SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 229-240, DOI: 10.59190/stc.v6i3.373

Interpretative comparative analysis of LSTM and random forest for multi-label classification of English Qur'an translation

Nur Delifah, Nazruddin Safaat Harahap, Okfalisa, Elvia Budianita
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

The rapid growth of digital Qur'anic resources has created a need for automated systems capable of accurately categorizing verses by thematic content. The thematic complexity of Qur'anic text, in which a single verse may simultaneously convey multiple moral, spiritual, and social messages, presents a significant challenge for automated classification systems. This study conducts a comparative and explainable evaluation of long short-term memory (LSTM) and random forest (RF) for multi-label classification of English Qur'an translations across six thematic categories: arkanul Islam, iman, amal, human and community relations, akhlak, and history and story. To address severe class imbalance, synthetic minority over-sampling technique (SMOTE) was applied per label, expanding the training set from 4,489 to 19,658 samples. LSTM captured sequential contextual relationships through integer token embeddings, while RF relied on TF-IDF vector representations. Evaluated on 1,248 unseen test verses, RF achieved a higher macro F1-score (0.2748) compared to LSTM (0.2432), while LSTM retained marginally higher accuracy (79.61% vs. 79.55%). Per-label analysis revealed that both models performed best on lexically explicit labels such as arkanul Islam and iman, but consistently failed on abstract categories such as akhlak, where LSTM recorded near-zero recall of 0.61% and RF only 6.10%. This study contributes empirical evidence that TF-IDF-based SMOTE interpolation is more effective for minority-class augmentation than token-sequence interpolation, and demonstrates that macro F1-score is a more appropriate evaluation metric than accuracy for imbalanced multi-label religious text classification.

ARTICLE INFO

Article history:

Received May 26, 2026
Revised Jun 7, 2026
Accepted Jun 8, 2026

Keywords:

LSTM
Multi-Label Classification
Qur'an Translation
Random Forest
SMOTE

* **Corresponding Author**
E-mail address: nazruddin.safaat@uin-suska.ac.id

1. INTRODUCTION

The Qur'an is the holy book of Muslims and is the main and first source of Islamic teachings according to the beliefs of Muslims, and its truth is recognized. This holy book contains the words (revelations) of Allah which were conveyed by the angel Gabriel to the Prophet Muhammad as the messenger of Allah in stages [1]. The main purpose of the Qur'an is to guide human life, helping Muslims live in accordance with moral and spiritual values so they can achieve well-being both in this world and the hereafter [2]. Because of this, the Qur'an is not only read but also studied and reflected upon in daily life.

The Qur'an is used as the main guideline in the lives of Muslims. In addition, the Qur'an is also a miracle given by Allah SWT to the Prophet Muhammad SAW. This holy book consists of 6236 verses spread across 114 surahs and 30 juz [3]. With more than two billion Muslims reading the Qur'an daily as their sacred religious text. This shows how deeply the Qur'an is embedded in the daily lives of Muslims, not only as a source of religious teachings, but also as guidance in moral values,

social interactions, and personal reflection [4]. With the rapid development of digital Islamic resources, many English translations of the Qur'an are now widely accessible through online platforms and digital applications. In addition, several studies and institutions have categorized Qur'anic verses into thematic topics to help readers better understand the meaning and context of the verses [5]. Thematic classification can support educational activities, Qur'anic studies, information retrieval systems, and intelligent Islamic learning applications.

The importance of translating the Qur'an into various languages, especially English, lies in its role in helping Muslims around the world understand Islamic teachings more clearly. Research conducted by Ananda Pane & Syahrul Mubarok (2018) states that one example is the translation of the Qur'an which classifies its verses into 15 topics [6]. This kind of classification can help readers focus on certain themes, making the learning process more practical and efficient. It also supports educational activities, both in formal and informal settings. To address these challenges, machine learning and deep learning methods have increasingly been applied in Qur'anic text classification research [7]. Due to the large number of verses and the complex meaning of the text, a computational method that is able to process text data effectively is required [8]. Long Short-Term Memory (LSTM), a deep learning architecture derived from Recurrent Neural Network (RNN), is widely used in text processing because of its ability to capture sequential and contextual information within sentences, making it suitable for use in the processing of Qur'anic texts that have complex structures and meanings [9].

Previous studies on the detection of hate speech in Urdu Roman texts from Facebook comments have shown that both machine learning and deep learning methods can be used for text classification. A wide range of supervised machine learning methods are used as comparisons but the findings generally indicate that deep learning methods tend to perform better. In particular, the CNN model yielded an accuracy of 95.1% and LSTM reached 96.2%, which is higher than the classical machine learning method [10]. This suggests that deep learning models are more capable of understanding patterns in textual data, especially when dealing with complex language structures. In another study, Akbar et al. (2024) examined the classification of Qur'an translations and found that the LSTM model was able to achieve an accuracy above 90%, along with a relatively stable F1-score across different labels [11]. In the context of more complex multi-label classification, research on the multi-label toxic comment dataset (2023) showed that the CNN-BiLSTM model with an attention mechanism was able to achieve an accuracy of 78.92% and a weighted F1-score of 0.86, suggesting that the LSTM is highly effective in capturing relationships between labels while understanding the linguistic context of the text in depth [12].

Even though LSTM shows strong performance, classical machine learning methods such as Random Forest (RF) remain important to consider [13]. One of the most widely used methods is Random Forest (RF), which is an ensemble learning-based algorithm with high ability to handle large-dimensional data and reduce the risk of overfitting [14]. However, its application in multi-label classification, especially in Quranic text data, is still limited. This method has advantages in terms of simplicity, computational efficiency, and ease of interpretation of results, and often performs well on text datasets with limited sizes [15]. Several studies also demonstrate that RF has a very competitive performance in multi-label classification[16, 17]. A recent study in text classification on the 20 Newsgroups dataset shows that RF has a competitive performance compared to other algorithms such as Naive Bayes, SVM, and Logistic Regression with an accuracy of 89.3% and an F1-score of 88.1%, These results confirm that RF has superior capabilities in handling high-dimensional and sparse text data, as well as providing a balance between performance and computational efficiency [18].

