

Multi-Label Theme-Based Classification of Quranic Verses for Urdu Translations A Case Study for Surah Al-Baqarah and Aal-e-Imran

Rao Muhammad Waseem Ashraf

Department of Computer Science, Allama Iqbal Open University (AIOU), Islamabad, Pakistan. raowaseem.biit@gmail.com

Dr. Muhammad Arshad Awan

Department of Computer Science, Allama Iqbal Open University (AIOU), Islamabad, Pakista

Afrasiab Kaikobad

Department of Computer Science, Barani Institute of Information Technology (BIIT), Rawalpindi, Pakistan

Dr. Muhammad Zeeshan Muzaffar

Department of Computer Science, Barani Institute of Information Technology (BIIT), Rawalpindi, Pakistan

ABSTRACT

The Holy Quran stands as our timeless source of guidance, which explores multiple deep subjects for understanding all aspects of human life and moral behavior. The research focuses on the Quranic verses from the Urdu translation of Surah Al-Baqarah and Aal-e-Imran into different themes to help with better topic analysis. The research aims to develop a thematic classification system for Quranic verses by utilizing machine learning techniques and evaluating the system's performance through standard classification metrics. Using XLM-Roberta Classifier is the powerful tool model to translate in Urdu language by Junagharri from the Tanzil website, it translates it with standard measure of performance. This model give produces the 0.9501 F1 score along with 0.8936 accuracy, 0.9726 precision, and 0.9286 recall rates. The result shows the performance and excellent working when they handling multiple labels for analyzing religious texts. This research provides a detail knowledge to scholars and students of the Quranic verses by different useful tools and open the doors for the progress and progress in using AI for religious studies.

Keywords— *Holy Quran; Quranic themes classification; Urdu translation; Multi-label classification; XLM-RoBERTa; Machine learning; Religious text analysis.*

Introduction

Introduction

The Holy Quran is the central religious text of Islam, believed by Muslims to be the literal word of God revealed to the Prophet Muhammad (PBUH). It encompasses a wide range of themes, including theology, ethics, law, personal conduct, and social interactions. For Muslims, it serves as an everlasting source of guidance and forms the foundation of their spiritual, moral, and social framework. However, understanding the Quran deeply requires more than a surface reading; it calls for thematic interpretation to uncover interlinked concepts such as morality, worship, commandments, and stories embedded across different Surahs.

Traditionally, scholars have dedicated years to studying the Quran to organize its themes for educational and research purposes. Yet, the complexity of its structure and the depth of its meanings pose challenges in systematically categorizing and retrieving specific information. With the advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), there is growing potential to apply computational techniques to religious texts for deeper exploration and accessibility.

The motivation behind this research stems from recognizing that, although many Muslims interact with the Quran daily, the ability to systematically search and understand its themes, especially in non-Arabic languages like Urdu, remains limited. Urdu is spoken by a vast population, particularly in South Asia, where millions engage with the Quran through its Urdu translation. Despite this, existing research has largely focused on Arabic and, to some extent, English translations, leaving Urdu translations underexplored.

Moreover, most Quranic classification models developed so far adopt a single-label approach, where each verse is assigned only one theme, ignoring the multifaceted nature of many verses. A single verse might simultaneously touch on spiritual teachings, historical narratives, and moral values. To address this limitation, the present research proposes a multi-label classification approach that better reflects the thematic richness of the Quranic text. This model enhances interpretative understanding and facilitates easier identification of verses relevant to specific topics or questions.

The primary objective of this study is to develop an effective NLP-based framework for multi-label thematic classification of Urdu-translated Quranic verses, using state-of-the-art deep learning models. Specifically, this research targets Surah Al-Baqarah and Surah Aal-e-Imran due to their diverse thematic content, including discussions on faith, law, stories of past nations, social guidance, and divine commandments.

The Urdu translation utilized is derived from the authentic translation by Maulana Junagadhi, obtained from Tanzeel.net. The thematic categorization is based on the 15 themes outlined in Mus'haf Al-Tajweed by Dar-al-Ma'rifah, covering a comprehensive range of topics such as beliefs, ethics, worship, moral relations, politics, economy, society, and historical accounts. A manually curated dataset was created by aligning each verse in the Urdu translation with appropriate themes, resulting in a multi-label format where each verse may be associated with multiple binary labels.

