CLASSIFICATION OF ENGLISH TRANSLATED HOLY QURAN & GENERATING VERSES BASED ON SEARCH INDEX

BY

HADIKA HASAN
ID: 193-15-13522

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MD. SAZZADUR AHAMED
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

MS. SHARMIN AKTER
Lecturer (Sr. Scale)
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

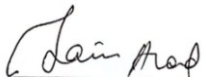
DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "**Classification of English Translated Holy Quran & Generating Verses Based on Search Index**", submitted by **Hadika Hasan** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15-07-2024.

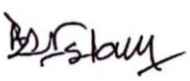
BOARD OF EXAMINERS



Dr. Md. Taimur Ahad — **Board Chairman**
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Ms. Shayla Sharmin — **Internal Examiner 1**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Mayen Uddin Mojumdar — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Manowarul Islam — **External Examiner**
Associate Professor
Department of CSE
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Ms. Sharmin Akter
Lecturer (Sr. Scale)
Department of CSE
Daffodil International University

Submitted by:

Hadika Hasan
ID: -193-15-13522
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Md. Sazzadur Ahamed**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Deep Learning & *Natural Language Processing*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori, Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

The English translation of the Holy Quran serves as an essential guide for spiritual enlightenment and wisdom globally. This research emphasizes the application of advanced deep learning techniques for classifying and generating verses from this translated Quran based on search indexes. This study explores the profound significance of making these teachings accessible worldwide and highlights the transformative impact of deep learning advancements in facilitating deeper understanding and interpretation. Utilizing models such as RNNs, CNNs, and our own Hybrid model, our research achieves significant classification accuracy. Specifically, RNN models all kinds and CNNs achieved accuracy above 85% while the Hybrid model surpassed them with 99.91%. These models enable efficient categorization and also this study focuses on retrieval of Quranic verses based on search queries, thereby enhancing accessibility and usability. This research underscores the significance of integrating religious teachings with cutting-edge deep learning innovations. By integrating deep learning into the study of religious texts, we facilitate scholarly research, educational tools, and personalized spiritual exploration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Overview

The Holy Quran, the central religious text of Islam, holds profound significance for Muslims around the world. The Qur'an is essentially a comprehensive code of conduct that addresses every facet of life, including intellectual, political, social, economic, and spiritual.[1] It is thought to be the exact words of God as they were sent to Prophet Muhammad by the angel Gabriel. The Quran is written in classical Arabic, and its teachings encompass all aspects of life, providing guidance on spiritual, moral, and legal matters. However, for non-Arabic speakers, understanding the Quran requires translations, which, while not equivalent to the original text, strive to convey its meanings as accurately as possible.[2] In Bangladesh, as in many other parts of the world, English translations of the Quran are widely used. These translations are essential for both Muslim and non-Muslim communities seeking to understand the Quran's teachings. The availability of English translations has facilitated a broader understanding and accessibility, making the Quran's messages reachable to a global audience.[3] However, given the complexity and depth of the Quranic text, ensuring the accuracy and relevance of these translations is a challenging task that often relies on manual processes.

The advent of deep learning and Natural Language Processing (NLP) has revolutionized the field of text analysis and information retrieval.[4] Deep learning (DL), a subset of machine learning, uses multi-layered neural networks to simulate complicated patterns in data. When applied to NLP, deep learning techniques can significantly enhance the detection and classification of texts, making it possible to handle large volumes of data with greater accuracy and efficiency than traditional methods.[5]

Our key application of deep learning using RNN (Traditional RNN, LSTM, BILSTM, GRU) & CNN in the context of the Quran is the automatic classification of verses based on thematic content. This involves training models to understand and categorize the verses according to various themes such as Allah, Mohammad, and Prayer etc. Such models can help in organizing the text in ways that make it more accessible to readers with specific interests. Moreover, deep learning can be used to generate verses or suggest relevant Quranic verses based on search queries. By developing search indices powered by NLP using Sentence Transformer, users can retrieve verses that are contextually relevant to their queries, thereby facilitating deeper

engagement with the text. This approach not only saves time compared to manual searching but also ensures a higher degree of precision in finding relevant content. The importance of using deep learning and NLP for Quranic text analysis lies in their ability to process and understand natural language with a level of nuance that traditional methods lack.[6] These technologies can recognize patterns, understand context, and provide insights that are crucial for accurate interpretation. This is especially important for religious texts where the meaning is deeply contextual and multifaceted. This study contributes to the better understanding of English translated quran by easily locating and focusing on specific topics needed by users and proposing a novel deep learning approach. In this case,

- The dataset used in this study was collected from English Translated Holy Quran and labeled manually by using reliable sources. The dataset contains 21 topics appointed. Allah, Quran, Muhammad, Belief, Disbelief, Evils, Prayer, Charity, Fasting & patience, Hajj & Purification, Marriage & Divorce, Heaven & Hell, Judgement & Justice, Hereafter, Unseen, Principles, Stories, Struggles, Lawful, Creations and Repentance. The dataset has a 1276 verses classified into multi labels by 21 topics.
- Implemented preprocessing techniques like data augmentation, lowercasing, removing punctuation, special characters, punctuations and stop-words, lemmatization, tokenization, text representation to handle verses as per our need.
- Applied 6 different deep learning techniques and achieved satisfying results.

  The integration of deep learning and NLP into the study and dissemination of the Quran represents a significant advancement over manual methods. These technologies enhance the accessibility and understanding of the Quranic text, ensuring that its teachings can be more easily and accurately accessed by a global audience.

1.2 Background and Present State

The Quran is a sophisticated and comprehensive book that offers guidance on a wide range of topics, including morality, religion, the law, and interpersonal behavior. Since an ayah (verse) might cover more than one theme or topic, multilabel classification is a good method to use while classifying these verses. In multilabel classification, every instance—in this case, every verse of the Quran—is given numerous pertinent labels.

