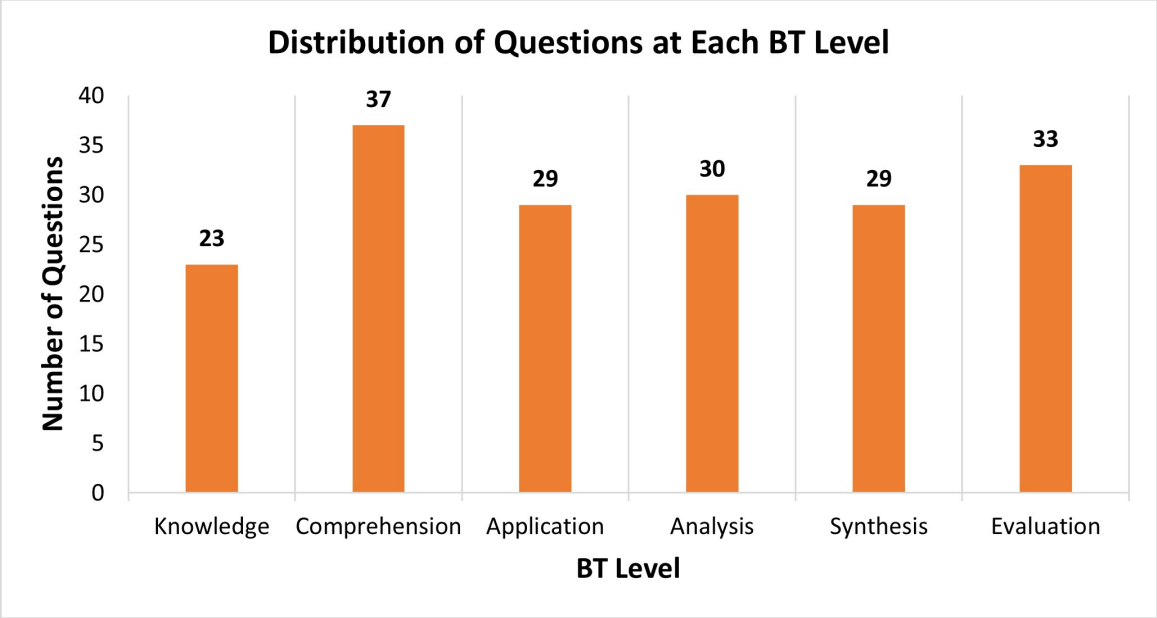

Figure 2: Distribution of Questions at Each BT Level

Table 3: Sample Questions at Each BT Level.

BT Level	Sample Question
Knowledge	State FOUR (4) basic business activities that are performed in the revenue cycle.
	Define brand audit.
Comprehension	Discuss any THREE (3) ways by which an organization can benefit from e-commerce.
	Explain the concept of clicks-and-bricks model in e-commerce.
Application	Apply Porter's five competitive forces analysis to examine the summer job industry for your uncle.
	Demonstrate email and social media approaches to create effective marketing plan.
Analysis	Differentiate between a wholesaler and retailer.
	Compare FOUR (4) point of views of entrepreneurs with FOUR (4) for managers the way they look at the things.
Synthesis	Suggest any TWO (2) efforts that organization may perform in order to discourage unethical behavior.
	Prepare a research proposal on a study that you have to conduct on the purchasing behavior of teenagers in the Klang Valley.
Evaluation	Evaluate the three specific effects caused by the applications of information technology on the nature of competition.
	Critically review the strengths, weakness, opportunities and threats of Associated Meats Sdn Bhd in light of the forecast trends and developments.

breviations such as SWOT, CRM that bring insignificant meaning to question classification is created manually to remove unnecessary tokens. Next, WordNet Lemmatizer in NLTK toolkit, one of the earliest and popular lemmatizer is used to perform the lemmatization task by convert each token into its original form as lemma [31]. Each question is checked with the presence of BT keyword. Only those questions that contain at least one BT keyword able to move further to the feature extraction phase.

Phase 2:
Phase 2 involved two tasks which are feature extraction and term weighting. Feature extraction converted the initial dataset into a set of features that will be used in the next process. In this study, the feature extraction method used is Bag-of-Words (BoW). BoW is an easy and high flexible Natural Language Processing technique to extract the feature from a text document [24]. This model extracted a feature set based on the occurrence of known words that exist in each question. After a feature set is extracted successfully, the term weighting method is applied by calculated and assigning the weight for each feature. The next section presents the variants of unsupervised term weighting proposed in the study.

3.3 Variants of Term Weighting

Term weighting method defined the weight for each term that exists in question. TF and TF-IDF features are two general unsupervised term weighting methods used in exam question classification. There are several TF and TF-IDF variants are being proposed in this study. The TF variants of unsupervised term weighting are represented by equations (1) to (3):

TF - Variant 1

$$TF(t,q) = n_q^t \tag{1}$$

where n_q^t indicates to the number of times of term t occurs in a question q, this variant used in these past researches [36, 41].

TF - Variant 2

$$TF(t,q) = \frac{n_q^t}{\sum_k n_q^k} \tag{2}$$

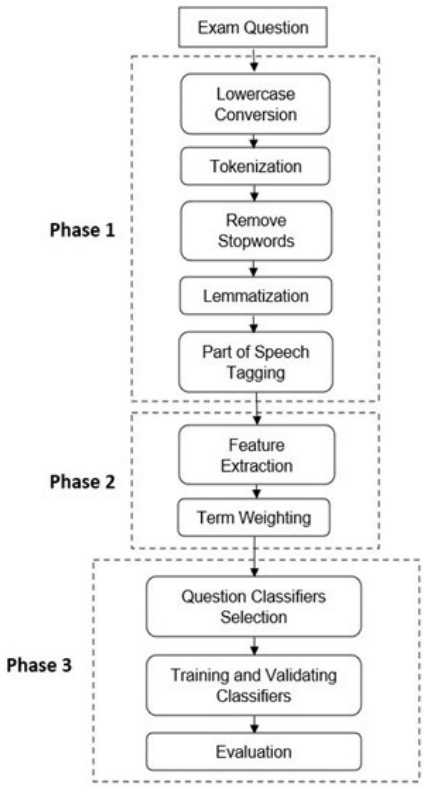

Figure 3: Proposed Question Classification Model

here n_q^t is the number of occurrences of term t in a question q, $\sum_k n_q^k$ is total occurrences of all terms in a question q, this variant used in these past researches [2, 47].