Previous studies on Qur'anic text classification mostly focused on using a single model or standard classification methods [19]. Although some studies have discussed multi-label classification, only a few have compared deep learning and ensemble learning methods on imbalanced data. In addition, the use of SMOTE in different feature representations, such as LSTM token sequences and TF-IDF vectors in Random Forest, is still rarely discussed. Most previous studies also focused mainly on accuracy, even though accuracy alone is not sufficient for evaluating imbalanced multi-label classification.

This study addresses these challenges by implementing and comparing LSTM and Random Forest (RF) models for multi-label classification of English Qur'an translations, with the addition of Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. LSTM is selected

for its capability to capture sequential and contextual relationships in text, while RF with TF-IDF is included as a computationally efficient baseline that performs well on high-dimensional sparse text features. In addition, this research analyzes how different feature representations influence the effectiveness of SMOTE in handling imbalanced multi-label religious text classification.

2. RESEARCH METHODS

2.1. Research Stages

This study conducted a classification comparison of the English Qur'an translation dataset using two approaches: LSTM and RF with SMOTE. The implementation used Python with TensorFlow/Keras for LSTM and Scikit-learn along with imbalanced-learn for Random Forest and SMOTE. The general pipeline consisted of: (1) data loading and exploration, (2) text preprocessing, (3) train/validation/test splitting with an 72-8-20 ratio, (4) feature extraction, (5) SMOTE oversampling per label on training data, (6) model training, and (7) evaluation on unseen test data.

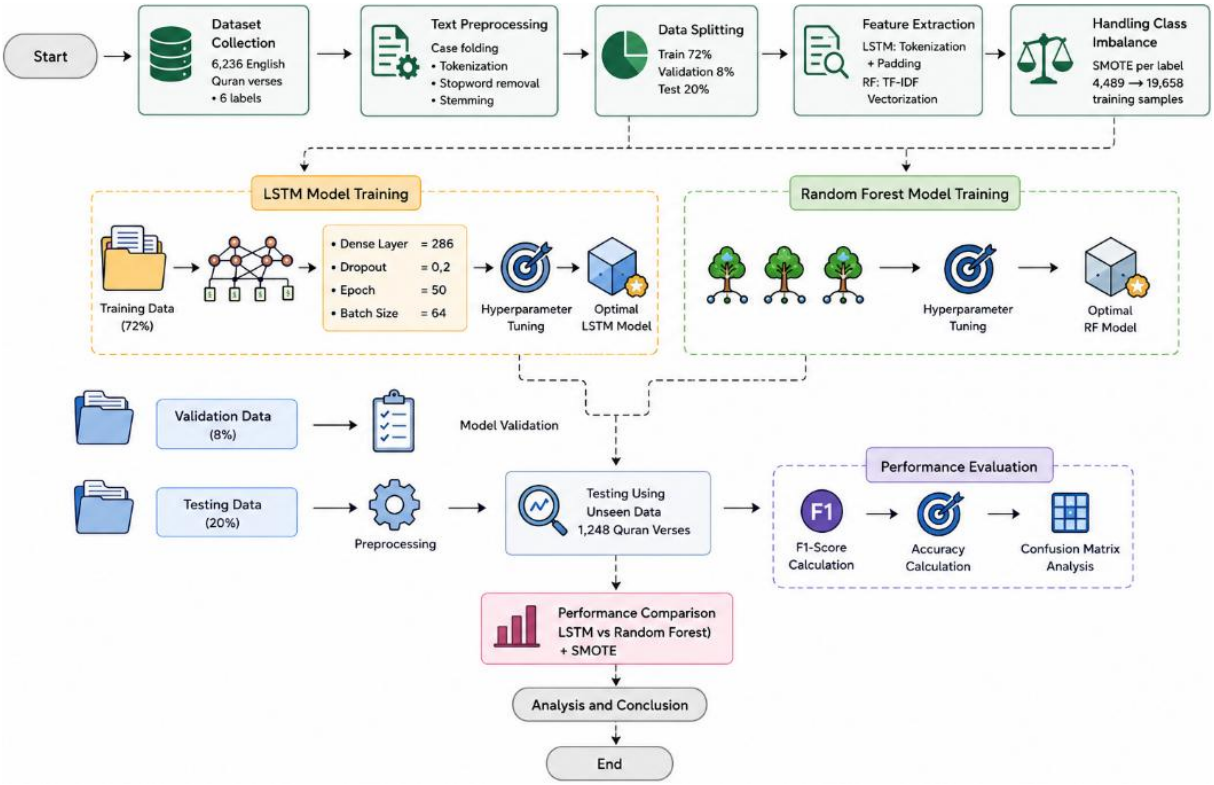

Figure 1. Research stages.

Figure 1 illustrates the research workflow used in this study, starting from the collection of 6,236 English Quran translation verses with six labels. The data then underwent preprocessing, data splitting, feature extraction, and class imbalance handling using SMOTE. Furthermore, the LSTM and Random Forest models were trained and optimized through hyperparameter tuning to obtain the best model. Finally, the models were evaluated using unseen test data with F1-score, accuracy, and confusion matrix analysis to compare the classification performance of both methods.