Traditional Approaches

For text categorization tasks at first, conventional machine learning techniques like Support Vector Machines (SVM) and Decision Trees were employed. Nevertheless, these methods frequently needed considerable feature engineering and had limitations in capturing the semantic intricacy of texts.

Deep Learning Methodologies

More complex models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have been applied to text categorization problems as a result of deep learning advances, providing notable improvements over traditional approaches.

Recurrent Neural Networks (RNNs)

Neural networks that are designed to process sequence input and retain contextual information over time steps are called recurrent neural networks, or RNNs. To solve problems like the vanishing gradient problem, variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are used, enabling the model to capture long-range dependencies in the text. RNNs can take use of the verses' sequential structure to comprehend context and semantics in order to classify Quranic verses.

CNNs, or convolutional neural networks

CNNs have been adapted for text classification, while they are mostly employed in image processing. Convolutional filters are used to extract text's local patterns and features, like words or phrases that are crucial to understanding a certain subject. CNNs are appropriate for handling massive text corpora such as the Quran due to their computational efficiency and ability to process data in parallel.

Present State

Transformer-based models and sophisticated embedding approaches have been included into the field in recent years to improve text categorization performance even more.

Embeddings

Word embeddings that capture the semantic associations between words are represented in continuous vector spaces by programs like Word2Vec, GloVe, and fastText. To capture meanings relevant to a certain domain, these embeddings can be adjusted on the text of the Quran.

Models of Transformers

Natural language processing (NLP) has been revolutionised by transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers). Compared to RNNs and

CNNs, these models are more successful at capturing contextual relationships in text by utilizing self-attention techniques. An extension of BERT called sentence transformers may produce embeddings for sentences or passages, allowing for the execution of activities such as semantic search and verse production in response to search queries.

Generating Verses Based on Search Index

Generating verses or retrieving relevant verses based on a search index involves using models that understand semantic similarity. Sentence transformers can be fine-tuned on Quranic text to generate embeddings that encapsulate the meaning of each verse. When a search query is provided, the model can compare the query's embedding with those of the verses to find the most relevant matches.

Current Challenges and Future Directions

Despite the progress, several challenges remain in multilabel classification and verse generation:

1. **Imbalanced Data**: Certain topics may be overrepresented, while others are underrepresented, which can skew model performance.
2. **Interpretability**: Deep learning models, especially transformers, can be complex and opaque, making it difficult to interpret their decisions.
3. **Domain Specificity**: Quranic text has unique linguistic and thematic characteristics that require domain-specific adaptations of standard NLP models.

Future research may focus on addressing these challenges by developing more interpretable models, better handling imbalanced data, and creating embeddings that capture the unique semantics of Quranic text. Integrating domain knowledge and leveraging advancements in NLP can further improve the accuracy and applicability of these models in understanding and categorizing Quranic verses.

1.3 Problem Statement

Despite English translations being available, accessing and comprehending the Holy Quran can be a daunting task for non-Arabic speakers & sometimes also helpful for Muslims. The large content of the Quran often makes it challenging for users to find verses that resonate with their spiritual and contextual needs. Current methods lack personalization and may not consider individual interpretations and applications which brings the necessity for a structured framework that

classifies English-translated Quranic verses and produces required verses tailored to users' spiritual and contextual requirements based on their search index. By bringing solution of this problem, I'll be able to promote inclusivity, improve accessibility & inspire people beyond cultural sensitivity in sharing the teachings of the Quran and use it for our various purposes. It'll enhance the spiritual journey of users, enabling a deeper understanding and practical application of the timeless wisdom found in the Quran.

1.4 Objectives

The motivation behind this project is to deliver tailored Quranic verses faster and in a less complicated way to meet individual users' spiritual, emotional, or educational needs. So that it can enhance accessibility of Quran. It'll Provide educational support & facilitate understanding of Quranic content with different faiths of people as well as promote the utility of Quranic knowledge. It aims to build a structured system for classifying English-translated Quranic verses based on their themes and topics and generating verses as per users need. It focuses on implementing algorithms for producing personalized Quranic verses based on users` spiritual and contextual requirements. Also, it'll ensure cultural sensitivity throughout the process, respecting various backgrounds and religious sensitivities. By applying advanced Deep Learning & NLP approaches, I'll try to get accurate classification and verse generation. This project adheres strictly to ethical guidelines to maintain the sanctity and integrity of the Quranic text. It aims to offer accurate, impactful, and tailored Quranic verses to enrich user satisfaction and spirituality.

1.5 Scope and Limitations

This scope of this research is to develop and evaluate a deep learning multilabel classification model for English-translated Quranic verses. The main goal is to categorize verses into several relevant topics by applying sophisticated neural network architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). To capture semantic subtleties unique to the Quran, the research applies transformer-based models (BERT, sentence transformers) and word embeddings (Word2Vec, GloVe, fastText). The scope also includes developing a system that uses sentence transformers to produce precise verse embeddings and performs semantic search and verse retrieval in response to user queries. Model performance will be thoroughly evaluated using Hamming Loss, ROC AUC, F1-score, precision, and recall

measures. The aim of the study is to further Islamic scholarship by offering a useful and accessible method for examining the various topics found in the Quran.

The study contains a number of shortcomings in spite of its broad objectives. The problem of managing imbalanced data—a situation in which some Quranic topics are underrepresented relative to others—possibly compromises classification accuracy and is a major constraint. The complexity of deep learning models, which can be computationally demanding and need significant resources for training and fine-tuning, is another drawback. One drawback of models is that they can be hard to interpret, particularly transformer-based models where the decision-making procedures can be confusing and opaque. The quality and accurateness of the English translations of the Quranic verses poses a constraint on the research as well, as they may include biases or misinterpretations that are not present in the original Arabic text. Finally, the specificity and kind of user searches may have an impact on the relevancy of the recovered verses, hence influencing the system's efficacy in semantic search and verse retrieval.

