QUBAHAN ACADEMIC JOURNAL
VOL. 5, NO. 1, March 2025
https://doi.org/10.48161/qaj.v5n1a1569

TWTFPOS-IDF: Thematic Term Weighting Scheme for Enhanced Question Classification Using Bloom's Taxonomy

Sucipto [1, 2], Didik Dwi Prasetya [1*], and Triyanna Widiyaningtyas [1]

[1] Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang 65145, Indonesia;
[2] Department Information System, Universitas Nusantara PGRI Kediri, Kediri 64112, Indonesia.

* **Corresponding author:** didikdwi@um.ac.id.

ABSTRACT: Creating question text using a cognitive approach based on Bloom's Taxonomy (BT) is essential for maintaining question quality in learning assessment. Various studies have explored term weighting schemes to improve BT-based question classification accuracy. However, achieving higher accuracy in classifying cognitive levels requires more than just analyzing verbs—it must also incorporate thematic terms relevant to BT. Existing approaches primarily assign weights to verbs and supporting verbs, often neglecting thematic terms that provide crucial context for classification. This study introduces a novel thematic term weighting scheme, TWTFPOS-IDF, which assigns the highest weight to thematic terms compared to verbs and other supporting words. Thematic terms are identified using the BT word database, with feature extraction, selection, and model tuning optimized to enhance classification accuracy. To ensure robustness, the model is evaluated using a newly constructed, larger dataset that includes a diverse set of educational questions across multiple domains. Machine Learning (ML) and Deep Neural Networks (DNN) are employed for classification, with performance assessed using standard metrics and ANOVA statistical testing. The experimental results demonstrate that the proposed model significantly outperforms previous schemes, achieving an average accuracy of 0.905 and a k-fold value of 0.886. The highest-performing ML algorithm recorded an accuracy of 0.977 and a k-fold value of 0.970. The use of a larger dataset ensures greater generalizability and stability of the model across different question structures. The ANOVA test confirms that model optimization and the expanded dataset significantly improve classification accuracy compared to prior research. This research addresses key challenges in automated question classification, enhancing the precision of cognitive level identification in educational assessment. Future studies will focus on automating weight identification and leveraging deep learning techniques to further refine classification performance and scalability.

Keywords: Bloom's Taxonomy (BT), thematic, question classification, TF-IDF, term weighting.

I. INTRODUCTION

Exam questions are indicators to determine students' understanding and learning ability [1]. The preparation of exam questions is sourced from the basic competencies of curriculum standards and materials delivered by teachers [2]. High-quality exam questions can provide an accurate picture of learning outcomes in the learning process [3]. The proper learning process can make it easier for teachers to evaluate teaching methods [3]. On the other hand, the quality of the exam questions is not by the curriculum and the material that has been delivered. The lack of appropriate quality can result in inaccurate evaluations in teaching [4, 5]. Therefore, a teacher needs to improve the quality of exam questions. Exam questions consist of various questions that must be classified according to educational achievements in the curriculum standards. In this regard, Artificial Intelligence (AI) can significantly support educators by automating the classification and

analysis of exam questions, ensuring better alignment with curriculum standards [6]. A popular and widely adopted classification in many studies using the Bloom Taxonomy (BT) theory approach [7-9].

BT is a learning classification framework with three domains: cognitive, psychomotor, and affective. BT was first developed by Benjamin Bloom in 1956 and revised several times. The last version used was published in 2001 [10]. Each BT domain category, specifically the Cognitive domain, is divided into High Order Thinking Skills (HOTS) and Lower Order Thinking Skills (LOTS). Some uses of BT include functioning as a metacognitive framework to encourage the development of prospective teachers' pedagogical content knowledge [11, 12]. Another benefit of the BT approach is that it can promote metacognitive expertise in students to achieve learning achievement [13]. BT can map thinking skills at LOTS and HOTS levels [14, 15]. Mapping thinking skills can improve effective teaching and can provide success in achieving learning achievement at school [14]. Figure 1 displays the BT domain map.

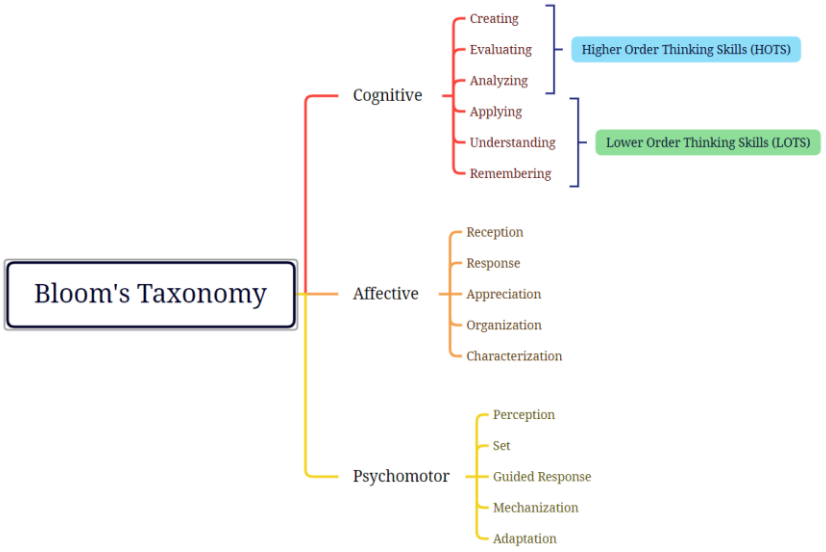

FIGURE 1. Bloom's taxonomy.

The BT cognitive domain consists of six levels of achievement, as shown in Figure 2. This level is sorted into two parts: low to high-level thinking. Each question should be tailored to one of these cognitive level achievements. The use of frameworks in the preparation of questions requires special skills because they are not only adjusted based on verbs but must also be by the meaning of BT. In addition, accurate BT classification requires a team of experts; if done manually, it will take a long time. Datamining research in education studies a lot about the classification of questions within the scope of the BT cognitive domain [16, 17]. Question classification research using data mining aims to extract the meaning of question sentences according to the BT cognitive domain.

The challenge in BT classification lies in the multiclass label. A multiclass of labels consists of 6 domains, as presented in Figure 2. In BT research, data is in sentences [18], whereas text classification uses data in paragraphs [19]. The purpose of BT classification is to see how appropriate the meaning of the sentence of the question text is by the provisions of the meaning of BT. In this provision, the previous researchers [18, 20, 21] have never used thematic meaning as an element of BT classification assessment used in the main proposal of this study.

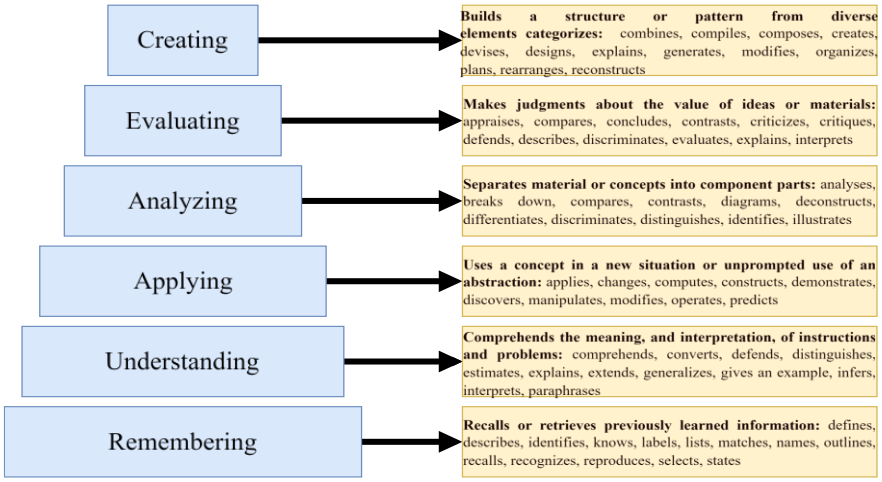

FIGURE 2. Adapted from Anderson's revisions on Bloom's taxonomy verbs.

Data mining in BT classification can automatically classify queries [22]. The methods in the classification are divided into two, namely the Machine Learning (ML) and Deep Learning (DL) methods [23, 24]. ML methods still depend on the stages of extraction features, but in DL, such as its derivative, Deep Neural Networks (DNN), the extraction feature can be used to improve model performance [23]. The initial stage of the classification model consists of a dataset process, which in this case is in the form of question text. In the case of classification in the realm of BT research, it refers to public datasets that several researchers have used before [25-27]. The next stage is the preprocessing stage, which then continues to the classification method stage and ends with the evaluation stage—an overview of the classification process is shown in Figure 3.

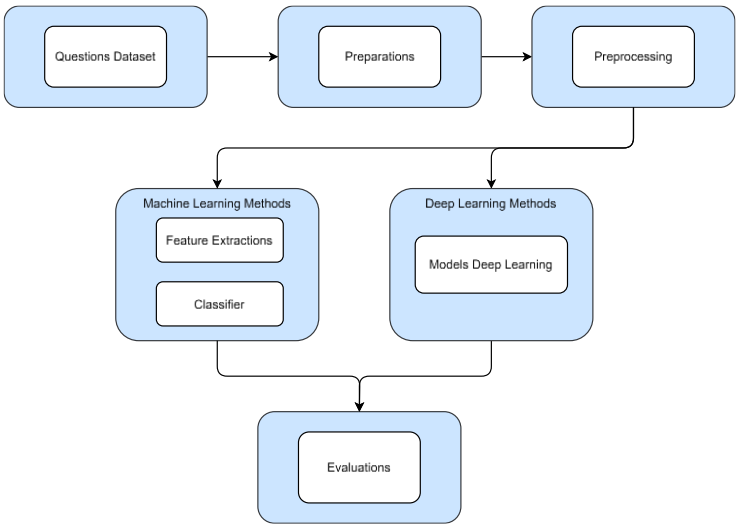

FIGURE 3. Classification methods.