2.2. Dataset

The dataset used in this study is an English translation of the Quran [20]. Each verse in the dataset was labeled with one or more topics from a total of 15 categories, namely: Arkanul Islam, Iman, Al-Qur'an, Sciences, Amal, Dakwah, Jihad, Human and Community Relations, Akhlak, Rules Relating to Property, Matters Relating to the Law, Country and Society, Agriculture and Trade, History and Story, and Religions [6]. However, this study only focuses on six selected labels: Arkanul

Islam (c1), Iman (c2), Amal (c5), Human and Community Relations (c8), Akhlak (c9), and History and Story (c14). Table 1 presents the distribution of positive and negative samples per label, illustrating the degree of class imbalance present in the dataset.

Table 1. Dataset distribution per label.

Label	Positive (1)	Negative (0)	Imbalance ratio
Arkanul Islam (c1)	3,425	2,811	0.82
Iman (c2)	2,288	3,948	1.73
Amal (c5)	718	5,518	7.69
Human and community relations (c8)	624	5,612	8.99
Akhlak (c9)	769	5,467	7.11
History and story (c14)	676	5,560	8.22

As shown in Table 1, the imbalance ratio ranges from 0.82 (c1, already near-balanced) to 8.99 (c8), confirming the need for oversampling strategies before training. The classification approach used in this study is multi-label classification [21]. where each verse can belong to more than one class depending on its content [22]. This approach is suitable for representing the complexity of Quranic text, as a single verse may contain multiple themes. Therefore, multi-label classification is used to assign several relevant labels to each verse simultaneously [23].

2.3. Text Preprocessing

Text preprocessing is an important initial stage in document summarization and grouping because it aims to make data more structured and easy for the model to process. This process includes several steps, starting from tokenization to breaking text into words [24]. Then case folding (turning all letters into lowercase), punctuation removal, removal of stopwords, and the stemming process to change the word to its basic form [25].

The entire preprocessing process is applied to both the training data and the test data. Effective preprocessing can significantly enhance the performance of classification models by reducing noise and improving the quality of the input data [26]. The resulting clean text was used as input for both the LSTM tokenizer and TF-IDF vectorizer.

2.4. Feature Extraction

Feature extraction of raw data is an important stage that requires careful observation to produce optimal data representation [27]. Feature extraction is the process of extracting important features from preprocessed text data [28]. In the context of Qur'anic text classification, feature extraction aims to capture relevant semantic and syntactic characteristics. In the LSTM model, feature extraction is done by converting text into a sequence of numbers using a tokenizer. Each word is indexed based on its occurrence, then converted into a sequence and equated with the padding process. This representation helps LSTM understand the pattern of word order in the text.

Meanwhile, in the Random Forest model, feature extraction is carried out through Term Frequency–Inverse Document Frequency (TF-IDF). This method converts text into a numerical vector based on the frequency of occurrence of words in the document and the level of uniqueness of the word in the entire corpus [29].

2.5. Data Splitting

The dataset was randomly shuffled using a fixed seed (random_state=42) to ensure reproducibility, then split into three subsets: 72% training (4,489 verses), 8% validation (499 verses), and 20% testing (1,248 verses). The TF-IDF vectorizer and LSTM tokenizer were fitted exclusively on training data to prevent data leakage into validation and test sets.

2.6. Handling Class Imbalance with SMOTE

Given the significant class imbalance across labels (Table 1), Synthetic Minority Over-sampling Technique (SMOTE) was applied per label on the training data. Since the task involves

multi-label classification, SMOTE cannot be applied globally; instead, it was applied independently for each label using k_neighbors=5. Labels with fewer than 6 positive samples were skipped, and labels already near-balanced (c1) were also excluded. Table 2 summarizes the SMOTE results applied to both LSTM and RF training data.

Table 2. SMOTE application results on training data.

Label	Original positive	SMOTE result	Final positive
Arkanul Islam (c1)	2,439	SKIP (already balanced)	2,439
Iman (c2)	1,642	+1,205 synthetic	2,847
Amal (c5)	524	+3,441 synthetic	3,965
Human and community relations (c8)	458	+3,573 synthetic	4,031
Akhlak (c9)	536	+3,417 synthetic	3,953
History and story (c14)	478	+3,533 synthetic	4,011

As a result, the training set expanded from 4,489 to 19,658 samples after SMOTE, providing a more balanced representation for minority labels such as Amal, Human and Community Relations, Akhlak, and History and Story.

2.7. Long Short Term Memory

Long Short Term Memory (LSTM) is a model proposed by Hochreiter and Schmidhuber in 1997 [30]. LSTM is one of the methods in deep learning developed from the Recurrent Neural Network (RNN), with the addition of cell memory to store information in the long term, which has been proven to work more accurately than RNN [31].

This algorithm consists of several units, including cell state, gate units, and gate output. Cell state is used to store information that will be forwarded to the next stage. Gate units are tasked with processing stored information, whether it will be forwarded or discarded. Inside gate units, there are input gates and forget gates. The input gate serves to determine the input value to be channeled to the cell state, while the forget gate is used to process the information in the cell state. In the final stage, the output gate determines the output value generated [32].

The LSTM architecture consisted of an embedding layer with 100 dimensions, a single LSTM layer with 100 hidden units, dropout regularization of 0.2, and a sigmoid-activated dense output layer for multi-label prediction. The model was trained using the Adam optimizer and binary cross-entropy loss with 50 epochs and a batch size of 64, while ModelCheckpoint was used to save the best model by validation accuracy and EarlyStopping with patience=5 to prevent overfitting.

2.8. Random Forest

Random Forest (RF) is a bagging method introduced by Breiman [33], which combines many Decision Tree models to produce better predictions. Each tree is built using randomly selected data and features, resulting in a more diverse model [34]. This approach reduces errors and results in more stable classification performance than using just one Decision tree [35].