1.6 Report Organization

The proposed report is meticulously structured to guide readers through the research process, findings, and implications of the study. Every chapter is written with a specific objective in mind, adding to the document's overall cohesion and depth.

Chapter 1: Introduction- The introductory chapter sets the stage for the research, providing background on the significance of classifying English-translated Quranic verses and generating verses based on search queries. It highlights the application of deep learning methods, such as RNNs and CNNs, and the use of sentence transformers for search indexing.

Chapter 2: Background- In this chapter, a detailed exploration of relevant literature and prior research is presented. This section offers an in-depth understanding of the theoretical foundations, methodologies, and technological advancements related to Quranic verse classification and retrieval using deep learning.

Chapter 3: Research Methodology- The research methodology chapter outlines the approach taken to conduct the study. It provides insight into the research design, data collection methods, model

architecture, and the rationale behind choosing specific methodologies like RNNs, CNNs, and sentence transformers.

Chapter 4: Implementation- This chapter details the steps taken to implement deep learning techniques for classifying Quranic verses. It includes data preprocessing (cleaning, tokenization, and embedding generation), selection of model architectures (RNNs, CNNs, transformers), and training procedures.

Chapter 5: Results & Analysis- This chapter presents the experimental results of classifying Quranic verses and generating relevant verses based on user queries. It evaluates model performance using metrics like hamming loss, precision, recall, F1-score, and ROC AUC, and compares different model architectures.

Chapter 6: Impact on Society, Environment and Sustainability- The societal impact chapter explores the broader implications of the research findings on Quranic studies and the accessibility of religious texts. It discusses how improved methods of verse classification and retrieval can enhance understanding, promote personalized learning, and positively influence the study of the Quran.

Chapter 7: Conclusion and Future Work- This final chapter serves as a synthesis of the entire report. It summarizes the key findings, reiterates the conclusions drawn from the research, and offers recommendations for practical applications and further studies. The implications for future research are discussed, providing a roadmap for researchers interested in expanding upon the current study.

This structured layout ensures a logical flow of information, guiding the reader through the research journey from the introduction to the broader implications and recommendations. It allows for a comprehensive understanding of the research process and its potential impact on both the academic and practical aspects of Quranic verse classification and retrieval using deep learning techniques.

1.7 Summary

The study's focus, which focuses on classifying English-translated Quranic passages using cutting-edge deep learning algorithms, is introduced in the introduction. It discusses the current state of text classification technology today, including CNNs, RNNs, and transformers. Improving semantic comprehension and data imbalance are two among the primary problems addressed. Creating a system for efficient verse retrieval based on semantic similarity, utilizing embeddings such as Word2Vec, and building a strong classification model are among the goals. The scope describes the report's organizational structure and admits difficulties with model interpretability.

CHAPTER 2
Background

2.1 Overview

The literature review for this research explores existing studies and methodologies related to the multilabel classification of Quranic verses using deep learning techniques. It begins by reviewing foundational works in text classification, focusing on the application of CNNs, RNNs, and transformer models in capturing semantic nuances within religious texts. Studies addressing challenges like data imbalance and model interpretability are analyzed, highlighting various strategies and advancements. The review also examines the use of word embeddings such as Word2Vec, GloVe, and fastText specifically tailored for Quranic text, emphasizing their role in enhancing classification accuracy. Furthermore, recent developments in semantic search and verse retrieval using sentence transformers and BERT models are discussed. The literature review aims to provide a comprehensive understanding of current methodologies and gaps in research, guiding the development and evaluation of the proposed deep learning model for Quranic verse classification.

2.2 Related Works

There are existing works about this study and the researchers tried different methods to classify Quranic Verses. This section reviews the related research work aiming at different models that can provide better accuracy for Classification & generation of Quranic verses.

Mohamed et al. (2022) developed a concept-driven Quranic search system by manually annotating Quran verses and calculating cosine similarity between the topic and input query vectors to retrieve the most relevant verses. They used a CSV file containing Arabic Quranic verses, additional information from "Mushaf Al Tajweed," Surah titles, part IDs, verse texts, IDs, and text without diacritics. Their classification method relied on this book, encompassing four phases: building the dataset with manual annotations, implementing word embedding using Continuous Bag of Words (CBOW), generating feature vectors for words, and calculating feature vectors for input inquiries and Quranic topics to find the most relevant verses. They achieved an accuracy of

91.95% with QSST, but comparison with Mushaf Al-Tajweed was limited due to potential human errors in the manual selection process.[7]

Zouaoui et al. (2021) aimed to create an ontology-based semantic search engine to index Quranic terms, linking them to user queries to find relevant verses. They divided the Holy Quran into 6236 text files, each containing a single verse, and used NLP techniques for ontology and query processing. They constructed an ontology called OurOnto, enhancing user queries with this ontology. They achieved 67% accuracy, compared to 56% for other works, but noted the need for further information on Arabic grammar verbs to avoid extracting incorrect relationships within verses.[8]

Beirade et al. (2021) utilized Quranic ontology to create a semantic search engine for the Quran's text, breaking words into affixes and nuclei, identifying main words in the DISCORAN database, and using Apache Lucene for semantic search. They added morphological processing and evaluated their system against Alfanous Quran search engine and Corpus Quran, achieving an average precision of 70%. They suggested expanding vocabulary for more precise results.[9]

Bsoul et al. (2021) addressed the difficulties of categorizing Arabic texts from the Quran according to themes . They used a dataset of approximately 1,680 documents across various domains and another with 383,872 Arabic documents from Agence France Press (1994-2000). Using the black hole (BH) method for feature extraction and clustering, they identified essential words and starting centroids for Quranic themes. However, they did not provide specific accuracy results or quantitative evaluation metrics.[10]

Wahdan et al. (2020) analyzed recent Arabic text classification findings using neural networks, dealing with 12 datasets from various sources. They used different types of neural networks, with Long Short-Term Memory (LSTM) networks showing high efficiency in most studies.[11]