Exam questions consist of text-based questions. Classifying text on questions, which was widely used in previous research, consists of two main methods. ML techniques and DL techniques. Techniques for using ML algorithm models such as Naïve Bayes (NB), k-Nearest Neighbour's (k-NN), Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR), Fuzzy, and Rocchio [19]. The use of DL includes

Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GANs) [28]. The two algorithmic engineering models differ in handling the amount of data and complexity but have the same goal: improving accuracy [24].

Table 1. Past research on text classification.

Work	Year	Extraction	Selection	Classification
[29]	2019	✓		ML
[30]	2020	✓		ML
[31]	2021	✓		DL
[27]	2022	✓		ML
[32]	2022	✓		DL
[33]	2023		✓	ML
[34]	2023		✓	ML
[35]	2024		✓	ML
Proposed		✓	✓	ML & DL

Table 1 presents previous studies that serve as references for text classification. Text classification is a crucial foundation for handling question text classification cases, as it shares several structural and contextual similarities. These similarities present an opportunity to introduce a novel approach to question classification by leveraging techniques from general text classification research. Previous studies have applied ML [29, 30-34, 36, 37] and DL [31, 32] for text classification. Additionally, feature extraction and selection techniques have been employed to enhance term weighting, with varying advantages depending on data characteristics, problem complexity, and available computational resources.

Despite these advancements, existing research on BT question classification has primarily focused on feature extraction through term weighting [26, 30]. The dominant approach involves numerical and verb-based weighting [7, 27], which remains limited in capturing the deeper semantic relationships within a question. While verbs play a crucial role in identifying cognitive levels, thematic words—key terms that define the overall meaning of a question are often overlooked. The omission of thematic terms in the weighting process reduces the effectiveness of classification, as questions often contain critical words beyond verbs that define their intended cognitive category.

This study addresses this gap by introducing a new thematic word weighting model, TWTFPOS-IDF, which prioritizes thematic words over verbs and supporting words in feature extraction. Thematic understanding refers to identifying words that encapsulate the main idea of a sentence [38, 39]. Prior research has shown that sentences containing more thematic words align more accurately with their intended meaning and classification labels [40]. Thematic word identification is widely used in text comprehension analysis [41, 42], as it plays a crucial role in determining sentence importance [43]. Additionally, counting the number of thematic words that appear frequently in a sentence can reveal words with maximum possible relativity, further strengthening the sentence's alignment with its intended meaning and cognitive label [44].

The primary baseline for this research is Gani et al. [27, 36]. which modified feature extraction by incorporating verb weighting and supporting verbs. However, their approach has limitations in capturing deeper semantic structures, as it does not consider the broader thematic context. To address this issue, this study proposes a novel combination of extraction and selection features by assigning primary weight categories to thematic words alongside verbs. This new weighting scheme aims to optimize BT question classification while enhancing feature selection through performance tuning.

The weighting of numerical terms and verbs is one of the essential components of ML classification. However, existing methods primarily focus on verb-based weighting, which may overlook other important components in text classification. In this study, optimization is performed by enhancing extraction features with semantic components, specifically thematic understanding. This approach helps capture the deeper semantic meaning of the question text. Additionally, performance tuning in feature selection is applied to refine the classification process [45, 46]. By optimizing both extraction features and feature selection, this

research aims to improve the accuracy and efficiency of BT question classification using ML and DL techniques.

Furthermore, this research contributes by implementing a more rigorous evaluation framework. Unlike previous studies, it utilizes a newly constructed, larger dataset to improve model generalizability. The study also incorporates the k-fold cross-validation test to assess algorithm stability and the ANOVA significance test to measure the statistical impact of performance improvements. These evaluation techniques, which were absent in prior research, ensure a more reliable assessment of the model's effectiveness beyond numerical accuracy improvements.

II. RELATED WORK

This section focuses on the literature on the thematic techniques of extraction features, datasets, and classification in the TF-IDF scheme to understand text classification. Table 1 shows the previous study on text classification in general, including extraction features, selection features, and classifications, while Table 2 focuses on the TF-IDF scheme.

1. WORK ON FEATURE EXTRACTION, FEATURE SELECTION AND CLASSIFICATION

Some text-processing techniques include unigram, bigram, trigram, n-gram, Word2Vec, GloVe, and fastText. Unigram, bigram, trigram, and n-gram techniques are ways to perform text processing where a combination of word sequences is used as a feature. Research using text processing (n-gram) can effectively overcome algorithm weaknesses [47]. The text processing weighting approach is simple, so it's easy to get the feature set from the question [48]. This feature will improve performance when the data is large enough [36]. Word2Vec is a technique in text processing that can help generate vector representations of words from text. Using word2vec merging with the extraction feature on word vectors as a word embedding layer can optimize better text prediction results [32, 49]. GloVe (Global Vectors for Word Representation) is a method in text processing that can produce vector representations of words based on co-occurrence statistics—combination with the use of GloVe on word embedding results in improved classification accuracy [50].

Some of the techniques in text selection include Chi-square, PCA (Principal Component Analysis), SMOTE (Synthetic Minority Over-sampling Technique), and PSO (Particle Swarm Optimization). Chi-square is a statistical method that can test the relationship between two categorical variables in a contingency table [45]. The use of Chi-square can increase the effectiveness of the word weighting term scheme to increase the accuracy of the evaluation value [45, 46]. PCA analysis techniques can be used to find discrimination features and identify key features for classification [51]. Combining PCA with classification algorithms can also improve accuracy [51, 52]. SMOTE is a method that solves the problem of unbalanced data to deal with class imbalances [53]. BT is a multiclass that can be handled using SMOTE [53]. On the other hand, PSO can produce the optimum value of the optimized function [54, 55]. Using PSO can further improve the performance of text classification [54].

The list of ML algorithms that are often used as text classification includes Naïve Bayes (NB), k-Nearest Neighbour's (k-NN), Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR), Fuzzy, and Rocchio. The use of ML in the case of question texts aims to classify the text, according to BT, based on its features [19]. Techniques other than ML, namely DL, have the advantage of being able to model and understand complex data. Some examples of DL algorithms include Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GANs) [28]. In addition to using ML and DL to obtain data understanding results, there are additional techniques in the extraction feature and selection feature that can be used to improve data understanding so that a higher level of accuracy is obtained.

2. WORK ON TERM WEIGHTING

Table 2 represents previous studies that served as references for discovering novel feature extraction methods using the TF-IDF technique. The dataset used in BT classification, particularly for question sentences in past research, was relatively small, impacting the classification techniques, which showed minimal

differences between ML and DL algorithms. The previous studies primarily focused on word weighting using TF-IDF, emphasizing verb-based weighting as the main factor in classification. However, no prior research has explored thematic word weighting, which could play a crucial role in understanding the overall context and meaning of questions. Further optimization through feature selection and comparative evaluation of ML and DL classification techniques is necessary. The datasets used in these previous studies are publicly available, as summarized in Table 2. From the data in Table 2, it is evident that earlier research primarily employed variations of TF-IDF, with gradual improvements in dataset size and classification techniques over time. The studies from 2019 to 2022 show a progression from simple TF-IDF to more refined schemes such as TFPOS-IDF and ETFPOS-IDF, incorporating additional linguistic features.

The study conducted by Aninditya (2019) [29] represents the earliest application of standard TF-IDF with machine learning for BT classification, utilizing the smallest dataset. In 2022, Liang [32] introduced a modification of TF-IDF using LSTM with a significantly larger dataset consisting of 1,000 instances. Alammary (2021) [31] initiated modifications in word weighting through M-TF-IDF by incorporating language-specific characteristics, particularly in Arabic datasets. Researchers Mohammed [30] and Gani [27] further refined the weighting techniques by implementing TFPOS-IDF and ETFPOS-IDF, respectively, using public datasets and focusing on verb-based weighting. The proposed method, TWTFPOS-IDF, introduces a significantly larger dataset (1,771 instances) and emphasizes thematic word weighting rather than verb-based weighting. Additionally, it enhances feature extraction through parameter tuning and employs a more comprehensive evaluation process, including *k*-fold cross-validation and ANOVA. This approach aims to bridge the gap by incorporating both ML and DL techniques, thereby addressing the limitations of previous research in terms of dataset size and classification criteria.

Table 2. Past research of question classification on term weighting.

