TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

AI-Assisted Labeling for Indonesian Hadith Classification using Four Thematic Categories: Evaluating TF-IDF+SVM and IndoBERT

Domi Sepri[1,*], Ahmad Fauzi[1], Firman[2], Muhammad Nabil[1]

[1] Faculty of Science and Technology, Information System, Universitas Islam Negeri Imam Bonjol Padang, Padang, Indonesia
[2] Faculty of Ushuluddin and Religious Studies, Science of Hadith, Universitas Islam Negeri Imam Bonjol Padang, Padang, Indonesia
Email: [1*] domisepri@uinib.ac.id, [2]ahmadfauzi@uinib.ac.id, [3]firman@uinib.ac.id, [4]2317020021@uinib.ac.id
Correspondence Author Email: domisepri@uinib.ac.id

Abstract–Hadith is the second primary source of Islamic law after the Qur'an, and its large volume makes manual thematic classification time-consuming and inefficient. This study proposes an AI-assisted labeling approach to construct an Indonesian thematic hadith dataset and evaluates the performance of two text classification methods, namely TF-IDF + Support Vector Machine (SVM) and IndoBERT. The dataset consists of 6,600 Indonesian-translated Sahih Bukhari hadiths collected from the Hadith API and categorized into four thematic classes: aqidah, ibadah, akhlak, and muamalah. The annotation process employed Gemini 2.5 Flash with a structured prompt and JSON-based output format, followed by validation performed by a hadith researcher validation on a randomly selected 5% sample, achieving an overall agreement of 73.3%. The annotated data were divided into training and testing sets using an 80:20 stratified split. Model performance was evaluated using Accuracy, Macro F1-score, and Weighted F1-score. Experimental results show that TF-IDF + SVM achieved an Accuracy of 73.1%, a Macro F1-score of 70.1%, and a Weighted F1-score of 73.0%, while IndoBERT achieved an Accuracy of 72.3%, a Macro F1-score of 69.7%, and a Weighted F1-score of 72.2%. The results indicate that the conventional TF-IDF + SVM approach slightly outperformed the Transformer-based IndoBERT model on the proposed dataset. The main contributions of this study are the construction of an Indonesian thematic hadith dataset through AI-assisted labeling and a comparative evaluation of conventional and Transformer-based methods for Indonesian hadith classification.

Keywords: AI-assisted Labeling; Hadith Classification; IndoBERT; Support Vector Machine; Natural Language Processing

1. INTRODUCTION

Hadith is the second primary source of Islamic law after the Qur'an and serves as a fundamental guideline for various aspects of Muslim life, including aqidah (faith), ibadah (worship), muamalah (social and economic transactions), and akhlak (morality) (Ni'mah. SM et al., 2024). With the rapid advancement of information technology, numerous hadith collections have been digitized and made available through databases, web platforms, and mobile applications, enabling easier access to Islamic literature (Susanti et al., 2025). Nevertheless, the large volume of hadiths and their diverse thematic contents make manual searching, organization, and classification increasingly inefficient, requiring considerable time and hadits researcher (Nabiilah et al., 2021). This challenge highlights the importance of applying Natural Language Processing (NLP) techniques to automatically process hadith texts, particularly for text classification, thereby enabling faster, more accurate, and systematic organization of hadith information(Rasenda et al., 2024). Furthermore, the availability of a well-structured hadith dataset is essential for supporting the development of artificial intelligence applications, including information retrieval, text classification, recommendation systems, and intelligent Islamic information services.

Numerous studies have proposed hadith classification systems to facilitate the organization and retrieval of hadith information for various applications. Most existing studies have focused on classifying hadiths into functional categories, such as recommendation, prohibition, and information, while others have explored multi-label classification and topic clustering to group hadiths according to semantic similarity (Handayani et al., 2023). These studies demonstrate that Natural Language Processing (NLP) techniques can effectively support the organization of digital hadith collections and improve the accessibility of Islamic knowledge. Furthermore, different feature extraction methods, text representation techniques, and classification algorithms have been investigated to improve classification accuracy and automate thematic information retrieval from hadith texts. Despite these advances, the adopted classification schemes generally depend on the specific objectives of each study, resulting in considerable variation in annotation standards and label definitions. Consequently, the resulting datasets and classification models are often difficult to compare or reuse across different studies.

In contrast, hadith classification based on thematic categories, such as aqidah, ibadah, akhlak, and muamalah, remains relatively underexplored, despite these categories representing the major dimensions of Islamic teachings and providing a more systematic framework for organizing hadith knowledge. A standardized thematic classification scheme is expected to facilitate the development of intelligent Islamic information systems, semantic search applications, educational technologies, and knowledge management systems by organizing hadiths according to their primary religious themes rather than application-specific objectives.

Several studies have employed different machine learning approaches for hadith classification. Taufiqurrahman et al. applied a combination of Chi-Square feature selection and Support Vector Machine (SVM) for multi-label classification of Indonesian-translated hadiths into recommendation, prohibition, and information categories (Rasenda et al., 2024). Ramadhani et al. compared TF-IDF + SVM and Long Short-Term Memory (LSTM) for multi-label classification of Sahih Bukhari hadiths and demonstrated that classification performance depends on both the selected

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

method and dataset characteristics (Ramadhan et al., 2024). Lutfiah utilized TF-IDF and One-vs-Rest SVM for multi-label classification of Sahih Muslim hadiths and identified class imbalance as one of the primary challenges affecting classification performance (Lutfiah et al., 2026).

