Automated Authentication of Quranic Verses Using BERT (Bidirectional Encoder Representations from Transformers) based Language Models

Khubaib Amjad Alam[1,*] , **Maryam Khalid**[2] , **Syed Ahmed Ali**[2] , **Haroon Mahmood**[1]
Qaisar Shafi[2] , **Muhammad Haroon**[2] and **Zulqarnain Haider**[2]
[1] College of Engineering, Al Ain University, UAE
[2] National University of Computer and Emerging Sciences (FAST-NUCES), Pakistan
khubaib.alam@aau.ac.ae

Abstract

The proliferation of Quranic content on digital platforms, including websites and social media, has brought about significant challenges in verifying the authenticity of Quranic verses. The inherent complexity of the Arabic language, with its rich morphology, syntax, and semantics, makes traditional text-processing techniques inadequate for robust authentication. This paper addresses this problem by leveraging state-of-the-art transformer-based Language models tailored for Arabic text processing. Our approach involves fine-tuning three transformer architectures **BERT-Base-Arabic**, **AraBERT**, and **MarBERT** on a curated dataset containing both authentic and non-authentic verses. Non-authentic examples were created using sentence-BERT, which applies cosine similarity to introduce subtle modifications. Comprehensive experiments were conducted to evaluate the performance of the models. Among the three candidate models, **MarBERT**, which is specifically designed for handling Arabic dialects demonstrated superior performance, achieving an F1-score of 93.80%. **BERT-Base-Arabic** also showed competitive F1 score of 92.90% reflecting its robust understanding of Arabic text. The findings underscore the potential of transformer-based models in addressing linguistic complexities inherent in Quranic text and pave the way for developing automated, reliable tools for Quranic verse authentication in the digital era.

1 Introduction

With the rapid spread of Islamic content online, particularly on social media and websites, verifying the authenticity of Quranic verses has become a challenge. There are many verses that are mistakenly shared, without proper checks, which makes it hard for the users to confirm their authenticity. Traditional methods of identifying Quranic verses, such as checking for diacritics or specific phrases, are time-consuming and inefficient (Hakak, 2018). There is a clear need for an automated system that can accurately and quickly authenticate Quranic verses. The complexity of the Arabic language, including its morphological nature and diacritics (Eksell, 1995), further complicates this task. Previous attempts to automate Quranic verse authentication using word embeddings and deep learning have shown promising results, but they often struggle with accurately capturing the full context and intricate meanings of the Quranic text. In contrast, transformer models have proven to be better in dealing with these complexities (Vaswani, 2017). Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), are designed to capture long-range dependencies in text and model complex linguistic structures through an attention mechanism. This enables them to focus on relevant parts of the input sequence, making them ideal for context-sensitive tasks like Quranic verse authentication, where understanding the meaning behind specific words and phrases is crucial (Shreyashree et al., 2022). In this project, we propose a solution for Quranic verse authentication through the use of models based on the Transformer architecture. We experimented with advanced large language models including **BERT-Base-Arabic**, **AraBERT** and **MarBERT** that are already pre-trained using Arabic datasets. Additionally, to create a more robust non-authentic dataset, we automated the generation of non-authentic verses using **Sentence-BERT** and cosine similarity thresholds, improving the quality of the dataset for training. This study aims to offer a more reliable method for Quranic verse authentication, addressing the limitations of existing techniques and providing an efficient, scalable solution for verifying digital Quranic content. Given the high sensitivity of the Quranic text, it is imperative that authentication tools achieve perfect accuracy. This paper sets the goal of developing error-proof tool to ensure that Quranic content can be reliably verified. While current models show promising results, future work will focus on further improving the precision and reliability of the system, reducing false positives and false negatives, and extending the methods to handle more diverse datasets.