One major contribution of this work is the manual curation and annotation of this dataset, addressing a vital gap in the availability of labeled data for Urdu Quranic analysis. The dataset is organized into 19 columns, including Surah and Ayah identifiers, Urdu text, and 15 binary theme labels, enabling scalable model development and evaluation.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

The linguistic differences between Arabic and Urdu introduce unique challenges in translation interpretation, tokenization, and semantic representation. Urdu, being morphologically rich and syntactically different from Arabic, presents additional complexities for NLP tasks, especially for models pre-trained predominantly on other languages.

To address these challenges, this study employs the XLM-RoBERTa transformer model, a multilingual contextual language model excelling in capturing semantic relationships across diverse languages, including Urdu. The model processes each Urdu-translated verse by tokenizing the text into subwords using SentencePiece tokenization, applying padding and attention masks to manage sequence variations, and feeding it through 12 transformer layers. The [CLS] token's final embedding is passed through a fully connected layer with sigmoid activation to predict the probability of each theme. A threshold of 0.5 is applied to determine theme assignments.

Due to class imbalance — where some themes occur more frequently than others — oversampling techniques were used during training to ensure fair learning and to reduce bias toward dominant classes. The model was trained on 80\% of the dataset and evaluated on the remaining 20\% using standard multi-label classification metrics: F1-score, precision, recall, and accuracy. The results show exceptional performance, achieving an F1-score of 0.9501, precision of 0.9726, recall of 0.9286, and an overall accuracy of 89.36\%.

The significance of this research lies not only in its technical achievements but also in its potential impact on Islamic studies, computational linguistics, and religious education. By enabling thematic exploration of Quranic verses in Urdu, the model makes the text more accessible to Urdu-speaking users interested in studying specific topics such as justice, worship, morality, or history. It provides a valuable tool for students, researchers, educators, and scholars, and paves the way for the development of intelligent Quranic search engines, question-answering systems, and educational platforms for Urdu speakers. Furthermore, it contributes to the broader goal of multilingual NLP research, extending the benefits of AI to underrepresented languages. At the same time, this study emphasizes the ethical use of AI in religious domains. Since Quranic texts are sacred, computational analysis must ensure accuracy, transparency, and scholarly integrity. Careful attention was given to maintaining the authenticity of translations and annotations, and the system was designed to support learning rather than replace traditional methods of Quran.

Section II presents existing literature related to Multi-Label Theme-Based Classification, and Section III describes the proposed technique. Section IV, presents the experimental setup and discusses the results. Finally, Section V concludes the given work.

Related Work

Several research studies have explored text analysis and thematic classification of Quranic verses using Natural Language Processing (NLP) techniques and algorithms. Mohamed and ElBehaidy [1] introduced a novel strategy for structuring Quranic verses using the Word2Vec model, focusing on the text analysis within Surah Al-Baqarah. Similarly, Nurfikri et al. [2] demonstrated that shared content and information gain significantly improve the performance of Naive Bayes in Quranic verse categorization

when compared with ANN and SVM, achieving a Hamming Loss of 0.0938. Their study highlighted the importance of feature selection in refining classification accuracy. Al-Hawarat [3] utilized Latent Dirichlet Allocation (LDA) for topic modeling to enhance the search functionality of the Arabic text of Surah Yousaf. This study explored the complexity of Arabic text and the difficulties in extracting meaningful themes, emphasizing the need for advancements in topic modeling for religious texts. Afzal and Mukhtar [4] introduced QuEST, a concept-based retrieval technique using a structured Quranic English WordNet, which outperformed conventional search tools in locating relevant verses and addressed the challenges of query-based searches.

Zouaoui and Rezeg [5] designed an ontology-based Quranic search engine capable of identifying keywords within a verse, improving user access to relevant results. Their approach combined manual and automated methods for ontology construction, surpassing traditional Quran search tools. Al Haromainy et al. [6] improved Quranic information retrieval through query expansion techniques, achieving enhanced recall, accuracy, and precision.