In the process, Random Forest combines the predicted results of all trees using the majority voting method, which is to select the most results as the final output. If the Random Forest consists of multiple trees, then the class of a data is determined based on the class that appears most often from the prediction results of each treeN [36], which is formulated as follows:

$$A = arg\ max_c(\sum_{n=1}^{N} I\ (h_n(y) = c)) \quad (1)$$

Description:
I : Indicator function
h_n : the decision tree in the Random Forest n
c : class
$l(y)$: class prediction results for data y

Random Forest has several advantages, such as being able to improve accuracy despite missing data and outliers, and efficient data management [37]. The RF model used Multi Output

Classifier wrapping a Random Forest Classifier (n_estimators=100, class_weight='balanced', random_state=42) to natively handle multi-label output. The class_weight parameter was set to 'balanced' as an additional safeguard alongside SMOTE.

2.9. Evaluation

At this stage, the trained model is tested using data that has never been used before, namely testing data of 1,248 verses. The goal is to see how well the model is in grouping the translation of Qur'an verses into predetermined categories. Model performance was evaluated per label using Precision (2), Recall (3), Accuracy (4), and F1-Score (5).

$$Precision = \frac{\sum_{i=1}^{l} TPi}{\sum_{i=1}^{l} FPi + TPi} x\ 100\% \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{l} TPi}{\sum_{i=1}^{l} TPi + FNi} x\ 100\% \quad (3)$$

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l} \times 100\% \quad (4)$$

$$F1 - Score = \frac{2 \times recall \times precision}{recall + precision} \quad (5)$$

Remarks:

- TPi (True Positive) is the amount of positive data that has been successfully correctly classified for class i.
- TNi (True Negative) is the amount of negative data that has been correctly classified by the system.
- FPi (False Positive) is the amount of negative data that has been misclassified as positive by the system.
- FNi (False Negative) is the amount of positive data that has been misclassified as negative by the system.

3. RESULTS AND DISCUSSIONS

This section presents and discusses the test set performance of two competing approaches: LSTM as a sequential deep learning model, and Random Forest (RF) as an ensemble method, both trained with SMOTE augmentation. The analysis goes beyond aggregate metrics by examining each model's behavior across six Qur'anic thematic labels, revealing how architectural differences translate into classification outcomes on semantically complex religious text.

Table 3. LSTM test results.

Classes	Data test			
	Precision	Recall	F1-Score	Accuracy
Arkanul Islam (c1)	0.7143	0.4178	0.5272	56.89%
Iman (c2)	0.6221	0.2316	0.3375	66.35%
Amal (5)	0.4857	0.1181	0.1899	88.38%
Human and community relations (8)	0.3810	0.0650	0.1111	89.74%
Akhlak (9)	1.0000	0.0061	0.0121	86.94%
History and story (14)	0.5000	0.1955	0.2811	89.34%
Average	0.6172	0.1724	0.2432	79.61%

Table 3 presents the LSTM classification results on 1,248 unseen test verses. The model recorded an average accuracy of 79.61% and a macro F1-score of 0.2432. semantically explicit labels such as Arkanul Islam (c1) and Iman (c2) yielded the strongest F1-scores at 0.5272 and 0.3375 respectively, reflecting that these categories carry more explicit and recurring vocabulary patterns. In contrast, Akhlak (c9) collapsed to an F1-score of just 0.0121 — a near-total classification failure. The

model achieved 100% precision but only 0.61% recall on c9, meaning it almost never predicted the Akhlak label, committing only when highly certain. This behavior persisted despite SMOTE generating 3,417 synthetic c9 training samples, indicating that interpolating integer token sequences does not produce semantically coherent representations for abstract moral concepts — a fundamental limitation of applying SMOTE in the discrete token space used by LSTM.

Table 4 presents the RF classification results on the same 1,248 test verses. RF achieved an average accuracy of 79.55% and a macro F1-score of 0.2748, marginally surpassing LSTM in F1. As with LSTM, labels with greater positive support performed comparatively better: Arkanul Islam (c1: F1=0.4807) and Iman (c2: F1=0.4350) led among all categories. Notably, RF managed 6.10% recall on Akhlak (c9) — a modest but meaningful gain over LSTM's near-zero 0.61% recall on the same label. This improvement stems from SMOTE operating in the continuous TF-IDF vector space, where feature interpolation is geometrically meaningful. Across minority labels, RF consistently yielded higher precision: c5 (0.7059 vs. 0.4857), c8 (0.6875 vs. 0.3810) — indicating that when RF commits to a positive prediction, its TF-IDF-based decision boundaries are more discriminative than LSTM's token-based boundaries.

Table 4. RF test results.

Classes	Data test			
	Precision	Recall	F1-score	Accuracy
Arkanul Islam (c1)	0.6751	0.3733	0.4807	53.61%
Iman (c2)	0.5911	0.3442	0.4350	66.91%
Amal (5)	0.7059	0.0833	0.1491	89.02%
Human and community relations (8)	0.6875	0.0894	0.1583	90.62%
Akhlak (9)	0.7143	0.0610	0.1124	87.34%
History and story (14)	0.5577	0.2180	0.3135	89.82%
Average	0.6553	0.1949	0.2748	79.55%

Table 5 presents a direct metric-by-metric comparison between LSTM and RF on the test set, distilling the core trade-offs between the two approaches.

Table 5. Comparative performance: LSTM vs. RF on test data.

Metric	LSTM	RF	Best model
Avg. F1-score (test)	0.2432	0.2748	RF
Avg. accuracy (test)	79.61%	79.55%	LSTM
Avg. precision (test)	0.6172	0.6553	RF
Avg. recall (test)	0.1724	0.1949	RF

The results indicate that RF slightly outperforms LSTM in terms of average macro F1-Score on test data (0.2748 vs. 0.2432), while LSTM achieves marginally higher accuracy (79.61% vs. 79.55%). A critical finding of this study is that both models applied SMOTE identically—expanding training from 4,489 to 19,658 samples—yet produced dramatically different training behavior: RF memorized the augmented data nearly perfectly (F1 train: 99.69%), while LSTM, restrained by EarlyStopping at Epoch 5, achieved only 35.78% F1 on training data. This difference stems from the nature of each learning paradigm: RF decision trees partition the feature space exhaustively until leaves are pure, naturally fitting any training distribution including synthetic samples; LSTM with a fixed architecture and early stopping inherently regularizes itself. The result is that RF suffers more acute overfitting compared to LSTM.