2 Related Work

In the last decade, several methods have been put forward for verifying Quranic texts through various techniques, especially in the context of preserving their integrity in the digital space. These studies offer diverse methodologies, including word embeddings and machine learning models for automated classification. Kamsin et al. (Kamsin et al., 2014) acknowledged the importance of both Quran and Hadith authentication by creating a central authenticated repository for verifying digital texts. Their approach relies on manual validation, which limits scalability and efficiency. Jarrar et al. (Jarrar et al., 2018) have suggested an Arabic word matching technique based on the use of diacritic marks, mainly emphasizing phonetic shaping. While effective in handling diacritics, this approach emphasizes phonetic accuracy over deeper contextual understanding. Touati-Hamad et al. (Touati-Hamad et al., 2021) used Long Short-Term Memory (LSTM) models to detect Quranic verses that have been altered or reordered. Their model achieved high accuracy using a dataset from the Tanzil website but the only center of attention was sequence reordering, neglecting the semantic variations in the verses. Later, Touati-Hamad et al. (Touati-Hamad et al., 2022) came up with a deep learning method that employs CNN and LSTM models which were used to classify Quranic verses, the specific text is drawn from an Arabic learner corpus for non-authentic text. While effective in distinguishing Quranic verses from regular Arabic text, the simpler Arabic corpus used leaves room for further exploration in addressing subtle variations that closely resemble authentic verses. Muaad et al. (Muaad et al., 2023) performed a survey on Arabic text detection, revolving around the morphology problem of Arabic, which is one of the major challenges, and the necessity of context-aware models. Their review of deep learning techniques underscores the limitations of traditional methods in processing complex Arabic texts. However, these previous techniques are less efficient in dealing with Arabic complexities, like the text bidirectionality and long-range dependencies, in particular, when recognizing verses with slight but significance among variations in the non-authentic sample which is very close to the actual Quranic verses. In contrast to the approaches outlined above, **string-based comparison methods** such as hashing algorithms or text similarity checks are traditional techniques that compare a reference validated source of Quranic text with a target test source. These methods, while effective in identifying exact textual matches, **struggle to detect subtle variations** such as paraphrasing, semantic shifts, or alterations that do not involve significant structural changes. Methods like hashing focus primarily on exact matches.

Despite their effectiveness, the techniques mentioned above (LSTM, CNN, word matching) often rely on simpler models or datasets that do not fully capture the complexities of Arabic text. Similarly, while **sequence reordering** and **CNN-LSTM hybrid models** achieve good performance in detecting certain types of alterations, they may still struggle with the nuanced variations that closely resemble authentic verses. Transformer-based models, particularly BERT-based architectures, address these gaps by being **context-aware** and capable of distinguishing subtle semantic differences. This makes them a **better fit** for the task of authenticating Quranic verses, especially when the alterations are not immediately visible but affect the **underlying meaning**.

In recent years, **deep learning models**, particularly **transformer-based architectures** (Gillioz et al., 2020) emerged as powerful tools for text analysis. These models surpass traditional methods by not only detecting structural changes but also understanding the **context** and **semantics** of the text. The ability of transformer models to process **long-range dependencies** and handle **bidirectional relationships** in text allows them to detect even **subtle semantic alterations** that might evade simpler techniques. Unlike traditional string-based comparison, these models are trained to capture complex relationships between words and phrases, making them more robust for **identifying non-authentic Quranic verses** where minor variations in wording might otherwise go unnoticed.

3 Proposed Methodology

QuranAuthentic aims to address the critical challenge of authenticating Quranic verses by utilizing advanced deep learning techniques, specifically leveraging transformer-based language models. This section outlines the methodology adopted in the development of the QuranAuthentic tool. Our system is designed to identify subtle variations and manipulations in Quranic verses, ensuring the sanctity of the text. The methodology is organized into key components: dataset collection, preprocessing, text representation, and classification. We employed 3 different BERT models such as **BERT-Base-Arabic**, **AraBERT** and **MarBERT**, and explored fine-tuning techniques to ensure accurate detection of non-authentic Quranic verses. Figure 1 illustrates the workflow followed in this approach.