More recently, Transformer-based language models have demonstrated superior performance across a wide range of Indonesian NLP tasks by learning contextual semantic representations that capture linguistic information beyond surface-level lexical features (Darfiansa et al., 2025). Nevertheless, previous hadith classification studies generally employ different annotation schemes, datasets, and evaluation settings. As a result, it remains unclear whether Transformer-based models consistently outperform conventional machine learning approaches for thematic hadith classification, particularly when trained on datasets constructed through AI-assisted annotation.

Based on the existing literature, several research gaps remain. First, relatively few studies have investigated hadith classification using the four thematic categories of aqidah, ibadah, akhlak, and muamalah, despite their importance as the fundamental dimensions of Islamic teachings. Second, the construction of hadith datasets still relies predominantly on manual annotation, which is labor-intensive, time-consuming, and highly dependent on hadits researcher, whereas the use of AI-assisted labeling for large-scale Indonesian hadith datasets has received limited attention. Third, previous comparative studies have primarily evaluated conventional machine learning or deep learning models using manually annotated datasets, while systematic comparisons between TF-IDF + SVM and Transformer-based models such as IndoBERT on AI-assisted labeled datasets remain scarce.

To address these gaps, this study proposes an AI-assisted approach for constructing a thematic Indonesian hadith dataset consisting of four major categories: aqidah, ibadah, akhlak, and muamalah. The resulting dataset is subsequently used to evaluate and compare two text classification approaches, namely TF-IDF + SVM as a conventional machine learning method and IndoBERT as a Transformer-based language model. Model performance is evaluated using Accuracy, Macro F1-score, and Weighted F1-score to provide a comprehensive assessment of classification effectiveness.

2. RESEARCH METHODOLOGY

2.1 Research Workflow

The overall methodology consists of seven stages: data collection, data preparation, AI-assisted labeling, Hadits researcher validation, text preprocessing, text classification, and performance evaluation. The complete workflow is presented in Figure 1.

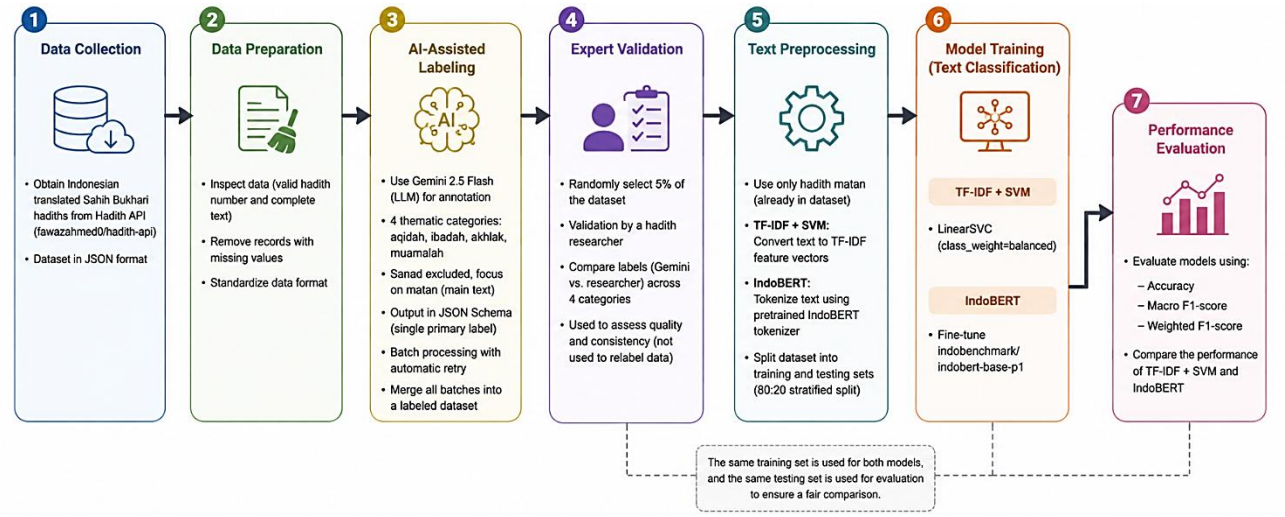

Figure 1. Overall Research Workflow

2.2 Data Collection

The dataset used in this study consists of 6,600 Indonesian-translated Sahih Bukhari hadiths obtained from the Hadith API repository in JSON format. The data were accessed through the fawazahmed0/hadith-api repository, distributed via the jsDelivr CDN service, and used as the primary source for annotation and classification. Prior to annotation, each record was inspected to ensure the availability of a valid hadith number and complete hadith text. Records containing missing values were removed, and the data format was standardized to ensure consistency throughout the annotation process.

The selected dataset was considered suitable for this study because Sahih Bukhari is one of the most widely recognized and authentic hadith collections in Islamic scholarship (Zakiah et al., 2025). Furthermore, the availability of high-quality Indonesian translations facilitates the application of Natural Language Processing (NLP) techniques for

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

thematic text classification. The use of a publicly accessible repository also enhances the reproducibility of the proposed methodology (Cahyawijaya et al., 2023).