Work	Year	Scheme	Dataset	Classification
[29]	2019	TF-IDF	300	ML
[30]	2020	TFPOS-IDF	600	ML
[31]	2021	M-TF-IDF	610	DL
[27]	2022	ETFPOS-IDF	600	ML
[32]	2022	TF-IDF	1000	DL
Proposed		**TWTFPOS-IDF**	**1771**	**ML & DL**

3. RESEARCH GAP IN TERM WEIGHTING

From the discussion above, it is clear that previous research using the TF-IDF scheme primarily emphasized word weighting based on verbs, without considering thematic word weighting related to the BT cognitive domain. These studies relied on relatively small datasets and focused on verb-based weighting as the key factor in classification. However, no prior research has explored the use of thematic words as a primary weighting approach. This study aims to address this gap by proposing thematic word weighting as a more effective method for determining the BT cognitive domain. The proposed scheme in this study is presented in Figure 4. This research proposal begins with the selection of datasets that have a difference of almost 3 times from the previous study; it is intended to determine the effectiveness of the performance of using ML algorithm engines. The dataset has the same number between BT domains, which differs from some previous studies in Table 1 and Table 2. The preprocessing stage uses techniques generally carried out by previous research, namely case folding, stopword, and lemmatization. The main novelty proposal is at the stage of extraction features with the weighting of thematic words composed from the word BT. Thematic words are identified based on the BT word database. The extraction feature also adds n-grams to optimize text weighting. Another proposal is the addition of optimization tuning of the selection feature, which was also rarely used in previous research due to the amount of data that was considered insufficient. In this study, n-gram and chi-square were added to optimize accuracy in the classification machine.

This study uses two models, namely the ML (SVM, NB) and DL (ANN, MLP) algorithms. The main algorithm proposed was the ML algorithm (SVM) compared to other algorithms. The SVM algorithm was chosen because of its reliability in classifying BT text with feature extraction modifications and tuning feature

selection. The evaluation uses a complete evaluation, namely precision, f1-score recall, and accuracy. Another test is to test the performance stability with k-fold cross-validation with k=10. In addition, a significance test using ANOVA was also added to determine the significance of the results of the evaluation of algorithm metrics. Therefore, this study aims to introduce the thematic word weighting scheme and add selection features as illustrated in the flow of the classification model in Figure 4.

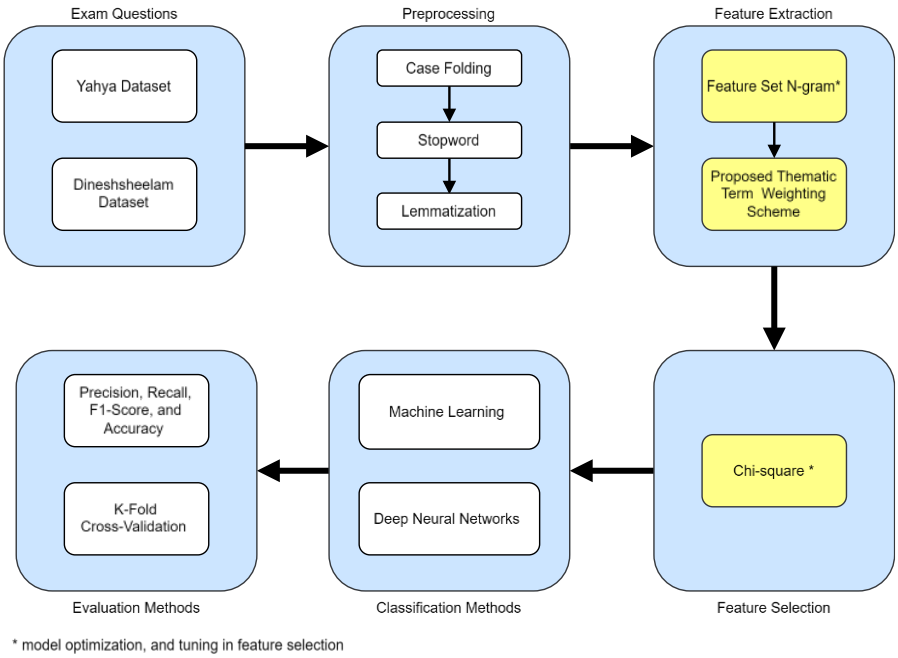

FIGURE 4. Stages of the proposed question classification model.

III. MATERIAL AND METHOD

Figure 4 illustrates all the stages in the question classification model with the proposed BT domain. The steps in classifying the question text are dataset capture, pre-processing, extraction features, selection features, classification, and model evaluation. Proposal of novelty at the feature extraction feature stage by creating a set of feature functions in calculating word weight, especially in thematic words.

1. DATASET

This study leverages two publicly available datasets that have been used in previous research to ensure comparability and consistency in the evaluation of the proposed method. The first dataset, consisting of 600 questions, has been widely utilized in prior studies [25, 27, 36, 56, 57], and is based on English language questions. By using this dataset, the current study allows for a direct comparison of the performance of the proposed model with those in the existing literature, ensuring alignment in the evaluation approach across different studies.

To address concerns about dataset size and diversity, this study incorporates a second, larger dataset, comprising 1,771 questions [58]. This new dataset not only increases the overall number of questions but also provides a more balanced distribution of questions across the six Bloom's Taxonomy (BT) cognitive domains, with an average of 300 questions per domain (except for the "Creating" domain, which contains 271 questions). This expanded dataset offers a better foundation for performance stability testing through *k*-fold cross-validation than previous studies with smaller datasets [7]. By utilizing a larger and more diverse set of questions, the study improves the reliability and robustness of the model evaluation.

While the datasets used are in English and contain general topic questions, they were selected to establish a solid baseline for performance evaluation. Figure 5 presents a visualization of the dataset used in this study. By including the larger second dataset, the study contributes to the field by providing more robust performance metrics and a stronger foundation for model validation [25, 28].

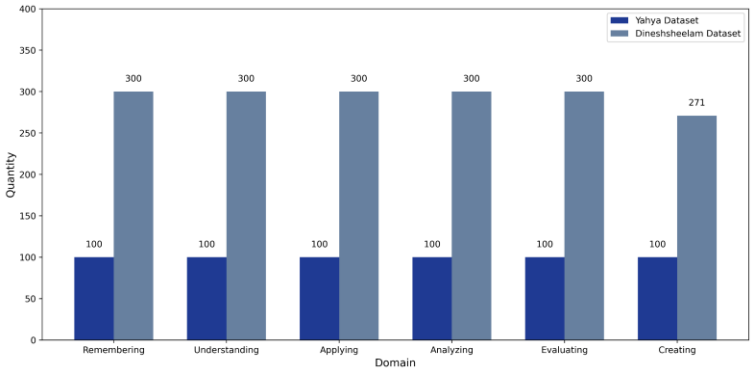

FIGURE 5. Visualization of bloom's taxonomy dataset questions.

2. PREPROCESSING

Several ways to clean and tidy up raw text to improve model understanding. Some methods generally include tokenization, lowering casing, cleaning text, stopword removal, stemming, lemmatization, and normalization [16, 28]. Some words need to be identified, and generally, the verbs in the question have a high weight on the BT assessment. However, identification needs to be carried out due to verb changes, for example, the word "do," which can change to verbs, auxiliaries, nouns, or abbreviations. Another proposed technique is the weighting of thematic words in BT as the main identification of BT.

The preprocessing step in this study begins with creating a case folding function by changing the text to be too small and removing URLs, numbers, and punctuation characters. After that, stop removal by including several words that have the potential to become the main verb, such as Modal (MD), Wh-pronoun (WP), Wh-determiner (WDT), etc. Next is the lemmatization process that uses wordNet to convert words into basic word forms [59, 60]. The final preprocessing step is creating a text preprocessing pipeline function to combine the text preprocessing steps.

3. FEATURE EXTRACTION

3.1. FEATURE SET

This text-processing technique can provide more information about the context of the text and can improve understanding of the relationships between words in the text [47]. This study uses a combination of n-grams to generate a feature set that includes all the unique terms in the dataset. Using n-grams can increase the complexity and size of text representations [32, 50].

3.2. TERM WEIGHTING

This study applies a scheme other than that listed in Table 2 for weighing words in BT proposed by the previous study on question classification. The scheme is ETFPOS-IDF [27]. ETFPOS-IDF is the latest scheme proposed in question classification for term weighting schemes that give BT a higher weight than supporting verbs. This study provides a thematic word weighting scheme as the primary reference in determining BT.

3.3. PROPOSED SCHEME TWTFPOS-IDF

The scheme proposed by TWTFPOS-IDF is another proposed version of the wording of the ETFPOS-IDF technique proposed by Gani [27]. ETFPOS-IDF distinguishes between verb types and gives higher weight to BT verbs than supporting verbs. However, ETFPOS-IDF ignores word weighting in verb changes, and there is

no weighting in thematic word schemes. The scheme proposed in this study distinguishes the types of words in the question, determines thematic words, and gives higher weight to BT thematic words than other supporting words. TWTFPOS-IDF is discussed in equations (1) to (4).

$$Tw_{\text{pos}}(t) = \begin{Bmatrix} w1, \text{ if } t \text{ is } Thematic \\ w2, \text{ if } t \text{ is } Verb \\ w3, \text{ if } t \text{ is } Noun \text{ or } Adjective \\ w4, otherwise \end{Bmatrix} \quad (1)$$

In contrast, the proposed TWTFPOS-IDF scheme enhances the term weighting process by distinguishing different word types in the question text and assigning higher weights to thematic words. Thematic words are key terms that directly reflect the cognitive level of the question and play a significant role in the classification process. In the TWTFPOS-IDF model, thematic words are assigned the highest weight, w1 = 4, followed by verbs (w2 = 3), nouns or adjectives (w3 = 2), and other words (w4 = 1). This distinction between word types is outlined in Equation (1).

$$\text{Score}(S_i) = \frac{\text{no.Thematic word in } S_i}{\text{no.word occuring in } S_i} \quad (2)$$

Thematic words are identified using semantic relationships, which are extracted from the BT word database and further refined with tools like WordNet. The identification process is crucial because thematic words encapsulate the core meaning of the sentence, which may not be fully conveyed by verbs alone. This distinction allows the scheme to better reflect the question's intent, improving the overall accuracy of the BT classification.

