JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Transformer-Based Support for Content-Validity Pre-Screening in Educational Materials

Safuan [1], Dhendra Marutho [2,*], Ahmad Ilham [2], Muhammad Munsarif [3], Wendy Sarasjati [2], Edy Winarno [2], Arnold Adimabua Ojugo [4], and De Rosal Ignatius Moses Setiadi [5]

[1] Department of Artificial Intelligence, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : safuan@unimus.ac.id;
[2] Master's Program in Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : dhendra@unimus.ac.id; ahmadilham@unimus.ac.id; wendysar@unimus.ac.id; edywin@unimus.ac.id
[3] Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia; e-mail : m.munsarif@unimus.ac.id
[4] Department of Computer Science, Federal University of Petroleum Resources, Effurun330102, Nigeria; e-mail : ojugo.arnold@fupre.edu.ng
[5] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id
* Corresponding Author : Dhendra Marutho

Abstract: Content validity assessment is essential for determining whether educational materials adequately represent intended learning outcomes. However, conventional assessment procedures require substantial expert time and may produce inconsistent decisions across large item collections. This study develops a transformer-based framework to support content-validity pre-screening through two complementary tasks: predicting expert-derived Aiken's V coefficients and classifying instructional-item essentiality. The final dataset comprised 652 Indonesian-language educational text items independently evaluated by four subject-matter experts. To reduce information leakage, identical and normalized-equivalent texts were grouped before applying a group-aware 70:15:15 training–validation–test split. Classical TF-IDF-based baselines were compared with IndoBERT, multilingual BERT, XLM-RoBERTa, and multilingual DeBERTa-v3. For Aiken's V regression, multilingual BERT achieved the lowest MAE of 0.0501, the lowest RMSE of 0.0625, and the highest R^2 of 0.5239, whereas multilingual DeBERTa-v3 achieved the highest Spearman correlation of 0.7532. For essentiality classification, XLM-RoBERTa achieved the highest accuracy of 0.8557 and Macro-F1 of 0.8161, whereas multilingual BERT achieved the highest balanced accuracy of 0.8135 and ROC-AUC of 0.9111. Error analysis showed that the models captured textual patterns associated with expert-derived outcomes but remained limited when judgments depended on broader curricular context, competency hierarchies, prerequisite relationships, or relationships among instructional items. The findings support the use of transformer models as human-in-the-loop decision-support tools for prioritizing uncertain or potentially problematic educational items. However, the framework should be interpreted as a pre-screening mechanism rather than a replacement for expert judgment, and external validation across institutions and disciplines remains necessary.

Keywords: Aiken's V; Content Validity; Educational Materials; Educational NLP; Essentiality Classification; Expert Annotation; Group-Aware Evaluation; Transformer Models.

Received: June, 26th 2026
Revised: August, 16th 2026
Accepted: August, 17th 2026
Published: August, 31st 2026

1. Introduction

In outcome-based education, instructional materials must adequately represent the competencies and learning objectives that students are expected to achieve [1]–[3]. Content validity is therefore an important component of curriculum development because it provides evidence that learning outcomes, instructional content, assessment items, and related educational documents are relevant to the intended domain [4]. Content validity is commonly assessed by subject-matter experts who evaluate individual items in terms of relevance, clarity,

representativeness, or essentiality [5]. Their judgments can then be summarized using quantitative indicators such as Aiken's V for graded relevance ratings and the Content Validity Ratio (CVR) for expert essentiality decisions [6]–[8].

Although expert-based validation provides an established basis for evaluating instructional content, its practical application can be difficult to scale. Each item must be examined by qualified experts, while the availability of evaluators may be limited when assessment requires specific disciplinary knowledge [9]–[11]. Differences in interpretation may also lead to variation in ratings even when the same evaluation criteria are applied [12]. These challenges become more pronounced when institutions manage large collections of curriculum documents or conduct frequent curriculum revisions. Thus, the practical challenge is not the validity of expert assessment itself, but the time and resources required to apply expert review consistently across large numbers of instructional items. Computational methods may help reduce this initial screening burden; however, they should not be interpreted as replacements for formal expert validation. Content-validity judgments depend on disciplinary knowledge, curriculum structure, and professional interpretation, which may not be fully observable from the wording of an isolated instructional item.

Natural language processing (NLP) offers a potential mechanism for addressing this scalability problem. Transformer-based models can represent contextual and semantic relationships and have been applied to text classification, regression, semantic similarity, and information extraction [13]–[15]. In educational settings, NLP has been used for automated essay scoring, question classification, feedback generation, learning-resource recommendation, and curriculum alignment [5], [16]–[20]. However, these applications do not directly address the methodological problem considered in this study: predicting expert-derived content-validity outcomes from instructional text. Automated essay scoring generally models learner-generated responses, whereas curriculum-alignment methods typically examine relationships between two or more textual components. In contrast, the present task takes a single instructional statement as input and predicts an aggregated outcome derived from multiple expert judgments. This formulation introduces several methodological challenges, including inter-rater disagreement, restricted target distributions, and judgments influenced by contextual factors that may not be fully represented in the text. Consequently, strong performance on conventional educational NLP tasks does not necessarily imply that the same models can reliably approximate expert-derived content-validity judgments.

A further methodological challenge concerns repeated instructional statements. Learning outcomes, content descriptions, and competency formulations may be reused across courses or curriculum documents. Under conventional random partitioning, identical or normalized-equivalent items may appear in both the training and test sets. In such cases, a model may achieve overly optimistic performance because it has encountered the same or effectively identical textual formulation during training. This issue is particularly important for the present task because the objective is to estimate performance on previously unseen instructional items rather than to recognize previously observed formulations. Reliable evaluation therefore requires a partitioning strategy that keeps repeated text groups within a single experimental subset.

Against this background, this study investigates whether classical lexical models and pretrained transformer models can approximate expert-derived content-validity outcomes from instructional text under a leakage-controlled evaluation protocol. Two complementary tasks are examined. The first predicts item-level Aiken's V scores as a regression problem, while the second classifies instructional items as essential or non-essential. The study compares TF-IDF-based baselines with IndoBERT, multilingual BERT, XLM-RoBERTa, and multilingual DeBERTa-v3 using a group-aware training, validation, and held-out test split. The following research questions guide the study:

- To what extent can classical and transformer-based models predict expert-derived Aiken's V scores from educational text?
- To what extent can these models classify educational items as essential or non-essential?
- How do Indonesian-specific and multilingual transformer models compare with TF-IDF-based approaches under a leakage-controlled evaluation protocol?
- What limitations and prediction errors arise when these models are used as preliminary support tools for educational content assessment?

The contribution of this study does not lie in proposing a new transformer architecture. Instead, it lies in formulating expert-derived relevance estimation and essentiality prediction as supervised educational NLP tasks, verifying the reliability of the expert-derived targets, and applying group-aware partitioning to prevent leakage from repeated instructional texts. The study also provides a consistent comparison of naïve baselines, TF-IDF-based models, an Indonesian-specific transformer, and multilingual transformer architectures. Beyond aggregate performance, item-level errors are examined to identify conditions under which textual models agree or disagree with expert-derived outcomes. The resulting framework is positioned as a pre-screening and decision-support mechanism, while final content-validity decisions remain the responsibility of subject-matter experts. The remainder of this paper is organized as follows. Section 2 reviews content-validity assessment and relevant educational NLP studies. Section 3 describes the dataset, expert annotation, group-aware partitioning, models, and evaluation protocol. Section 4 presents the regression and classification results, error analysis, comparative implications, and limitations. Section 5 concludes the study and outlines directions for future research.

2. Related Work

2.1. Content Validity in Educational Instrument Development

Content validity concerns the extent to which educational materials or assessment items adequately represent their intended learning objectives and disciplinary domains [1], [2], [4]. It is commonly established through structured evaluations by subject-matter experts, who assess dimensions such as relevance, representativeness, clarity, and essentiality [5]. Aiken's V is commonly used to summarize ordinal relevance ratings, whereas the Content Validity Ratio (CVR) is traditionally used to quantify experts' judgments regarding whether an item is essential [6]–[8]. These indicators have been applied in the validation of educational modules, questionnaires, rubrics, and assessment instruments.

Although expert-based validation provides an established methodological basis for evaluating instructional content, its practical implementation can be difficult to scale. The process depends on the availability of qualified experts, coordination among evaluators, and repeated assessment when educational materials are revised [9]–[12]. Differences in expertise and interpretation may also produce disagreement among raters. Computational models could therefore support the initial screening process by prioritizing items that may require further examination. Such predictions, however, should be interpreted as preliminary decision support rather than substitutes for formal expert validation.