Both models demonstrated significant difficulty in generalizing to unseen data, rooted in two compounding structural challenges inherent to this dataset. First, severe label imbalance: labels such as Amal (c5), Human and Community Relations (c8), Akhlak (c9), and History and Story (c14) each contain fewer than 15% positive samples in the original dataset, making them structurally difficult to learn even after SMOTE augmentation. Second, semantic overlap: Qur'anic verses often simultaneously address multiple themes — a single verse may contain both moral guidance (Akhlak)

and community instruction (c8) — making clean decision boundaries elusive regardless of the modeling approach. These two factors together explain why both models struggle to achieve high F1 performance on minority labels despite producing reasonable overall accuracy.simultaneously belong to multiple categories, making clear decision boundaries difficult for any model to learn regardless of SMOTE augmentation.

A key methodological distinction between the two models lies in the feature space where SMOTE operated. For RF, SMOTE interpolated within a 4,551-dimensional TF-IDF vector space, where each dimension carries interpretable term-frequency meaning and interpolation produces geometrically plausible synthetic samples. For LSTM, SMOTE interpolated over 250-dimensional integer token sequences, where averaging two token indices does not necessarily produce a linguistically coherent word or context. This fundamental incompatibility between discrete token representations and continuous interpolation-based oversampling may be a key reason why SMOTE provided more tangible benefits to RF than to LSTM. Nevertheless, neither model fully overcame the gap between training distribution and test performance, confirming that class imbalance mitigation alone is insufficient when the primary challenge is semantic complexity.

These findings align with prior research on multi-label Qur'anic text classification, which consistently documents difficulty in generalizing across semantically complex and thematically overlapping label spaces. The present study adds to this body of evidence by demonstrating that the choice of feature representation — TF-IDF versus token sequences — directly influences how effectively oversampling techniques such as SMOTE can mitigate class imbalance. Furthermore, the results affirm that macro F1-score, rather than accuracy, should be the primary evaluation criterion in multi-label classification tasks with severe label imbalance.

Figure 2 presents the confusion matrix generated from the LSTM model using SMOTE-balanced training data. The visualization illustrates the classification capability of the LSTM architecture in understanding sequential and contextual Quranic text representations.

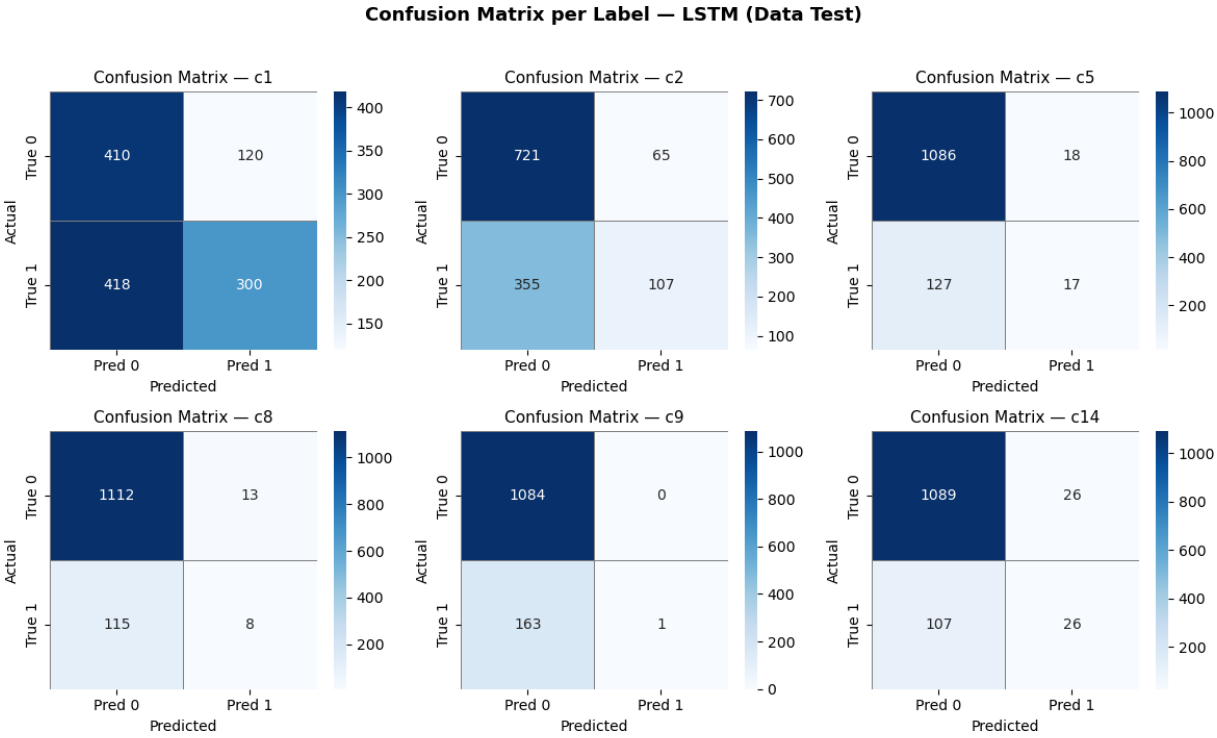

Figure 2. Confusion matrix per label for LSTM.