El Mouatasim and Oudaani [7] emphasized the relevance of Quranic information retrieval for both academic and public use. They developed a framework using MATLAB's Text Analytics Toolbox for data mining and NLP analysis of Arabic Quranic text. Similarly, Pain et al. [8] studied multi-label thematic classification of English translations of Quranic verses using the Multinomial Naive Bayes algorithm. Their work revealed that stemming reduced processing time without enhancing classification accuracy and reported a Hamming Loss of 0.1247.

Aiman et al. [9] examined thematic classification using word centrality analysis, identifying *ALLAH* as the most central term. Their findings indicated that Support Vector Machines (SVM) outperformed Naive Bayes, KNN, and Decision Trees in thematic classification, especially after applying stop-word removal and closeness centrality.

Huda et al. [10] explored Back Propagation Neural Networks (BPNN) for classifying Quranic verses and compared their performance with TAN and Naive Bayes. They addressed class imbalance using SMOTE and highlighted the importance of stemming and optimization techniques like stochastic gradient descent and Adam optimizer. Mubarak et al. [11] analyzed Mutual Information (MI) and Tree Augmented Naive Bayes (TAN) for feature selection in multi-label classification, showing that MI significantly reduced Hamming Loss.

Adeleke et al. [12] applied text classification techniques on 200 Quranic verses, showing how feature selection can improve classification performance. Elmitwally and Alsayat [13] utilized Support Vector Classifiers (SVC) and Gaussian Naive Bayes for multi-class classification across the first six Surahs, demonstrating the applicability of ML in Quranic text classification.

Mohammed et al. [14] reviewed various classification algorithms for Arabic Quranic verses and identified a gap in automated topical classification, calling for more diverse thematic datasets. Adeleke et al. [15] showed that group-based feature selection achieved 94.5% accuracy with the TF-IDF technique over 451 verses from Surahs Al-Baqarah and Al-Anaam. In a follow-up study [16], they reached 93.6% accuracy using CFS feature selection and chi-square filtering within 4.17 seconds.

Arkok and Zeki [17] focused on segmenting Tawheed/Shirk verses and distinguishing Meccan and Medinan chapters, applying SMOTE for dataset rebalancing and introducing the concept of Imbalanced Learning in Quranic topic classification.

Rahman et al. [18] implemented SVM and the J48 Decision Tree algorithm for Quranic topic classification, achieving over 85\% accuracy. Their research emphasized the value of computational techniques in extracting knowledge from Quranic texts and recommended extending these approaches to Hadith classification. Mohamed and Shokry [19] proposed QSST, a concept-oriented search engine for Quranic texts, and evaluated its performance against other retrieval tools.

The team of Hammo et al. [20] introduced QMT, a Quranic text classification tool that achieved 79\% precision. Their system began with direct word matching and data restructuring and planned to incorporate ontology-based representations in future developments.

Proposed Technique

In this work, we propose a transformer-based multi-label classification approach using XLM-RoBERTa for thematic categorization of Urdu-translated Quranic verses from Surah Al-Baqarah and Aal-e-Imran. This section outlines the complete mathematical formulation of the model pipeline, embedding mechanism, transformer operations, multi-label classification layer, and optimization process.

Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ be the dataset containing

N Urdu-translated verses, where $x^{(i)}$ denotes the input verse and $y^{(i)} \in \{0,1\}^{15}$ denotes the binary multi-label vector indicating the presence or absence of each of the 15 predefined themes.

Each verse $x^{(i)}$ first tokenized into a sequence of subword tokens using Sentence Piece tokenization:

$$x^{(i)} = [\omega_1^{(i)}, \omega_2^{(i)}, \omega_3^{(i)}, \dots, \omega_{n_i}^{(i)}] \quad (1)$$

Where n_i is the number of tokens in the i th verse.

Each token $\omega_j^{(i)}$ mapped to a unique integer ID using the

XLM-RoBERTa vocabulary:

$$t_j^{(i)} = \text{vocab}(\omega_j^{(i)}), \quad t_j^{(i)} \in \mathbb{Z} \quad (2)$$

To fit into a fixed-length input required by the transformer, each token sequence is padded or truncated to a maximum length L :

$$T^{(i)} = [t_1^{(i)}, t_2^{(i)}, t_3^{(i)}, \dots, t_L^{(i)}] \quad (3)$$

An attention mask $M^{(i)} \in \{0,1\}^L$ created such that:

$$M_j^{(i)} = \begin{cases} 1 & \text{if } j \leq n_i, \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Each token ID $t_j^{(i)}$ mapped to an embedding vector using an embedding matrix $\mathbb{R}^{|V| \times d}$, where $|V|$ is the vocabulary size and $d = 768$ is the hidden size:

$$\mathbf{e}_j^{(i)} = \mathbf{E}[t_j^{(i)}] \quad (5)$$

The input embeddings then passed through $L_T = 12$ layers of XLM-RoBERTa transformer blocks. Each transformer block includes multi-head self-attention (MHSA) and position-wise feed-forward networks (FFN).