2.3 AI-Assisted Labeling

The dataset annotation process was performed using an AI-assisted labeling approach with Gemini 2.5 Flash as the Large Language Model (LLM). Since the quality and consistency of LLM-generated annotations largely depend on prompt design, a structured prompt was carefully developed to define the operational criteria for the four thematic categories: aqidah (faith), ibadah (worship), akhlak (morality), and muamalah (social and economic transactions). The prompt also specified explicit labeling rules, including single-label assignment and tie-breaking criteria, to guide the model in selecting the most representative thematic category based solely on the hadith matan (main text), while excluding the sanad (chain of narrators). To assess the reliability of the AI-generated annotations, a randomly selected subset of the labeled dataset was subsequently reviewed by a hadith researcher.

Gemini 2.5 Flash was selected because it provides strong multilingual language understanding capabilities and supports structured output generation through JSON Schema. These features enable the model to produce consistent annotations while minimizing formatting errors during large-scale batch processing. The use of structured prompts further improves annotation consistency by explicitly defining the thematic categories and classification criteria before the annotation process begins (Geng et al., 2025).

To ensure output consistency, Gemini was configured to generate annotations in a predefined JSON Schema format, enabling each hadith to receive a single label following a standardized structure. The annotation process was executed incrementally using a batch-processing mechanism. Whenever the model failed to produce a valid JSON output, the system automatically retried the request until the predefined maximum number of attempts was reached. All successfully annotated batches were subsequently merged into a single labeled dataset, which was used for the classification stage (Irugalbandara, 2024).

Batch processing was employed to improve computational efficiency and ensure stable execution when annotating thousands of hadiths. This approach also facilitates automatic recovery from temporary API failures through the retry mechanism, thereby reducing the need for manual intervention during dataset construction (Sahoo et al., 2025).

As part of the annotation quality assessment, 5% of the entire dataset was randomly selected for validation by a hadith researcher. The validation process compared the labels generated by Gemini with the labels assigned by the hadith researcher across the four thematic categories. The validation results were used to assess the quality and consistency of the AI-assisted annotation process; however, they were not used to modify or relabel the dataset employed for model training (Gilardi et al., 2023).

2.4 Text Preprocessing

The labeled dataset was subsequently subjected to a text preprocessing stage before being used for model training(Kowsari et al., 2019). Since the annotation process was performed using only the hadith matan (main text), the dataset at this stage already contained the textual content required for classification. Therefore, the preprocessing focused on preparing the text according to the requirements of each classification method (Minaee et al., 2021).

For the TF-IDF + SVM approach, the hadith text was transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, allowing each document to be represented as a weighted vector of terms. For the IndoBERT approach, the hadith text was processed using the pretrained IndoBERT tokenizer, which converted the text into token representations compatible with the Transformer architecture. Finally, the dataset was divided into training and testing sets using an 80:20 stratified split, ensuring that the distribution of each thematic category was preserved in both subsets (Joseph, 2022).

Different preprocessing strategies were applied because the two classification approaches require different input representations. TF-IDF relies on sparse numerical vectors derived from term frequencies, whereas IndoBERT directly processes tokenized text to learn contextual representations through the Transformer architecture (Sunendar & Saputra, 2025). Consequently, the preprocessing stage was designed to preserve compatibility with the input requirements of each model.

2.5 Text Classification

This study employed two text classification approaches: TF-IDF + Support Vector Machine (SVM) and IndoBERT. TF-IDF + SVM represents a conventional machine learning approach based on sparse text representations, whereas IndoBERT is a Transformer-based language model capable of learning contextual representations of Indonesian text (Adigunawan & Fathoni, 2023). To ensure a fair comparison, both models were trained on the same training dataset and evaluated on the same testing dataset (Rainio et al., 2024). The implementation details of each approach are presented in the following subsections.

2.5.1 TF-IDF + Support Vector Machine

For the conventional approach, the TF-IDF feature vectors were used as input to a Linear Support Vector Machine (LinearSVC) classifier. The TF-IDF vectorizer was configured with a maximum of 20,000 features, unigram and

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

bigram (1–2 grams) representations, and sublinear term frequency weighting (Taha et al., 2024). To reduce the impact of class imbalance, the classifier employed class_weight = balanced during training.

2.5.2 IndoBERT

For the Transformer-based approach, the pretrained indobenchmark/indobert-base-p1 model was fine-tuned using the labeled hadith dataset. The maximum input length was set to 256 tokens, with a learning rate of 2×10^{-5}, a training batch size of 16, and an evaluation batch size of 32. The model was trained for a maximum of 10 epochs using class-weighted cross-entropy loss to mitigate class imbalance. In addition, an early stopping strategy was employed to terminate training automatically when the Macro F1-score on the validation set no longer improved (Bartlett et al., 2023).

2.6 Performance Evaluation

The performance of the classification models was evaluated using three metrics: Accuracy, Macro F1-score, and Weighted F1-score. Accuracy measures the proportion of correctly classified instances over the total number of testing samples. Macro F1-score evaluates the classification performance by assigning equal importance to each class regardless of its size, making it suitable for assessing the overall performance across all thematic categories. In contrast, Weighted F1-score considers the number of instances in each class, providing an evaluation that reflects the class distribution within the dataset. These three metrics were employed to compare the performance of the TF-IDF + SVM and IndoBERT models for classifying Indonesian hadiths into the four thematic categories (Grandini et al., 2020).

Accuracy alone may provide an overly optimistic evaluation when class distributions are imbalanced. Therefore, Macro F1-score and Weighted F1-score were also employed to provide complementary perspectives on classification performance across individual thematic categories and the overall dataset.