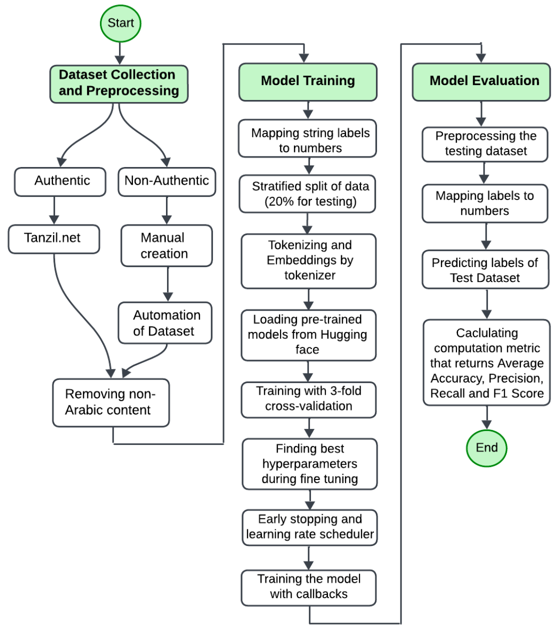
Start
Dataset Collection and Preprocessing
Authentic
Non-Authentic
Tanzil.net
Manual creation
Automation of Dataset
Removing non-Arabic content
Model Training
Mapping string labels to numbers
Stratified split of data (20% for testing)
Tokenizing and Embeddings by tokenizer
Loading pre-trained models from Hugging face
Training with 3-fold cross-validation
Finding best hyperparameters during fine tuning
Early stopping and learning rate scheduler
Training the model with callbacks
Model Evaluation
Preprocessing the testing dataset
Mapping labels to numbers
Predicting labels of Test Dataset
Caclulating computation metric that returns Average Accuracy, Precision, Recall and F1 Score
End

Figure 1: Step-by-step procedure of QuranAuthentic

3.1 Research Questions

This study addresses the following primary research question: **How effective is the QuranAuthentic system in accurately identifying and authenticating Quranic verses, especially when distinguishing between authentic and subtly altered texts using transformer-based language models?**. This research question focuses on evaluating the performance of the QuranAuthentic tool in detecting non-authentic Quranic verses, particularly those with minor alterations, which could resemble authentic verses closely. The study aims to measure the accuracy and reliability of the system compared to traditional methods of Quranic verse verification. Specifically, it explores the use of advanced deep learning models, such as BERT-based architectures (Koroteev, 2021), and how well these models handle the complexities of the Arabic language, including subtle semantic and contextual differences (Alammary, 2022). The goal is to create a more efficient, scalable, and automated solution for Quranic verse authentication in the digital age. This work has the potential to significantly improve the verification of Quranic verses within digital contexts, thus protecting both their authenticity and integrity in light of increasing online distribution (Hakak et al., 2019).

3.2 Dataset Collection

The dataset for our proposed methodology, QuranAuthentic, consists of two main classes: authentic Quranic verses and non-authentic, altered verses. While previous studies have focused on classifying Quranic verses against general Arabic text, no research has specifically addressed distinguishing authentic verses from subtly altered versions that closely resemble the original. In our data collection step, Sentence-BERT (sBE) was employed to compare the similarity between authentic verses and their altered counterparts. A cosine similarity threshold ensured minimal alterations, preserving structural similarity while introducing meaningful contextual differences. This approach aligns with recent advances in semantic textual similarity using transformer-based models (Reimers, 2019). The automated process altered words, phrases, or sentence structures within verses, keeping the difference between authentic and non-authentic text within a controlled range. Natural language processing techniques, such as TF-IDF and WordNet antonym replacement (Fellbaum, 1998), were applied to introduce alterations. The system replaced high-importance words or modified verse structures, and in cases where antonym replacement failed, selective word removal was used. This method ensured non-authentic verses retained close resemblance to their authentic counterparts while introducing subtle but detectable differences. This automated approach significantly enhanced the quality and scalability of the non-authentic dataset, offering a controlled method for generating text that closely mimics authentic Quranic verses with minor, contextually distinct variations.

- **Authentic dataset:** Sourced from tanzil.net[1],

[1] tanzil.net is a Quran initiative founded in 2007 to generate

providing verified Quranic texts.

- **Non-authentic dataset:** Initially, 500 verses were manually altered to create the dataset. This was then expanded through automation using sentence-BERT [2] and cosine similarity, followed by quality checks, resulting in a total of 2068 non-authentic verses.

Given the sensitive nature of the Quranic text, the non-authentic verses generated for this research are labeled as non-authentic and are **kept strictly for internal use.** These altered verses will not be published or shared publicly in any form. They are solely utilized for the purpose of model training and evaluation, and all data handling complies with ethical guidelines to ensure the protection of the Quran's integrity.