In the present study, Aiken's V is treated as a continuous prediction target for relevance estimation, while essentiality is formulated as a binary classification target derived from expert judgments. The latter should be distinguished from directly predicting a CVR value: although CVR is a conventional measure for summarizing essentiality judgments, the present classification task predicts the underlying expert-derived essentiality label. This distinction allows the study to examine whether textual information alone can approximate two complementary aspects of expert-based content validity while retaining expert judgment as the basis for final assessment.

2.2. NLP and Transformer-Based Models in Educational Evaluation

Natural language processing (NLP) has increasingly been used to analyze educational text and support academic decision-making [15]. Early approaches commonly relied on lexical representations such as bag-of-words and term frequency-inverse document frequency (TF-IDF), which convert textual documents into numerical representations for comparison or use as inputs to machine-learning algorithms. TF-IDF has been applied to measure similarity between educational documents, including the alignment between course descriptions and learning outcomes [19]. Cosine similarity and related methods have also been used to compare curriculum components and identify overlapping or insufficiently aligned statements [21]. These approaches are computationally efficient and provide interpretable lexical baselines.

However, lexical representations primarily capture word occurrence rather than contextual meaning. Educational statements may express similar competencies using different vocabulary, while identical terms may convey different meanings depending on their context. This limitation is particularly relevant in curriculum documents, where technical terminology,

abbreviations, bilingual expressions, and variations in sentence structure are common. More recent educational NLP research has therefore adopted contextual representations for tasks such as automated essay scoring [22], question classification [23], feedback generation [24], learning-resource recommendation [18], and learning-outcome alignment [21]. Recent work has also explored cognitive-level educational question answering using transformer-based and large language model approaches, including structured reasoning mechanisms designed to align educational responses with Bloom's Taxonomy [25]. These studies demonstrate the potential of language models to capture semantic and contextual patterns that are difficult to represent using purely lexical features. Beyond educational applications, transformer-based sentence representations and attention mechanisms have also been applied to aspect-based sentiment analysis, aspect-term extraction, and sentiment classification [26], [27].

Nevertheless, these applications primarily address learner-generated content, semantic similarity, document alignment, or educational question answering rather than the prediction of expert-derived content-validity indicators. Automated essay scoring, for example, evaluates learner writing based on characteristics such as coherence, grammar, organization, and relevance [22]. Curriculum-alignment studies typically estimate the semantic relationship between two text components, such as a course learning outcome and a program learning outcome [21]. Similarly, cognitive-level question-answering approaches focus on generating or evaluating responses according to predefined cognitive requirements [28]. The present study addresses a different prediction setting: the model receives a single instructional text item and predicts an aggregated outcome derived from multiple expert judgments. This requires the model to learn textual patterns associated with perceived relevance and essentiality rather than direct correspondence between paired documents or predefined cognitive categories.

This distinction is also important because specialized educational texts may contain domain-specific terminology and discourse patterns. Evidence from other specialized domains indicates that general-purpose transformer models may require domain-specific pretraining or adaptation when applied to such texts [29]. Accordingly, the present study compares an Indonesian-specific transformer with multilingual transformer models and classical lexical baselines to examine whether contextual representations provide an advantage for predicting expert-derived validity outcomes. The comparison is conducted under a leakage-controlled evaluation protocol because repeated instructional statements may otherwise lead to overly optimistic estimates of predictive performance.

2.3. Research Gap and Positioning of the Present Study

Table 1 summarizes the methodological differences between representative studies in educational assessment and NLP and the present study. Previous research has demonstrated the use of expert-based content-validity assessment, automated essay scoring, instructional-text similarity, and curriculum alignment [5], [18], [19], [21], [22]. These studies establish the relevance of computational methods and expert-derived indicators in educational assessment; however, they address different prediction objectives and do not directly investigate whether aggregated expert judgments can be predicted from individual instructional text items. The present study therefore focuses on the prediction of expert-derived content-validity indicators under a leakage-controlled evaluation protocol.

The studies summarized in Table 1 illustrate several ways in which computational methods can support educational assessment. Kania et al. [20], for example, demonstrate the use of Aiken's V to quantify expert judgments in content-validity assessment, but their study does not examine whether such expert-derived scores can be inferred from the corresponding item text. Similarly, Yang et al. [23] and Amalia et al. [25] use textual information to approximate human assessment, but their tasks focus on learner-generated essay responses. Zaki et al. [4] address a different problem by using NLP to assess the semantic relationship between course and program learning outcomes. These studies are therefore relevant to the broader use of computational methods in educational assessment, but their objectives are not directly equivalent to predicting expert-derived content-validity outcomes from individual instructional statements.

Three methodological gaps motivate the present study. First, limited evidence is available on the direct prediction of aggregated expert judgments from instructional text, particularly for continuous relevance scores such as Aiken's V and binary essentiality judgments. Second, the comparative behavior of classical lexical representations, Indonesian-specific transformers, and multilingual transformer models for this prediction setting remains insufficiently

examined. Third, evaluation leakage caused by repeated or normalized-equivalent instructional statements has received limited attention, even though such repetitions may occur across courses and curriculum documents and can produce overly optimistic performance estimates under conventional random partitioning.

Table 1. Methodological comparison with representative previous studies

Study	Main task	Method	Key outcome	Relation to the present study
Kania et al. [22]	Content-validity assessment of a mathematical problem-solving instrument using five experts	Aiken's V	Aiken's V coefficients ranged from 0.817 to 0.884	Demonstrates expert-based content-validity assessment but does not predict expert-derived scores from item text
Yang et al. [18]	Automated essay scoring using the ASAP dataset	R²BERT with regression and ranking losses	Average QWK = 0.794	Predicts learner-generated writing scores rather than expert-derived content-validity indicators
Amalia et al. [19]	Automated evaluation of Indonesian-language essay responses	Latent Semantic Analysis (LSA) with text preprocessing	Accuracy = 83.3% relative to teachers' assessments	Uses textual information to approximate human assessment but addresses learner-generated responses
Zaki et al. [5]	Mapping course learning outcomes (CLOs) to program learning outcomes (PLOs)	AI-based NLP system	Mapping precision = 83.1% and 88.1% for two educational programs	Addresses semantic alignment between educational components rather than prediction of expert-derived validity indicators
Present study	Aiken's V regression and essentiality classification from instructional text	TF-IDF-based models, IndoBERT, mBERT, XLM-RoBERTa, and multilingual DeBERTa-v3 with group-aware splitting	Predicts expert-derived relevance and essentiality outcomes from instructional text under a leakage-controlled evaluation protocol	Extends previous work by formulating expert-derived relevance and essentiality outcomes as supervised prediction tasks and controlling leakage from repeated instructional texts

The present study addresses these gaps by formulating content-validity pre-screening as two complementary supervised-learning tasks: Aiken's V regression and expert-derived essentiality classification. TF-IDF-based baselines and transformer models are evaluated using the same group-aware data partitions, with normalized duplicate statements assigned to the same group before dataset splitting. This design allows the study to assess the extent to which textual models can approximate expert-derived validity outcomes while providing a more conservative estimate of generalization to previously unseen instructional text.

3. Proposed Method

This study develops and evaluates a natural language processing framework to support the preliminary assessment of educational content validity. The framework consists of two complementary supervised-learning tasks: regression for predicting expert-derived Aiken's V scores and binary classification for predicting whether an instructional item is considered essential. Classical TF-IDF-based models are used as lexical baselines and compared with four pretrained transformer models: IndoBERT, multilingual BERT (mBERT), XLM-RoBERTa, and multilingual DeBERTa-v3 (mDeBERTa-v3).

The methodological framework comprises dataset preparation and quality screening, expert annotation, target construction, duplicate-aware text grouping and data partitioning, model training, and performance evaluation. A key consideration in the experimental design is the presence of identical or normalized-equivalent instructional statements. To reduce potential information leakage, such statements are assigned to the same group before the training, validation, and test sets are constructed. This group-aware design provides a more

conservative evaluation of model performance by preventing repeated textual formulations from being distributed across different data subsets.

As illustrated in Figure 1, the process begins with the collection of Indonesian-language educational text items, followed by data-quality screening to obtain eligible instructional records. The retained items are independently evaluated by four subject-matter experts. Their relevance ratings are aggregated into Aiken's V scores, while their essentiality judgments are converted into binary essentiality labels. The texts are subsequently normalized and grouped according to identical or normalized-equivalent formulations. These groups are then partitioned into training, validation, and test sets using a 70:15:15 group-aware split. The same partitions are used for all classical and transformer-based models to ensure a controlled comparison. Finally, the trained models generate regression and classification predictions, which are evaluated using task-specific performance metrics and item-level error analysis..