3. RESULTS AND DISCUSSION

3.1 AI-Assisted Labeling Results

The AI-assisted labeling process using Gemini 2.5 Flash successfully produced an Indonesian hadith dataset consisting of 6,600 hadiths categorized into four thematic classes: aqidah (faith), ibadah (worship), akhlak (morality), and muamalah (social and economic transactions). Each hadith was assigned a single primary label based on the semantic content of the hadith matan (main text) through a structured prompting strategy with the output formatted according to a predefined JSON Schema. The resulting labeled dataset was subsequently used as the input for training and evaluating the text classification models.

Table 1. Distribution of Hadith Dataset

Category	Number of Hadiths	Percentage (%)
Muamalah	2,485	37.65
Ibadah	1,921	29.11
Aqidah	1,252	18.97
Akhlak	942	14.27
Total	6,600	100.00

The dataset distribution shows that *Muamalah* represents the largest category (37.65%), whereas *Akhlak* is the smallest category (14.27%). This indicates that the class distribution is moderately imbalanced, which may affect the learning process of the classification models. Therefore, model performance was evaluated not only using Accuracy but also using Macro F1-score and Weighted F1-score to provide a more comprehensive assessment of classification performance across all thematic categories.

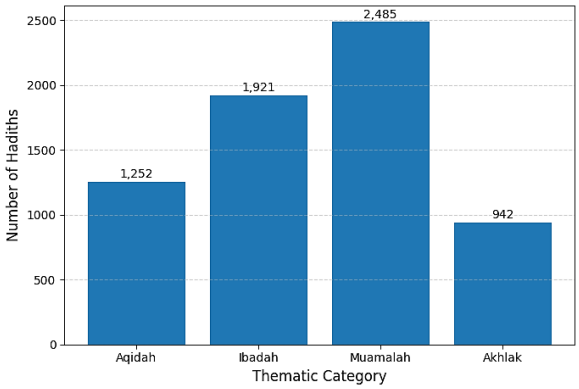

Figure 2. Distribution of the Indonesian Hadith Dataset

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

Figure 2 illustrates the distribution of the annotated hadith dataset across the four thematic categories. The Muamalah category contains the largest number of hadiths (2,485), followed by Ibadah (1,921), Aqidah (1,252), and Akhlak (942). Although the class distribution is moderately imbalanced, all categories contain a sufficient number of instances for training and evaluating the classification models. This distribution also justifies the use of Macro F1-score and Weighted F1-score in addition to Accuracy to provide a more representative evaluation of model performance.

The AI-assisted labeling approach substantially reduced the manual effort required to construct the thematic hadith dataset. By employing a structured prompt and a predefined JSON Schema, the annotation process produced standardized outputs that could be directly integrated into the dataset with minimal manual post-processing. To assess annotation quality, a randomly selected subset comprising 5% of the entire dataset was validated by a hadith researcher. This validation provided an assessment of the consistency between the AI-generated labels and Hadits researcher judgment while preserving the scalability advantages of LLM-assisted annotation (Gilardi et al., 2023).

Table 2. Validation Results of AI-Assisted Hadith Annotation

Category	Total Hadiths	Validation Sample (5%)	Valid	Invalid	Agreement (%)
Muamalah	2,485	124	63	61	50.8
Ibadah	1,921	96	83	13	86.5
Aqidah	1,252	63	55	8	87.3
Akhlak	942	47	41	6	87.2
Total	6,600	330	242	88	73.3

The validation results indicate that the agreement between the AI-assisted annotation and the hadith researcher varied across the four thematic categories. The Aqidah, Ibadah, and Akhlak categories achieved agreement rates above 85%, indicating that hadiths belonging to these categories generally contain distinctive semantic characteristics that enable consistent label assignment (Gilardi et al., 2023). In contrast, the Muamalah category achieved the lowest agreement rate (50.8%), suggesting that this category presents greater challenges for single-label annotation.

One possible explanation is the broad conceptual scope of Muamalah, which encompasses commercial transactions, family matters, and various aspects of Islamic legal rulings (Taha et al., 2024). Consequently, many hadiths categorized as Muamalah also contain moral teachings (Akhlak) or discussions related to acts of worship (Ibadah), resulting in overlapping semantic characteristics. Under the single-label annotation scheme adopted in this study, both the AI model and the hadith researcher were required to identify only one dominant theme, even when multiple themes were simultaneously present within the same hadith. Such overlap naturally increases the likelihood of annotation disagreement(Tsoumakas & Katakis, 2009).

From the perspective of classical Islamic scholarship, Ibadah and Muamalah are commonly regarded as subdomains of Syariah, whereas Aqidah and Akhlak constitute separate major dimensions of Islamic teachings. This conceptual relationship suggests that the boundary between Ibadah and Muamalah is inherently less distinct than the boundaries involving Aqidah or Akhlak. Therefore, the lower agreement observed in the Muamalah category should not necessarily be interpreted as a weakness of the AI-assisted labeling approach, but rather as evidence of the inherent complexity of thematic hadith classification. Future research may investigate alternative annotation schemes, such as a broader three-category taxonomy (Aqidah–Syariah–Akhlak) or multi-label classification, to better represent hadiths containing multiple thematic dimensions (Handayani et al., 2023).

3.2 Performance of TF-IDF + SVM

The TF-IDF + SVM model achieved an Accuracy of 73.1%, a Macro F1-score of 70.1%, and a Weighted F1-score of 73.0% on the testing dataset. These results indicate that the conventional feature-based approach using TF-IDF remains effective for classifying Indonesian hadiths into the four thematic categories.