For a given input $\mathbf{H}^{(l)} = [\mathbf{h}_1^{(l)}, \dots, \mathbf{h}_L^{(l)}]$ at layer l , the self-attention mechanism computes:

$$\mathbf{Q} = \mathbf{H}^{(l)}\mathbf{W}_Q, \mathbf{K} = \mathbf{H}^{(l)}\mathbf{W}_K, \mathbf{V} = \mathbf{H}^{(l)}\mathbf{W}_V \quad (6)$$

Where $\mathbf{W}_Q, \mathbf{W}_K$ and $\mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices.

The scaled dot-product attention then computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (7)$$

The output from multi-head attention is passed through a Feed-forward network:

$$\mathbf{z} = \text{ReLU}(\mathbf{h}\mathbf{W}_1 + \mathbf{b}_1) \quad (8)$$

$$\mathbf{h}' = \mathbf{z}\mathbf{W}_2 + \mathbf{b}_2 \quad (9)$$

Where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_f}$, $\mathbf{W}_2 \in \mathbb{R}^{d_f \times d}$ and $d_f = 3072$.

After the final transformer layer L_T , the representation of the [CLS] token is extracted as the verse-level embedding:

$$\mathbf{h}_{\text{CLS}}^{(i)} = \mathbf{H}^{(L_T)}[0] \quad (10)$$

This embedding fed into a fully connected output layer with sigmoid activations to produce multi-label outputs:

$$\mathbf{z}^{(i)} = \mathbf{h}_{\text{CLS}}^{(i)}\mathbf{W}_{FC} + \mathbf{b}_{FC} \quad (11)$$

$$\hat{y}_j^{(i)} = \sigma(z_j^{(i)}) = \frac{1}{1 + e^{-z_j^{(i)}}}, j = 1, \dots, 15 \quad (12)$$

The final predicted labels are obtained by thresh-holding the output probabilities:

$$\hat{y}_j^{(i)} = \begin{cases} 1 & \text{if } \hat{y}_j^{(i)} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

The model trained using the binary cross-entropy loss Summed over all labels:

$$\mathcal{L}^{(i)} = -\sum_{j=1}^{15} [y_j^{(i)} \log(\hat{y}_j^{(i)}) + (1 - y_j^{(i)}) \log(1 - \hat{y}_j^{(i)})] \quad (14)$$

The overall loss over the dataset averaged looks like:

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{i=1}^N \mathcal{L}^{(i)} \quad (15)$$

To handle class imbalance, we apply oversampling in the dataset such that rare labels are sufficiently represented. Let π_j be the class frequency of label j and define class weights:

$$\omega_j = \frac{1}{\pi_j + \epsilon} \quad (16)$$

Where ϵ is a small constant to prevent division by zero.

These weights incorporated into the loss:

$$\mathcal{L}_{\text{weighted}}^{(i)} = -\sum_{j=1}^{15} \omega_j [y_j^{(i)} \log(\hat{y}_j^{(i)}) + (1 - y_j^{(i)}) \log(1 - \hat{y}_j^{(i)})] \quad (17)$$

Optimization carried out using Adam Optimizer. Given gradient $\nabla \mathcal{L}$, the parameters θ are updated as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla, v_t = \beta_2 v_{t-1} + (1 - \beta_2) \nabla \mathcal{L}^2 \quad (18)$$

$$\theta_t = \theta_{t-1} - \eta \frac{m_t}{\sqrt{v_t} + \epsilon} \quad (19)$$

Where η is the learning rate.