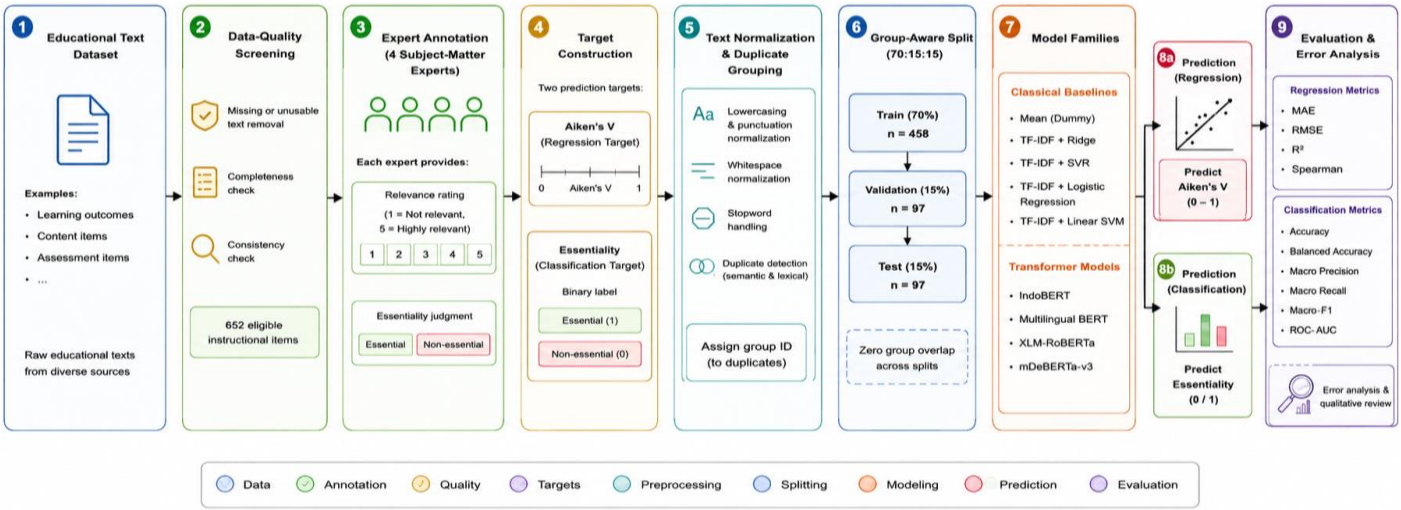

Figure 1. Proposed framework for data-quality screening, expert-derived target construction, duplicate-text grouping, group-aware data partitioning, model training, prediction, and leakage-controlled evaluation.

3.1. Dataset Preparation and Expert Annotation

The initial dataset comprised 777 Indonesian-language instructional records collected from undergraduate computer science course materials, including Web Programming, Artificial Intelligence, and Natural Language Processing. The records represented several types of educational content, including learning outcomes, content descriptions, and assessment-related statements. Each record was treated as an individual unit of analysis because the objective was to predict its expert-derived content-validity outcome directly from its textual representation.

Before expert annotation, a data-quality screening procedure was applied to identify records that did not satisfy the predefined completeness and validity requirements. Records containing unusable or incomplete instructional text were excluded from further analysis. Following this screening, 652 eligible instructional items were retained for expert annotation, target construction, and subsequent model evaluation.

Each eligible item was independently evaluated by four subject-matter experts. The experts assessed the relevance of each instructional item using a five-point ordinal scale, ranging from 1 (not relevant) to 5 (highly relevant). They also independently determined whether each item was essential within its intended instructional context. The evaluations were conducted independently to minimize potential influence among raters, and the ratings were aggregated only after the expert evaluations had been completed.

The resulting expert annotations provided the basis for two prediction targets. Relevance ratings were aggregated into continuous Aiken's V coefficients for the regression task, whereas the expert essentiality judgments were used to derive binary essentiality labels for the classification task. The final dataset contained 441 items classified as essential and 211 items classified as non-essential. The item-level Aiken's V scores had a mean of 0.7049 and a standard deviation of 0.0986, with values ranging from 0.5625 to 0.8750.

The dataset also contained repeated instructional statements. To identify these repetitions without altering the substantive content of the original texts, textual representations were normalized with respect to letter case, whitespace, punctuation, and other non-semantic formatting differences. This process identified 455 unique normalized texts among the 652 eligible items, while the remaining 197 records represented additional occurrences of repeated or normalized-equivalent statements. These records were retained as valid observations because they contained expert-derived outcomes. However, their normalized-text identities were subsequently used to construct groups for leakage-controlled data partitioning, as described in Section 3.3. Table 2 summarizes the characteristics of the dataset after data-quality screening, including the number of eligible items, expert raters, class distribution, normalized-text groups, and descriptive statistics of the Aiken's V scores.

Table 2. Dataset characteristics after data-quality screening.

Characteristic	Value
Initial records	777
Eligible records after screening	652
Number of expert raters	4
Essential items	441
Non-essential items	211
Unique normalized texts	455
Repeated text instances	197
Mean Aiken's V	0.7049
Standard deviation of Aiken's V	0.0986
Minimum Aiken's V	0.5625
Maximum Aiken's V	0.8750

The original instructional texts were retained as model inputs, whereas the normalized representations were used only for duplicate detection and group construction. Thus, normalization did not replace or modify the textual inputs used by the predictive models. The resulting dataset and annotation structure provided a common basis for the subsequent target construction, group-aware partitioning, and comparative evaluation described in the following sections.

3.2. Target Construction

The expert evaluations were transformed into two complementary prediction targets: a continuous Aiken's V coefficient representing item relevance and a binary essentiality label representing whether an instructional item was considered essential by the expert panel. Because these targets capture different aspects of content validity, they were formulated as separate supervised-learning tasks for regression and classification, respectively.

Aiken's V Regression Target

Each expert evaluated the relevance of an instructional item using a five-point ordinal scale, where 1 indicated not relevant and 5 indicated highly relevant. The item-level Aiken's V coefficient was calculated using:

$$V_i = \frac{\sum_{j=1}^{n}(r_{ij} - l)}{n(c-1)} \tag{1}$$

where r_{ij} denotes the relevance rating assigned to item i by expert j, l is the lowest possible rating, c is the number of response categories, and n is the number of expert raters. In this study, $l = 1$, $c = 5$, and $n = 4$. Thus, the coefficient for each instructional item was calculated as:

$$V_i = \frac{\sum_{j=1}^{4}(r_{ij} - 1)}{4(5-1)} \tag{2}$$

Higher Aiken's V values indicate stronger expert agreement regarding the relevance of an instructional item to its intended instructional context. In the final dataset, the coefficients

ranged from 0.5625 to 0.8750, with a mean of 0.7049 and a standard deviation of 0.0986. These item-level coefficients were used as continuous target values for the regression task.

Essentiality Classification Target

In addition to relevance, the four experts independently determined whether each instructional item was essential within its intended instructional context. Their judgments were aggregated using the Content Validity Ratio (CVR), calculated as:

$$CVR_i = \frac{N_{e,i} - \frac{N}{2}}{\frac{N}{2}} \tag{3}$$

where $N_{e,i}$ denotes the number of experts who rated item i as essential and N represents the total number of experts. With four expert raters, the possible CVR values are -1.0, -0.5, 0.0, 0.5, and 1.0. An item was assigned to the essential class when at least three of the four experts rated it as essential, corresponding to $CVR_i > 0$. Otherwise, it was assigned to the non-essential class. The resulting binary target was defined as:

$$y_i = \begin{cases} 1, & CVR_i > 0 \\ 0, & CVR_i \leq 0 \end{cases} \tag{4}$$

where y_i represents an essential item and $y_i = 0$ represents a non-essential item. This procedure resulted in 441 essential items and 211 non-essential items.

In this formulation, CVR is used as an aggregation mechanism for deriving the expert-based essentiality decision rather than as a numerical prediction target. The classification task therefore predicts the resulting expert-derived essentiality label, whereas the regression task directly predicts the continuous Aiken's V coefficient. This distinction ensures that the two tasks remain conceptually and statistically separate while representing complementary dimensions of the expert assessment. The two prediction tasks can consequently be expressed as:

$$\hat{V}_i = f_{\text{reg}}(x_i) \tag{5}$$

$$\hat{y}_i = f_{\text{cls}}(x_i) \tag{6}$$

where x_i denotes the original instructional text of item i, $\hat{V}_i$ denotes the predicted Aiken's V coefficient, and $\hat{y}_i$ denotes the predicted essentiality class. Separate models were trained for the regression and classification tasks; the framework therefore does not employ a joint or multitask output architecture.

3.3. Duplicate-Aware Grouping and Data Partitioning

Repeated or normalized-equivalent instructional statements may occur across courses and curriculum documents. If conventional row-level random partitioning is applied, records containing identical or effectively equivalent text may be distributed across the training, validation, and test sets. In such cases, a model may achieve artificially optimistic performance by recognizing textual formulations encountered during training rather than learning relationships that generalize to previously unseen instructional statements.