3.3 Text Pre-processing

The Quranic dataset comprises 2000 verses labeled as "Authentic", and the non-authentic dataset consists of 2068 lines labeled as "non-Authentic". Since our dataset was sourced from authentic Quranic content which serves as a reliable source provided by **tanzil.net** (Tanzil, 2023), and the non-authentic dataset was manually curated, and further rechecked after automation, it was inherently clean and required minimal preprocessing. However, to ensure that the QuranAuthentic tool can handle real-world data, preprocessing was primarily applied to the unseen test data. This preprocessing was designed to clean and standardize content that might be encountered in online settings. For the unseen real world data, non-Arabic content such as numbers and foreign text, was removed to focus solely on relevant Arabic text. Normalization was conducted to unify variations in the script, particularly different forms of Alef, to reduce ambiguity and improve text consistency. Special symbols, including URLs, emoticons, and extraneous characters, were eliminated to maintain a clean dataset. Additionally, Kashida (Tatweel), often used for text decoration, was removed to prevent interference with text analysis. Such steps are important in handling the complexities of Arabic text, as highlighted in studies like preprocessing pipelines for Arabic NLP (Awajan, 2007). Finally, tokenization was carried out to separate the text into tokens which allows the model to process particular words or characters with a greater efficiency (Alkaoud and Syed, 2020). These techniques made sure that the text data utilized in real-life scenarios was standardized and appropriate for proper model processing.

3.4 Model Training

In recent years, models based on transformer architecture have set the benchmark for multiple natural language processing tasks. Based on a thorough analysis of the existing body of knowledge (Alammary, 2022), we considered experimenting with the most outperforming BERT models for Arabic text which include **BERT-Base-Arabic**, **AraBERT** and **MarBERT**. Our aim was to utilize their sophisticated language understanding capabilities for accurately distinguishing between authentic and altered Quranic verses. **BERT-Base-Arabic** was chosen for its general applicability and strong performance in a wide range of Arabic text processing tasks. **BERT-Base-Arabic** is pre-trained on a large Arabic corpus, making it well-suited for understanding context and semantics in Arabic. It serves as an excellent baseline for Quranic verse authentication by detecting semantic shifts and structural changes, which are essential for distinguishing authentic from altered verses. **AraBERT**, trained on a large corpus of Modern Standard Arabic (MSA) text, was selected due to its strong performance in various Arabic NLP tasks (Antoun et al., 2020). While **AraBERT** does not have the same specialized focus on Quranic text as other models, its understanding of standard Arabic syntax and semantics makes it an important baseline model in this study. We employed **MarBERT** for its specialized focus on Arabic text, particularly in identifying textual discrepancies. Given its training on a large Arabic corpus and different Arabic dialects (AlKhamissi et al., 2021), **MarBERT** is highly effective at recognizing slight deviations in text structure and style, allowing it to detect alterations in Quranic verses that may compromise their authenticity. All models were fine-tuned through hyperparameter tuning, including adjustments to learning rates, batch sizes, and the number of training epochs to achieve optimal performance. Training was conducted using the balanced dataset of authentic and non-authentic Quranic verses described in the previous section. The results of this model training and testing are evaluated in the subsequent sections.

a thoroughly Unicode Quran text to be utilized in Quranic websites and apps

[2] https://huggingface.co/sentence-transformers

4 Experimental Setup

This section is divided into different subsections that explain the collection of datasets, preparation of dataset for authentic and non-authentic classes, the evaluation metrics used for finding and working out the results and then the implementation and experimentation of the models. These steps ensure a robust and accurate evaluation of the Quranic verse classification system.

4.1 Data Preparation

For this study, the dataset consists of two classes: authentic Quranic verses and non-authentic, altered verses. The authentic dataset was sourced from *Tanzil.net*, a verified repository of Quranic texts. The non-authentic dataset was generated automatically by introducing controlled alterations to the authentic verses using **sentence-BERT** and cosine similarity threshold. The dataset statistics are as follows:

Authentic Dataset: 2000 Quranic verses from *Tanzil.net*.

Non-authentic Dataset: 2068 Non-Authentic text generated by modifying authentic verses, ensuring a cosine similarity threshold that maintains structural resemblance but introduces contextual changes.