TF - Variant 3

$$TF(t,q) = 1 + \log[f(t,q)] \tag{3}$$

where $f(t,q)$ is number of term t appearing in a question q, this variant used in the past research [40].

While equations (4) to (7) represent the TF-IDF variants proposed in this study:

TF-IDF - Variant 1

$$TF\text{-}IDF(t,q) = \frac{n_q^t}{\sum_k n_q^k} \cdot \log \frac{N}{n_Q^t} \tag{4}$$

where n_q^t is the number of occurrences of term t in a question q, $\sum_k n_q^k$ is the total number of terms in a question q, N is the total number of questions in the dataset, n_Q^t is the number of question q that contained term t exists in the whole dataset Q, this variant used in these past researches [7, 38].

TF-IDF - Variant 2

$$TF\text{-}IDF(t,q) = f(t,q) \cdot \log\left(\frac{N}{n_Q^t}\right) + 1 \tag{5}$$

here $f(t,q)$ is the frequency of term t exists in a question q, N is the total number of questions in the dataset, n_Q^t is the number of question q that contained term t exists in the whole dataset Q, this variant used in these past researches [9, 21].

TF-IDF - Variant 3

$$TF\text{-}IDF(t,q) = \frac{n_q^t}{\sum_k n_q^k} \cdot \log\left(\frac{N}{n_Q^t}\right) + 1 \tag{6}$$

where n_q^t is the number of occurrences of term t in a question q, $\sum_k n_q^k$ is the total number of terms in a question q, N is the total number of questions in the dataset, n_Q^t is the number of question q that contained term t exists in the whole dataset Q, this variant used in these past researches [22].

TF-IDF - Variant 4

$$TF\text{-}IDF(t,q) = (1 + \log[f(t,q)]) \cdot \log \frac{N}{n_Q^t} \tag{7}$$

here $f(t,q)$ is the number of term t appearing in a question q, N is the total number of questions in the dataset, n_Q^t is the number of question q that contained term t exists in the whole dataset Q, this variant used in the past research [20].

Finally, the normalized TF-IDF was proposed based on the normalization of TF-IDF variants to ensure the weightings for each feature are in the range between 0 to 1. In this study, L2 norm is used to normalize all TF-IDF variants proposed above, since it is a popular and commonly used norm [22]. The following equation demonstrated Normalized TF-IDF:

Normalized TF-IDF - Variant 1, 2, 3, 4

$$\text{Normalized } TF\text{-}IDF(t,q) = \frac{TF\text{-}IDF(t,q)}{\sqrt{\sum TF\text{-}IDF(t,q)^2}} \tag{8}$$

where $TF\text{-}IDF(t,q)$ is the $TF\text{-}IDF$ value obtained for term t in question q.

Table 4 presents some commonly used variants of unsupervised term weighting in text and question classifications. These existing works shown in the table support the proposed variants used in this study, except the normalized TF-IDF variant.

Phase 3:

In this phase, machine learning classifiers were used to define the BT cognitive level that belongs to each question in the dataset. There are two machine learning classifiers selected and used in this study, which are Support Vector Machine (SVM) and Naïve Bayes (NB).

Support Vector Machine (SVM) is a widely used classifier in the text classification process [32]. SVM aims to generate a suitable hyperplane that splits two sets of data from each other, by maximizing the width of margin among the hyperplane and the set of data points closest to it [21]. Compared to other classifiers, SVM can perform better and offer a higher accuracy result [25]. In this study, an SVC model of SVM with linear kernel in SVM was used.

Table 4: Supporting Past Research Works for Proposed Variants.

Article No.	Author / Year	Variants Used
[41]	Utomo and Bijaksana (2016)	$TF\,(t,q) = n_q^t$
[36]	Shimomoto et al. (2018)	$TF\,(t,q) = n_q^t$
[17]	Liu et al. (2018)	$TF\,(t,q) = \frac{n_q^t}{\sum_k n_q^k}$
[2]	Abdulrahman and Baykara (2020)	$TF\,(t,q) = \frac{n_q^t}{\sum_k n_q^k}$
[40]	Tongman and Wattanakitrungroj (2018)	$TF\,(t,q) = 1 + \log[f\,(t,q)]$
[38]	Sundus, Fatimah and Hammo (2019)	$TF\text{-}IDF\,(t,q) = \frac{n_q^t}{\sum_k n_q^k} \cdot \log \frac{N}{n_Q^t}$
[7]	Dalaorao, Sison and Medina (2019)	$TF\text{-}IDF\,(t,q) = \frac{n_q^t}{\sum_k n_q^k} \cdot \log \frac{N}{n_Q^t}$
[21]	Mohammed and Omar (2018)	$TF\text{-}IDF\,(t,q) = f\,(t,q) \cdot \log\left(\frac{N}{n_Q^t}\right) + 1$
[9]	Djajadinata et al. (2020)	$TF\text{-}IDF\,(t,q) = f\,(t,q) \cdot \log\left(\frac{N}{n_Q^t}\right) + 1$
[22]	Mohammed and Omar (2020)	$TF\text{-}IDF\,(t,q) = \frac{n_q^t}{\sum_k n_q^k} \cdot \log\left(\frac{N}{n_Q^t}\right) + 1$
[20]	Meng and Xu (2018)	$TF\text{-}IDF\,(t,q) = (1 + \log\left[f\,(t,q)\right]) \cdot \log\left(\frac{N}{n_Q^t}\right)$

The second classifier used to classify exam question is Naïve Bayes. Naïve Bayes is a probabilistic machine learning model that assumes the rear possibility of the word or term, considering the existence of the word either is independent or connected to existing entity class [45]. Based on the rear possibility obtained for different categories, the word or term is assigned to the category that has the highest value [27]. In this study, Multinomial NB was used since it is popular for document classification problems [10]. After the selection of question classifiers, both classifiers chosen are trained and validated in order to perform the question classification. The performance of both classifiers are then evaluated after the question classification process completed.