To calculate the thematic score of a sentence, we use Equation (2), which measures the ratio of thematic words to the total number of words in the sentence. This score indicates how closely the sentence aligns with its intended cognitive category based on the frequency of thematic words it contains. The thematic score indicates how many thematic words are contained in a sentence [61]. Thematic words are grouped according to semantic relationships, which are essential for extracting relevant cognitive cues for classification.

Additionally, to optimize BT identification further, special attention is given to specific words like "do," which can function as different parts of speech (e.g., noun or verb) depending on context

- Sentence 1: Can you tell me do for solving this math problem
  tag: [('Can', 'MD'), ('you', 'PRP'), ('tell', 'VB'), ('me', 'PRP'), ('the', 'DT'), ('do', 'NN'), ('for', 'IN'), ('solving', 'VBG'), ('this', 'DT'), ('math', 'NN'), ('problem', 'NN')]
- Sentence 2: Can you tell me how to solve this math problem
  tag: [('Can', 'MD'), ('you', 'PRP'), ('tell', 'VB'), ('me', 'PRP'), ('how', 'WRB'), ('to', 'TO'), ('do', 'VB'), ('solving', 'VBG'), ('this', 'DT'), ('math', 'NN'), ('problem', 'NN')]

In Sentence 1, "do" is tagged as a noun (NN), while in Sentence 2, it is tagged as a verb (VB). The distinction is crucial for correctly identifying its role in the cognitive structure of the question. These improvements in the TWTFPOS-IDF scheme ensure that thematic words receive appropriate weighting, ultimately enhancing the classification of BT questions. The combination of extraction feature optimization and performance tuning leads to more accurate and efficient BT question classification. Algorithm 1 is created to determine the semantic proximity of a sentence assessed as part of BT after the preprocessing stage is carried out.

Algorithm 1 Process in Identifying BT Thematic

1: *q*: A Question
2: *d*: BT Thematic Database
3: **function** *Identify*(*q*, *d*)
4: *sentences* ← split question
5: *new list* ← []
6: **for** *sentence* in *sentences*, **do**

```
7:     words ← split sentence
8:     x ← Find the word of the sentence
9:     if x is in d, then
10:          new list.insert((x, "BT"))
11:    else
12:          new list.insert((x, "non-BT"))
13:    end if
14:  for word the remaining words, do
15:    if the word is in d, then
16:       y ← previous word
17:       if y == "and" then
18:          new list.insert((word, "BT"))
19:       else if y is (Verb, Adverb, Wh, other), then
20:          new list.insert((word, "BT"))
21:       else
22:          new list.insert((word, "non-BT"))
23:       end if
24:    else
25:          new list.insert((word, "non-BT"))
26:    end if
27:    end for
28:  end for
29:  return new list
30: end function
```

Calculate $TW_{pos}(t)$ from equation formula (1) is used to calculate TWTFPOS(t, q) , as shown in the equation of the formula (3).

$$\text{TWTFPOS } (t,q) = \frac{C(t,q) \times Tw_{pos}(t)}{\sum_i C(t_i,q) \times Tw_{pos}(t_i)} \tag{3}$$

Where $C(t,q)$ symbolizes frequency t on the question q and $\sum_i C(t_i,q)$ is the sum of terms in the question q. Last, *TWTFPOS - IDF(t,q)* was calculated using the formula equation (4).

$$\text{TWTFPOS - IDF}(t,\ q) = \text{TWTFPOS}(t,\ q) \cdot \text{IDF}(t) \tag{4}$$

TWTFPOS - IDF(t, q) is the multiplication of TWTFPOS(t, q)T and IDF(t), as shown in equation formula (4). Normalization techniques are used to prevent the complexity of numerical calculations during the model training process, as stated in previous studies [26, 27, 62]. This study normalizes the weight value of the proposed TWTFPOS-IDF scheme using the L2 normalization technique. As a result, all weight values are converted between 0 and 1. The TWTFPOS-IDF formula equation has also been normalized by referring to previous research [26, 27]. The formula equation (5) obtains the normalized term weight value.

$$\text{normalized term weight value} = \frac{TWTFPOS - IDF(t,q)}{\sqrt{\sum\ TWTFPOS - IDF(t,q)^2}} \tag{5}$$

In the equation of the formula (5), $\text{TWTFPOS} - \text{IDF}(t,q)$ is used as a term for the value of the weight obtained for t on the question q.

4. FEATURE SELECTION

This selection feature technique can be used with several different approaches. Chi-square can be used to determine the selection of category data features [45, 63]. PCA can reduce the dimension of features in complex datasets [64, 65]. SMOTE is not specific to the selection feature but can affect the distribution of features and can be used as part of the feature selection process [66-68]. This study used only chi-square for the word features most related to the target class, especially with large datasets.

4.1. CLASSIFICATION AND EVALUATION

This study uses two techniques, namely ML and DL. ML uses machine learning classification algorithms, namely Support Vector Machine (SVM) and Naïve Bayes (NB) [69]. DL uses its derivative, namely the DNN algorithm. The DNN algorithms used include Artificial Neural Networks (ANN) and Multi-layer Perceptron (MLP) [18, 70]. The tools used are Anaconda version 24.3.0 and VSCode version 1.87.2 with the Python programming language version 3.11.7 [27, 71, 72] to process text data and test models. The three classification algorithms include machine learning models with a Supervised Learning approach to obtain equivalent comparative values [73]. SVM is a machine learning algorithm used for class separation and regression. A previous study improved the Term Weighting TF-IDF by using the SVM algorithm with optimal results compared to other algorithms with an accuracy of 73.3% [73]. In addition, other studies have also improved TF-IDF on the SVM algorithm with an optimal accuracy result of 89.7% [30]. Another study describes short-answer questions with TF-IDF that provide accuracy values to show stable performance on SVM algorithms [74]. Therefore, SVM is the main choice for classification algorithms.

NB is a classification algorithm based on Bayes' theorem, assuming independence between features [29]. The classification study on BT with NB for TF-IDF optimization combined with n-gram had a significant result of 85% [29]. Another study presents an approach to automatically classifying items according to BT, showing that NB has free parameters and is a suitable candidate for exploratory studies [70]. The free parameter method has high flexibility and can handle unstructured data or distributions that are not well-known [70]. By improving the TF-IDF method, the research utilizes part-of-speech markers that get significant results in the NB algorithm with an accuracy of 85% [26].

Artificial Neural Networks (ANNs) are artificial neural network algorithms consisting of neurons connected in various layers. ANNs can be simple (have one hidden layer) or complex (have many hidden layers), and ANNs are considered deep neural networks (DNNs) [27, 56, 75]. Multilayer Perceptron (MLP) is a type of ANN with at least three layers (input, hidden layer, and output). An MLP with more than one hidden layer is called a deep MLP, an example of a deep learning model [36]. The evaluation technique uses precision, recall f1-score, and accuracy metrics. Previous research's evaluation metric approach to question classification also extensively uses this metric [76-78]. Here's a metric evaluation equation:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \text{TP}_i}{N} \tag{6}$$

$$\text{Precision} = \frac{\sum_{i=1}^{N} \text{TP}_i}{\sum_{i=1}^{N} (\text{TP}_i + \text{FP}_i)} \tag{7}$$

$$\text{Recall} = \frac{\sum_{i=1}^{N} \text{TP}_i}{\sum_{i=1}^{N} (\text{TP}_i + \text{FN}_i)} \tag{8}$$

Where N is the total number of samples or observations, TPi is the number of cases in which the prediction is correct, and FNi is the number of cases where the prediction is wrong negative for sample i. This formula represents the number of correct predictions divided by the total number of cases the model should have predicted positively. F1-measure is an evaluation metric that measures the balance between precision and recall in classification. F1-measure combines the two metrics into a single value that comprehensively reflects the model's performance. The equation of the F1-measure formula is as follows:

$$F_{1-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

In addition, it also uses performance with *k*-fold cross-validation. In the evaluation of *k*-fold cross-validation, data is divided into k different subsets, and the model is trained and tested *k* times using a different subset as test data [29, 66, 79]. The *k*-fold technique can see the stability of the model's performance against possible data variations [36, 80]. If the model's performance varies significantly between folds (subsets), this can indicate that the model is unstable to dataset variations. Conversely, if the model's performance is relatively consistent across folds, this can be taken as an indication that the model is fairly stable [81]. The *k*-fold formula equation is as follows:

$$k\text{-fold} = \frac{1}{k}\sum_{i=1}^{k} \text{error}_i \quad (10)$$

Where k-fold is the error estimation of the model using *k*-fold cross-validation, *k* is the number of folds. error*i* is a model error on the fold *i*. ANOVA is used to determine whether there is a significant difference between the mean of more than two groups. The F-statistic test formula is:

$$\text{F} = \frac{\frac{SSB}{df_B}}{\frac{SSW}{df_W}} \quad (11)$$

Where: *SSB* = Sum of Squares Between, df_B = Degree of Freedom Between Groups, *SSW* = Sum of Squares Within, df_W = Degree of Freedom in Groups.