Table 3. Overall Performance of the TF-IDF + SVM Model

Metric	Value
Accuracy	0.731
Macro F1-score	0.701
Weighted F1-score	0.730

Based on the classification report on Table 4, the Ibadah category achieved the highest F1-score (0.810), whereas the Akhlak category obtained the lowest F1-score (0.538). The high performance of the Ibadah category indicates that hadiths related to worship practices tend to contain distinctive lexical patterns and terminology, making them easier for the TF-IDF + SVM model to distinguish. In contrast, the lower performance observed for the Akhlak category suggests that its textual characteristics overlap with those of other thematic categories, resulting in a higher likelihood of misclassification.

These findings are consistent with the characteristics of TF-IDF, which relies primarily on word frequency rather than contextual semantics. When different thematic categories share similar vocabulary, TF-IDF-based representations become less discriminative, making it more difficult for the classifier to separate semantically related classes. Similar observations have been reported in previous studies on hadith classification using TF-IDF and SVM, where overlapping

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

textual features and class imbalance were identified as factors affecting classification performance (Ramadhan et al., 2024). Overall, the results indicate that the TF-IDF + SVM approach remains effective for thematic hadith classification despite these challenges.

Table 4. Classification Report of the TF-IDF + SVM Model

Category	Precision	Recall	F1-Score	Support
Aqidah	0.718	0.684	0.701	250
Ibadah	0.799	0.820	0.810	384
Muamalah	0.750	0.763	0.756	497
Akhlak	0.546	0.529	0.538	189

The relatively strong performance achieved by TF-IDF + SVM suggests that lexical features remain highly informative for thematic hadith classification. Many hadiths belonging to the Ibadah category contain explicit worship-related terminology, allowing the linear classifier to separate them effectively. In contrast, Akhlak-related hadiths frequently share vocabulary with Muamalah and Aqidah, resulting in overlapping feature representations and lower classification performance.

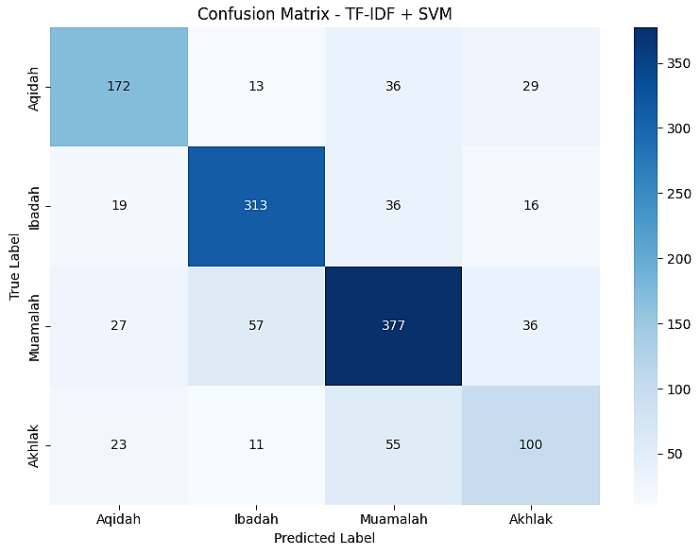

Figure 3. Confusion Matrix of the TF-IDF + SVM Model

Figure 3 illustrates the confusion matrix of the TF-IDF + SVM model. Most correctly classified instances are concentrated along the main diagonal, indicating that the model was able to distinguish the four thematic categories reasonably well. The Ibadah category achieved the highest number of correct predictions (313 out of 384 instances), followed by Muamalah (377 out of 497 instances). In contrast, the Akhlak category recorded the lowest number of correct predictions (100 out of 189 instances), which is consistent with its lower F1-score presented in Table 3.

The confusion matrix also reveals that the most frequent misclassification occurred between the Akhlak and Muamalah categories, where 55 Akhlak hadiths were incorrectly classified as Muamalah. Similarly, 57 Muamalah hadiths were predicted as Ibadah, while 36 Aqidah hadiths were classified as Muamalah. These results suggest that several thematic categories share similar lexical characteristics, making them difficult to distinguish using TF-IDF features based solely on term frequency.

Overall, the confusion matrix confirms that the TF-IDF + SVM model performs well in recognizing hadiths with distinctive thematic vocabulary, particularly in the Ibadah category. However, categories with greater semantic overlap, especially Akhlak and Muamalah, remain more challenging to classify accurately, indicating that lexical feature representations alone are insufficient to completely separate semantically related thematic categories.

3.3 Performance of IndoBERT

The fine-tuned IndoBERT model achieved an Accuracy of 72.3%, a Macro F1-score of 69.7%, and a Weighted F1-score of 72.2% on the testing dataset. As shown in Table 5, the overall performance of IndoBERT was comparable to that of TF-IDF + SVM. The relatively small difference between the Macro F1-score and Weighted F1-score indicates that the model maintained a reasonably balanced performance across the four thematic categories despite the moderate class imbalance in the dataset.

Table 5. Overall Performance of the IndoBERT Model

Metric	Value
Accuracy	0.723

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

Metric	Value
Macro F1-score	0.697
Weighted F1-score	0.722

Based on the classification report shown in Table 6, the Ibadah category achieved the highest F1-score (0.793), whereas the Akhlak category recorded the lowest F1-score (0.583). This result suggests that hadiths related to worship practices are relatively easier for IndoBERT to classify, whereas the semantic boundaries between the Akhlak category and other thematic categories remain less distinct. Consequently, some hadiths belonging to the Akhlak category were more likely to be classified into other categories.