3.4 Evaluation Metrics

To measure the performance of the classification model and identify the most optimal unsupervised term weighting variant, the experiment outcomes are calculated with two evaluation metrics, accuracy, and F1-measure. To calculate the F1-measure, the value of recall and precision are required first. The recall metric measures the level of completeness while the precision metric measures the exactness [39]. Commonly, the value of accuracy and F1-measure is in the range between 0 to 1. As the value obtained closer to 1, it implies a good performance. Besides that, the cross-validation method was used to validate the classifiers since it is a method that is used to predict the effectiveness of machine learning models on a data sample [5]. In this study, k-fold cross-validation is applied. A parameter k is required to split the dataset into certain groups, and the k indicates the number of groups. When the k is defined, the training dataset and testing dataset were split out based on the k. To evaluate the machine learning classifiers, the experiment was conducted with the k-fold values that in the range of 3 to 10.

For each k-fold value, accuracy metric is calculated by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

where TP is the outcome of classifier correctly classified the question to suitable class, FP and TN is the outcome of classifier incorrectly classified the question to unsuitable class, FN is the number of questions that have not been classified by the classifier.

To calculate the F1-measure, recall and precision metric required and computing by these formulae (10), (11):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{F1-measure} = \frac{2 \cdot (\text{Recall} \cdot \text{Precision})}{\text{Recall} + \text{Precision}} \tag{12}$$

4 Result and Discussion

4.1 Experimental Steps

To implement the comparative study, several experiments have been conducted with different term weighting variants and tested with two classifiers. The variants used were classified into three types of term weighting, which are TF, TF-IDF, and Normalized TF-IDF. The classifiers that used for question classification are SVM and Naïve Bayes. The default setting of kernel in SVM is linear and the parameter C is 1.0. The model of Naïve Bayes used is Multinomial NB. We developed a small-scale prototype to perform the question classification by using NLTK library, Sckit-learn library and PyCharm IDE. Besides that, we used k-fold cross-validation method to validate the question classifiers. We experimented with several k-fold values ranging from 3 to 10 and obtain the average accuracy of each k-fold value. The average accuracy obtained

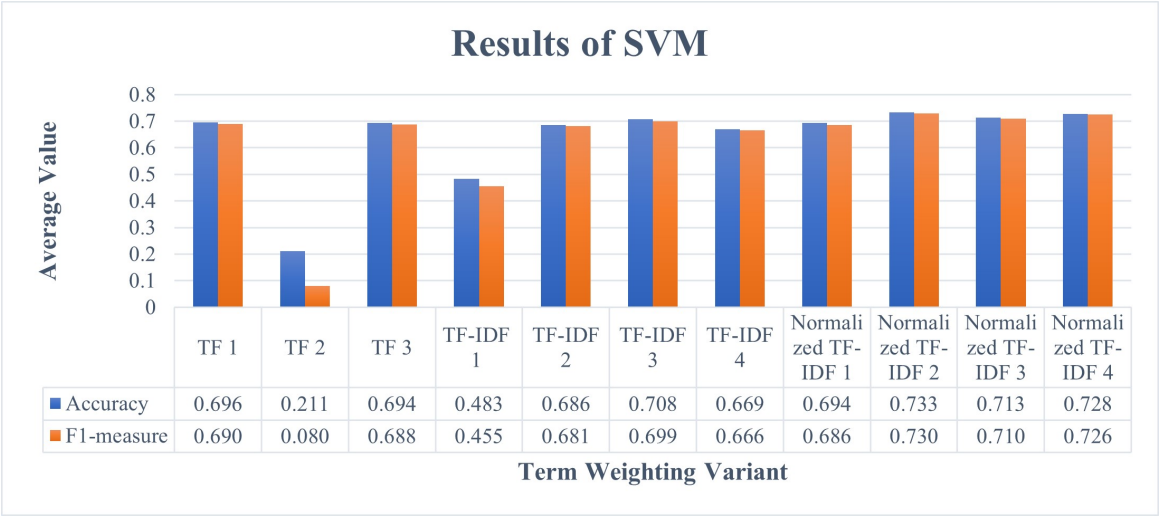

Figure 4: Overall Average Results of SVM.

from each k-fold value was then summed together and divided to get the overall average accuracy for each term weighting variant.

4.2 Results of SVM

This section discusses the experiment result obtained with the SVM classifier in terms of the accuracy and F1-measure for the question dataset. The results of TF, TF-IDF, and normalized TF-IDF variations are shown in the Tables 5, 6, and 7.

Table 5: Accuracy (Acc) and F1-measure (F1) of SVM with TF variants.

	TF 1		TF 2		TF 3	
K-Fold	Acc	F1	Acc	F1	Acc	F1
3	0.680	0.676	0.204	0.069	0.674	0.669
4	0.701	0.700	0.210	0.080	0.696	0.694
5	0.702	0.698	0.216	0.089	0.696	0.696
6	0.685	0.677	0.204	0.070	0.702	0.695
7	0.707	0.697	0.210	0.079	0.702	0.693
8	0.691	0.687	0.215	0.088	0.686	0.680
9	0.707	0.701	0.210	0.078	0.702	0.697
10	0.696	0.684	0.216	0.088	0.690	0.677
Avg	0.696	0.690	0.211	0.080	0.694	0.688

Table 6: Accuracy (Acc) and F1-measure (F1) of SVM with TF-IDF variants.

	TF-IDF 1		TF-IDF 2		TF-IDF 3		TF-IDF 4	
K-Fold	Acc	F1	Acc	F1	Acc	F1	Acc	F1
3	0.365	0.328	0.641	0.640	0.696	0.694	0.641	0.636
4	0.470	0.450	0.690	0.685	0.713	0.702	0.674	0.667
5	0.487	0.463	0.680	0.679	0.718	0.719	0.658	0.658
6	0.492	0.466	0.668	0.661	0.707	0.699	0.647	0.643
7	0.503	0.483	0.690	0.682	0.719	0.713	0.669	0.665
8	0.525	0.494	0.707	0.700	0.702	0.690	0.691	0.684
9	0.514	0.485	0.707	0.703	0.724	0.706	0.685	0.686
10	0.509	0.473	0.701	0.697	0.685	0.665	0.690	0.686
Avg	0.483	0.455	0.686	0.681	0.708	0.699	0.669	0.666

The TF1 yielded the highest average value of 0.696 for the accuracy, according to the results shown in Table 5. The average accuracy value obtained using TF2, on the other hand, is the lowest, at 0.211. It is because the equation for TF2 involved the division of numbers,

Table 7: Accuracy (Acc) and F1-measure (F1) of SVM with Normalized TF-IDF variants.