To reduce this risk, duplicate grouping was performed before data partitioning. Each instructional statement was converted into a normalized representation by standardizing letter case, whitespace, punctuation, and other non-semantic formatting differences. Each distinct normalized representation was then assigned a unique group identifier. Consequently, records sharing the same normalized text were treated as a single group, regardless of whether their expert-derived target values were identical. Let x_i denote the original instructional text of item i, and let g_i denote its normalized-text group identifier. The normalization and grouping process can be represented as:

$$g_i = \text{GroupID}\big(\text{Normalize}(x_i)\big) \tag{7}$$

where $\text{Normalize}(\cdot)$ applies the predefined text-standardization procedure and $\text{GroupID}(\cdot)$ assigns a unique identifier to each distinct normalized representation. Thus, two items belong to the same group when their normalized representations are identical:

$$g_i = g_j \iff \text{Normalize}(x_i) = \text{Normalize}(x_j) \tag{8}$$

The duplicate analysis identified 455 unique normalized-text groups among the 652 eligible records, with 197 records representing additional occurrences of repeated or normalized-equivalent statements. These records were retained because they constituted valid observations with expert-derived outcomes. However, all records sharing the same normalized-text group were assigned exclusively to one data subset. A group present in the training set therefore could not appear in either the validation or test set. The group-aware partitioning constraint can be expressed as:

$$G_{\text{train}} \cap G_{\text{val}} = G_{\text{train}} \cap G_{\text{test}} = G_{\text{val}} \cap G_{\text{test}} \tag{9}$$

where G_{train}, G_{val}, and G_{test} denote the sets of normalized-text group identifiers assigned to the training, validation, and test subsets, respectively. This constraint ensures zero group overlap across the three partitions.

The dataset was partitioned using a target ratio of 70:15:15 for training, validation, and test data. The training subset was used for model parameter estimation, the validation subset for model selection and early stopping, and the test subset was reserved exclusively for final performance evaluation. Because complete groups rather than individual records were assigned to the subsets, the resulting number of records in each partition could differ slightly from the nominal proportions. The final partition contained 458 training items, 97 validation items, and 97 test items.

The same group-aware partitions were used for both the Aiken's V regression and essentiality-classification tasks and were kept fixed across the TF-IDF-based baselines and transformer models. This controlled partitioning ensured that differences in predictive performance were attributable to the modeling approaches rather than differences in the evaluated samples.

The grouping procedure reduces direct leakage caused by identical or normalized-equivalent instructional statements across data subsets. However, it does not eliminate all forms of semantic overlap. Differently worded statements may still express closely related instructional concepts and can therefore remain similar across partitions. The resulting test performance should consequently be interpreted as a leakage-controlled estimate under normalized exact-match grouping, rather than as evidence that all forms of semantic similarity or curricular overlap have been eliminated.

3.4. Model Training Configuration

Two categories of predictive models were evaluated: classical machine-learning models based on TF-IDF representations and pretrained transformer models. Separate models were trained for the Aiken's V regression and essentiality-classification tasks. All models used the fixed group-aware training, validation, and test partitions described in Section 3.3 to ensure a controlled comparison.

TF-IDF-Based Baselines

Classical models were included to determine whether contextual transformer representations provide measurable advantages over conventional lexical representations. The original instructional texts were converted into TF-IDF vectors using unigram and bigram features. Unigrams capture individual terms, whereas bigrams represent short local word sequences.

For Aiken's V regression, three baselines were considered: a mean predictor, TF-IDF with Ridge regression, and TF-IDF with support vector regression (SVR). The mean predictor provides a naïve reference by assigning the training-set mean Aiken's V to every test item. Ridge regression provides a regularized linear baseline, while SVR evaluates whether a nonlinear margin-based regression approach can improve prediction from lexical features.

For essentiality classification, three baselines were used: a majority-class predictor, TF-IDF with logistic regression, and TF-IDF with a linear support vector machine (SVM). The majority-class predictor assigns every item to the most frequent class, providing a reference for evaluating whether the trained classifiers offer meaningful improvement. Logistic regression estimates the probability of essentiality from the TF-IDF representation, while the linear SVM provides an alternative discriminative classifier. The TF-IDF-based regression and classification formulations can be represented as:

$$\hat{V}_i = \mathbf{w}^T \mathbf{x}_i + b \tag{10}$$

and

$$P(y_i = 1|\mathbf{x}_i) = \sigma(\mathbf{w}^T \mathbf{x}_i + b) \tag{11}$$

where $\mathbf{x}_i$ denotes the TF-IDF representation of instructional item i, $\mathbf{w}$ is the learned parameter vector, b is the intercept, and $\sigma(\cdot)$ is the sigmoid function. The corresponding regularized or margin-based variants were trained using the Ridge, SVR, and linear SVM formulations described above.

Transformer Models

Four pretrained transformer architectures were evaluated:

1. IndoBERT, an Indonesian-specific language model;
2. multilingual BERT (mBERT), a multilingual version of BERT;
3. XLM-RoBERTa, a multilingual transformer trained on cross-lingual corpora; and
4. multilingual DeBERTa-v3 (mDeBERTa-v3), a multilingual model based on the DeBERTa-v3 architecture.

This selection enables comparison between an Indonesian-specific model and multilingual models with different pretraining strategies and representation characteristics. The comparison examines whether language-specific pretraining or broader multilingual representations are more effective for modeling Indonesian instructional text.

Each instructional statement was tokenized using the tokenizer associated with its respective pretrained model. The resulting sequences were padded or truncated to a maximum length of 256 subword tokens. No stemming or stop-word removal was applied because such transformations could discard contextual information used by pretrained encoders. For each task, the encoded representation of the complete instructional sequence was passed to a task-specific prediction head. For Aiken's V regression, the prediction head produced a single continuous output:

$$\hat{V}_i = f_{\text{reg}}(\mathbf{h}_i) \tag{12}$$

where $\mathbf{h}_i$ denotes the contextualized representation of instructional item i produced by the transformer encoder and $f_{\text{reg}}(\cdot)$ denotes the regression head.

For essentiality classification, the prediction head produced logits for the essential and non-essential classes. The probability of class k was calculated using the softmax function:

$$P(y_i = k|x_i) = \frac{\exp(z_{ik})}{\sum_{j=1}^{K} \exp(z_{ij})} \tag{13}$$

where z_{ik} denotes the logit produced for class k, and $K = 2$ is the number of classes. Separate instances of each transformer architecture were fine-tuned for regression and classification. Thus, Aiken's V prediction and essentiality classification were not jointly optimized in a multitask architecture.

Model Training

The regression models were optimized by minimizing mean squared error:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B} \sum_{i=1}^{B} \left(V_i - \hat{V}_i\right)^2 \tag{14}$$

where B is the batch size, V_i is the expert-derived Aiken's V coefficient, and $\hat{V}_i$ is the corresponding model prediction.

The classification models were optimized using cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} y_{ik} \log P(y_i = k|x_i) \tag{15}$$

where y_{ik} is the binary indicator that item i belongs to class k, and $P(y_i = k|x_i)$ is the predicted probability for that class.

Transformer fine-tuning was conducted using the AdamW optimizer. The models were trained for a maximum of 10 epochs with early stopping after two consecutive epochs without improvement in the task-specific validation metric. Model selection was based on validation MAE for regression and validation Macro-F1 for classification. The held-out test set was not used for model selection, early stopping, or parameter optimization.

All transformer experiments were implemented in Python using PyTorch and the Hugging Face Transformers library and were executed in a GPU-enabled Google Colab environment with an NVIDIA L4 GPU. A random seed of 42 was used to improve reproducibility. The training batch size was 8 and the evaluation batch size was 16, with a weight-decay coefficient of 0.01. The same group-aware partitions were maintained across all comparable experiments.

After model selection, only the best validation checkpoint for each model was evaluated on the held-out test set. Separate model instances were trained for the two prediction tasks. Multilingual DeBERTa-v3 produced a valid regression result but exhibited numerical instability during classification training. Its classification run was therefore excluded from the final classification comparison rather than treated as a valid performance estimate.