Elmitwally et al. (2020) developed a machine learning model to classify Quranic verses and identify key features of the first six Surahs. Using a dataset of 954 verses, they employed Support Vector Classifier and Gaussian Naive Bayes, achieving 80% and 60% accuracy, respectively. The limited dataset did not fully capture the Quran's complexity.[12]

Utomo et al. (2020) explored the effect of stemming on instance classification results using datasets that included the Quran's interpretation and translation into Indonesian. Using several Quran Surahs and thematic topics, they employed BPNN, KNN, and SVM classifiers with TF-IDF and stemming. For various test data sizes, SVM using TF-IDF/BoW/Stem achieved 70.75% and 67.30% accuracy. Stemming negatively impacted certain classification approaches.[13]

Arkok et al. (2020) assigned matching themes to Quranic verses based on content, addressing imbalanced learning. They used two datasets (Tawheed and shirk, Meccan and Medinan) and employed SMOTE for preprocessing. Random Forest, Decision Tree (J48), Naive Bayes, KNN, LibSVM, and Bayes Net were used, with Bayes Net and Naive Bayes achieving 74% and 75% accuracy, respectively, before SMOTE. SMOTE was applied, and Random Forest produced 96% accurate results.[14]

Choirulfikri et al. (2022) developed a method for multi-label classification of Quran verses using ensemble methods with naive Bayes algorithms. Using an English dataset of over 6000 text data points and 15 labels, they evaluated classification methods, achieving a hamming loss value of 0.1167 but lacked a wider range of datasets or domains.[15]

Aiman et al. (2022) used word centrality measurements for multi-label classification of English Quranic translations into ten categories. Their dataset included 4054 verses and 5672 distinct words, with preprocessing stages like text cleaning and tokenization. SVM, Naive Bayes, KNN, and Decision Tree were used, with SVM achieving the best hamming loss score between 0.1110 and 0.14. This approach focused on English translations, potentially excluding non-English speakers.[16]

Abdullahi et al. (2021) automated the labeling of Quranic verses using multi-label classification (MLC). Their dataset included 1098 instances with labels like Faith, Worship, Etiquette, and combinations thereof. They pre-processed the text for machine learning tasks and used SVMs, Naive Bayes, kNN, and J48 classifiers, with SVM achieving the highest accuracy and lowest hamming loss (0.103).[17]

Rostam et al. (2021) proposed a text classification approach using English texts from the Hadith and Quran corpus. They used Naive Bayes, SVM, and KNN with TF-IDF to manually classify verses into Hajj, Prayer, and Zakat categories. SVM showed a 10-20% improvement in accuracy compared to other methods.[18]

Alashqar (2023) developed a multi-label classification technique using deep learning that reliably classifies Quranic verses into 12 predetermined major subjects. The dataset, based on "Quran by Subject" and additional Arabic information, used Word2Vec for semantic text analysis. RNN and CNN models achieved 90.38% and 89.74% accuracy, respectively.[19]

Borhani (2020) presented a quasi-Newton updating approach for an artificial neural network used for multi-label multi-class text categorization of Quranic verses. Using a dataset of 6236 verses and 71,568 labels with English translations, they employed the ML4BFGS algorithm, achieving

88% precision and 84% hamming loss compared to other algorithms.[20]

Yulianto et al. (2021) examined the impact of the Quran's word centrality measure on verse topic classification, using Arabic verses from Tanzil.net. They covered eight themes and applied centrality measurements like degree, betweenness, and closeness centrality. Feature extraction methods like TW-IDF and TF-IDF were used, with Naive Bayes and SVM classifiers. Betweenness centrality with SVM achieved the best hamming loss score of 0.15.[21]

2.3 Comparison between existing works

The comparative analysis in this research evaluates various deep learning models for classifying English-translated Quranic verses and generating relevant verses based on search queries. By examining the performance of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), the study highlights the strengths and limitations of each approach in handling complex religious texts. RNNs are assessed for their ability to capture contextual relationships within verses, while CNNs are evaluated for their pattern recognition capabilities across translated texts. The research also explores the application of sentence transformers to enhance search functionality through semantic understanding. The findings contribute valuable insights into the most effective strategies for these tasks and guide future studies and implementations in the analysis of religious texts. This analysis sets the stage for further exploration into the societal impact and implications for future research in this critical area.

TABLE 2.3.1: Comparative Analysis Table

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Mohamed et al. [7]	Word Embedding (Word2vec)	91.95%
2.	Zouaoui et al. [8]	Natural Language Processing (NLP)	67%
3.	Beirade. et al. [9]	Search algorithm + Apache Lucene	70%
4.	Bsoul. et al. [10]	Black hole optimization	_____
5.	Wahdan et al. [11]	CNN, LSTM, RNN, FFNN, MLP, DBN	_LSTM

6.	Elmitwally. et al. [12]	SVC, Gaussian naïve Bayes	SVC = 80% GaussianNB = 60%
7.	Utomo. et al. [13]	BPNN, SVM, KNN with TF-IDF/BoW/Stem operation	SVM + TF-IDF + Stem = 70.75%
8.	Arkok et al. [14]	Decision Tree (J48), Random Forest, KNN, LibSVM, Naïve Bayes, and Bayes Net	Without smote, Bayes net = 74%, Naïve Bayes = 75% With smote, Random Forest = 96%
9.	Choriulfikri et al. [15]	Gaussian NB, Multinomial NB, Bernoulli NB, Complement NB	Hamming loss of combo of Bernoulli & multinomial NB = 0.1167
10.	Aiman et al. [16]	SVM, Naïve Bayes, KNN, Decision Tree	Hamming loss of SVM = 0.1110-0.14
11	Abdullahi et al.[17]	SVMs, NB, kNN, and J48	Hamming loss of SVM = 0.103
12	Rostam et al.[18]	Naïve Bayes, SVM, KNN	SVM ~ 71%
13	Alashqar[19]	RNN, CNN	RNN = 90.38% CNN = 89.74%
14	Borhani[20]	L3BFGS, L-BFGS-B, BFGS, ITCG, SVM, ML-kNN, ML4BFGS	Hamming loss of ML4BFGS = 84%
15	Yulianto et al. [21]	Naïve Bayes, SVM	Hamming loss of SVM = 0.15