The evaluation of model performance is done using micro- averaged Precision, Recall, and F1-score. Let

TP_j , FP_j , and FN_j be the true positives, false positives, and false negatives for label j :

$$\text{Precision}_{\text{micro}} = \frac{\sum_j TP_j}{\sum_j (TP_j + FP_j)} \quad (20)$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_j TP_j}{\sum_j (TP_j + FN_j)} \quad (21)$$

$$F1_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \quad (22)$$

This section provides a representation of the model's working pipeline, showing how Urdu-translated Quranic verses are processed and mapped to multiple thematic labels using deep learning principles based on XLM-RoBERTa transformer architecture.

Results And Performance Evaluation

Dataset Description

The dataset used in this study consists of Urdu translations of Quranic verses from Surah Al-Fatiha, Al-Baqarah, and Aal-e-Imran, translated by Maulana Muhammad Junagharri, obtained from the Tanzil Quran Project. This translation is widely trusted by scholars for its linguistic accuracy and adherence to classical Islamic interpretations. Each record in the dataset represents a single verse, annotated with one or more thematic labels. A total of 15 themes were used for multi-label classification, as derived from Mushaf Al-Tajwid by Dar-Al-Maarifah. These themes include:

1. Pillars of Islam (اسلام کے ستون)
2. Faith (ایمان)
3. The Call for Allah (اللہ کی طرف دعوت)
4. The Holy Quran (قرآن پاک)
5. Jihad (جہاد)
6. Action (Work) ((عمل(کام))
7. Man and Moral Relations (انسان اور اخلاقی تعلقات)
8. Man and Social Relations (انسان اور سماجی تعلقات)
9. Regulating Financial Relations (مالیاتی تعلقات کو منظم کرنا)
10. Trade, Agriculture, and Industry (تجارت، زراعت اور صنعت)
11. Juridical Relations (عدالتی تعلقات)
12. Political and General Relations (سیاسی اور عمومی تعلقات)
13. Sciences and Arts (سائنس اور آرٹس)
14. Religions (مذہب)
15. Stories and History (کہانیاں اور تاریخ)

The dataset contains the following fields:

id: Unique identifier for each verse

sura: Surah number

aya: Verse number within the Surah

translation: Urdu translation of the verse (by Maulana Junagharri)

15 binary columns: One for each theme (1 = verse belongs to the theme, 0 = otherwise)

Experimental Setup

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

The proposed model uses the XLM-RoBERTa transformer model for classification. The dataset was divided into training and testing sets, with 80% allocated for training and 20% for testing. The model is trained on a variety of themes and evaluated based on predicted theme distribution, evaluation metrics, and confusion metrics.

Performance Evaluation Metrics

To evaluate the performance of the model, we use key metrics such as recall, precision, and F1-Score. Recall (or Sensitivity) is the fraction of relevant themes retrieved from all possible relevant themes, and it is used to assess the coverage of the model. Precision (or Specificity) is the ratio of relevant themes retrieved over all retrieved themes, including both relevant and irrelevant tags. F1-Score provides a balance between precision and recall, capturing both the correctness and completeness of the model's predictions.

Results And Discussion

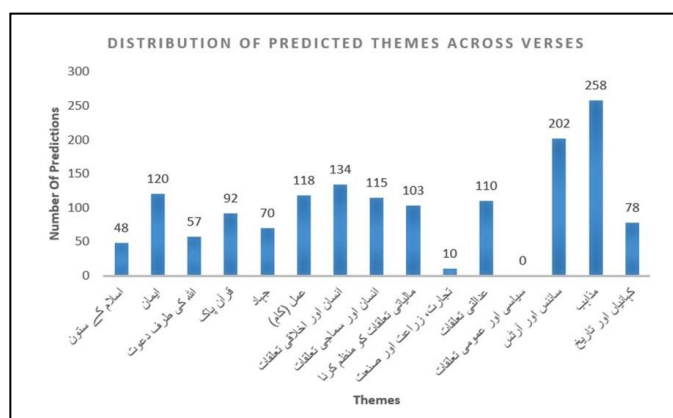


Figure 1: Predicted Themes

The XLM-RoBERTa model was evaluated on the Urdu translation of the Quran. Figure 1 shows the predicted themes distribution for verses, illustrating that themes such as "Religions" and "Science and Arts" are more frequently predicted in the test set. This suggests that these themes are more prevalent in the selected Quranic verses.