Table 6. Classification Report of the IndoBERT Model

Category	Precision	Recall	F1-Score	Support
Aqidah	0.706	0.624	0.662	250
Ibadah	0.737	0.859	0.793	384
Muamalah	0.796	0.708	0.750	497
Akhlak	0.555	0.614	0.583	189

Unlike TF-IDF, IndoBERT learns contextual representations through the Transformer architecture, enabling the model to capture semantic relationships beyond word frequency. However, the results indicate that this capability did not lead to a substantial improvement in classification performance on the dataset used in this study. This outcome may be attributed to the relatively limited dataset size for fine-tuning, the moderate class imbalance, and the semantic overlap among thematic categories. The performance of Transformer-based models is influenced by several factors, including dataset size, annotation quality, and domain characteristics (Taufiqurrahman et al., 2021). Overall, the results demonstrate that IndoBERT achieved competitive performance for thematic hadith classification, although it did not outperform the conventional TF-IDF + SVM approach under the experimental setting adopted in this study.

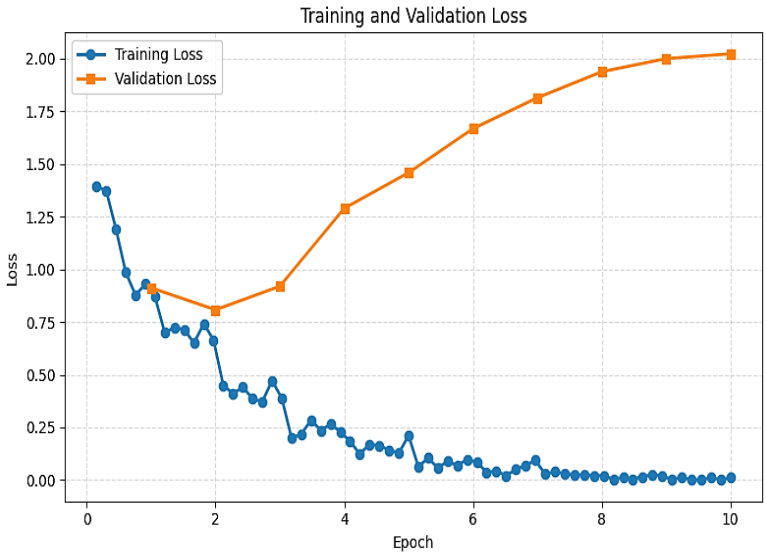

Figure 4. Training and Validation Loss during IndoBERT Fine-Tuning

Figure 4 illustrates the training and validation loss during the fine-tuning process of IndoBERT. The training loss decreased steadily throughout training, while the validation loss improved during the early epochs before increasing in later epochs. Therefore, the best-performing model selected based on the Macro F1-score was retained for the final evaluation.

Although Transformer-based models generally outperform conventional approaches on large-scale NLP tasks, their effectiveness strongly depends on the availability of sufficient training data and high-quality annotations. With only 6,600 labeled hadiths available for fine-tuning, the contextual representation learned by IndoBERT may not have reached its full potential. Consequently, the performance gain over the conventional TF-IDF + SVM approach was limited under the experimental setting adopted in this study.

Figure 5 presents the confusion matrix of the fine-tuned IndoBERT model. Similar to the TF-IDF + SVM model, most correctly classified instances are located along the main diagonal, indicating that the model successfully identified a large proportion of hadiths in each thematic category. The Ibadah category achieved the highest number of correct predictions (330 out of 384 instances), followed by Muamalah (352 out of 497 instances). In contrast, the Aqidah category recorded the lowest number of correct predictions (156 out of 250 instances), while the Akhlak category correctly classified 116 out of 189 instances.

TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

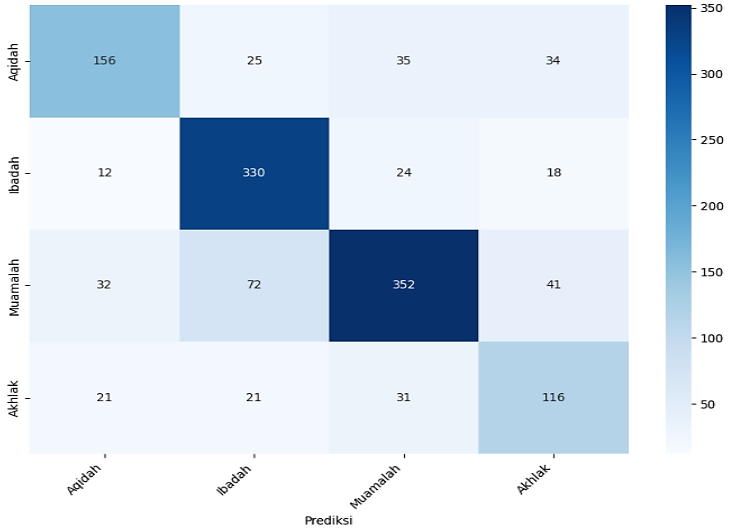

Figure 5. Confusion Matrix - IndoBERT (fine-tuned)

The confusion matrix further shows that the most frequent misclassification occurred between the Muamalah and Ibadah categories, where 72 Muamalah hadiths were incorrectly predicted as Ibadah. In addition, 35 Aqidah hadiths were classified as Muamalah, and 34 Aqidah hadiths were misclassified as Akhlak. These results indicate that several thematic categories exhibit semantic similarities, making them difficult to distinguish even when contextual representations learned by the Transformer architecture are employed.