2.4 Open Issues

The literature review identifies several open issues in the application of deep learning to Quranic text analysis:

- Dataset Quality and Size: Challenges include imbalanced labels and limited annotated data, affecting model robustness.
- Interpretability: Deep learning models often lack transparency, making it difficult to understand their decision-making processes, especially in sensitive religious contexts.
- Ethical Considerations: Addressing biases in model predictions, ensuring user privacy, and maintaining cultural sensitivity are critical.
- Generalization: Extending findings beyond English-translated texts to other languages and cultural contexts remains underexplored.

2.5 Summary

The literature review highlights the progress in using deep learning models like RNNs, CNNs, LSTM, BiLSTM, and hybrids for Quranic text analysis, showing improvements in accuracy and other metrics. However, challenges persist in dataset quality, model interpretability, ethical concerns, and generalization across languages. Addressing these issues is crucial for advancing the field and ensuring responsible, effective applications of deep learning in Quranic studies.

CHAPTER 3

Research Methodology

3.1 Overview

The study's research methodology involves gathering and preparing Quranic verses that have been translated into English. This includes cleaning, tokenizing, eliminating stop words, and creating word embeddings using Word2Vec. This data is used to create and train deep learning models, namely Recurrent Neural Networks (RNNs) using LSTM and GRU and Convolutional Neural Networks (CNNs), hybrid model employing an 80-10-10 split for training, validation, and testing. Hyperparameters are adjusted and optimization strategies like dropout and learning rate scheduling are used. Techniques including oversampling, class weighting, and are used to overcome data imbalance. Semantic search and verse retrieval embeddings are produced using transformer-based models, such sentence transformers. Hamming loss, precision, recall, F1-score, and ROC AUC are used to assess the performance of the model; comparative analyses are used to show the advantages and disadvantages of each method. By improving the accuracy and efficacy of Quranic verse categorization and retrieval, this all-encompassing approach seeks to provide insightful information for religious text study.

3.2 Proposed Methodology

The multilabel classification of Quranic verse involves a few specific actions. These specifics are listed below and explain how it works. The suggested method's working flow chart is displayed in Figure 3.2.1.

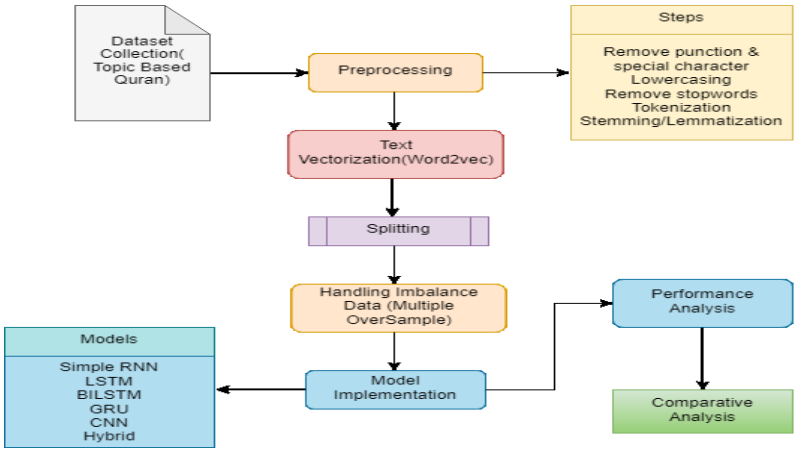


Figure 3.2.1 Methodology

A. Dataset Collection

The dataset used in this study was collected from Holy Quran which was translated in English by Sahih International Book and labeled manually by using web scraping from reliable sources like Quran By Subject and Topic based Quran Verses. The dataset contains 3 columns called Chapter-Verse, Verse & Topic. The dataset contains 15 topics appointed. Topics are Allah, Disbelief, Worship, Heaven/Hell, Love/Marriage/Divorce, Quran, Muhammad, Unseen, Hope, Depression, Haram/Forbidden, Belief, Halal/Permissible, Grave. The dataset have 1596 verses classified into multi labels by 15 topics. Each verse is denoted to minimum 1 topic.

Allah groups the verses about Allah; **Disbelief** is about disbelievers and having disbelief towards Islam. **Worship** is all about prayer, charity, hajj, salah, fasting etc. **Heaven/Hell** classifies the verses about heaven/hell. **Love/Marriage/Divorce** is the verses about love, marriage or divorce rules in Islam. **Quran** is all about the Holy book. **Muhammad** is about Our Prophet Muhammad Saw. **Stories** classify the verses about our prophets stories, about their life, themselves, their family etc. **Hope** is all about having hope about hope that are given by Allah to us and **Depression** talks about removing sadness, anxiety & grief. Unseen is all about the unseen things we don't know or which are beyond our watch. **Haram** gives the verses about all the haram things. Whereas **Halal** gives the verses about the name suggest itself. **Belief** groups the verses that are about having faith in Allah and about believers. **Grave** classifies the verses about after death or about grave. Here is the data set link.[22]

Table 3.2.1 Distribution of verses

Topic	**Number of Verses**	**Topic**	**Number of Verses**
Allah	974	Stories	78
Disbelief	430	Hope	78
Worship	181	Unseen	64
Heaven/Hell	177	Depression	63
Love/Marriage/Divorce	142	Haram/Forbidden	50
Quran	126	Belief	46
Muhammad	97	Halal/Permissible	16