As shown in Figure 2, LSTM exhibits a comparatively more balanced distribution between class 0 (negative) and class 1 (positive) predictions across several labels. This reflects LSTM's capacity to capture contextual and sequential relationships within Qur'anic text through its gating mechanism, enabling it to detect positive-class patterns with greater sensitivity than a bag-of-words

approach. On labels such as Arkanul Islam (c1) and History and Story (c14), LSTM achieves noticeable True Positive rates, suggesting that sequential context provides relevant discriminative signals for thematically distinct categories.

Nevertheless, LSTM still generates substantial False Negative counts for abstract labels such as Akhlak (c9) and Human and Community Relations (c8). These categories are characterized by diffuse, context-dependent moral or social themes that do not manifest through consistent surface-level vocabulary — a challenge that neither sequential learning alone nor SMOTE augmentation fully resolves.

Figure 3 shows the confusion matrix visualization generated from the Random Forest model after applying SMOTE oversampling. The matrix illustrates prediction performance for each thematic label in the test dataset.

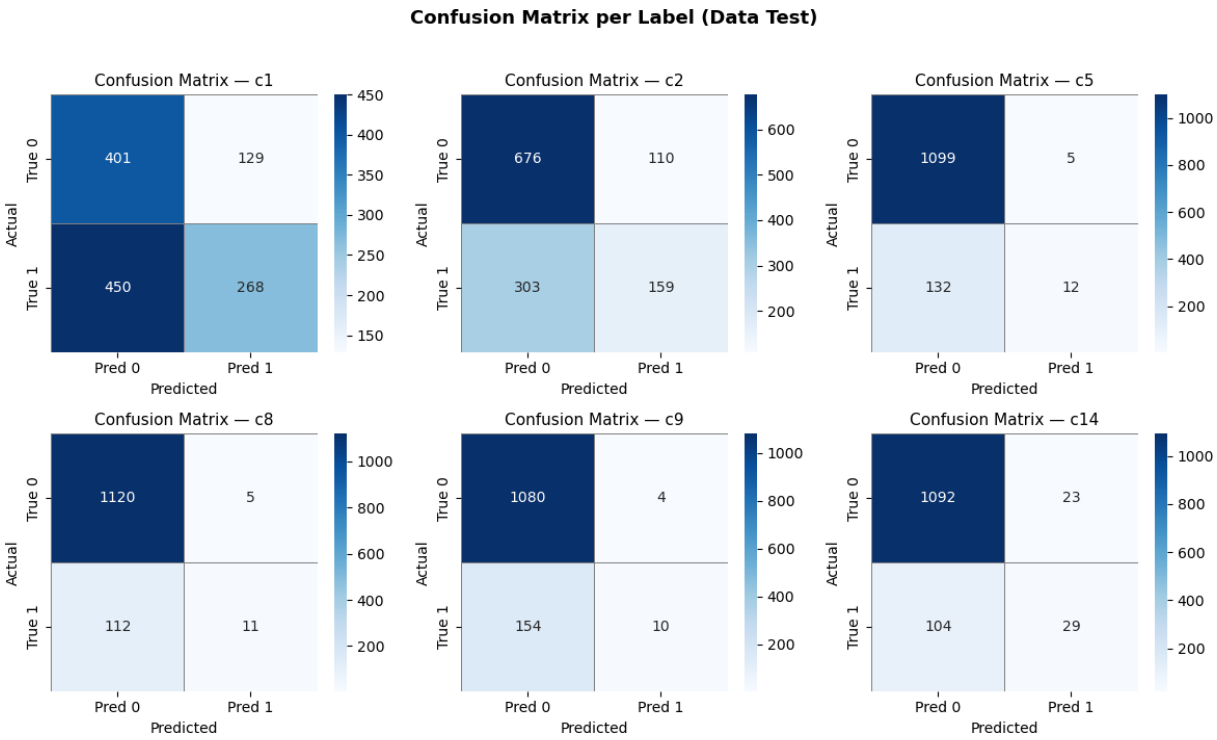

Figure 3. Confusion matrix per label for RF.

Figure 3 reveals that Random Forest exhibits a clear skew toward predicting the majority class (class 0). Large True Negative counts are evident across all six labels, particularly in Akhlak (c9), Human and Community Relations (c8), and History and Story (c14). While this majority-class dominance sustains relatively high accuracy, it comes at the cost of minority-class detection: False Negative counts remain substantial across nearly all labels, indicating that many relevant Qur'anic verses are systematically missed. This pattern confirms that despite SMOTE expanding positive-class training samples from a few hundred to over 4,000 per label, RF's TF-IDF-based decision trees still gravitate toward majority-class partitions at inference time.

RF performs comparatively better on Arkanul Islam (c1) and Iman (c2), where thematic vocabulary is more explicit and repetitive — terms related to pillars of Islam and matters of faith tend to cluster in identifiable TF-IDF dimensions. This underscores a key strength of RF with TF-IDF: it excels when thematic signal is lexically concentrated but weakens when meaning is distributed across abstract contextual cues, as in Akhlak and Human and Community Relations.

Table 6 presents selected false negative cases from the RF model's classification of Akhlak (c9). Each verse carries a ground-truth label of 1 (positive Akhlak) but was predicted as 0 by the model. These misclassifications reveal a consistent pattern: verses expressing moral guidance through prayer, patience, or supplication are systematically overlooked, likely because their vocabulary overlaps with other thematic categories such as Iman and Human and Community Relations.

Table 6. Classification errors in data test in akhlak class with RF method.

Translation	Data label	Prediction result
And their saying was no other than that they said: Our Lord! forgive us our faults and our extravagance in our affair and make firm our feet and help us against the unbelieving people.	1	0
And seek assistance through patience and prayer, and most surely it is a hard thing except for the humble ones,	1	0
Those who believe and whose hearts are set at rest by the remembrance of Allah; now surely by Allah's remembrance are the hearts set at rest.	1	0

This indicates that the model struggles to capture abstract moral semantics. In addition, these verses may also be related to other categories, so there is an overlap of meanings that makes the model less accurate in classification.