3.5. Evaluation Metrics

Model performance was evaluated separately for the Aiken's V regression and essentiality-classification tasks using the held-out test set. For regression, four metrics were used: mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination R^2, and Spearman's rank correlation coefficient ρ. For classification, accuracy, balanced accuracy, macro-precision, macro-recall, Macro-F1, and ROC-AUC were used. For Aiken's V regression, MAE measures the average absolute difference between the expert-derived and predicted scores:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|V_i - \hat{V}_i\right| \tag{16}$$

where V_i and $\hat{V}_i$ denote the observed and predicted Aiken's V scores, respectively, and n is the number of test instances. RMSE places greater weight on larger prediction errors:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(V_i - \hat{V}_i\right)^2} \tag{17}$$

The coefficient of determination measures the proportion of variance in the observed Aiken's V scores explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(V_i - \hat{V}_i\right)^2}{\sum_{i=1}^{n}(V_i - \bar{V}_i)^2} \tag{18}$$

where $\bar{V}_i$ is the mean observed Aiken's V score. Spearman's rank correlation coefficient (ρ) evaluates whether the model preserves the relative ordering of instructional items:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{19}$$

where d_i is the difference between the ranks of the observed and predicted values for item i. Lower MAE and RMSE indicate smaller prediction errors, whereas higher R^2 and Spearman's ρ indicate stronger agreement with the expert-derived outcomes. For essentiality classification, accuracy was calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision and recall were calculated for each class as:

$$Precission = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN} \tag{21}$$

Macro-precision, macro-recall, and Macro-F1 were obtained by averaging the corresponding class-specific measures:

$$Macro - F1 = \frac{1}{K}\sum_{k=1}^{K} F1_k, \qquad F1_k = \frac{2P_k R_k}{P_k + R_k} \tag{22}$$

where $K = 2$ represents the essential and non-essential classes, and P_k and R_k denote precision and recall for class k. Balanced accuracy was calculated as the average recall across the two classes:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{23}$$

ROC-AUC measures the ability of the model to rank essential items above non-essential items across classification thresholds and is defined as the area under the receiver operating characteristic curve:

$$ROC - AUC = \int_0^1 TPR(FPR)d(FPR) \tag{24}$$

where TPR and FPR denote the true-positive and false-positive rates, respectively. Higher values indicate better classification performance. Because accuracy alone may obscure class-specific behavior, Macro-F1 and balanced accuracy were emphasized for class-balanced performance assessment, while ROC-AUC was used to evaluate threshold-independent discrimination.

4. Results and Discussion

This section presents the results obtained from the fixed, group-aware held-out test set. All models were trained using the same training subset, selected using the validation subset, and evaluated on the same test subset of 97 instructional items. Because identical and normalized-equivalent texts were assigned exclusively to a single subset, the evaluation reduces direct leakage from repeated instructional statements and provides an estimate of performance on previously unseen normalized-text groups.

The evaluation addressed two complementary prediction tasks. The first treated expert-derived Aiken's V as a continuous regression target, while the second classified instructional items according to the expert-derived essentiality label. Classical TF-IDF-based baselines and pretrained transformer models were evaluated under the same data partitions and evaluation criteria. The results are presented in terms of predictive performance, model differences, and error patterns, followed by their practical implications and limitations for text-based content-validity pre-screening.

4.1. Aiken's V Regression Performance

Table 3 presents the test-set performance of the classical baselines and transformer models for predicting expert-derived Aiken's V coefficients. The mean predictor provides a naïve reference, while TF-IDF with Ridge regression and support vector regression (SVR) represent lexical baselines. These models were compared with IndoBERT, multilingual BERT (mBERT), XLM-RoBERTa, and multilingual DeBERTa-v3 (mDeBERTa-v3) using the same group-aware held-out test set.

Among the classical baselines, TF-IDF with SVR produced the strongest regression performance, with an MAE of 0.0623, RMSE of 0.0835, R^2 of 0.1510, and Spearman's ρ of 0.4949. Its improvement over the mean predictor and TF-IDF with Ridge indicates that lexical information contains predictive signals associated with expert-derived relevance. However, its relatively limited explained variance and rank correlation suggest that lexical features alone do not fully capture the contextual factors reflected in expert judgments.

The transformer models generally achieved lower prediction errors and higher R^2 values than the classical baselines, although their performance differed across metrics. mBERT

obtained the lowest MAE (0.0501), lowest RMSE (0.0625), and highest R^2 (0.5239). Thus, among the evaluated models, mBERT provided the strongest numerical approximation of the expert-derived Aiken's V coefficients on the held-out test set. Its R^2 indicates that approximately 52.4% of the observed variance was explained by the model under the present evaluation setting. mDeBERTa-v3 produced a slightly higher MAE of 0.0514 and RMSE of 0.0670, with an R^2 of 0.4534, but achieved the highest Spearman's ρ at 0.7532. This indicates stronger preservation of the relative ordering of instructional items, despite slightly less accurate numerical prediction than mBERT. IndoBERT also performed competitively, obtaining an MAE of 0.0565, RMSE of 0.0707, R^2 of 0.3913, and Spearman's ρ of 0.7198. XLM-RoBERTa showed weaker regression performance than the other transformer models, with an MAE of 0.0661, RMSE of 0.0792, R^2 of 0.2360, and Spearman's ρ of 0.5374.

Table 3. Test-set performance for Aiken's V regression.

Model	MAE ↓	RMSE ↓	R^2 ↑	Spearman's ρ ↑
Mean predictor	0.0772	0.0917	−0.0246	–
TF-IDF + Ridge	0.0729	0.0892	0.0300	0.4588
TF-IDF + SVR	0.0623	0.0835	0.1510	0.4949
IndoBERT	0.0565	0.0707	0.3913	0.7198
mBERT	**0.0501**	**0.0625**	**0.5239**	0.7024
XLM-RoBERTa	0.0661	0.0792	0.2360	0.5374
mDeBERTa-v3	0.0514	0.0670	0.4534	**0.7532**

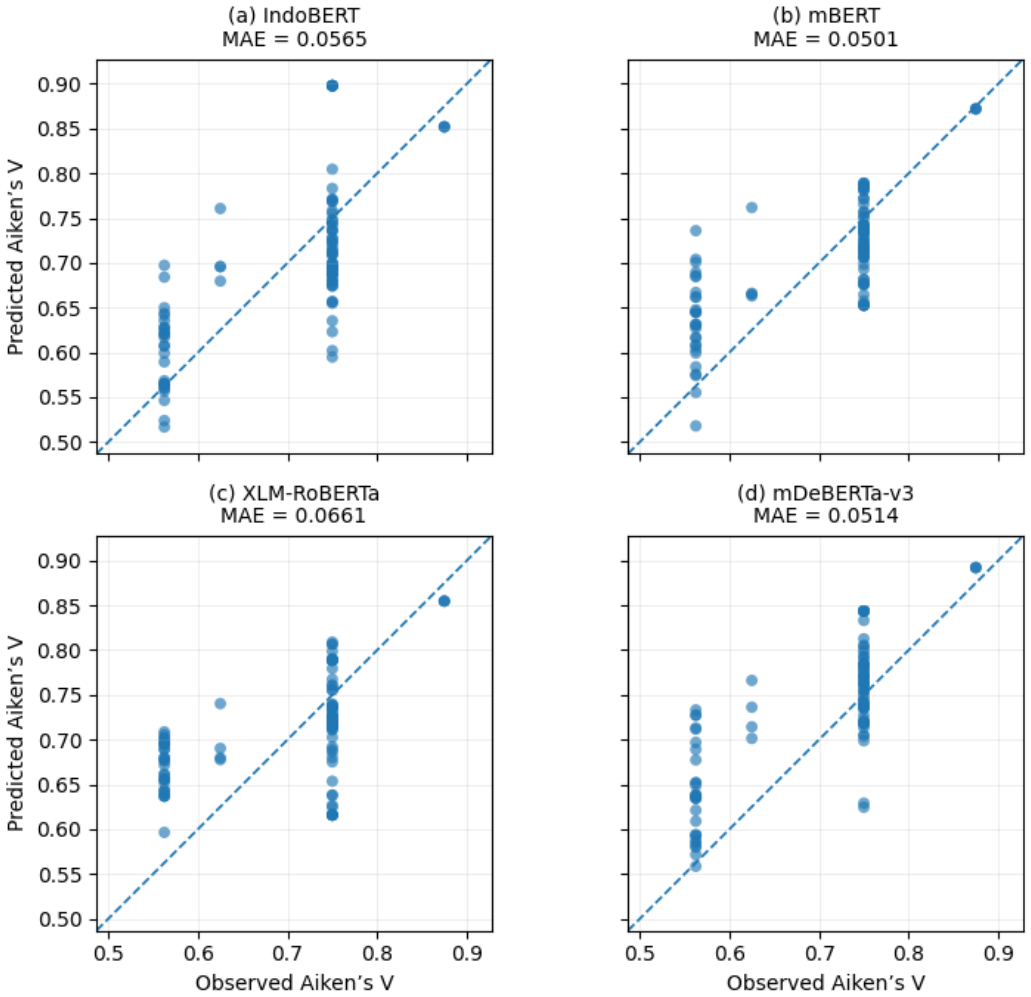

Figure 2. Predicted versus observed Aiken's V coefficients for the evaluated transformer models. The dashed diagonal line represents perfect agreement, with identical axis ranges across panels.