	N TF-IDF 1		N TF-IDF 2		N TF-IDF 3		N TF-IDF 4	
K-Fold	Acc	F1	Acc	F1	Acc	F1	Acc	F1
3	0.680	0.680	0.702	0.706	0.680	0.683	0.696	0.698
4	0.702	0.696	0.718	0.718	0.724	0.723	0.718	0.718
5	0.663	0.657	0.712	0.712	0.679	0.681	0.701	0.703
6	0.680	0.670	0.729	0.724	0.718	0.713	0.718	0.713
7	0.702	0.697	0.746	0.742	0.718	0.713	0.729	0.725
8	0.707	0.699	0.751	0.743	0.718	0.714	0.751	0.745
9	0.718	0.709	0.762	0.759	0.746	0.742	0.768	0.764
10	0.696	0.678	0.745	0.735	0.718	0.712	0.745	0.738
Avg	0.694	0.686	0.733	0.730	0.713	0.710	0.728	0.726

which produced a lower term weighting value for each term where classifier such as SVM may not work well with a very low term value. Whereas for TF3, the average value gained is 0.694 and it is quite similar to the TF1 result. The average accuracies involving TF1, TF2, and TF3 follow the same pattern as the average F1-measure.

Among the TF-IDF variants, TF-IDF3 had the highest average accuracy value of 0.708, which was higher than the other TF-IDF versions. In contrast to the TF-IDF3, TF-IDF1 yielded the lowest average accuracy value of 0.483. Based on the results, TF-IDF3 can improve the classifier performance more effectively in classifying exam questions. The average accuracies involving TF-IDF1, TF-IDF2, TF-IDF3 and TFIDF-4 follow the same pattern as the average F1-measure. The reason TF-IDF1 recorded the lowest average accuracy and F1-measure could be due to the effect of TF2. But what is noticeable is the impact of the IDF version used in TF-IDF2 and TF-IDF3 on the classification accuracy. This will explain why the average accuracies of TF-IDF1 and TF-IDF4 are lower than TF-IDF2 and TF-IDF3.

In Table 7, the classification results obtained in terms of accuracy metric for each normalized TF-IDF variant that range between 0.694 and 0.733 are higher than unnormalized TF-IDF variants. The highest average result is obtained when using Normalized

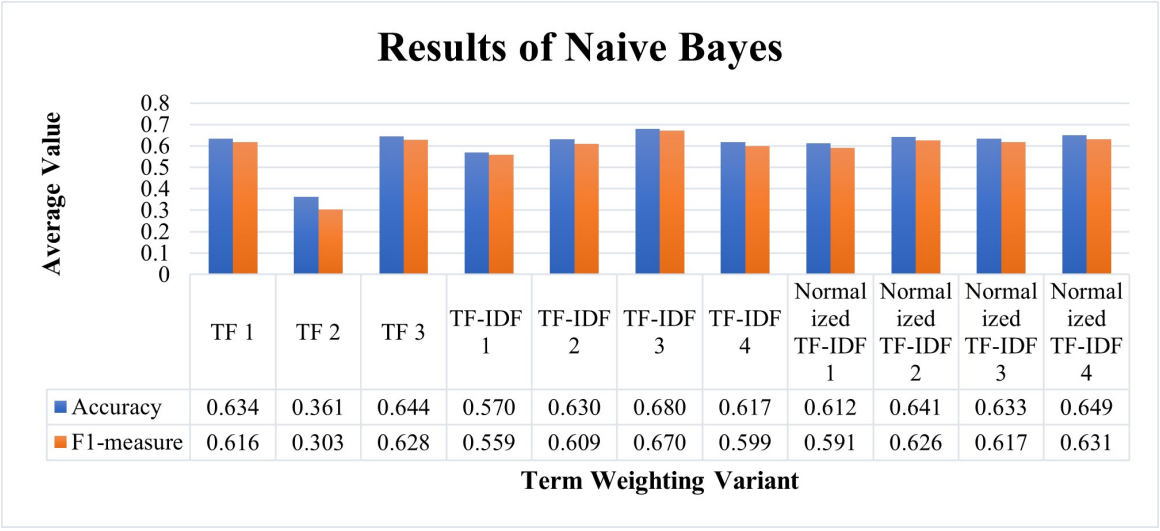

Figure 5: Overall Average Results of Naïve Bayes.

TF-IDF2, which is 0.733 while Normalized TF-IDF1 recorded the lowest average accuracy value of 0.694. Normalizing had the effect of making the intrinsic differences in accuracy and F1-measure between TF-IDF variants before normalization insignificant. Fig. 4 shows the average accuracy and F1-measure results for all proposed variants.

Results from Fig. 4 show that the normalized TF-IDF outperforms the traditional TF and TF-IDF. All Normalized TF-IDF variants generate accuracy and F1-measure value that ranges between 0.686 to 0.733. By comparing the classification results between TF-IDF variants with Normalized TF-IDF variants, the average value for accuracy and F1-measure metrics increases significantly when using normalized TF-IDF variants, especially for the TF-IDF1 variant. The value of the average accuracy measure obtained from TF-IDF1 is 0.483, however, when it is being normalized, the average accuracy value obtained is 0.694. The value obtained by using Normalized TF-IDF1 increased obviously, which means that the normalization of the TF-IDF variant brings a positive impact on the performance of SVM in classifying exam questions.

4.3 Results of Naïve Bayes

The section analyses the results that were obtained from Naïve Bayes classifier, in terms of accuracy and F1-measure. Tables 8, 9, and 10 show the result of TF, TF-IDF and normalized TF-IDF variants respectively.

In Table 8, the results show that TF3 yielded the highest average value of 0.644 for the accuracy measure. The average accuracy value acquired by TF1 is 0.634, which is comparable to the accuracy value obtained by TF3. However, the average accuracy value obtained by using TF2 is the lowest, which is 0.361.