Table 7. Classification errors in data test in akhlak class with RF method.

Translation	Data label	Prediction result
He said: My Lord knows best what you do.	0	1
Will they not then turn to Allah and ask His forgiveness? And Allah is Forgiving, Merciful.	0	1
And there were in the city nine persons who made mischief in the land and did not act aright.	0	1

Table 7 presents false positive cases: verses that carry no Akhlak label yet were predicted as Akhlak by RF. These errors arise because certain vocabulary items — such as words related to forgiveness, moral conduct, or supplication — are semantically adjacent to the Akhlak category and trigger positive predictions. This pattern exposes the core challenge of TF-IDF-based classification: it cannot distinguish whether a word signals a theme or merely appears in context, leading to misclassification when vocabulary overlaps across label boundaries.

Across both models, a consistent gradient of classification difficulty emerges: labels with explicit and recurring thematic vocabulary — Arkanul Islam (c1) and Iman (c2) — are learned more readily, while semantically diffuse labels such as Amal (c5), Akhlak (c9), and Human and Community Relations (c8) resist accurate classification regardless of the modeling approach. This gradient reflects an inherent property of Qur'anic text: its most abstract guidance is also its most contextually variable, making it the hardest target for any feature-based classification system. Both LSTM and RF must contend with this reality, and neither, in the present configuration, fully overcomes it.

4. CONCLUSION

This study compared LSTM and Random Forest (RF) with SMOTE augmentation for multi-label classification of English Qur'anic translations across six thematic labels. On the test set of 1,248 verses, RF achieved a slightly higher macro F1-score of 0.2748 compared to LSTM's 0.2432, while LSTM retained marginally higher accuracy at 79.61% versus RF's 79.55%. Across both models, classification performance was strongest on thematically explicit labels — Arkanul Islam (c1) and Iman (c2) — and weakest on semantically abstract labels such as Akhlak (c9), Human and Community Relations (c8), and Amal (c5). Despite SMOTE expanding training data from 4,489 to 19,658 samples, neither model achieved satisfactory F1-scores on minority labels, indicating that the primary bottleneck is semantic complexity rather than class imbalance alone.

Four key findings emerge from this study. First, SMOTE successfully balanced the training distribution and enabled both models to learn some positive-class patterns, particularly on labels such as Arkanul Islam (c1) and History and Story (c14) where SMOTE-generated samples were most linguistically coherent. Second, RF benefited more directly from SMOTE augmentation than LSTM, owing to the geometric meaningfulness of TF-IDF vector interpolation compared to integer sequence interpolation used by LSTM. Third, Akhlak (c9) proved the most resistant label for both models — LSTM nearly refused to predict it (0.61% recall), while RF managed only 6.10% — confirming that

abstract moral semantics lie beyond the reach of current feature representations. Fourth, the marked divergence between accuracy (~79%) and macro F1-score (< 0.28) across both models underscores that accuracy is an unreliable primary metric for this imbalanced multi-label task.

These results reinforce that macro F1-score is the appropriate primary metric for imbalanced multi-label classification tasks, as accuracy consistently overestimates model effectiveness when negative-class predictions dominate. Future work should prioritize contextual embedding models such as BERT or transformer-based architectures capable of capturing the deep semantic structure of Qur'anic text, alongside multi-label-aware oversampling techniques that go beyond per-label SMOTE. Exploring label correlation mechanisms and attention-based classification heads may further improve the recognition of overlapping and abstract thematic categories.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our colleagues and mentors at Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia, for their valuable support, insights, and constructive feedback throughout this research. We also gratefully acknowledge the anonymous reviewers and the editorial team for their helpful suggestions, which have contributed to improving the quality of this article.

REFERENCES

[1] Zahraini, H. & Muslehuddin, M. (2021). *Studi Al-Qur'an dan Hadis*. Sanabil.

[2] Daulay, S. S., Suciyandhani, A., Sofian, S., Julaiha, J., & Ardiansyah, A. (2023). Pengenalan Al-Quran. *Jurnal Ilmiah Wahana Pendidikan*, **9**(5), 472–480.

[3] Yasir, M. & Jamaruddin, A. (2016). *Studi Al-Quran*. Asa Riau (CV. Asa Riau).

[4] bin Ahmad, K. & Huda, D. M. S. (2023). The role of reading the Al-Quran on peace of mind. *Focus*, **4**(1), 39–44.

[5] M Alashqar, A. (2024). A classification of Quran verses using deep learning. *International Journal of Computing and Digital Systems*, **16**(1), 1041–1053.

[6] Pane, R. A., Mubarok, M. S., & Huda, N. S. (2018). A multi-lable classification on topics of quranic verses in english translation using multinomial naive bayes. *2018 6th International Conference on Information and Communication Technology (ICoICT)*, 481–484.

[7] Bashir, M. H., Azmi, A. M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2021). Arabic natural language processing for Qur'anic research: a systematic review. *Artif. Intell. Rev.*, **56**(7), 6801–6854.

[8] Zulkarnaen, I. & Lhaksmana, K. M. (2025). Klasifikasi Multilabel pada Topik ayat Al-Qur'an Menggunakan Random Forest dan Naïve Bayes. *E-Proceeding Eng.*, **12**(2), 3231.

[9] Akbar, I. (2023). *Klasifikasi multi-label terjemahan Al-Qur'an bahasa Indonesia menggunakan model Long Short-Term Memory*. Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim.

[10] Naseeb, A., Zain, M., Hussain, N., Qasim, A., Ahmad, F., Sidorov, G., & Gelbukh, A. (2025). Machine learning-and deep learning-based multi-model system for hate speech detection on Facebook. *Algorithms*, **18**(6), 331.