These results indicate that transformer-based contextual representations can provide additional predictive information beyond the lexical baselines for this dataset, but the results do not establish a universally superior architecture. In particular, mBERT was strongest for numerical estimation, whereas mDeBERTa-v3 was strongest for rank preservation. The difference also suggests that model selection depends on how the predicted Aiken's V values are intended to be used in a pre-screening workflow.

Figure 2 provides a visual comparison between the observed and predicted Aiken's V coefficients for the four transformer models. The predictions of mBERT and mDeBERTa-v3 generally show closer alignment with the diagonal reference than those of XLM-RoBERTa, consistent with their lower regression errors. IndoBERT also shows relatively close correspondence between observed and predicted values, although some dispersion remains. The target distribution also imposes an important constraint on the regression task. Because each Aiken's V coefficient was derived from four expert ratings on a five-point ordinal scale, the resulting coefficients occupy a limited set of possible values. The models consequently tend to produce predictions toward the center of the observed distribution: lower observed coefficients may be overestimated, whereas higher coefficients may be underestimated. This pattern is consistent with regression toward the mean and suggests that the predicted Aiken's V values are more appropriate for preliminary estimation and item prioritization than as exact substitutes for expert-derived coefficients. The regression results provide evidence that the evaluated transformer models can approximate expert-derived relevance scores under the leakage-controlled setting used in this study. However, the remaining prediction errors and restricted target distribution reinforce the intended role of the models as decision-support tools for preliminary screening rather than replacements for expert assessment.

4.2. Essentiality Classification Performance

Table 4 presents the test-set performance of the classical baselines and transformer models for classifying instructional items as essential or non-essential. Because essential items constituted the majority class, accuracy alone may not adequately reflect performance across both classes. Balanced accuracy, macro-precision, Macro-F1, and ROC-AUC were therefore also considered to provide a more comprehensive assessment.

Table 4. Test-set performance for essentiality classification.

Model	Accuracy ↑	Macro-Precision ↑	Balanced Accuracy ↑	Macro-F1 ↑	ROC-AUC ↑
Majority-class predictor	0.6701	0.3351	0.5000	0.4012	–
TF-IDF + Logistic Regression	0.7216	0.7001	0.7209	0.7035	0.7457
TF-IDF + Linear SVM	0.6907	0.6660	0.6820	0.6687	0.7538
IndoBERT	0.7732	0.7657	0.8070	0.7660	0.8909
mBERT	0.8351	0.8135	**0.8135**	0.8135	**0.9111**
XLM-RoBERTa	**0.8557**	**0.8906**	0.7892	**0.8161**	0.8887

The majority-class predictor achieved an accuracy of 0.6701 by assigning all test items to the essential class. However, its balanced accuracy was 0.5000 and its Macro-F1 was 0.4012, confirming that the relatively high accuracy primarily reflected the class distribution rather than effective discrimination between the two classes. Among the classical baselines, TF-IDF with logistic regression performed better than the linear SVM on accuracy, balanced accuracy, and Macro-F1, while the SVM obtained a slightly higher ROC-AUC (0.7538 versus 0.7457). These results indicate that lexical representations contain useful signals for essentiality prediction, although their performance remained below that of the transformer models.

The three valid transformer models consistently outperformed the classical baselines on the principal class-balanced measures. IndoBERT achieved an accuracy of 0.7732, balanced accuracy of 0.8070, Macro-F1 of 0.7660, and ROC-AUC of 0.8909. mBERT achieved an accuracy of 0.8351, balanced accuracy of 0.8135, Macro-F1 of 0.8135, and the highest ROC-AUC of 0.9111. These results indicate that mBERT provided the strongest discrimination when both classes and different classification thresholds were considered.

XLM-RoBERTa achieved the highest accuracy (0.8557), macro-precision (0.8906), and Macro-F1 (0.8161), but its balanced accuracy was lower (0.7892) than those of mBERT and

IndoBERT. This difference is important because accuracy and Macro-F1 alone do not fully capture the model's ability to recognize the minority non-essential class. Thus, XLM-RoBERTa produced the strongest classification performance at the applied decision threshold according to accuracy and Macro-F1, whereas mBERT provided more balanced class discrimination and the strongest threshold-independent performance.

Figure 3 provides the confusion matrices for the three valid transformer models and illustrates the differences in their class-specific predictions. IndoBERT correctly classified 29 of the 32 non-essential items but misclassified 19 essential items as non-essential. mBERT produced a more balanced error pattern, with eight misclassifications in each direction. XLM-RoBERTa correctly classified 64 of the 65 essential items but identified only 19 of the 32 non-essential items correctly. This explains why XLM-RoBERTa achieved the highest overall accuracy while obtaining a lower balanced accuracy than mBERT and IndoBERT.

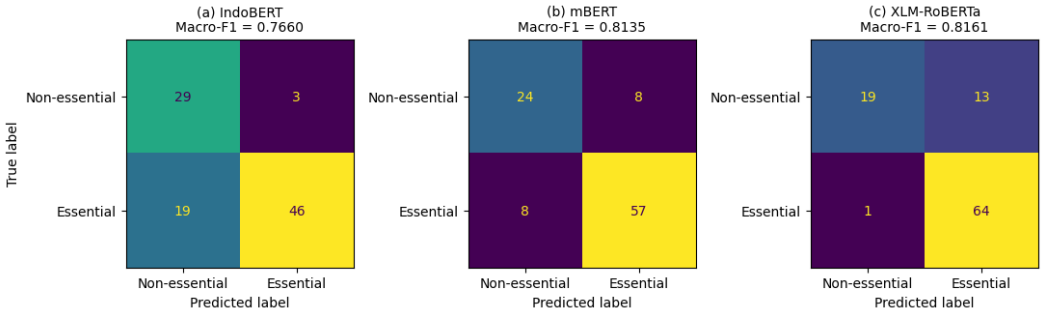

Figure 3. Confusion matrices for IndoBERT, mBERT, and XLM-RoBERTa on the group-aware held-out test set. Class 0 represents non-essential items and Class 1 represents essential items.

The results therefore suggest that model preference depends on the intended use of the classification output. XLM-RoBERTa may be preferable when the primary objective is accurate class assignment at the applied decision threshold. mBERT may be more appropriate when balanced recognition of essential and non-essential items or threshold-independent discrimination is more important. For a content-validity pre-screening workflow, the latter considerations are particularly relevant because a model that performs well on the majority class may still overlook a substantial proportion of non-essential items.

The classification results should consequently be interpreted as evidence of predictive support rather than autonomous content-validity assessment. In practice, predicted class probabilities can be used to prioritize items for expert review, particularly when the model is uncertain or when different models produce conflicting predictions. Final essentiality judgments should remain subject to review by qualified subject-matter experts, who can consider curricular context and other information not represented in the textual input.

4.3. Error Analysis

Aggregate performance metrics indicate that transformer models can capture textual patterns associated with expert-derived content-validity outcomes. However, these metrics do not explain the conditions under which predictions deviate from expert judgments. An item-level error analysis was therefore conducted using the group-aware held-out test set. For regression, the analysis focused on mBERT because it achieved the lowest MAE, while representative classification errors were examined for mBERT and XLM-RoBERTa.

For Aiken's V regression, most predictions remained within a relatively narrow error range, although larger deviations were observed for short or context-dependent instructional statements. A recurring pattern was the overestimation of items with relatively low expert-derived Aiken's V coefficients. This tendency is consistent with the restricted target distribution resulting from four expert ratings on a five-point scale, which provides limited resolution at the item level. Table 5 presents representative prediction errors selected to illustrate these patterns rather than to represent their overall frequency.

The regression examples suggest that short statements containing recognizable technical terminology may receive relatively high predicted relevance despite lower expert-derived scores. For example, "Time dan global states" was overestimated by 0.1735, while "Indirect communication" was overestimated by 0.1416. Such cases suggest that textual cues alone may

not provide sufficient information to determine whether a concept is appropriate for a particular course, competency level, or intended learning outcome.

A similar limitation appears in essentiality classification. False positives occurred when technically specific statements were predicted as essential even though the expert-derived label was non-essential. Conversely, concise statements such as "Pengantar Aljabar Linier" and "Representasi pengetahuan" were predicted as non-essential despite being labeled essential by the experts. These examples illustrate that linguistic completeness or the presence of domain-specific terminology does not necessarily correspond to expert judgments of essentiality. An item may require broader curricular knowledge to determine whether it is foundational, redundant, outside the intended scope, or appropriately aligned with the learning objectives.