As for the TF-IDF variants, the results presented in Table 9 indicated that TF-IDF3 achieved the highest average accuracy value of 0.680. TF-IDF1, on the other hand, had the lowest average accuracy value of 0.570. Hence, TF-IDF3 is the most optimal TF-IDF variant that can generate a satisfactory classification result among other TF-IDF variants. The average accuracies involving TF-IDF1, TF-IDF2, TF-IDF3 and TFIDF-4 follow the same pattern as the average F1-

Table 8: Accuracy (Acc) and F1-measure (F1) of Naïve Bayes with TF variants.

	TF 1		TF 2		TF 3	
K-Fold	Acc	F1	Acc	F1	Acc	F1
3	0.658	0.654	0.309	0.239	0.652	0.649
4	0.646	0.632	0.354	0.301	0.651	0.639
5	0.624	0.613	0.337	0.280	0.635	0.627
6	0.624	0.607	0.370	0.319	0.640	0.624
7	0.618	0.596	0.392	0.341	0.635	0.619
8	0.635	0.610	0.370	0.317	0.646	0.625
9	0.635	0.613	0.370	0.306	0.651	0.631
10	0.629	0.602	0.387	0.321	0.640	0.613
Avg	0.634	0.616	0.361	0.303	0.644	0.628

Table 9: Accuracy (Acc) and F1-measure (F1) of Naïve Bayes with TF-IDF variants.

	TF-IDF 1		TF-IDF 2		TF-IDF 3		TF-IDF 4	
K-Fold	Acc	F1	Acc	F1	Acc	F1	Acc	F1
3	0.497	0.500	0.630	0.620	0.652	0.652	0.613	0.604
4	0.586	0.583	0.613	0.593	0.679	0.671	0.613	0.596
5	0.580	0.577	0.630	0.612	0.685	0.679	0.619	0.604
6	0.575	0.564	0.635	0.615	0.674	0.669	0.619	0.602
7	0.563	0.543	0.618	0.597	0.669	0.657	0.602	0.588
8	0.581	0.562	0.630	0.602	0.691	0.669	0.613	0.588
9	0.602	0.591	0.647	0.626	0.702	0.689	0.624	0.607
10	0.575	0.552	0.635	0.608	0.685	0.672	0.630	0.601
Avg	0.570	0.559	0.630	0.609	0.680	0.670	0.617	0.599

Table 10: Accuracy (Acc) and F1-measure (F1) of Naïve Bayes with Normalized TF-IDF variants.

	N TF-IDF 1		N TF-IDF 2		N TF-IDF 3		N TF-IDF 4	
K-Fold	Acc	F1	Acc	F1	Acc	F1	Acc	F1
3	0.586	0.575	0.618	0.614	0.608	0.605	0.624	0.614
4	0.602	0.585	0.646	0.632	0.641	0.627	0.646	0.629
5	0.624	0.607	0.613	0.605	0.619	0.607	0.635	0.620
6	0.602	0.571	0.641	0.629	0.630	0.615	0.630	0.611
7	0.607	0.588	0.630	0.613	0.630	0.614	0.641	0.624
8	0.624	0.600	0.657	0.640	0.652	0.632	0.668	0.646
9	0.641	0.623	0.679	0.657	0.652	0.634	0.701	0.683
10	0.608	0.577	0.641	0.617	0.630	0.600	0.646	0.619
Avg	0.612	0.591	0.641	0.626	0.633	0.617	0.649	0.631

Table 11: Weighting Results for Each Term (Question 1).

	suggest	justify	strategy	company	consider	market	product	internationally
TF 1	1	1	1	1	1	1	1	1
TF 2	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
TF 3	1	1	1	1	1	1	1	1
TF-IDF 1	0.132	0.169	0.195	0.169	0.195	0.207	0.195	0.245
TF-IDF 2	2.054	2.355	2.559	2.355	2.559	2.656	2.559	2.957
TF-IDF 3	0.257	0.294	0.320	0.294	0.320	0.332	0.320	0.370
TF-IDF 4	1.054	1.355	1.559	1.355	1.559	1.656	1.559	1.957
N TF-IDF 1	0.244	0.314	0.361	0.314	0.361	0.383	0.361	0.453
N TF-IDF 2	0.288	0.331	0.359	0.331	0.359	0.373	0.359	0.415
N TF-IDF 3	0.288	0.331	0.359	0.331	0.359	0.373	0.359	0.415
N TF-IDF 4	0.244	0.314	0.361	0.314	0.361	0.383	0.361	0.453

Table 12: Weighting Results for Each Term (Question 2).

	identify	discuss	issue	employee	productivity	problem	company	face
TF 1	1	1	1	1	1	1	1	1
TF 2	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
TF 3	1	1	1	1	1	1	1	1
TF-IDF 1	0.169	0.085	0.245	0.245	0.245	0.207	0.169	0.245
TF-IDF 2	2.355	1.678	2.957	2.957	2.957	2.656	2.355	2.957
TF-IDF 3	0.294	0.210	0.370	0.370	0.370	0.332	0.294	0.370
TF-IDF 4	1.355	0.678	1.957	1.957	1.957	1.656	1.355	1.957
N TF-IDF 1	0.288	0.144	0.415	0.415	0.415	0.352	0.288	0.415
N TF-IDF 2	0.315	0.224	0.395	0.395	0.395	0.355	0.315	0.395
N TF-IDF 3	0.315	0.224	0.395	0.395	0.395	0.355	0.315	0.395
N TF-IDF 4	0.288	0.144	0.415	0.415	0.415	0.352	0.288	0.415

measure.

Based on the results shown in Table 10, the highest average value in terms of the accuracy metric is obtained when using Normalized TF-IDF4, which is 0.649. But the Normalized TF-IDF1 recorded the lowest average accuracy value of 0.612. Table 8, Table 9, and Table 10 findings are consistent with the SVM classification results. Fig. 5 shows the result of average accuracy and F1-measure value for all proposed variants.

Based on Fig. 5, the normalization of TF-IDF variants can improve the accuracy of question classification. For example, the average accuracy value obtained when using TF-IDF1 is 0.570, but the value increased to 0.612 when TF-IDF1 is being normalized. The overall results indicate that the normalization of TF-IDF generally brings a positive impact to improve the performance of Naïve Bayes classifier in classifying exam questions based on Bloom's taxonomy cognitive domain. In general, the normalization effect on SVM has a higher impact on classification accuracy than the Naive Bayes classifier.