This study adopts an evaluation approach using precision, recall, f1-score, accuracy, and *k*-fold cross-validation metrics. The evaluation approach aims to maximize the evaluation of questions with the model that has been built and as a differentiator from previous research, which only used a few evaluations [7, 25-27]. This study also used the ANOVA significance test to test the significance of the evaluation algorithm metric.

IV. DATA ANALYSIS

The results of this study are presented in the TF-IDF model initially introduced by the researcher, who became a reference for BT research [25] other comparisons with recent research on the classification of BT weighting techniques. TF-IDF [25], ETF-IDF [26] and ETFPOS-IDF Model [27] is a previous research model and a scheme model (TWTFPOS-IDF) is the scheme model proposed in this study. The results of this study used an evaluation matrix and *k*-fold, and at the end of the test, the ANOVA statistical test was used.

1. EXPERIMENT RESULTS OF SVM

Tables 3 and 4 are the evaluation results using the formula equation (6-10). The model scheme in this experiment uses TF-IDF modification with the SVM algorithm. Dataset 1 contains 600 BT questions, and Dataset 2 contains 1771 BT questions. Both models were evaluated using the SVM algorithm.

Table 3. SVM experimental results on Dataset 1.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.749	0.746	0.742	0.742	0.730
ETF-IDF	0.785	0.742	0.746	0.742	0.693
ETFPOS-IDF	0.756	0.723	0.729	0.733	0.707
TWTFPOS-IDF	**0.856**	**0.850**	**0.847**	**0.850**	**0.777**

Table 3 is the result of the SVM evaluation using dataset 1. The proposed scheme model (TWTFPOS-IDF) in Table 3 received the highest score in all evaluations. In dataset 1, all models have a k-fold performance lower

than the accuracy performance. However, the proposed model (TWTFPOS-IDF) still has an advantage over other models, and the k-fold value of the proposed model is higher than the accuracy value of other models. The average difference in accuracy performance with the proposed model is 0.111, and the k-fold is 0.067. The highest difference value in the recall evaluation was 0.113.

The evaluation results using dataset 2, contained in Table 4, are not much different from those in Table 3. The proposed scheme model (TWTFPOS-IDF) received the highest evaluation of all. All model performances with Dataset 2 are better than those with Dataset 1. The average difference in the accuracy performance of the proposed model is 0.029, and the k-fold is 0.013. In the SVM trial, dataset 2 still performs better than the accuracy results of other models.

Table 4. SVM experimental results on Dataset 2.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.953	0.952	0.952	0.952	0.966
ETF-IDF	0.936	0.935	0.935	0.935	0.945
ETFPOS-IDF	0.958	0.957	0.958	0.958	0.961
TWTFPOS-IDF	**0.978**	**0.977**	**0.977**	**0.977**	**0.970**

2. EXPERIMENT RESULTS OF NB

The second evaluation used ML on the TF-IDF modified scheme model using NB. The results were tested using Datasets 1 and 2.

Table 5. NB experimental results on Dataset 1.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.665	0.656	0.644	0.650	0.678
ETF-IDF	0.747	0.749	0.747	0.750	0.760
ETFPOS-IDF	0.800	0.799	0.798	0.800	0.790
TWTFPOS-IDF	**0.832**	**0.830**	**0.829**	**0.833**	**0.825**

The same evaluation results on the SVM algorithm with the NB algorithm in dataset 1 in Table 5 show that the evaluation value in the proposed scheme model (TWTFPOS-IDF) has a higher evaluation performance value in all evaluation values. The NB algorithm of dataset 1 has a slightly lower performance in the range of 0.8 compared to SVM, but the *k*-fold performance has a better value than the SVM trial of dataset 1. The average difference in accuracy performance of the proposed model is 0.100, and k-fold is 0.082.

Table 6 presents the model performance on the NB algorithm dataset 2. The proposed scheme model (TWTFPOS-IDF) on dataset 2 had the highest score among all evaluation models. NB dataset 2 has a lower average score than SVM dataset 2 but a better k-fold value than its accuracy performance. The average difference in the proposed model's accuracy performance is 0.025, and the fold is 0.019.

Table 6. NB experimental results on Dataset 2.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.881	0.875	0.873	0.876	0.888
ETF-IDF	0.891	0.887	0.886	0.887	0.913
ETFPOS-IDF	0.899	0.893	0.893	0.893	0.919
TWTFPOS-IDF	**0.913**	**0.909**	**0.910**	**0.910**	**0.926**

3. EXPERIMENT RESULTS OF ANN

Evaluation with DL uses 2 DNN models: the ANN algorithm and MLP. The confusion metric assessment results are presented in Tables 7 and 8.

Table 7. ANN experimental results on Dataset 1.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.696	0.513	0.481	0.483	0.532
ETF-IDF	0.790	0.781	0.782	0.783	0.777
ETFPOS-IDF	0.862	0.856	0.856	0.858	0.803
TWTFPOS-IDF	**0.894**	**0.894**	**0.891**	**0.892**	**0.838**

Table 7 is the result of the ANN evaluation using dataset 1. The proposed scheme model (TWTFPOS-IDF) in Table 7 with the DNN group algorithm received the highest evaluation score in all evaluations. The proposed model's average difference in accuracy performance is 0.184, and the k-fold is 0.134. The DNN performance of the ANN model dataset 1 is different in accuracy compared to ML performance.

The evaluation results using dataset 2 contained in Table 8 show results that are not much different from those in Table 7. The proposed scheme model (TWTFPOS-IDF) received the highest evaluation of all. In ANN Dataset 2, k-fold performance is better than accuracy performance. The average difference in accuracy performance of the proposed model is 0.027, and k-fold is 0.024. The performance of ANN dataset 1 is superior to ML, but in ANN dataset 2 ML chooses superior results.

Table 8. ANN experimental results on Dataset 2.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.894	0.893	0.892	0.893	0.906
ETF-IDF	0.956	0.955	0.955	0.955	0.967
ETFPOS-IDF	0.956	0.955	0.955	0.955	0.967
TWTFPOS-IDF	**0.959**	**0.958**	**0.958**	**0.958**	**0.968**

4. EXPERIMENT RESULTS OF MLP

The final evaluation of the BT question uses the MLP algorithm. The MLP algorithm is a multi-layer model selected to obtain differentiating results from the previous algorithm evaluation DL models. The results of the confusion metric evaluation are presented in Table 9 and Table 10.

Table 9. MLP experimental results on Dataset 1.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.657	0.639	0.637	0.642	0.640
ETF-IDF	0.812	0.806	0.806	0.808	0.777
ETFPOS-IDF	0.851	0.851	0.850	0.850	0.803
TWTFPOS-IDF	**0.864**	**0.855**	**0.856**	**0.858**	**0.818**

The MLP results in dataset 1, as shown in Table 9 of the other scheme model on dataset 1, still cannot excel in all metric evaluations. In MLP, dataset 1 has slightly lower accuracy than ANN dataset 1, possibly due to the lack of dataset count. The average difference in the accuracy performance of the proposed model is 0.091, and the k-fold is 0.078.

Overall, the proposed scheme model (TWTFPOS-IDF) received the highest scores in all datasets. In MLP dataset 2 Table 10, the proposed scheme (TWTFPOS-IDF) is also superior to other models. The average difference in accuracy performance of the proposed model was 0.016, and the fold was 0.013. In contrast to the

lower MLP of dataset 1 with ANN, in dataset 2, the MLP performance is better than that of ANN. Larger data can improve the performance of MLP algorithms that have more layers.

Table 10. MLP experimental results on Dataset 2.

Model	Precision	Recall	F1-Score	Accuracy	K-Fold
TF-IDF	0.919	0.918	0.918	0.918	0.937
ETF-IDF	0.958	0.957	0.957	0.958	0.962
ETFPOS-IDF	0.959	0.957	0.958	0.958	0.962
TWTFPOS-IDF	**0.962**	**0.960**	**0.960**	**0.961**	**0.967**

5. EXPLORE DOMAIN BT WITH THE PROPOSED METHOD

After evaluating the ML (SVM and NB) and DNN (ANN and MLP) algorithms, it was followed by evaluating the main algorithm. The main algorithm proposed in this study is the SVM algorithm. The SVM algorithm has the highest advantage compared to other algorithms, with an average accuracy value of 0.914 and a k-fold of 0.874.

This evaluation aims to explore the data from the evaluation results on each BT. 6 domains in BT can also be called classes. The evaluation expository is presented in Table 11 and Table 12, where Table 11 is an experiment with a smaller dataset compared to Table 12. The evaluation exploring 6 BT domains uses all evaluation models: precision, recall, f1-score, and accuracy. The order of the domains is shown in figure 2: 1: Remembering, 2: Understanding, 3: Applying, 4: Analyzing, 5: Evaluating, and 6: Creating.

Table 11. BT experimental results on Dataset 1.

Domain	Precision	Recall	F1-Score	Accuracy
1	0.815	**0.957**	0.880	0.884
2	0.938	0.750	0.833	0.840
3	0.895	0.895	**0.895**	**0.895**
4	0.824	0.824	0.824	0.824
5	0.714	0.882	0.789	0.795
6	**0.950**	0.792	0.864	0.869

Table 11 presents the details of the SVM evaluation that has been presented in Table 3, with a more detailed presentation of the metric values on each domain using Dataset 1. The average precision value is 0.856, with the highest precision value in the Creating domain of 0.950 and the lowest in the Evaluating domain of 0.714. The average recall value was 0.850, with the highest recall value in the Remembering domain of 0.957 and the lowest in the Understanding domain of 0.750.