Table 5. Representative prediction errors on the group-aware held-out test set.

Task	Model	Instructional item	Expert-derived target	Model prediction	Probability / absolute error	Error type
Regression	mBERT	"Time dan global states."	Aiken's V = 0.5625	Aiken's V = 0.7360	0.1735	Overestimation
Regression	mBERT	"Indirect communication."	Aiken's V = 0.5625	Aiken's V = 0.7041	0.1416	Overestimation
Classification	mBERT	"Time dan global states."	Non-essential	Essential	0.9913	False positive
Classification	mBERT	"*Pengantar Aljabar Linier*."	Essential	Non-essential	0.0077	False negative
Classification	XLM-RoBERTa	"*Mampu menerapkan kebijakan keamanan pada sistem server*."	Non-essential	Essential	0.9463	False positive
Classification	XLM-RoBERTa	"*Representasi pengetahuan*."	Essential	Non-essential	0.0594	False negative

Note: The instructional items are presented in their original Indonesian-language form to preserve the textual characteristics of the evaluated dataset.

Several classification errors were also made with high confidence. For example, mBERT assigned an essential-class probability of 0.9913 to "Time dan global states" although the expert-derived label was non-essential. Conversely, it assigned a probability of only 0.0077 to "Pengantar Aljabar Linier," which experts classified as essential. These cases indicate that prediction confidence alone is not sufficient to identify all problematic items. A model can be highly confident while still making an incorrect prediction, particularly when textual patterns learned during training do not adequately represent the contextual factors considered by experts.

The observed errors highlight a central limitation of text-only content-validity pre-screening. The models primarily processed instructional statements as isolated text and did not directly incorporate course objectives, competency hierarchies, prerequisite relationships, instructional sequence, student level, or relationships among multiple instructional items. Subject-matter experts, in contrast, may use these factors when assessing relevance and essentiality. Consequently, some prediction errors may result from contextual information that is unavailable in the model input rather than from deficiencies in language representation alone.

These findings support the use of the models as human-in-the-loop pre-screening tools. Predicted Aiken's V values, essentiality probabilities, and model disagreement can help prioritize items for expert examination. However, the presence of confident errors demonstrates that even apparently reliable predictions should not be accepted automatically. Final content-validity decisions should remain with qualified subject-matter experts who can evaluate the broader instructional and curricular context.

4.4. Model Performance and Practical Implications

The results across the two prediction tasks indicate that model performance depends on the specific objective of the pre-screening process. Rather than identifying a single universally superior architecture, the findings suggest that different models provide different strengths. For Aiken's V regression, mBERT was most effective for numerical estimation, while mDeBERTa-v3 provided the strongest preservation of item ranking. For essentiality

classification, XLM-RoBERTa produced the highest accuracy and Macro-F1 at the applied decision threshold, whereas mBERT provided stronger balanced discrimination across the two classes.

This distinction is relevant to how the framework may be used in practice. When the objective is to estimate the relative relevance of instructional items, the predicted Aiken's V values can provide an initial indication of which items may warrant further examination. When the objective is to identify potentially non-essential items, classification probabilities provide an additional signal that can be used to prioritize cases for expert review. These outputs should be considered complementary rather than interchangeable because Aiken's V represents graded relevance, whereas essentiality classification reflects a binary decision derived from expert judgments.

A practical pre-screening workflow could therefore use model outputs to prioritize items rather than automatically accept or reject them. Items with relatively low predicted Aiken's V, probabilities close to the classification decision threshold, or disagreement between models can be flagged for closer examination. The error analysis in Section 4.3 further shows that this strategy should not rely solely on prediction confidence, since some incorrect predictions were made with high confidence.

The results also suggest that contextual transformer representations can provide useful predictive information beyond the evaluated lexical baselines. However, the advantage should not be interpreted as evidence that transformer models can independently determine content validity. The models operate primarily on the textual formulation of individual instructional items, whereas expert judgments may incorporate curriculum objectives, competency levels, prerequisite relationships, instructional sequence, and other contextual information that is not explicitly provided as model input.

Accordingly, the most appropriate role of the proposed framework is as a human-in-the-loop pre-screening and prioritization mechanism. It can help reduce the initial screening burden by identifying items that appear relatively clear as well as those that may require closer expert attention. Formal content-validity decisions, however, should remain under the responsibility of qualified subject-matter experts. This positioning also provides a more appropriate interpretation of the reported results: the models demonstrate the feasibility of predicting expert-derived indicators from instructional text under a leakage-controlled evaluation setting, rather than replacing the underlying expert assessment process.

4.5. Limitations and Generalizability

Although the proposed framework achieved competitive predictive performance on the held-out test set, several limitations constrain the interpretation and generalizability of the findings.

First, the dataset was derived from a specific collection of undergraduate computer science instructional materials. Its linguistic characteristics, technical terminology, and curricular structure may differ from those of other institutions, disciplines, or educational levels. Moreover, the study did not include external validation using an independent dataset. The results therefore demonstrate generalization to previously unseen normalized-text groups within the same source dataset, rather than cross-institutional or cross-disciplinary generalization.

Second, the expert-derived targets were based on evaluations from four subject-matter experts. Although the experts evaluated the items independently, the resulting targets represent the judgments of a relatively small panel. In addition, Aiken's V values were derived from four ratings on a five-point ordinal scale, resulting in a limited number of possible coefficient values. This restricted target resolution may contribute to the regression behavior observed in Section 4.1. The binary essentiality target also simplifies the underlying expert judgments and does not explicitly preserve the degree of disagreement among raters.

Third, the evaluation relied on a single group-aware 70:15:15 training–validation–test split. Grouping identical and normalized-equivalent texts reduced direct leakage between subsets, but a single split remains sensitive to the particular composition of the held-out data. The grouping strategy also addresses exact or normalized textual overlap rather than broader semantic similarity. Future studies should therefore assess robustness using repeated group-aware splits, multiple random seeds, or group-aware cross-validation.

Fourth, the models received instructional statements primarily as isolated textual inputs. They did not directly incorporate course identity, learning-objective hierarchies, competency levels, prerequisite relationships, instructional sequence, or relationships among multiple

instructional items. As illustrated by the error analysis, such information may be important for determining whether an item is relevant or essential. Future work could therefore combine transformer representations with structured curricular metadata and inter-item relationships.

Fifth, optimization stability differed across the evaluated architectures. Multilingual DeBERTa-v3 produced a valid regression result but did not provide a sufficiently stable classification result for inclusion in the final classification comparison. Its exclusion limits the conclusions that can be drawn about its classification capability. Additional experiments with multiple random seeds and alternative optimization or stabilization strategies would be required to evaluate this model more reliably for essentiality classification.

Finally, the model comparisons are descriptive because formal statistical significance testing and confidence intervals were not performed. Some differences between the strongest models were relatively small, particularly for Macro-F1. The reported rankings should therefore be interpreted as observed performance differences under the present experimental setting rather than evidence of statistically significant superiority. Probability calibration was also not evaluated, despite the potential use of predicted probabilities for prioritizing items for expert review. Future studies should assess the robustness of these comparisons using confidence intervals or paired statistical procedures and evaluate probability calibration using measures such as the Brier score and reliability diagrams.

Taken together, these limitations indicate that the current findings should be interpreted as evidence of the feasibility of text-based content-validity pre-screening within the evaluated setting, rather than as evidence of broad generalizability or autonomous content-validity assessment. Future research should prioritize independent datasets from different institutions and disciplines, larger and more diverse expert panels, repeated group-aware evaluation, and integration of curricular context to determine whether the observed predictive patterns remain consistent across settings.

5. Conclusions

This study investigated the feasibility of predicting expert-derived content-validity outcomes from Indonesian instructional text through two complementary supervised-learning tasks: Aiken's V regression and essentiality classification. Classical TF-IDF-based approaches were compared with IndoBERT, multilingual BERT, XLM-RoBERTa, and multilingual DeBERTa-v3 using a group-aware 70:15:15 partition. Grouping identical and normalized-equivalent statements before partitioning reduced direct information leakage and enabled evaluation on previously unseen text groups. For Aiken's V regression, mBERT achieved the lowest MAE (0.0501), RMSE (0.0625), and highest R^2 (0.5239), whereas mDeBERTa-v3 achieved the highest Spearman's ρ (0.7532). For essentiality classification, XLM-RoBERTa obtained the highest accuracy (0.8557) and Macro-F1 (0.8161), while mBERT achieved the highest balanced accuracy (0.8135) and ROC-AUC (0.9111). These results indicate that model selection depends on the intended use of the predictions, including numerical estimation, item ranking, class assignment, or balanced discrimination.