4.4 Term Weighting Results

This section shows the result of the weighting value for two questions chosen from the dataset. These questions were selected randomly, and the weighting value of each term that exists in the question was calculated with all proposed term weighting variants and shown in Tables 11 and 12.

Question 1: [suggest, justify, strategy, company, consider, market, product, internationally]

Question 2: [identify, discuss, issue, employee, productivity, problem, company, face]

Based on Table 11 and Table 12, the weighting for each term by using TF variants is consistent, For example, the weighting results for TF1 and TF3 is 1, since it calculates only the occurrences of each term that exist in the question, and each term only exists once in the question.

On the other hand, the weighting results obtained for each term by using TF-IDF variants generally has higher accuracy compared to TF variants. It is because TF-IDF variants have discriminating power, thus it can differentiate well the term. Therefore, each term in the question contained a different weighting value. For example, in Table 12, for TF-IDF2, the weighting for the verb term "discuss" is 1.678, while for the noun term "issue" is 2.957.

According to the results presented in Table (11-12), the weighting value obtained from all Normalized TF-IDF is precise, since each value is in the range between 0 to 1. Besides that, the results showed that the accuracy of weighting value when using Normalized TF-IDF variants is slightly higher compared to TF-IDF variants. It is because the normalization of TF-IDF variants can reduce the bias of each feature when using TF-IDF variants, which minimize its numerical contri-

bution to become lower [37]. Moreover, the normalization of TF-IDF variants can guarantee each feature has an equal numerical distribution before it is fed into machine learning algorithms [26]. In consequence, a better distribution of words can be reached. Under this situation, SVM and Naïve Bayes classifiers may be able to work properly in classifying exam questions.

4.5 Statistical Analysis

In this study, the type of statistical test used is the t-test since it was the popular statistical method [30]. T-test implementation aims to explore the significance of two data samples[24]. In each two-sample t-test, the accuracy and F1-measure values were used. Tables 13-16 present the result of the t-test by comparing each proposed TF, TF-IDF and Normalized TF-IDF variants separately and two classifiers SVM and Naïve Bayes. Besides that, Table 16 shows the t-test result with the comparison of the average value gained from three different term weighting variant types used in the study, which are TF, TF-IDF and Normalized TF-IDF variants. After running the statistical analysis, the result obtained for t-test indicates the significance level between these two variants. Tables 17-20 show the significance results for t-test.

To check the significance between the two samples, a null hypothesis was set. Before that, the setting of parameter alpha value was 0.05. Therefore, if the P-value obtained for the t-test run was less than the alpha value, it indicated that the null hypothesis was rejected and there was a significant difference among the two samples. On the contrary, if the P-value obtained was more than or equal to the alpha value, it meant that the null hypothesis was accepted and there was no significant difference among the two samples.

Table 13: P-value of t-test for TF variants.

Classifier	TF 1 vs. TF 2		TF 2 vs. TF 3	
	Acc	F1	Acc	F1
SVM	1.05E-13	4.95E-14	4.72E-13	3.51E-13
Naïve Bayes	1.70E-7	3.76E-7	3.39E-8	9.06E-8

Table 14: P-value of t-test for TF-IDF variants.

Classifier	TF-IDF 1 vs. 2		TF-IDF 2 vs. 3		TF-IDF 3 vs. 4	
	Acc	F1	Acc	F1	Acc	F1
SVM	4.61E-7	5.70E-7	3.05E-2	1.28E-1	2.53E-3	1.54E-2
Naïve Bayes	1.01E-3	2.96E-3	1.80E-5	4.00E-6	3.00E-6	3.03E-7

Table 15: P-value of t-test for Normalized TF-IDF variants.

Classifier	N TF-IDF 1 vs. 2		N TF-IDF 2 vs. 3		N TF-IDF 3 vs. 4	
	Acc	F1	Acc	F1	Acc	F1
SVM	5.70E-5	3.20E-5	3.22E-3	2.37E-3	1.28E-2	1.05E-2
Naïve Bayes	2.12E-3	8.22E-4	5.56E-2	2.03E-2	1.67E-2	4.11E-2

Based on the results shown in Table 17, it indicates that TF2 is statistically significant compared to TF1. Besides that, TF3 is considered statistically significant compared to TF2.

Table 16: P-value of t-test for TF vs TF-IDF vs Normalized TF-IDF Variants.

Classifier	TF vs. TF-IDF		TF-IDF vs. N TF-IDF	
	Acc	F1	Acc	F1
SVM	5.81E-3	4.64E-3	6.00E-6	5.00E-6
Naïve Bayes	8.12E-3	8.30E-3	1.82E-1	3.43E-1

Table 17: Significance results for TF variants.

Classifier	TF 1 vs. TF 2		TF 2 vs. TF 3	
	Acc	F1	Acc	F1
SVM	Significant	Significant	Significant	Significant
Naïve Bayes	Significant	Significant	Significant	Significant

Table 18: Significance results for TF-IDF variants.

Classifier	TF-IDF 1 vs. 2		TF-IDF 2 vs. 3		TF-IDF 3 vs. 4	
	Acc	F1	Acc	F1	Acc	F1
SVM	Significant	Significant	Significant	Insignificant	Significant	Significant
Naïve Bayes	Significant	Significant	Significant	Significant	Significant	Significant

Table 19: Significance results for Normalized TF-IDF variants.

Classifier	N TF-IDF 1 vs. 2		N TF-IDF 2 vs. 3		N TF-IDF 3 vs. 4	
	Acc	F1	Acc	F1	Acc	F1
SVM	Significant	Significant	Significant	Significant	Significant	Significant
Naïve Bayes	Significant	Significant	Significant	Significant	Significant	Significant

Table 20: Significance results for TF vs TF-IDF vs Normalized TF-IDF Variants.

Classifier	TF vs. TF-IDF		TF-IDF vs. N TF-IDF	
	Acc	F1	Acc	F1
SVM	Significant	Significant	Significant	Significant
Naïve Bayes	Significant	Significant	Insignificant	Insignificant

For the results presented in Table 18, it is clear that TF-IDF2 is statistically significant compared to TF-IDF1, which means that TF-IDF2 performed better in improving the effectiveness of the question classification model compared to TF-IDF1. For the comparison between TF-IDF2 and TF-IDF3, the significance level of accuracy obtained for the SVM classifier is significant, which means that there was a significant difference between TF-IDF2 and TF-IDF3. However, different from the accuracy result, the significance level for F1-measure obtained when using SVM classifier showed that there was no significant difference between TF-IDF2 and TF-IDF3. Whereas for the Naïve Bayes classifier, the significance result for accuracy and F1-measure acquired among Normalized TF-IDF2 and Normalized TF-IDF3 is significant. In addition, the TF-IDF4 is statistically significant compared to TF-IDF3 for both classifiers.