The dataset was divided into training and testing sets, with 20% of the data reserved for testing. This separation ensures the model's performance is evaluated on unseen data, reflecting potential real-world performance. Further, the training data underwent Stratified K-Fold cross-validation with three splits, where approximately 1/3 of the training data serves as the validation set in each fold. This method preserves the percentage of samples for each class, crucial for maintaining class distribution consistency across training and validation subsets. The rotational use of data for validation in each fold facilitates effective hyperparameter tuning and model validation without a separate dedicated validation dataset.

Table 1: Dataset Division

Class	Train	Validation	Test	Total
Authentic	1067	533	400	2000
Non-Authentic	1101	553	414	2068
Total	2168	1086	814	4068

4.2 Evaluation Metrics

To evaluate the performance of the models, we utilized the following metrics:

- **Accuracy**: The proportion of correctly classified verses.
- **Precision**: The ratio of true positive predictions to the total predicted positives.
- **Recall**: The ratio of true positive predictions to the actual positives.
- **F1-Score**: The harmonic mean of precision and recall, balancing the trade-off between the two.

These metrics provide a comprehensive view of the models' effectiveness in distinguishing between authentic and non-authentic Quranic verses.

4.3 Experimental Evaluation

For the experimentation of the QuranAuthentic system, we employed three transformer-based models: **BERT-Base-Arabic**, **MarBERT**, and **AraBERT**. After preprocessing, we merged two datasets of Quranic verses, ensuring no missing values. We split the data into training (80%) and test (20%) set using stratified sampling to maintain class distribution. Tokenization was performed using the respective BERT tokenizer, and the pretrained models from Hugging Face were fine-tuned with hyperparameters given in Table 2. We implemented 3-fold stratified cross validation to evaluate model performance. Each fold involved tokenizing the data, compiling the model with the Adam optimizer, and applying early stopping and learning rate scheduling. We calculated precision, recall, F1-score, and accuracy after training on the validation sets. Finally, the model was tested on the unseen dataset to evaluate the model. The models were fine-tuned using varying hyperparameters. Table 2 shows the configurations used during the training process:

Table 2: Hyperparameter Configuration

Model	Learning Rate	Epochs	Batch Size
BERT-Base-Arabic	5e-5	10	8
MarBERT	3e-5	12	16
AraBERT	4e-5	15	8

It is worth mentioning that this process was repeated for several variations of BERT-based models. However, **MarBERT** and **BERT-Base-Arabic** both consistently exhibited the most promising results due to its specialized training on Arabic text. Each model's performance was measured based on average accuracy, precision, recall, and F1 score and the results are detailed in the following section.

5 Results and Analysis

To address the defined research question, we conducted a thorough systematic literature review followed by rigorous experimentation, implementation and analysis to select the best performing models for arabic language (Alrajhi et al., 2022). In order to evaluate the effectiveness of the QuranAuthentic system, we conducted a comprehensive set of experiments on each of the transformer-based models, namely **BERT-Base-Arabic**, **MarBERT**, and **AraBERT**. These experiments aimed to address the research question. The performance of each model was evaluated using average accuracy, precision, recall, and F1-score as the primary metrics. The primary research question revolved around assessing the performance of these models on a specialized dataset comprised of Quranic verse classifications. Our evaluation of models revealed nuanced differences in performance across these models. Notably, **BERT-Base-Arabic** and **MarBERT** demonstrated significantly better performances. In contrast, **AraBERT**, while showing lower precision and overall test accuracy, suggesting a greater tendency for false positives. The results underscore the importance of selecting appropriate models that align with the linguistic and syntactic demands of the text. Table 3 will detail the comparative results of these models under a specific set of configurations, highlighting their respective strengths and limitations in the context of Quranic verse classification.

Table 3: Model Performance on the Test Set

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
MarBERT	93.73%	91.25%	96.50%	93.80%
BERT-Base-Arabic	92.75%	89.56%	96.50%	92.90%
AraBERT	57.99%	53.92%	99.75%	70.00%