The findings support the use of transformer-based models as human-in-the-loop tools for preliminary content-validity pre-screening. Their outputs can help prioritize instructional items for expert examination, but they should not be interpreted as replacements for formal expert assessment. The error analysis further showed that text-only models have difficulty with cases requiring broader curricular information, including competency hierarchies, prerequisite relationships, instructional sequence, and relationships among instructional items. The findings should be interpreted within the scope of the dataset, four expert raters, single group-aware holdout split, and absence of external validation. Future research should examine the framework across multiple institutions and disciplines, use larger and more diverse expert panels, adopt repeated group-aware evaluation, incorporate structured curricular information, and evaluate calibrated prediction uncertainty. Such extensions are needed to determine whether the observed results generalize beyond the current dataset and support broader educational quality-assurance applications..

Author Contributions: Conceptualization: S. and D.M.; Methodology: S. and D.M.; Software: D.M.; Validation: S., D.M., and A.I.; Formal analysis: S. and D.M.; Investigation: S., D.M., M.M., and W.S.; Resources: S. and A.I.; Data curation: D.M.; Writing original draft

preparation: S., D.M., M.M., and W.S.; Writing review and editing: S., D.M., A.I., M.M., W.S., E.W., A.A.O., and D.R.I.M.S.; Visualization: D.M.; Supervision: S. and A.A.O.; Project administration: S.; Funding acquisition: S. and D.M. M.M. and W.S. contributed particularly to the literature review; E.W. contributed to language refinement and manuscript editing; A.A.O. conducted an internal scholarly review; and D.R.I.M.S. contributed to internal academic and editorial review. All authors have read and agreed to the published version of the manuscript

Funding: This research was funded by the Internal PKLN Research Grant of Universitas Muhammadiyah Semarang for the 2025 fiscal year.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The dataset is not publicly available because it contains internally developed educational materials and expert assessment data.

Acknowledgments: The authors would like to thank the subject-matter experts who contributed to the evaluation and annotation of the educational materials used in this study. The authors also acknowledge the administrative and technical support provided by Universitas Muhammadiyah Semarang. This research was conducted under the 2025 Internal International Research Collaboration Grant (Penelitian Kerjasama Luar Negeri PKLN) of Universitas Muhammadiyah Semarang, Contract No. 057/UNIMUS.L/PG/PKLN/PJ.INT/2025.

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

References

[1] W. G. Spady, *Outcome-Based Education: Critical Issues and Answers*. Arlington, VA: American Association of School Administrators, 1994. [Online]. Available: https://eric.ed.gov/?id=ED380910

[2] R. M. Harden, "Outcome-Based Education: the future is today," *Med. Teach.*, vol. 29, no. 7, pp. 625–629, Jan. 2007, doi: 10.1080/01421590701729930.

[3] J. Biggs and C. Tang, *Teaching for Quality Learning at University: What the Student Does*, 4th ed. Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press, 2011.

[4] M. S. B. Yusoff, "ABC of Content Validation and Content Validity Index Calculation," *Educ. Med. J.*, vol. 11, no. 2, pp. 49–54, Jun. 2019, doi: 10.21315/eimj2019.11.2.6.

[5] N. Zaki, S. Turaev, K. Shuaib, A. Krishnan, and E. Mohamed, "Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation," *Educ. Inf. Technol.*, vol. 28, no. 12, pp. 16723–16742, Dec. 2023, doi: 10.1007/s10639-023-11877-4.

[6] E. Tocto-Cano, S. Paz Collado, and J. L. López-Gonzales, "A Holistic Maturity Model for Quality Assessment and Innovation in Peruvian Universities," *Educ. Sci.*, vol. 15, no. 2, p. 142, Jan. 2025, doi: 10.3390/educsci15020142.

[7] M. Praseptiawan, A. N. Che Pee, M. H. Zakaria, and A. Noertjahyana, "Advancing the Measurement of MOOCs Software Quality: Validation of Assessment Tools Using the I-CVI Expert Framework," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 138–145, May 2025, doi: 10.52088/ijesty.v5i3.911.

[8] L. B. Mokkink, S. Herbelet, P. R. Tuinman, and C. B. Terwee, "Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument – a COSMIN perspective," *J. Clin. Epidemiol.*, vol. 185, p. 111879, Sep. 2025, doi: 10.1016/j.jclinepi.2025.111879.

[9] R. Doneva, S. Gaftandzhıeva, and G. Totkov, "Automated Quality Assurance of Educational Testing," *Turkish Online J. Distance Educ.*, vol. 19, no. 3, pp. 71–92, Jul. 2018, doi: 10.17718/tojde.444961.

[10] A. M. McCarthy, D. Maor, A. McConney, and C. Cavanaugh, "Digital transformation in education: Critical components for leaders of system change," *Soc. Sci. Humanit. Open*, vol. 8, no. 1, p. 100479, 2023, doi: 10.1016/j.ssaho.2023.100479.

[11] M. Kayyali, "The Evolution of Quality Assurance in Higher Education," in *Navigating Quality Assurance and Accreditation in Global Higher Education*, 2024, pp. 1–26. doi: 10.4018/979-8-3693-6915-9.ch001.

[12] J. Devlin, M.-W. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[13] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," *ArXiv*. Apr. 08, 2020. [Online]. Available: http://arxiv.org/abs/1911.02116

[14] S. Devaraju, "Natural Language Processing (NLP) in AI-Driven Recruitment Systems," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 555–566, Jun. 2022, doi: 10.32628/CSEIT2285241.

[15] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, "Natural language processing with transformers: a review," *PeerJ Comput. Sci.*, vol. 10, p. e2222, Aug. 2024, doi: 10.7717/peerj-cs.2222.

[16] L. Krstić, V. Aleksić, and M. Krstić, "Artificial Intelligence in Education: A Review," in *Proceedings TIE 2022*, 2022, pp. 223–228. doi: 10.46793/TIE22.223K.

[17] P. Premananthan and M. Fahim, "The Role of Machine Learning in Smart Education: Taxonomy, Challenges, and Use Cases," *EAI Endorsed Trans. Tour. Technol. Intell.*, vol. 1, no. 1, Sep. 2024, doi: 10.4108/eettti.6833.

[18] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1560–1569. doi: 10.18653/v1/2020.findings-emnlp.141.

[19] A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia, "Automated Bahasa Indonesia essay evaluation with latent semantic analysis," *J. Phys. Conf. Ser.*, vol. 1235, no. 1, p. 012100, Jun. 2019, doi: 10.1088/1742-6596/1235/1/012100.

[20] D. M. Adayilo, I. O. Oyefolahan, J. N. Ndunagu, C. Otuya, E. Malcalm, and K. Twabu, "AI-Powered Tutoring Systems for Personalised Learning Feedback in Developing Secondary Education Contexts," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 549–564, Dec. 2025, doi: 10.62411/faith.3048-3719-290.

[21] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," *J. Biomed. Inform.*, vol. 126, p. 103982, Feb. 2022, doi: 10.1016/j.jbi.2021.103982.

[22] N. Kania, Y. S. Kusumah, J. A. Dahlan, E. Nurlaelah, F. Gürbüz, and E. Bonyah, "Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills," *REID (Research Eval. Educ.*, vol. 10, no. 1, pp. 64–79, Jun. 2024, doi: 10.21831/reid.v10i1.71032.

[23] L. R. Aiken, "Three Coefficients for Analyzing the Reliability and Validity of Ratings," *Educ. Psychol. Meas.*, vol. 45, no. 1, pp. 131–142, Mar. 1985, doi: 10.1177/0013164485451012.

[24] D. M. Rubio, M. Berg-Weger, S. S. Tebb, E. S. Lee, and S. Rauch, "Objectifying content validity: Conducting a content validity study in social work research," *Soc. Work Res.*, vol. 27, no. 2, pp. 94–104, Jun. 2003, doi: 10.1093/swr/27.2.94.

[25] A. Muzli, R. N. E. Anggraini, and D. Purwitasari, "Cog-CoT: A Cognitive Chain-of-Thought Framework for Bloom's Taxonomy-Aligned Educational Question Answering," *J. Comput. Theor. Appl.*, vol. 4, no. 1, pp. 330–353, Aug. 2026, doi: 10.62411/jcta.17084.

[26] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 1, p. 100211, Mar. 2024, doi: 10.1016/j.joitmc.2024.100211.

[27] D. Marutho, Muljono, S. Rustad, and Purwanto, "Optimizing Aspect Term Extraction and Sentiment Classification through Attention Mechanism and Sparse Attention Techniques," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 5, pp. 1004–1015, Oct. 2024, doi: 10.22266/ijies2024.1031.75.

[28] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[29] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2898–2904. doi: 10.18653/v1/2020.findings-emnlp.261.