Grave	12

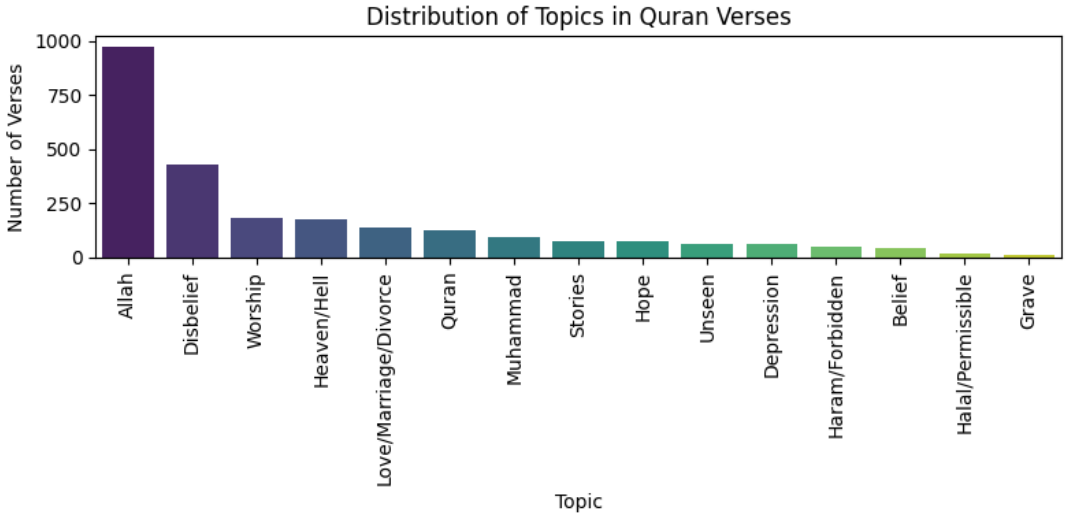


Figure 3.2.2 Distribution of Topics

B. Dataset Preprocessing

Preprocessing steps are necessary to enhance the quality of data to classify them easily.[23] The preprocessing steps we followed to enhance our quality of data to train models are given below.

Basic Preprocessing Steps

The basic preprocessing steps that we used in our dataset is removing punctuations & special characters, converting text to lower case, tokenize the verses, removing stopwords, stemming the tokens and use lemmatization.[24]

Table 3.2.2 Preprocessed Verse

Raw Verse	Preprocessed Verse
Then He causes his death and provides a grave.	[cause, death, provides, grave]
Those are the disbelievers, the wicked ones.	[disbeliever, wicked, one]

C. Text Vectorization

To train deep learning models, text data needs to be converted in numbers so that machine can understand.

Multilabel Binarizer

In this step, we convert our verses topic or labels into binary format using multilabel binarizer to learn machine.

Table 3.2.3 Multilable Binarizer

Verse	Topic
Then He causes his death and provides a grave.	Grave
Those are the disbelievers, the wicked ones.	Disbelief

Verse	Grave	Allah	Disbelief
Then He causes his death and provides a grave.	1	0	0
Those are the disbelievers, the wicked ones.	0	0	1

Word Embeddings

Words with similar meanings can have comparable representations thanks to a kind of word representation called word embeddings. They have greatly enhanced the performance of numerous natural language processing (NLP) tasks and represent a significant advancement in the discipline. Word embeddings, which capture the semantic links between words based on their usage in huge text corpora, portray words as dense vectors of real values. For our verses we used word2vec models CBow architecture.

Word2Vec

Developed by Google, Word2Vec uses shallow neural networks to learn word representations. There are two main architectures: CBOW and n-skip gram.[25] We used CBOW as our dataset is not huge.

Continuous Bag of Words (CBOW): Predicts the current word based on its context (surrounding words).[26] We split our verses into words and created word embeddings by training word2vec and created a matrix of embedding lookup. And converted the text into integer sequences and padded them.

D. Handling Imbalanced Data

In machine/deep learning and data processing, oversampling is a technique used to address class imbalance in a dataset. When there are substantially more or less instances in one class than in other classes, it is referred to as a class imbalance. Because of this imbalance, machine learning models may perform badly, particularly when applied to the minority class, as they may be biased

in favor of the dominant class. We used oversampling to each class independently to handle our imbalance data.

E. Splitting Dataset

Splitting dataset is necessary to train, test & validate our dataset. For splitting we split our dataset into training 80%, testing 10% and for validation 10%.

F. Training

To train the dataset, we have used 6 types of models from where 4 types are RNN types (Simple RNN, LSTM, BILSTM, GRU) , CNN and hybrid model using BILSTM & CNN with additional layers. We've used stratified k fold validation for training and train this dataset over 10 epochs.

G. Testing

For testing, we evaluate our models and got accuracy over them. Among them, Hybrid model got highest accuracy.

H. Performance Analysis

In this stage, we used Hamming loss which is perfect for multi-label classification. Also precision, F1 score, recall, roc-auc curve , confusion matrix were used.

Generating verses

For verse generation based on search index, the given figure approach we followed.