From the results shown in Table 19, it can be concluded that Normalized TF-IDF2 is statistically significant compared to Normalized TF-IDF1. Besides that, the results for SVM classifier also indicate that the Normalized TF-IDF3 variant is statistically significant compared to Normalized TF-IDF2. Whereas for the Naïve Bayes classifier, the result showed that there was no significant difference between Normalized TF-IDF2 and Normalized TF-IDF3 in terms of accuracy metric, but a significant difference in terms of F1-measure metric. For the comparison between Normalized TF-IDF3 and Normalized TF-IDF4, the results show that there is a significant difference between these two normalized term weighting variants.

Lastly, the result in Table 20 presents the significance between TF and TF-IDF variants, also TF-IDF and Normalized TF-IDF variants. The result indicates that TF-IDF variants are statistically compared to TF variants for both classifiers. Whereas for the comparison between TF-IDF and Normalized TF-IDF variants, the significant result obtained for SVM classifier is significant, while for Naïve Bayes classifier is insignificant. Therefore, it can be concluded that the normalization of TF-IDF variants can assist SVM classifier in improving the question classification result significantly.

5 Conclusion and Future Work

To identify the most optimal term weighting variant of TF-IDF, SVM and Naïve Bayes classifiers were used in this study. And results show that the Normalized TF-IDF2 is the most optimal variant among other normalized TF-IDF variants with the highest accuracy value of 73.3%. But, among TF-IDF variants, TF-IDF3 recorded the highest accuracy of 70.8%. In general, the normalized TF-IDF variants outperformed TF and TF-IDF variants. Based on the results obtained from T-test, it indicates that there is a significant difference between TF versus TF-IDF and TF-IDF versus Normalized TF-IDF, which means that TF-IDF can work better than TF in classifying exam question, and Normalized TF-IDF outperforms TF-IDF for SVM classifier. For the comparison between TF-IDF and Normalized TF-IDF, the statistical test indicates that Normalized TF-IDF can perform better than TF-IDF when using SVM classifier in classifying exam questions. However, some classifiers may perform well without normalising term weighting values, therefore whether or not to employ normalization depends on the classifiers used in exam question classification based on Bloom Taxonomy.

According to the findings of this study, the TF-IDF2 normalized variation should be used when a classifier favours normalisation and the TF-IDF3 variant should be used when a classifier does not in the context of question classification based on Bloom Taxonomy. Despite the current trend in exam question classification is to compare deep learning models and word embedding models against unsupervised term weighting techniques, identifying optimal variant of unsupervised term weighting is important so as to ensure the results are not bias and are instead more conclusive.

This study can be furthered in the future by expanding the dataset to include exam questions from other fields. Besides that, the experiments conducted in this study can be implemented with various normalization techniques instead of L2 norm.

Acknowledgement: The acknowledgment to the funder, stakeholder, co-researchers, or any other parties contribute directly or indirectly in producing the article and research.

References

[1] Abduljabbar, D. A., and Omar, N. Exam questions classification based on bloom's taxonomy cognitive level using classifiers combination. *Journal of Theoretical and Applied Information Technology 78*, 3 (2015), 447.

[2] Abdulrahman, A., and Baykara, M. Fake news detection using machine learning and deep learning algorithms. In *2020 International Conference on Advanced Science and Engineering (ICOASE)* (2020), IEEE, pp. 18–23.

[3] Alsaeedi, A. A survey of term weighting schemes for text classification. *International Journal of Data Mining, Modelling and Management 12*, 2 (2020), 237–254.

[4] Aninditya, A., Hasibuan, M. A., and Sutoyo, E. Text mining approach using tf-idf and naive bayes for classification of exam questions based on cognitive level of bloom's taxonomy. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)* (2019), IEEE, pp. 112–117.

[5] Brownlee, J. A gentle introduction to k-fold cross-validation [online, accessed 07 october, 2021], 2018. https://machinelearningmastery.com/k-fold-cross-validation.

[6] Chen, K., Zhang, Z., Long, J., and Zhang, H. Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications 66* (2016), 245–260.

[7] Dalaorao, G. A., Sison, A. M., and Medina, R. P. Integrating collocation as tf-idf enhancement to improve classification accuracy. In *2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA)* (2019), IEEE, pp. 282–285.

[8] Deller, J. Bloom's taxonomy levels of learning: The complete post [online, accessed 07 october, 2021], 2019. https://kodosurvey.com/blog/blooms-taxonomy-levels-learning-complete-post.

[9] Djajadinata, K., Faisol, H., Shidik, G. F., Fanani, A. Z., et al. Evaluation of feature extraction for indonesian news classification. In *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)* (2020), IEEE, pp. 585–591.

[10] Gandhi, R. Naive bayes classifier. what is a classifier? [online, accessed 02 october, 2021], 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.

[11] Jayakodi, K., Bandara, M., and Perera, I. An automatic classifier for exam questions in engineering: A process for bloom's taxonomy. In *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (2015), IEEE, pp. 195–202.

[12] Jiang, Z., Gao, B., He, Y., Han, Y., Doyle, P., and Zhu, Q. Text classification using novel term weighting scheme-based improved tf-idf for internet media reports. *Mathematical Problems in Engineering 2021* (2021).

[13] Jiang, Z.-Y., Gao, B., Tian, X., He, Y.-L., and Zhu, Q.-X. An improved term weighting method for content analysis on chinese internet media contents. In *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)* (2020), vol. 1, IEEE, pp. 48–52.

[14] Kusuma, S. F., Siahaan, D., and Yuhana, U. L. Automatic indonesia's questions classification based on bloom's taxonomy using natural language processing a preliminary study. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)* (2015), IEEE, pp. 1–6.

[15] Lan, M. A new term weighting method for text categorization. *PhD Theses, School of Computing, National University of Singapore* (2007).