The detailed performance metrics for each model reveal significant insights into their capabilities in processing Quranic text authentication. **MarBERT** and **BERT-Base-Arabic** both show exemplary performance, with **MarBERT** slightly leading in terms of precision and overall F1-score. **MarBERT**, with an accuracy of 93.73% and an F1-score of 93.80%, demonstrates a strong balance between precision and recall. This model's focused training on Arabic language content clearly enhances its ability to accurately authenticate Quranic verses while minimizing false positives. Such performance is essential for applications where the authenticity of the text is paramount, like in digital platforms for religious texts. **BERT-Base-Arabic**, while trailing slightly behind **MarBERT**, still presents impressive results with an accuracy of 92.75% and an F1-score of 92.90%. Its high recall rate, in particular, suggests it is highly effective at identifying authentic texts, though its precision indicates some challenges in minimizing false positives. Nonetheless, the improved metrics suggest that even general-purpose models, with appropriate tuning, can offer robust solutions for specialized tasks like Quranic text authentication. Conversely, **AraBERT** displays an unmatched recall of 99.75%, highlighting its strength in identifying relevant verses. However, its lower precision points to a high rate of false positives, limiting its practical utility in scenarios where both identification and precision are critical. This underperformance can be attributed to several factors. First, AraBERT, primarily trained on Modern Standard Arabic (MSA), is suitable for general Arabic tasks (Antoun et al., 2020). However, it does not fully capture the linguistic nuances of Quranic texts, which include classical Arabic features and unique stylistic elements. In contrast, **MARBERT**, trained on both MSA and Arabic dialects, benefits from its ability to handle linguistic diversity, making it better equipped to address text variations and subtle alterations often encountered in Quranic verse authentication. **BERT-Base-Arabic**, while also trained on MSA, performs better than **AraBERT** in this task due to its robust handling of formal Arabic text and stronger generalization capabilities in tasks requiring contextual understanding. Additionally, **AraBERT** struggles with dialectal variations, as its training corpus does not encompass the specific differences in pronunciation and word choice found in Quranic dialects, such as those associated with Qira'at. To enhance AraBERT's performance in Quranic verse authentication, future work should focus on fine-tuning the model on Quran-specific datasets, incorporating dialectal and Qira'at variations, and optimizing hyperparameters to improve precision and reduce false positives.

In summary, the performance of **MarBERT** and **BERT-Base-Arabic** underscores the effectiveness of specialized and appropriately tuned models in enhancing the accuracy and reliability of Quranic verse authentication systems. Their robust capabilities demonstrate significant potential for real-world applications, ensuring the integrity and authenticity of Quranic verses in digital formats. To enhance AraBERT's utility and address its underperformance in Quranic verse authentication tasks, several strategies can be implemented. First, incorporating dialectal Arabic and Qira'at variations in the training data would allow AraBERT to handle the **linguistic diversity** and recitation styles characteristic of Quranic Arabic. This approach would enhance its precision across diverse representations. Furthermore, data augmentation techniques, such as generating varied non-authentic verses and introducing subtle textual changes (e.g., paraphrasing or simulated typographical errors), would improve the model's robustness against real-world variations. In the future, we can employ these methods to provide a comprehensive framework to enhance AraBERT's performance in Quranic text-related tasks.

6 Limitations of the Current Study

Despite the promising outcomes of this research, several limitations warrant consideration:

Dependency on a Single Data Source: The authentic dataset is sourced exclusively from **tanzil.net**[3]. While this source is verified and widely regarded as reliable, it may not cover some of the canonical Quranic variations which are from the different Qira'at or recitation styles. This dependency limits the generalizability of the model to other verified Quranic texts.

Reproducibility Challenges: The fine-tuning process utilized in this study, including hyperparameter optimization, is highly dependent on computational resources. The results achieved may not be easily replicable without having similar hardware configurations and pretrained models, thus limiting the embracing of these methods.

Performance Under Resource Constraints: Despite excellent performance of BERT-based Language Model, their deployment in real environments is resource intensive. However, inference optimization need to be carried out in order to have efficiency and scalability in such environments which has not been dealt with in this paper. These limitations underscore the need of future research that addresses these limitations which relate to dataset diversity, reproducibility issues, and optimal model performance in practical applications.