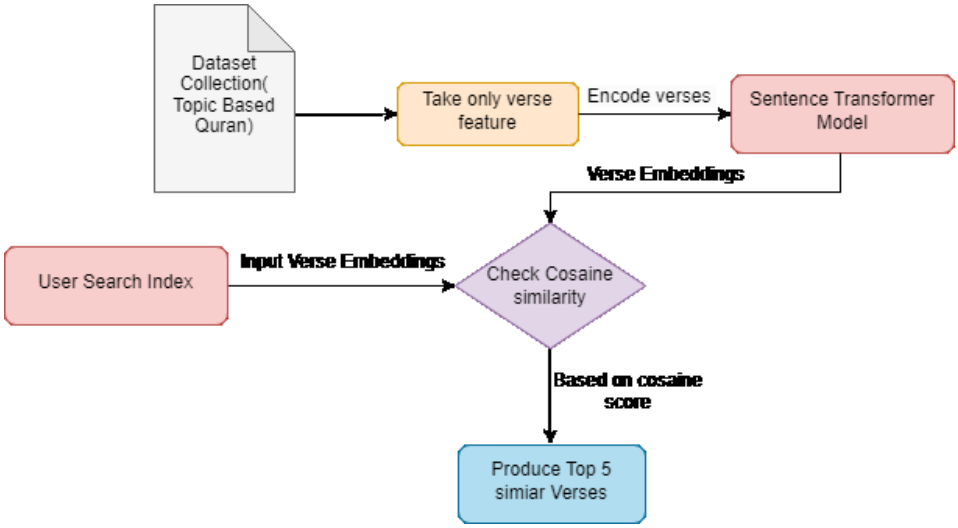


Figure 3.2.3 Methodology of generating verses

Based on built in bert models like sentence transformer, by checking cosaine similarity produce verses similar to search index which we've turned into a web app.

3.3 Hardware/Software Requirement

Hardware requirements:

- Hardware requirements for deep learning and text analysis include high-performance computing resources with enough processing power and memory to handle significant data processing and model training.
- A graphics processing unit (GPU) is necessary for accelerating the training of deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), resulting in a significant reduction in model development and execution times.

Software requirements:

- **Deep learning frameworks:** Popular deep learning frameworks such as TensorFlow and PyTorch are used to design, train, and evaluate models.
- **Python Programming Language:** Python is a versatile and frequently used platform for developing deep learning algorithms and performing data analysis.
- **Natural Language Processing (NLP) Libraries:** Use of NLP libraries such as NLTK, SpaCy, and Hugging Face Transformers for text preprocessing, tokenization, and transformer model implementation to improve semantic understanding and text classification.
- **Data Manipulation Libraries:** Use libraries such as pandas and NumPy to perform efficient data cleaning, manipulation, and numerical calculations.
- **Visualization Libraries:** The use of visualization libraries like as Matplotlib, Seaborn, and Plotly to display data, model performance, and analytic results.

3.4 Project Management & Financial Analysis

- **Project Objectives:**
    A. To categorize Quranic verses based on their themes and topics and generate verses as per users need
    B. To enable readers or users to easily find and read Quranic verses faster that are relevant to their interests and needs
    C. To help readers to better understand the Quran's teachings and apply them to their daily lives
- **Project Timeline:**
    A. Phase 1: Project Planning and Research (2-3 weeks)
    B. Phase 2: Data Collection and Preparation (3-6 weeks)

C. Phase 3: Model Development (7-12 weeks)
D. Phase 4: Integration and Testing (2-4 weeks)
E. Phase 5: Deployment and Evaluation (4-6 weeks)
F. Phase 6: Documentation and Reporting (2-3 weeks)
G. Phase 7: Review and Conclusion (1-2 weeks)

- **Resource Planning:**

A. Equipment and Tools:

- Powerful computer resources for deep learning model training
- Workstations and laptops with the applications and resources required for testing and development installed.
- Metadata, search index, and Quranic text data can all be stored using data storage technologies.

B. Software and Technologies:

- Programming languages (e.g., Python for development, Google Colab Pro version, PyTorch or TensorFlow for deep learning model implementation)
- Libraries for data preprocessing (e.g., pandas, Scikit-learn, Numpy, Matplotlib, spaCy, NLTK (Natural Language Toolkit) etc.)
- NLP Frameworks: (Transformer based models-GPT, BERT, seq2seq models, RNNs etc.)
- Version Control (Git for code versioning)
- Project management tools (Jira and Trello)

C. Data and Content:

- Data: accurate and authentic English translation of the Quran from reliable sources.
- Chapter names, verse numbers, and other information related to the text of the Quran.
- Quranic lexicons and dictionaries for verification and reference.
- Predefined themes/topics for classification
- User documentation: Prepared by the technical writing member.

D. Training and Skill Development:

- Offer training sessions or courses covering machine learning, natural language processing, and deep learning methodologies.
- Plan seminars on Quranic terminology and studies to improve team members' domain expertise.

- By using online resources, webinars, and conferences in related industries, promote lifelong learning and skill development.

- **Risk Management:**

Analyzing a project's SWOT entails determining its factors. The SWOT analysis for my project is as follows:

INTERNAL FACTORS

STRENGTHS +	WEAKNESSES –
➢ **Rich collection of Quranic literature translated into English.** ➢ **Application of deep learning and sophisticated machine learning methods.** ➢ **Possibility of having a significant effect on how people interpret and relate to the Quran.** ➢ **Multidisciplinary team with a wide range of skills.**	➢ **Difficulties with verifying the authenticity and correctness of data.** ➢ **Technical intricacy needing specific expertise and resources.** ➢ **Deep learning model training requires a lot of resources.** ➢ **Possible roadblocks to promoting user involvement and adoption.**

EXTERNAL FACTORS

OPPORTUNITIES +	THREATS –
➢ **Users' personalized learning experiences.** ➢ **Participation of Quranic experts and the community.** ➢ **Possibility of working along with religion groups.** ➢ **Prospects for using value-added services to achieve commercialization.**	➢ **Competition from already-existing platforms with comparable features.** ➢ **Ethical issues with privacy and religious sensitivities.** ➢ **Adherence to regulations pertaining to content control and data privacy.** ➢ **Technological hazards including algorithmic biases or system malfunctions.**

Figure 3.4.1: SWOT Analysis

- **Communication Plan:**

A. Meetings Control:

Purpose: Discuss and approve any changes to the project scope or requirements.

Participants: Supervisor and team member.

Frequency: As needed when change requests arise.