[11] Akbar, I., Faisal, M., & Chamidy, T. (2024). Multi-label classification of Indonesian qur'an translation using long short-term memory model. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 119–128.

[12] Belal, T. A., Shahariar, G. M., & Kabir, M. H. (2023). Interpretable multi labeled bengali toxic comments classification using deep learning. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6.

[13] Ezz, M., Sharaf, M. A., & Hassan, A. A. A. (2019). Classification of Arabic writing styles in ancient Arabic manuscripts. *International Journal of Advanced Computer Science and Applications*, **10**(10).

[14] Rahayu, K., Fitria, V., Septhya, D., Rahmaddeni, R., & Efrizoni, L. (2023). Klasifikasi Teks untuk Mendeteksi Depresi dan Kecemasan pada Pengguna Twitter Berbasis Machine Learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, **3**(2), 108–114.

[15] Alyasiri, O. M. & Cheah, Y. N. (2025). Multi-Class Text Classification using Machine Learning Techniques. *Engineering, Technology & Applied Science Research*, **15**(3), 22598–22604.
[16] Aftari, D. P. (2024). Perbandingan performa klasifikasi terjemahan Al-Qur'an menggunakan metode random forest dan long short term memory. *J. Comput. Syst. Inform. JoSYC*, **5**(3), 567.
[17] Ardiansyah, R., Yuliansyah, H., & Yudhana, A. (2025). Multi-Label Opinion Mining Based on Random Forest with SMOTE and ADASYN. *Compiler*, **14**(2), 65–75.
[18] Venkateshwarlu, G., Akhila, S., Kavyasree, V., Vishnu, S., & Prasad, V. S. (2024). Enhanced Text Classification Using Random Forest: Comparative Analysis and Insights on Performance and Efficiency. *Int. J. Comput. Eng. Res. Trends*, **11**, 1–8.
[19] Ezz, M., Sharaf, M. A., & Hassan, A. A. A. (2019). Classification of Arabic writing styles in ancient Arabic manuscripts. *International Journal of Advanced Computer Science and Applications*, **10**(10).
[20] Shakir, M. H. (2025). The Holy Quran. URL: https://www.theholyquran.org.
[21] Mediamer, G., Adiwijaya, & Faraby, S. A. (2019). Development of Rule-Based Feature Extraction in Multi-label Text Classification. *Int. J. Adv. Sci. Eng. Inf. Technol.*, **9**(4), 1460.
[22] Nurfikri, F. S. & Adiwijaya. (2019). A comparison of Neural Network and SVM on the multi-label classification of Quran verses topic in English translation. *Journal of Physics: Conference Series*, 1192(1), 012030.
[23] Adeleke, A., Azah Samsudin, N., Hisyam Abdul Rahim, M., Kamal Ahmad Khalid, S., & Efendi, R. (2021). Multi-label classification approach for Quranic verses labeling. *Indonesian Journal of Electrical Engineering and Computer Science*, **24**(1), 484–490.
[24] Hamed, S. K. & Ab Aziz, M. J. (2018). Classification of holy quran translation using neural network technique. *Journal of Engineering and Applied Sciences*, **13**(12), 4468–4475.
[25] Ahmed, M. A., Baharin, H., & Nohuddin, P. E. (2023). K-means variations analysis for translation of English Tafseer Al-Quran text. *Int. J. Electr. Comput. Eng.*, **13**(3), 3255–3265.
[26] Shrivash, B. K., Verma, D. K., & Pandey, P. (2025). A Novel Framework for Text Preprocessing using NLP Approaches and Classification using Random Forest Grid Search Technique for Sentiment Analysis. *Economic Computation & Economic Cybernetics Studies & Research*, **59**(2).
[27] Azim, K., Tahir, A., Shahroz, M., Karamti, H., Vazquez, A. A., Vistorte, A. R., & Ashraf, I. (2025). Ensemble stacked model for enhanced identification of sentiments from IMDB reviews. *Scientific Reports*, **15**(1), 13405.
[28] Armaya, A. M. R. (2024). Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan. *Jurnal Teknologi Informasi*, **3**(1), 13–23.
[29] Xu, Q. (2025). Application of an intelligent English text classification model with improved KNN algorithm in the context of big data in libraries. *Systems and Soft Computing*, **7**, 200186.
[30] Sari, W. K., Rini, D. P., Malik, R. F., & Azhar, I. S. (2017). Klasifikasi teks multilabel pada artikel berita menggunakan long short-term memory dengan Word2Vec. *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, **1**(3), 276–285.
[31] Yuspriyadi, F. (2023). Klasifikasi Sentimen Twitter Menggunakan Lstm. *METHODIKA: Jurnal Teknik Informatika dan Sistem Informasi*, **9**(1), 4–8.
[32] Leong, L. A. (2023). Forecasting SOXX with Long Short-Term Memory (LSTM) Neural Networks Based on Varying Sampling Periods. Doctoral dissertation, Selinus University of Sciences and Literature.
[33] Leo, B. (2001). Random Forests. *Kluwer Acad. Publ.*, 5–32.
[34] Sarawan, K., Polpinij, J., Somprasertsri, G., Rojarath, A., & Luaphol, B. (2025). Multiclass Classification Approach for Detecting Software Bug Severity Level from Bug Reports. ICIC Express Letters, **16**(5), 567–576.
[35] Afianto, M. F., Adiwijaya, & Al-Faraby, S. (2018). Text categorization on hadith Sahih Al-Bukhari using random forest. *Journal of Physics: Conference Series*, **971**(1), 012037.
[36] Saputro, D. (2023). Cable News Network (CNN) Articles Classification Using Random Forest Algorithm with Hyperparameter Optimization. *BAREKENG J. Ilmu Mat. Terap.*, 17(2), 0847.
[37] Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Repository UIN Sultan Syarif Kasim Riau*, **5**(1), 122–130.