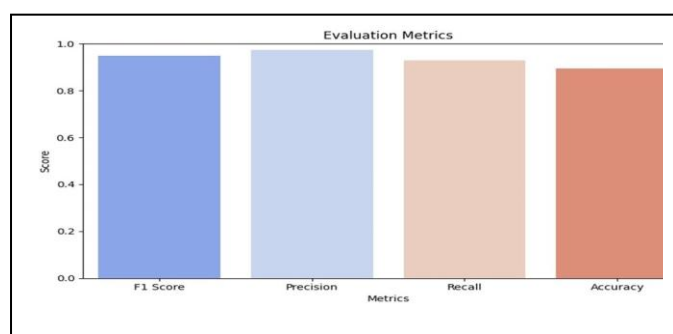


Figure 2: Evaluation Metrics

The evaluation metrics of the proposed model shown in Figure 2. The model achieves

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

an F1-Score of 0.95, accuracy of 89%, precision of 97%, and recall of 92%. These metrics demonstrate that the model performs exceptionally well, with particularly high precision and F1-scores. The pre-trained language representations used by XLM-RoBERTa contribute to the model's ability to achieve these high scores.

Confusion Matrix Analysis

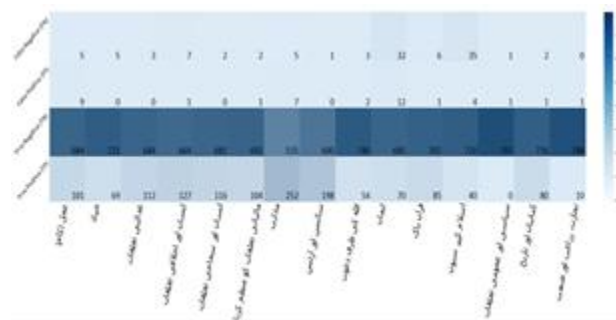


Figure 3: Confusion Matrix

Figure 3 shows the confusion metrics for all themes, providing a detailed breakdown of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The model exhibits strong performance in classifying themes such as "Religions" and "Man and the Moral Relations," with high TP and low FN rates. This indicates that the model successfully captures these themes.

In contrast, some themes such as "The Stories" and "The History" exhibit relatively lower TP rates and higher FN counts. This suggests that these themes may be difficult to distinguish due to their semantic similarity or the complexity of the verses associated with them.

Interestingly, the model shows consistently high TN values across all themes, indicating that it is precise in rejecting incorrect classifications. The low FP values suggest that the model rarely misclassifies verses into incorrect themes.

Insights And Challenges

While the model performs well in classifying frequently occurring themes, areas for improvement exist, particularly for themes with higher FN counts. These issues could arise from the difficulty in distinguishing between similar themes. To further improve the model's performance, strategies such as data augmentation, fine-tuned embeddings, or the incorporation of theme-specific attention mechanisms could be explored.

Conclusions

This research successfully developed and evaluated a multi-label theme-based classification system for Quranic verses in the Urdu translation of Surah Al-Baqarah and Aal-e-Imran, achieving exceptional performance metrics (F1-score: 0.9501, Accuracy: 0.8936, Precision: 0.9726, Recall: 0.9286). Through the combination of robust transformer based architectures (especially XLM-RoBERTa), and with the support of resampling strategies to reduce the effects of class imbalance, the present paper found that state-of-the-art methods in NLP are capable of labeling Quranic passages with various intersecting thematic topics with few errors. The main value

addition of the prospective study is the production of a tool to illustrate computational linguistics and Islamic studies by providing researchers, scholars, and students with an effective tool to estimating thematic analysis on religious texts, and thus turn into better understanding the interrelation of the related themes in the Quran. This work also contributes to the field of Quranic text analytics since a best practice pipeline that performs preprocessing, normalization, and annotation on religious text has been described here making machine learning data preparation of high quality. When looking at what can be done in the future, this work would be the solid base to extend the classification scheme to the full text of the Quran, perform cross-linguistic experiments with different translations of it, analyze larger transformer models such as T5 and GPT versions, and consider making semantic contextualization to better understand relationships between themes. Moreover, this will enable a more detailed interpretation of the concepts of the Quran, not only improvements on computational techniques but also religious studies. Overall, this paper not only establishes the fact that the use of transformer-based models could be effective when it comes to processing complicated multi-label classification in religious works, but it also creates new avenues when it comes to combining AI and Quranic studies, establishing a symbiotic relationship between natural language processing and Islamic science.