7 Conclusion

This paper highlights the effectiveness of transformer-based deep learning models in authenticating Quranic verses in Arabic text. We employed three models: **BERT-Base-Arabic**, **MarBERT**, and **AraBERT** which are fine-tuned to classify verses as authentic or non-authentic. Despite the inherent complexities of the Arabic language and subtle alterations in verses, the models achieved notable performance especially the two models **BERT-Base-Arabic** and **MarBERT**. From these experimental analysis, we can conclude that Transformer models can accurately differentiate between authentic and non-authentic Quranic verses, with **MarBERT** and **BERT-Base-Arabic** delivering the promising results. The findings indicate that the specialized models like **MarBERT** and **BERT-Base-Arabic** provide better context and semantic understanding, making them ideal for tasks like religious text authentication. This work offers a potential solution for Quranic verses verification, addressing concerns about unverified verses online. In the future, we aim to enhance the model's performance and create a real-time tool for authenticating Quranic verses. Furthermore, we plan to expand the dataset to include diverse sources and variations, particularly by incorporating multiple **Qira'at (canonical recitations)**, which will ensure that the model is more robust and comprehensive. This will be achieved by collaborating with specialized Qira'at databases and employing data augmentation techniques to simulate variations across different recitation styles. This approach can also be extended to authenticate religious texts in other languages or domains. By fine-tuning pre-trained transformer models like mBERT (Wang et al., 2019) or XLM-R (Goyal et al., 2021) on domain-specific datasets, such as the Bible or the Torah, these models can be adapted to handle the unique linguistic features and structural intricacies of other religious texts. Additionally, data augmentation and transfer learning (Torrey and Shavlik, 2010) can be employed to adapt the model for new languages or domains with limited data. By training on multilingual datasets and incorporating cultural and linguistic nuances, this approach can be extended beyond Quranic texts to other religious and historical contexts, ensuring

[3] https://tanzil.net/download/

robust performance across various languages and traditions.

References

Semantic textual similarity using sbert. Accessed: November 2024.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: a systematic review. *Applied Sciences*, 12(11):5720.

Mohamed Alkaoud and Mairaj Syed. 2020. On the importance of tokenization in arabic embedding models. In *Proceedings of the fifth Arabic natural language processing workshop*, pages 119–129.

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting marbert for improved arabic dialect identification: Submission to the nadi 2021 shared task. *arXiv preprint arXiv:2103.01065*.

Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. Assessing the linguistic knowledge in arabic pre-trained language models using minimal pairs. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Arafat Awajan. 2007. Arabic text preprocessing for the natural language processing applications. *Arab Gulf Journal of Scientific Research*, 25(4):179–189.

Kerstin Eksell. 1995. Complexity of linguistic change as reflected in arabic dialects. *Studia Orientalia Electronica*, 75:63–74.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on computer science and information systems (FedCSIS)*, pages 179–183. IEEE.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Saqib Hakak, Amirrudin Kamsin, Omar Tayan, Mohd Yamani Idna Idris, and Gulshan Amin Gilkar. 2019. Approaches for preserving content integrity of sensitive online arabic content: A survey and research challenges. *Information Processing & Management*, 56(2):367–380.

Saqib Iqbal Hakak. 2018. *Authenticating Sensitive Diacritical Texts Using Residual, Data Representation and Pattern Matching Methods*. Ph.D. thesis, University of Malaya (Malaysia).

Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–21.

Amirrudin Kamsin, Abdullah Gani, Ishak Suliaman, Salinah Jaafar, Rohana Mahmud, Aznul Qalid Md Sabri, Zaidi Razak, Mohd Yamani Idna Idris, Maizatul Akmar Ismail, Noorzaily Mohamed Noor, et al. 2014. Developing the novel quran and hadith authentication system. In *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, pages 1–5. IEEE.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Abdullah Y Muaad, Shaina Raza, Usman Naseem, and Hanumanthappa J Jayappa Davanagere. 2023. Arabic text detection: a survey of recent progress challenges and opportunities. *Applied Intelligence*, 53(24):29845–29862.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

S Shreyashree, Pramod Sunagar, S Rajarajeswari, and Anita Kanavalli. 2022. A literature review on bidirectional encoder representations from transformers. *Inventive Computation and Information Technologies: Proceedings of ICICIT 2021*, pages 305–320.

Tanzil. 2023. Tanzil quran navigator. Accessed: 2023-10-09.

Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Zineb Touati-Hamad, Mohamed Ridda Laouar, and Issam Bendib. 2021. Authentication of quran verses sequences using deep learning. In *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, pages 1–4. IEEE.

Zineb Touati-Hamad, Mohamed Ridda Laouar, Issam Bendib, and Saqib Hakak. 2022. Arabic quran verses authentication using deep learning and word embeddings. *The International Arab Journal of Information Technology*, 19(4):681–688.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.