[16] Lan, M., Tan, C. L., Su, J., and Lu, Y. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence 31*, 4 (2008), 721–735.

[17] Liu, C.-z., Sheng, Y.-x., Wei, Z.-q., and Yang, Y.-Q. Research of text classification based on improved tf-idf algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)* (2018), IEEE, pp. 218–222.

[18] Makhlouf, K., Amouri, L., Chaabane, N., and Nahla, E.-H. Exam questions classification based on bloom's taxonomy: Approaches and techniques. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)* (2020), IEEE, pp. 1–6.

[19] Mazyad, A., Teytaud, F., and Fonlupt, C. A comparative study on term weighting schemes for text classification. In *International Workshop on Machine Learning, Optimization, and Big Data* (2017), Springer, pp. 100–108.

[20] Meng, F., and Xu, L. An improved native bayes classifier for imbalanced text categorization based on k-means and chi-square feature selection. In *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)* (2018), IEEE, pp. 894–898.

[21] Mohammed, M., and Omar, N. Question classification based on bloom's taxonomy using enhanced tf-idf. *Int J Adv Sci Eng Inf Technol 8* (2018), 1679–1685.

[22] Mohammed, M., and Omar, N. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one 15*, 3 (2020), e0230442.

[23] Moreo, A., Esuli, A., and Sebastiani, F. Learning to weight for text classification. *IEEE Transactions on Knowledge and Data Engineering 32*, 2 (2018), 302–316.

[24] Mujtaba, H. An introduction to bag of words in nlp using python [online, accessed 07 october, 2021], 2020. https://www.mygreatlearning.com/blog/bag-of-words/#ed4.

[25] Navlani, A. Sklearn svm (support vector machines) with python - datacamp [online, accessed 07 october, 2021], 2019. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python.

[26] Nayak, S., Misra, B. B., and Behera, H. S. Impact of data normalization on stock index forecasting. *International Journal of Computer Information Systems and Industrial Management Applications 6*, 2014 (2014), 257–269.

[27] Nidaa, G. A., and Dhiyaa, S. H. Classifying exam questions based on bloom's taxonomy using machine learning approach. In *Technol. Dev. Inf. Syst. Tris-2019* (2019), pp. 260–269.

[28] Osadi, K., Fernando, M., Welgama, W., et al. Ensemble classifier based approach for classification of examination questions into bloom's taxonomy cognitive levels. *International Journal of Computer Applications 162*, 4 (2017), 1–6.

[29] Osman, A., and Yahya, A. Classifications of exam questions using linguistically-motivated features: a case study based on bloom's taxonomy. In *The Sixth International Arab Conference on Quality Assurance in Higher Education (IACQA'2016)* (2016), vol. 467, p. 474.

[30] Potochnik, A., Colombo, M., and Wright, C. *Recipes for Science*. Routledge, 2018, ch. Statistics and Probability, pp. 167–206.

[31] Prabhakaran, S. Lemmatization approaches with examples in python [online, accessed 07 october, 2021], 2018. https://www.machinelearningplus.com/nlp/lemmatization-examples-python.

[32] Ray, S. Support vector machine algorithm in machine learning [online, accessed 07 october, 2021], 2017. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code.

[33] Sangodiah, A., Ahmad, R., and WAN AHMAD, W. F. Taxonomy based features in question classification using support vector machine. *Journal of Theoretical & Applied Information Technology 95*, 12 (2017).

[34] Sangodiah, A., Fui, Y. T., Heng, L. E., Jalil, N. A., Ayyasamy, R. K., and Meian, K. H. A comparative analysis on term weighting in exam question classification. In *2021 5th International Symposium on Multidisciplinary Stud-*

ies and Innovative Technologies (ISMSIT) (2021), IEEE, pp. 199–206.

[35] Shaikh, S., Daudpotta, S. M., and Imran, A. S. Bloom's learning outcomes' automatic classification using lstm and pretrained word embeddings. *IEEE Access 9* (2021), 117887–117909.

[36] Shimomoto, E. K., Souza, L. S., Gatto, B. B., and Fukui, K. Text classification based on word subspace with term-frequency. In *2018 International Joint Conference on Neural Networks (IJCNN)* (2018), IEEE, pp. 1–8.

[37] Singh, D., and Singh, B. Investigating the impact of data normalization on classification performance. *Applied Soft Computing 97* (2020), 105524.

[38] Sundus, K., Al-Haj, F., and Hammo, B. A deep learning approach for arabic text classification. In *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)* (2019), IEEE, pp. 1–7.

[39] Taqi, M. K., and Ali, R. Automatic question classification models for computer programming examination: A systematic literature review. *Journal of Theoretical & Applied Information Technology 93*, 2 (2016).

[40] Tongman, S., and Wattanakitrungroj, N. Classifying positive or negative text using features based on opinion words and term frequency-inverse document frequency. In *2018 5th international conference on advanced informatics: Concept theory and applications (ICAICTA)* (2018), IEEE, pp. 159–164.

[41] Utomo, B. Y., and Bijaksana, M. A. Comprehensive comparison of term weighting method for classification in indonesian corpus. In *2016 4th International Conference on Information and Communication Technology (ICoICT)* (2016), IEEE, pp. 1–5.

[42] Waheed, A., Goyal, M., Mittal, N., Gupta, D., Khanna, A., and Sharma, M. Bloomnet: A robust transformer based model for bloom's learning outcome classification. *arXiv preprint arXiv:2108.07249* (2021).

[43] Wijaya, M. The classification of documents in malay and indonesian using the naive bayesian method uses words and phrases as a training set. *Mendel Journal 26*, 2 (Dec. 2020), 23–28.

[44] Yahya, A. A., and Osman, A. Automatic classification of questions into bloom's cognitive levels using support vector machines.

[45] Yahya, A. A., Osman, A., Taleb, A., and Alattab, A. A. Analyzing the cognitive level of classroom questions using machine learning techniques. *Procedia-Social and Behavioral Sciences 97* (2013), 587–595.

[46] Zelinka, I., and Dao, T. On voynich alphabet analysis with relation to the old indian dialects. *Mendel Journal 26*, 1 (Aug. 2020), 15–22.

[47] Zhang, S., Wang, Y., and Tan, C. Research on text classification for identifying fake news. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)* (2018), IEEE, pp. 178–181.