References

- E. H. Mohamed and W. H. El-Behaidy, "An ensemble multi-label themes based classification for holy qur'an verses using word2vec embedding," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3519- 3529, 2021.
- F. S. Nurfikri et al., "A comparison of neural network and svm on the multi-label classification of quran verses topic in english translation," in *Journal of Physics: Conference Series*, vol. 1192, no. 1. IOP Publishing, 2019, p. 012030.
- M. Alhawarat, "Extracting topics from the holy quran using generative models," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 12, pp. 288–294, 2015.
- H. Afzal and T. Mukhtar, "Semantically enhanced concept search of the holy quran: Qur'anic english wordnet," *Arabian Journal for Science and Engineering*, vol. 44, pp. 3953–3966, 2019.
- S. Zouaoui and K. Rezeg, "A novel quranic search engine using an ontology-based semantic indexing," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3653–3674, 2021.
- M. M. Al Haromainy, A. P. Sari, D. A. Prasetya, M. Subhan, A. Lisdiyanto, and T. Septianto, "Enhancing thematic holy quran verse retrieval through vector space model and query expansion for effective query answering," in *2023 IEEE 9th Information Technology International Seminar (ITIS)*. IEEE, 2023, pp. 1–6.
- A. El Mouatasim and J. Oudaani, "Topics classification of arabic text in quran by using matlab," in *International Conference on Big Data and Smart Digital Environment*. Springer, 2018, pp. 333–341.
- R. A. Pane, M. S. Mubarak, N. S. Huda et al., "A multi-lable classification on topics of quranic verses in english translation using multinomial naive bayes," in *2018 6th International Conference on Information and Communication Technology*

- (ICoICT). IEEE, 2018, pp. 481–484.
- A. S. Aiman, K. M. Lhaksmana et al., “Topic classification of quranic verses in english translation using word centrality measurement,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 803–809, 2022.
- N. S. Huda, M. S. Mubarak et al., “A multi-label classification on topics of quranic verses (english translation) using backpropagation neural network with stochastic gradient descent and adam optimizer,” in 2019 7th International conference on information and communication technology (ICoICT). IEEE, 2019, pp. 1–5.
- M. S. Mubarak, N. S. Huda et al., “A multi-label classification on topics of quranic verses in english translation using tree augmented naïve bayes,” in 2018 6th International Conference on Information and Communication Technology (ICoICT). IEEE, 2018, pp. 103–106.
- A. Adeleke, N. Samsudin, A. Mustapha, and S. A. Khalid, “Automating quranic verses labeling using machine learning approach,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 925–931, 2019.
- N. S. Elmitwally and A. Alsayat, “The multi-class classification for the first six surats of the holy quran,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 327–332, 2020.
- N. A.-K. Mohammed, B. Wahsheh, and I. Alsmadi, “A topical classification of quranic arabic text,” in *Proceedings of the NOORIC 2013: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 2013.
- A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawati, “A group-based feature selection approach to improve classification of holy quran verses,” in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018)*, Johor, Malaysia, February 06-07, 2018. Springer, 2018, pp. 282–297.
- A. Adeleke, N. Samsudin, Z. Othman, and S. A. Khalid, “A two-step feature selection method for quranic text classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 730–736, 2019.
- B. Arkok and A. Zeki, “Classification of holy quran verses based on imbalanced learning,” *International Journal on Islamic Applications in Computer in Science and Technology*, vol. 8, pp. 11–24, 2020.
- M. I. Rahman, N. A. Samsudin, A. Mustapha, and A. Abdullahi, “Comparative analysis for topic classification in juz al-baqarah,” *Indones. J. Electr. Eng. Comput. Sci*, vol. 12, no. 1, pp. 406–411, 2018.
- E. H. Mohamed and E. M. Shokry, “Qsst: A quranic semantic search tool based on word embedding,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 934–945, 2022.
- B. H. Hammo, A. Sleit, A. Baarah, and H. Abu-Salem, “A computational approach for identifying quranic themes,” *International Journal of Computer Processing of Languages*, vol. 24, no. 02, pp. 189–206, 2012.