Overall, the confusion matrix suggests that although IndoBERT is capable of capturing contextual semantic information, semantic overlap among thematic categories remains a significant challenge. This observation is consistent with the overall evaluation results, where IndoBERT achieved competitive classification performance but did not outperform the conventional TF-IDF + SVM approach under the experimental setting adopted in this study.

3.4 Discussion

Table 7 compares the overall performance of the two classification approaches evaluated in this study. The TF-IDF + SVM model achieved slightly higher scores than IndoBERT across all evaluation metrics, obtaining an Accuracy of 73.1%, Macro F1-score of 70.1%, and Weighted F1-score of 73.0%. In comparison, the fine-tuned IndoBERT model achieved an Accuracy of 72.3%, Macro F1-score of 69.7%, and Weighted F1-score of 72.2%. Although the performance gap between the two models is relatively small, the results indicate that the conventional TF-IDF + SVM approach slightly outperformed the Transformer-based model under the experimental setting adopted in this study.

Table 7. Performance Comparison of TF-IDF + SVM and IndoBERT

Model	Accuracy	Macro F1-score	Weighted F1-score
TF-IDF + SVM	0.731	0.701	0.730
IndoBERT	0.723	0.697	0.722

Several factors may explain these findings. First, the dataset used in this study consists of 6,600 Indonesian-translated Sahih Bukhari hadiths, which is relatively moderate in size for fine-tuning a Transformer-based language model. Transformer models generally require larger and more diverse datasets to fully exploit their contextual learning capabilities. In contrast, TF-IDF + SVM often performs well on medium-sized text classification datasets because it relies on discriminative lexical features rather than deep contextual representations. In addition, the dataset exhibits a moderate class imbalance, with the Muamalah category containing substantially more instances than the Akhlak category. Such imbalance may reduce the effectiveness of deep learning models during fine-tuning, particularly for minority classes.

From a practical perspective, TF-IDF + SVM also offers several advantages over Transformer-based models. The conventional model requires substantially lower computational resources, shorter training time, and simpler deployment while producing classification performance comparable to IndoBERT. These characteristics make TF-IDF + SVM particularly suitable for institutions with limited computational infrastructure.

Conversely, IndoBERT provides contextual representations that may become increasingly beneficial as larger and more diverse datasets become available. Therefore, the selection between conventional and Transformer-based approaches should consider not only model complexity but also dataset characteristics, computational resources, and application requirements.

Overall, the findings demonstrate that a more sophisticated model does not necessarily guarantee superior classification performance. Under the experimental conditions of this study, the conventional TF-IDF + SVM approach achieved slightly better performance than the Transformer-based IndoBERT model. These results highlight the


TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

importance of selecting classification methods based on dataset characteristics and practical deployment requirements rather than model complexity alone.

4. CONCLUSION

This study successfully constructed an Indonesian thematic hadith dataset consisting of 6,600 Sahih Bukhari hadiths using an AI-assisted labeling approach with Gemini 2.5 Flash. The dataset was systematically organized into four thematic categories, namely aqidah, ibadah, akhlak, and muamalah, and subsequently employed to evaluate the performance of two text classification approaches: the conventional TF-IDF combined with Support Vector Machine (SVM) and the Transformer-based IndoBERT model. Experimental results demonstrated that TF-IDF + SVM slightly outperformed the fine-tuned IndoBERT model, achieving an Accuracy of 73.1%, a Macro F1-score of 70.1%, and a Weighted F1-score of 73.0%, while IndoBERT obtained an Accuracy of 72.3%, a Macro F1-score of 69.7%, and a Weighted F1-score of 72.2%. These findings suggest that conventional feature-based methods remain highly competitive for thematic hadith classification, particularly when applied to a moderately sized and moderately imbalanced dataset. The primary contribution of this study is the development of an Indonesian thematic hadith dataset through an AI-assisted labeling strategy and the comprehensive comparison between conventional machine learning and Transformer-based classification methods. Nevertheless, the study is limited by the use of only Indonesian-translated Sahih Bukhari hadiths, which may reduce the generalizability of the findings to other hadith collections and thematic annotation frameworks. Future research should expand the dataset by incorporating additional hadith collections, improve annotation reliability through validation by multiple hadith scholars, and investigate more advanced Transformer architectures, ensemble learning, or data augmentation techniques to further enhance classification performance. Overall, this study demonstrates that AI-assisted labeling offers an efficient and practical approach for constructing thematic Indonesian hadith datasets, while highlighting that classification performance is influenced not only by model complexity but also by dataset characteristics and annotation quality.