The average f1-score is 0.848, with the highest f1-score in the Applying domain of 0.895 and the lowest f1-score in the Evaluating domain of 0.789. In the evaluation, the recall has the highest performance value of 0.957. In the evaluation, precision has the highest value in one of the low domains, namely Evaluating, with a value of 0.714. Still, precision has the highest average evaluation value among other evaluation values. The domain with the highest accuracy is applying, with a value of 0.895, and the lowest is evaluating, with 0.795.

Table 12. BT experimental results on Dataset 2.

Domain	Precision	Recall	F1-Score	Accuracy
1	0.962	0.943	0.952	0.952
2	0.938	0.952	0.945	0.945
3	0.985	**1.000**	0.992	0.992

4	**1.000**	0.982	0.991	0.991
5	0.984	0.984	0.984	0.984
6	**1.000**	**1.000**	**1.000**	**1.000**

Table 12 presents the SVM evaluation on each BT domain with a larger dataset. Evaluation results with big data provide better performance differences than smaller datasets. The average result was the same as obtained by the overall evaluation of the six BT domains. The average value of precision is 0.978. The highest value of precision in the Analyzing and Creating domain with a perfect value of 1.00. The lowest score in the Understanding domain is 0.938. The average recall value was 0.977, with the highest recall value in the Applying and Creating domain with a perfect value of 1.00 and the lowest in the Remembering domain of 0.943.

The average value of the f1-score is 0.977. The Creating domain obtained the highest score in the f1-score with a perfect score of 1.00. The Understanding domain obtained the lowest score in the f1-score with a value of 0.945. The average accuracy score was 0.977. The Creating domain obtained the highest accuracy score, with a perfect score of 1.00. The Understanding domain obtained the lowest value in accuracy with a value of 0.945.

In dataset 2, several domains have perfect performance values, including the Applying, Analyzing, and Creating domains. The performance of dataset 2 is superior to dataset 2 due to the amount of data and good data distribution in each domain [82, 83].

6. SUMMARY

The last step is to test the scheme models' evaluation value using the ANOVA statistical test. The statistical results are presented in Table 13.

Table 13. Test Results ANOVA

Metric	F-value	P-value
Precision	9.255	0.002
Recall	7.926	0.004
F1-Score	8.046	0.003
Accuracy	7.395	0.005
K-Fold 10	7.647	0.004

In Table 13 of the ANOVA test, the F value ranges from 7.395 to 9.255, which is a high value. This suggests that the differences between the groups were statistically significant for all metrics (precision, recall, f1-score, accuracy, and *k*-fold). The results of the F test show a comparison of the variation ratio between groups. The Accuracy Value has the highest F value (9.255), which indicates that the variation between the group averages is the greatest for accuracy compared to other metrics. The recall had the lowest F value (7.395) but was still high enough to show significant differences between groups.

In Table 13 of the ANOVA test, all P values are below 0.05, ranging between 0.002 and 0.004. This suggests that for all metrics, there are statistically significant differences between groups. The P value for Precision is the lowest (0.002), which means the result is the most statistically significant compared to other metrics. The P value for Accuracy is the highest (0.005), but it is still well below the 0.05 threshold, which indicates a significant result.

Figure 6 illustrates a summary of the overall performance of the previous model with the proposed model. The summary of performance performance in this study is the average value of the ML and DL classification results along with both small and large datasets. Based on the graph in Figure 6, it can be seen that the proposed scheme (TWTFPOS-IDF) consistently outperforms the previous research scheme model. The accuracy

difference with the nearest TF-IDF modification model in dataset 1 is 0.048, and the k-fold value is 0.039. The accuracy difference with the nearest TF-IDF modification model in dataset 2 with a value of 0.010 and a k-fold value of 0.005. In dataset 1, only TF-IDF has a higher k-fold value than the accuracy value. In dataset 2, only ETF-IDF has a lower k-fold value than the accuracy value. The proposed model (TWTFPOS-IDF) has the highest difference between accuracy and k-fold of 0.019.

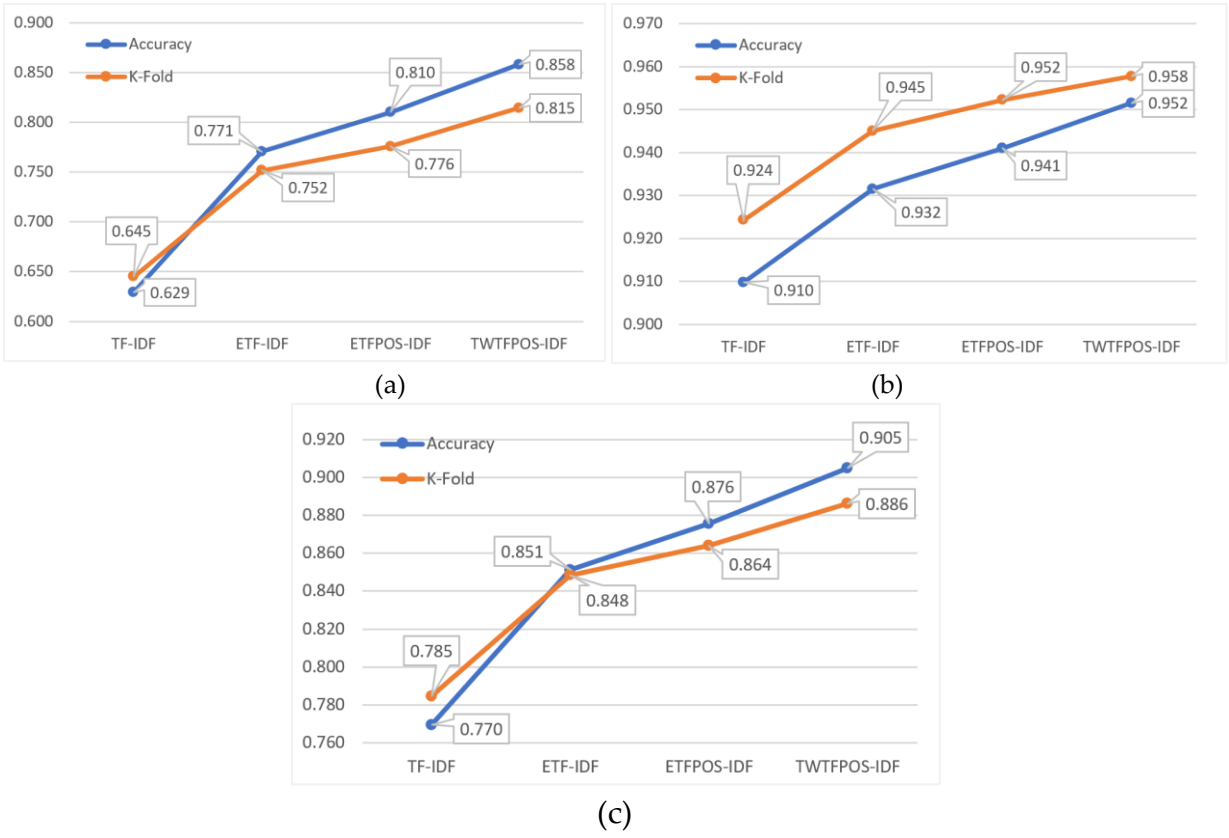

FIGURE 6. Performance of each of the schemes (a) Yahya, (b) Dineshsheelam, and (c) Average.

7. DISCUSSION

Based on the results of the experiment using two classifications of ML and DL with two datasets, it was found that the proposed model (TWTFPOS-IDF) was superior to both datasets, with Dataset 1 showing a larger performance difference compared to Dataset 2. The higher distribution difference in Dataset 1 suggests that the proposed model excels in BT (Behavioral Thematic) modeling, particularly when the data distribution is more varied. This could be attributed to the model's ability to better capture thematic word meanings and optimize feature extraction and selection, which are critical for handling diverse data distributions. In Dataset 2, all models exhibited improved BT performance, likely due to the more balanced data distribution, as supported by comparative research conducted by Althnian and Durden [82, 83]. However, despite the performance improvements in other models, none were able to surpass the proposed TWTFPOS-IDF scheme.

The average performance results, as shown in Figure 6(c), demonstrate that the proposed thematic model outperforms the evaluation averages from previous studies [25-27]. This improvement aligns with the model's design, which was specifically developed to address classification challenges in BT [11-12]. The model's focus on thematic word meaning, combined with enhanced feature extraction and tuning, contributes to its superior performance.

In Dataset 1, the proposed model performed best with the DL-based ANN algorithm, while in Dataset 2, ML algorithms, particularly SVM, showed superior performance despite improvements in DL. On average, the SVM classification algorithm achieved the best results across both datasets. The proposed SVM model consistently outperformed previous models, which only achieved the highest results in specific datasets. Other models demonstrated stable performance across tested algorithms but were unable to surpass the proposed model. The inclusion of k-fold evaluation further confirmed the stability and robustness of the proposed algorithm, consistently outperforming average ML and DL classification algorithms.

Figure 6 highlights that the proposed model's optimization of thematic word identification, feature extraction, and selection tuning leads to higher performance and stability compared to previous research schemes. The model's accuracy and stability are further validated by improved k-fold values. The proposed model's performance is consistent across both small and large datasets, with even better results observed in larger datasets. Statistical significance tests using ANOVA, with high F-values and very low P-values across all metrics (precision, recall, f1-score, accuracy, and k-fold), confirm the model's significant performance improvements.