REFERENCES

Adigunawan, & Fathoni, R. (2023). Machine Learning and Transformer-based Model for Sentiment Analysis of Indonesian E-Commerce Reviews. *Indonesian Journal of Computer Science*, *12*(2), 284–301. http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

Bartlett, C. W., Bossenbroek, J., Ueyama, Y., McCallinhart, P., Peters, O. A., Santillan, D. A., Santillan, M. K., Trask, A. J., & Ray, W. C. (2023). Invasive or More Direct Measurements Can Provide an Objective Early-Stopping Ceiling for Training Deep Neural Networks on Non-invasive or Less-Direct Biomedical Data. *SN Computer Science*, *4*(2), 1–12. https://doi.org/10.1007/s42979-022-01553-8

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I. H., Alfina, I., Putra, I. F., Rahmadani, S., … Purwarianti, A. (2023). NusaCrowd: Open Source Initiative for Indonesian NLP Resources. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 13745–13818. https://doi.org/10.18653/v1/2023.findings-acl.868

Darfiansa, L. S., Fitriyani, & Larasati, S. S. A. (2025). Optimizing IndoBERT for Revised Bloom's Taxonomy Question Classification Using Neural Network Classifier. *Journal of Information Systems Engineering and Business Intelligence*, *11*(2), 226–237. https://doi.org/10.20473/jisebi.11.2.226-237

Geng, S., Cooper, H., Moskal, M., Jenkins, S., Berman, J., Ranchin, N., West, R., Horvitz, E., & Nori, H. (2025). *JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models*. http://arxiv.org/abs/2501.10868

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), 1–3. https://doi.org/10.1073/pnas.2305016120

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756

Handayani, R. N., Najiyah, I., & Wisnuwardana, D. A. (2023). Classification of Bulughul Maraam Categories: Prohibitions, Recommendations, and Information Using Extreme Learning Machine and Fasttext. *Jurnal Online Informatika*, *8*(2), 242–251. https://doi.org/10.15575/join.v8i2.1205

Irugalbandara, C. (2024). *Meaning Typed Prompting: A Technique for Efficient, Reliable Structured Output Generation*. http://arxiv.org/abs/2410.18146

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, *10*(4), 1–68. https://doi.org/10.3390/info10040150

Lutfiah, D., Harahap, N. S., Yanto, F., & Cynthia, E. P. (2026). Klasifikasi Multi-Label Hadits Shahih Muslim Menggunakan Metode Support Vector Machine (SVM). *Bulletin of Data Science*, *5*(3), 158–169. https://doi.org/10.47065/bulletinds.v5i3.10147


TIN: Terapan Informatika Nusantara
Vol 7, No 2, July 2026, page 1182-1191
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v7i2.10743

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). *Deep Learning Based Text Classification: A Comprehensive Review*. *1*(1), 1–43. http://arxiv.org/abs/2004.03705

Nabiilah, G. Z., Al Faraby, S., & Dwifebri Purbolaksono, M. (2021). Classification of Hadith Topic of Indonesian Translation Using K-Nearest Neighbor and Chi-Square. *International Journal on Information and Communication Technology (IJoICT)*, *7*(2), 11–22. https://doi.org/10.21108/ijoict.v7i2.573

Ni'mah. SM, K., Arifi, A., & Abror, I. (2024). Hadith as a Source of Islamic Law: Its Role and Significance. *Studi Multidisipliner: Jurnal Kajian Keislaman*, *11*(2), 193–204. https://doi.org/10.24952/multidisipliner.v11i2.13302

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 1–14. https://doi.org/10.1038/s41598-024-56706-x

Ramadhan, T. I., Supriatman, A., & Kurniawan, T. R. (2024). Passage Retrieval untuk Question Answering Bahasa Indonesia Menggunakan BERT dan FAISS. *Jurnal Algoritma*, *21*(2), 156–163. https://doi.org/10.33364/algoritma/v.21-2.2100

Rasenda, R., Fabrianto, L., & Faizah, N. M. (2024). Optimizing Hadith Classification With Neural Networks: a Study on Bukhari and Muslim Texts. *JIKO (Jurnal Informatika Dan Komputer)*, *7*(2), 145–149. https://doi.org/10.33387/jiko.v7i2.8732

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. http://arxiv.org/abs/2402.07927

Sunendar, N. S., & Saputra, I. (2025). Comparative Performance of Transformer and Lstm Models for Indonesian Information Retrieval With Indobert. *Jurnal Pilar Nusa Mandiri*, *21*(2), 228–233. https://doi.org/10.33480/pilar.v21i2.6920

Susanti, S., Najiyah, I., Ramdhani, Y., Herliana, A., Muckti, M. K., & Oktaviani, F. R. (2025). Searching Sahih Hadiths Based on Queries using Neural Models and FastText. *Journal of Applied Data Sciences*, *6*(1), 272–285. https://doi.org/10.47738/jads.v6i1.467

Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*, *54*(August), 100664. https://doi.org/10.1016/j.cosrev.2024.100664

Taufiqurrahman, F., Faraby, S. Al, & Purbolaksono, M. D. (2021). Klasifikasi Teks Multi Label pada Hadis Terjemahan Bahasa Indonesia Menggunakan Chi Square dan SVM. *E-Proceeding of Engineering*, *8*(5), 10650–10659.

Tsoumakas, G., & Katakis, I. (2009). Multi-Label Classification: An Overview. *Database Technologies: Concepts, Methodologies, Tools, and Applications: Volumes 1-4*, *1*, 309–319. https://doi.org/10.4018/978-1-60566-058-5.ch021

Zakiah, R., Ahmad, N., Safaat Harahap, N., Agustian, S., Iskandar, I., Sanjaya, S., Studi, P., Informatika, T., Sains, F., & Teknologi, D. (2025). *MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison of Random Forest and Long Short-Term Memory Performance in Multilabel Text Classification of Bukhari Hadith Translation* . *5*(July), 862–874.