However, there are limitations to consider. While the DL-based MLP algorithm demonstrated good performance stability and high-performance differences in both datasets, the ML-based SVM algorithm outperformed DL in certain cases. This suggests that DL may require more data to achieve optimal performance, particularly in scenarios where data distribution is less balanced. Additionally, the computational complexity of the proposed model, especially when applied to larger datasets, could pose practical challenges in real-world applications where resources are limited. Future research should address these limitations by exploring ways to reduce computational overhead and improve DL performance with smaller datasets. Furthermore, the model's reliance on thematic word extraction may limit its applicability in domains where thematic analysis is less relevant or where data is highly unstructured.

In conclusion, while the proposed TWTFPOS-IDF model demonstrates significant advancements in BT modeling, its practical implementation may face challenges related to data requirements, computational resources, and domain applicability. Future work should focus on addressing these limitations to enhance the model's versatility and scalability.

V. CONCLUSION

This study proposes a novel weighting modification scheme in TF-IDF, termed TWTFPOS-IDF, for BT-based question classification. The proposed thematic model (TWTFPOS-IDF) distinguishes between different types of words in questions, identifies thematic words, and assigns higher weights to BT thematic words compared to other supporting words. The thematic scheme model is optimized through feature extraction and feature selection tuning. To evaluate the performance of the proposed scheme, two publicly available datasets were used. Machine learning (ML) and deep learning (DL) models were employed to assess the effectiveness, stability (using k-fold cross-validation), and significance (using ANOVA) of the proposed scheme. The k-fold evaluation metrics demonstrate that the proposed model achieves highly stable results compared to other models. Additionally, statistical tests reveal significant differences between the proposed scheme and baseline models.

These results indicate that the proposed approach can effectively identify the meaning of BT in question classification, performing well on both small and large datasets with significant differences in evaluation metrics. However, this study has certain limitations, such as reliance on manual weight adjustments and the lack of exploration into potential biases or ethical considerations in thematic term weighting. Future research should focus on automating weight determination, integrating advanced deep learning techniques (e.g., transformer-based models), and addressing biases and ethical implications in automated classification systems. Additionally, practical implementation strategies and robustness testing on diverse datasets should be explored to enhance scalability and generalizability.

Funding Statement

This research was funded by Directorate General of Higher Education, Research and Technology, Ministry of Education, Culture, Research and Technology of the Republic of Indonesia, grant number 0667/E5/AL.04 /2024.

Author Contributions

All authors made an equal contribution to the development and planning of the study.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data are available from the authors upon request.

Acknowledgments

Not applicable

REFERENCES

1. Jansen, T., & Möller, J. (**2022**). Teacher judgments in school exams: Influences of students' lower-order-thinking skills on the assessment of students' higher-order-thinking skills. *Teaching and Teacher Education, 111*, 103616.
2. Glaesser, J. (**2019**). Competence in educational theory and practice: A critical discussion. *Oxford Review of Education, 45*(1), 70–85.
3. Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (**2021**). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal of Armed Forces India, 77*, S85–S89.
4. Tomlinson, C. A., & Jarvis, J. M. (**2023**). Differentiation: Making curriculum work for all students through responsive planning & instruction. In *Systems and models for developing programs for the gifted and talented* (2nd ed., pp. 599–628).
5. Chiu, T. K. F., Meng, H., Chai, C. S., King, I., Wong, S., & Yam, Y. (**2022**). Creation and evaluation of a pretertiary artificial intelligence (AI) curriculum. *IEEE Transactions on Education, 65*(1), 30–39.
6. Lavidas, K., et al. (**2024**). Determinants of humanities and social sciences students' intentions to use artificial intelligence applications for academic purposes. *Information, 15*(6), 314.
7. Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (**2023**). Bloom's taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies, 28*(12), 15893–15914.
8. Awouda, A., Traini, E., Asranov, M., & Chiabert, P. (**2024**). Bloom's IoT taxonomy towards an effective Industry 4.0 education: Case study on Open-source IoT laboratory. *Education and Information Technologies*, 1–23.
9. West, J. (**2023**). Utilizing Bloom's taxonomy and authentic learning principles to promote preservice teachers' pedagogical content knowledge. *Social Sciences & Humanities Open, 8*(1), 100620.
10. Goh, T. T., Jamaludin, N. A. A., Mohamed, H., Ismail, M. N., & Chua, H. S. (**2022**). A comparative study on part-of-speech taggers' performance on examination questions classification according to Bloom's taxonomy. *Journal of Physics: Conference Series, 2224*(1), 012001.
11. West, J. (**2023**). Utilizing Bloom's taxonomy and authentic learning principles to promote preservice teachers' pedagogical content knowledge. *Social Sciences & Humanities Open, 8*(1), 100620.
12. Waite, L. H., Zupec, J. F., Quinn, D. H., & Poon, C. Y. (**2020**). Revised Bloom's taxonomy as a mentoring framework for successful promotion. *Currents in Pharmacy Teaching and Learning, 12*(11), 1379–1382.
13. Callaghan-Koru, J. A., & Aqil, A. R. (**2022**). Theory-informed course design: Applications of Bloom's taxonomy in undergraduate public health courses. *Pedagogy in Health Promotion, 8*(1), 75–83.
14. Muhayimana, T., Kwizera, L., & Nyirahabimana, M. R. (**2022**). Using Bloom's taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools. *Curriculum Perspectives, 42*(1), 51–63.
15. Lavidas, K., Papadakis, S., Manesis, D., Grigoriadou, A. S., & Gialamas, V. (**2022**). The effects of social desirability on students' self-reports in two social contexts: Lectures vs. lectures and lab classes. *Information, 13*(10), 491.
16. Makhlouf, K., Amouri, L., Chaabane, N., & El-Haggar, N. (**2020**). Exam questions classification based on Bloom's taxonomy: Approaches and techniques. *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*.
17. Masapanta-Carrión, S., & Velázquez-Iturbide, J. Á. (**2019**). Evaluating instructors' classification of programming exercises using the revised Bloom's taxonomy. *Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE)*, 541–547.
18. Sucipto, S., Prasetya, D. D., & Widiyaningtyas, T. (**2024**). A review questions classification based on Bloom taxonomy using a data mining approach. *ITEGAM-JETIA, 10*(48), 161–170.
19. Silva, V. A., Bittencourt, I. I., & Maldonado, J. C. (**2019**). Automatic question classifiers: A systematic review. *IEEE Transactions on Learning Technologies, 12*(4), 485–502.

20. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (**2021**). Deep learning-based text classification. *ACM Computing Surveys, 54*(3).
21. Selva Birunda, S., & Kanniga Devi, R. (**2021**). A review on word embedding techniques for text classification. *Lecture Notes on Data Engineering and Communications Technologies, 59*, 267–281.
22. Huang, J., et al. (**2021**). Automatic classroom question classification based on Bloom's taxonomy. *ACM International Conference Proceeding Series*, 33–39.
23. Li, Q., et al. (**2022**). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology, 13*(2), 31.
24. Wang, X., et al. (**2022**). Comparisons of deep learning and machine learning while using text mining methods to identify suicide attempts of patients with mood disorders. *Journal of Affective Disorders, 317*, 107–113.
25. Yahya, A. A., Toukal, Z., & Osman, A. (**2012**). Bloom's taxonomy-based classification for item bank questions using support vector machines. *Studies in Computational Intelligence, 431*, 135–140.
26. Mohammed, M., & Omar, N. (**2018**). Question classification based on Bloom's Taxonomy using enhanced TF-IDF. *International Journal of Advanced Science, Engineering and Information Technology, 8*(4–2).
27. Gani, M. O., Ayyasamy, R. K., Alhashmi, S. M., Sangodiah, A., & Fui, Y. T. (**2022**). ETFPOS-IDF: A novel term weighting scheme for examination question classification based on Bloom's Taxonomy. *IEEE Access, 10*, 132777–132785.
28. Sharma, H., Mathur, R., Chintala, T., Dhanalakshmi, S., & Senthil, R. (**2023**). An effective deep learning pipeline for improved question classification into Bloom's taxonomy's domains. *Education and Information Technologies, 28*(5).
29. Aninditya, A., Hasibuan, M. A., & Sutoyo, E. (**2019**). Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of Bloom's taxonomy. *Proceedings - 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 112–117.
30. Mohammedid, M., & Omar, N. (**2020**). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLoS ONE, 15*(3), e0230442.
31. Alammary, A. S. (**2021**). Arabic questions classification using modified TF-IDF. *IEEE Access, 9*, 95109–95122.
32. Liang, M., & Niu, T. (**2022**). Research on text classification techniques based on improved TF-IDF algorithm and LSTM inputs. *Procedia Computer Science, 208*, 460–470.
33. Kavi Priya, S., & Pon Karthika, K. (**2023**). An embedded feature selection approach for depression classification using short text sequences. *Applied Soft Computing, 147*, 110828.
34. Okkalioglu, M. (**2023**). TF-IGM revisited: Imbalance text classification with relative imbalance ratio. *Expert Systems with Applications, 217*, 119578.
35. Li, Q., Zhao, S., He, T., & Wen, J. (**2024**). A simple and efficient filter feature selection method via document-term matrix unitization. *Pattern Recognition Letters, 181*, 23–29.
36. Gani, M. O., Ayyasamy, R. K., Fui, T., & Sangodiah, A. (**2022**). USTW vs. STW: A comparative analysis for exam question classification based on Bloom's Taxonomy. *Mendel, 28*(2).
37. Tong, G., Shao, W., & Li, Y. (**2024**). ReverseGAN: An intelligent reverse generative adversarial networks system for complex image captioning generation. *Displays, 82*, 102653.
38. Mikko, M., Stein, Ø., & Jaakko, S. (**2022**). Machine learning and the identification of Smart Specialisation thematic networks in Arctic Scandinavia. *Regional Studies, 56*(9), 1429–1441.
39. Patel, D., & Chhinkaniwala, H. (**2018**). Fuzzy logic-based single document summarisation with improved sentence scoring technique. *International Journal of Knowledge Engineering and Data Mining, 5*(1/2), 125.
40. Widyassari, A. P., Noersasongko, E., Syukur, A., & Affandy. (**2022**). An extractive text summarization based on candidate summary sentences using fuzzy-decision tree. *International Journal of Advanced Computer Science and Applications, 13*(7).
41. Gupta, P., Nigam, S., & Singh, R. (**2023**). Automatic extractive text summarization using multiple linguistic features. *ACM Transactions on Asian and Low-Resource Language Information Processing*.
42. Oyebode, O., Alqahtani, F., & Orji, R. (**2020**). Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. *IEEE Access, 8*, 111141–111158.
43. Waseemullah, et al. (**2022**). A novel approach for semantic extractive text summarization. *Applied Sciences, 12*(9), 4479.
44. Zhou, H., Yip, W. S., Ren, J., & To, S. (**2022**). Thematic analysis of sustainable ultra-precision machining by using text mining and unsupervised learning method. *Journal of Manufacturing Systems, 62*, 218–233.
45. Wang, T., Cai, Y., Leung, H. F., Lau, R. Y. K., Xie, H., & Li, Q. (**2021**). On entropy-based term weighting schemes for text categorization. *Knowledge and Information Systems, 63*(9).
46. Listiowarni, I., & Dewi, N. P. (**2020**). Pemanfaatan klasifikasi soal biologi cognitive domain Bloom's taxonomy menggunakan KNN chi-square sebagai penyusunan naskah soal. *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi, 11*(2), 185–195.
47. Haris, S. S., & Omar, N. (**2015**). Bloom's taxonomy question categorization using rules and N-gram approach. *Journal of Theoretical and Applied Information Technology, 76*(3).

48. Sangodiah, A., Fui, Y. T., Heng, L. E., Jalil, N. A., Ayyasamy, R. K., & Meian, K. H. (**2021**). A comparative analysis on term weighting in exam question classification. *5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 199–206.
49. Razzaghnoori, M., Sajedi, H., & Jazani, I. K. (**2018**). Question classification in Persian using word vectors and frequencies. *Cognitive Systems Research, 47*, 16–27.
50. Chotirat, S., Meesad, P., & Unger, H. (**2022**). Question classification from Thai sentences by considering word context to question generation. *2022 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C)*, 9–14.
51. Antonio, T., & Paramita, A. S. (**2015**). Feature selection technique impact for internet traffic classification using Naïve Bayesian. *Jurnal Teknologi, 72*(5), 141–145.
52. Sucipto, S., Kusrini, K., & Taufiq, E. L. (**2016**). Classification method of multi-class on C4.5 algorithm for fish diseases. *2nd International Conference on Science in Information Technology (ICSITech)*.
53. Salma, F. S., Pratiwi, O. N., & Farifah, R. Y. (**2022**). Classification of high school history questions based on cognitive level revised Bloom's taxonomy using K-nearest neighbor method. *International Conference Advancement in Data Science, E-Learning and Information Systems (ICADEIS)*.
54. Gupta, V., & Rattan, P. (**2023**). Improving Twitter sentiment analysis efficiency with SVM-PSO classification and EFWS heuristic. *Procedia Computer Science, 230*, 698–715.
55. Jamil, F., & Hameed, I. A. (**2023**). Toward intelligent open-ended questions evaluation based on predictive optimization. *Expert Systems with Applications, 231*, 120640.
56. Ifham, M., Banujan, K., Kumara, B. T. G. S., & Wijeratne, P. M. A. K. (**2022**). Automatic classification of questions based on Bloom's taxonomy using artificial neural network. *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 311–315.
57. Thomas, B., & Chandra, J. (**2020**). Random forest application on cognitive level classification of e-learning content. *International Journal of Electrical and Computer Engineering, 10*(4), 4372–4380.
58. Sheelam, D. (**2024**). Blooms data set. *Kaggle*. Retrieved April 3, 2024, from https://www.kaggle.com/datasets/dineshsheelam/blooms-data-set
59. Jayakodi, K., Bandara, M., Perera, I., & Meedeniya, D. (**2016**). WordNet and cosine similarity based classifier of exam questions using Bloom's taxonomy. *International Journal of Emerging Technologies in Learning, 11*(4).
60. Goh, T. T., Jamaludin, N. A. A., Mohamed, H., Ismail, M. N., & Chua, H. (**2023**). Semantic similarity analysis for examination questions classification using WordNet. *Applied Sciences, 13*(14), 8323.
61. Widyassari, A. P., et al. (**2022**). Review of automatic text summarization techniques & methods.
62. Sudarma, M., Sulaksono, J., Informasi, R., & Intensif, T.-I. (**2020**). Implementation of TF-IDF algorithm to detect human eye factors affecting the health service system. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 4*(1), 123–130.
63. Mahan, F., Mohammadzad, M., Rozekhani, S. M., & Pedrycz, W. (**2021**). Chi-MFlexDT: Chi-square-based multi flexible fuzzy decision tree for data stream classification. *Applied Soft Computing, 105*, 107301.
64. Wadud, M. A. H., Kabir, M. M., Mridha, M. F., Ali, M. A., Hamid, M. A., & Monowar, M. M. (**2022**). How can we manage offensive text in social media - A text classification approach using LSTM-BOOST. *International Journal of Information Management Data Insights, 2*(2), 100095.
65. Singh, K. N., Devi, S. D., Devi, H. M., & Mahanta, A. K. (**2022**). A novel approach for dimension reduction using word embedding: An enhanced text classification approach. *International Journal of Information Management Data Insights, 2*(1), 100061.
66. Callista, A. S., Pratiwi, O. N., & Sutoyo, E. (**2021**). Questions classification based on revised Bloom's taxonomy cognitive level using Naive Bayes and Support Vector Machine. *Proceedings of the 4th International Conference on Computer and Informatics Engineering (IC2IE 2021)*, 260–265.
67. Khurana, A., & Verma, O. P. (**2023**). Optimal feature selection for imbalanced text classification. *IEEE Transactions on Artificial Intelligence, 4*(1), 135–147.
68. Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (**2021**). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access, 9*, 78621–78634.
69. Gunawan, I., Widyaningtyas, T., Wibawa, A. P., Haviluddin, Darusalam, D., & Pranolo, A. (**2018**). The performance of correlation-based support vector machine in illiteracy dataset. *Proceedings of the 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT 2018)*, 96–99.
70. Meissner, R., Jenatschke, D., & Thor, A. (**2021**). Evaluation of approaches for automatic e-assessment item annotation with levels of Bloom's taxonomy. *Lecture Notes in Computer Science (LNCS), 12511*, 57–69.
71. Sucipto, S., Prasetya, D. D., & Widiyaningtyas, T. (**2024**). Educational data mining: Multiple choice question classification in vocational school. *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, 23*(2), 367–376.
72. Prasetya, D. D., Wibawa, A. P., & Hirashima, T. (**2018**). The performance of text similarity algorithms. *International Journal of Advances in Intelligent Informatics, 4*(1), 63–69.
73. Sangodiah, A., San, T. J., Fui, Y. T., Heng, L. E., Ayyasamy, R. K., & Jalil, N. B. A. (**2022**). Identifying optimal baseline variant of unsupervised term weighting in question classification based on Bloom taxonomy. *Mendel, 28*(1).

74. Wang, P., et al. (**2020**). Classification of proactive personality: Text mining based on Weibo text and short-answer questions text. *IEEE Access*, *8*, 97370–97382.
75. Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (**2019**). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38.
76. Hasmawati, Romadhony, A., & Abdurohman, R. (**2022**). Primary and high school question classification based on Bloom's taxonomy. *Proceedings of the 10th International Conference on Information and Communication Technology (ICoICT 2022)*, 234–239.
77. Saifudin, I., & Widiyaningtyas, T. (**2024**). Systematic literature review on recommender system: Approach, problem, evaluation techniques, datasets. *IEEE Access*, 1–1.
78. Ilmawan, L. B., Muladi, M., & Prasetya, D. D. (**2023**). Feature space augmentation for negation handling on sentiment analysis. *ILKOM Jurnal Ilmiah*, *15*(2), 353–357.
79. Zhang, J., Wong, C., Giacaman, N., & Luxton-Reilly, A. (**2021**). Automated classification of computing education questions using Bloom's taxonomy. *ACM International Conference Proceeding Series*, 58–65.
80. Wong, T. T., & Yang, N. Y. (**2017**). Dependency analysis of accuracy estimates in k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, *29*(11), 2417–2427.
81. Zhang, X., & Liu, C. A. (**2023**). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics*, *235*(1), 280–301.
82. Althnian, A., et al. (**2021**). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, *11*(2), 796.
83. Durden, J. M., Hosking, B., Bett, B. J., Cline, D., & Ruhl, H. A. (**2021**). Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance. *Progress in Oceanography*, *196*, 102612.