Financial Analysis:

Table 3.4.1: Estimated Cost for Deep Learning based Project

SN	Components	Estimated Cost (BDT)
01.	Hardware/Infrastructure	4500-6000
02..	Software and Tools	2000-3000
03.	Data Collection and Processing	2500-3000
04.	Documentation and Report Writing	500-1000
05.	Miscellaneous (e.g.,English translated Holy Quran, Books related to Quranic verses , tools)	500-1000
06.	Contingency (10% of total)	1000-1500
	Total Estimated Cost	10,500-15,500

3.5 Summary

A systematic approach introduced for Classification of verses & generating them. By collecting data, labeling, preprocessing using algorithms like word2vec for better accuracy, training deep learning models with the dataset, testing & validating them evaluation matrix by obtaining accuracy. Also generating verses based on search index using NLP approaches so that user can easily access.

CHAPTER 4

Implementation

4.1 Overview

This project's implementation includes a thorough strategy to classifying English-translated Quranic verses and generating appropriate verses depending on user queries via advanced deep learning algorithms. Initially, the data is thoroughly preprocessed, including text cleaning, tokenization, stop word removal, and lemmatization. The Word2Vec model generates word embeddings, which comprise an embedding matrix for neural network initialization. To remedy class imbalance, minority classes are oversampled. The basic model is built using both Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), using their respective strengths in sequential data processing and pattern recognition. These models are trained and evaluated using metrics including hamming loss, precision, recall, F1-score, and ROC AUC. Transformer-based models are also fine-tuned to provide sentence embeddings for semantic search, which allows for accurate verse retrieval depending on user queries. The project focuses on model interpretability and uses visualization tools to provide classification findings and insights, resulting in a transparent and practical solution for Quranic text analysis.

4.2 Train Model

To train the dataset, we have used 6 types of models from where 4 types are RNN types (Simple RNN, LSTM, BILSTM, GRU) , CNN and hybrid model using BILSTM & CNN with additional layers. We've used stratified k fold validation for training and train this dataset over 10 epochs.

Deep Learning (DL)

As we know, Deep Learning models are now essential for handling text datasets. This part will give information about our application of 6 models and how to evaluate them to find better accuracy using evaluation metrics.

Recurrent Neural Network (RNNs)

RNNs work on sequence data. There are many types. I'll describe about the models I've used.

Simple RNNs

A Simple Recurrent Neural Network (RNN) is a form of neural network that processes sequential data by updating a hidden state at each time step.[27] This hidden state serves as a memory for

earlier inputs in the sequence, making RNNs especially effective for jobs that require context from preceding data points.

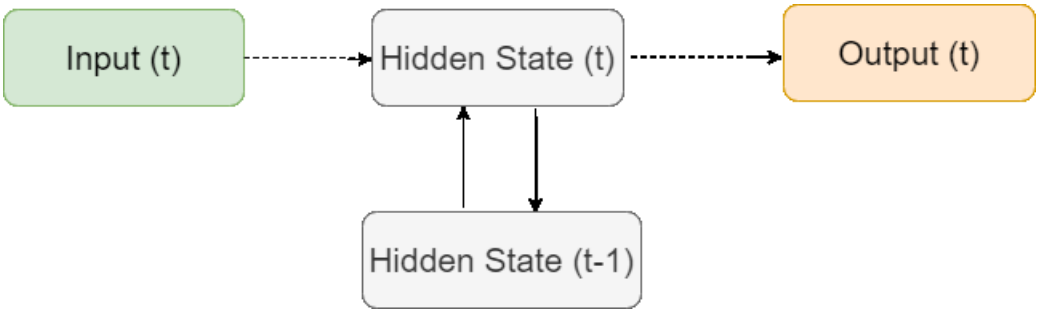


Figure 4.2.1 Simple RNN model

LSTM

Long Short-Term Memory (LSTM) networks are recurrent neural networks (RNNs) designed to detect long-term dependencies in sequential input.[28] Unlike traditional RNNs, LSTMs avoid the vanishing gradient problem, allowing them to store knowledge over longer time periods.

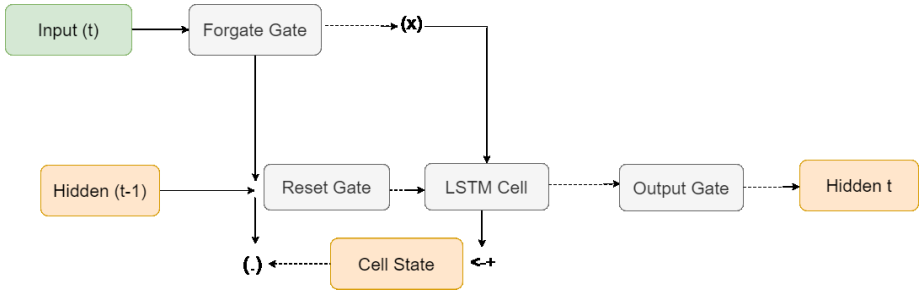


Figure 4.2.2 LSTM model

BILSTM

Bidirectional Long Short-Term Memory (BiLSTM) networks extend the capabilities of regular LSTMs by processing input sequences in both forward and reverse orientations.[29] This enables the network to gather context from both past and future input, making it particularly useful for jobs that necessitate a thorough comprehension of context.

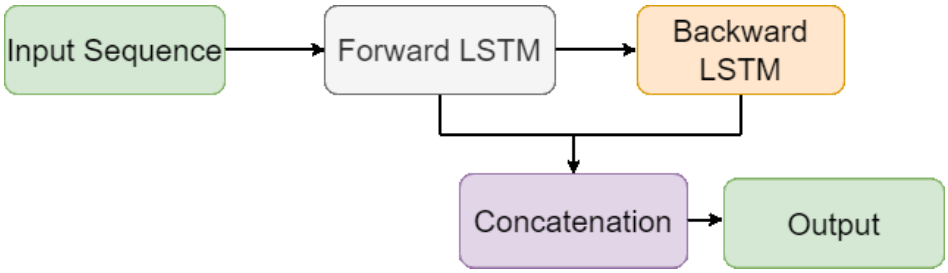


Figure 4.2.3 BILSTM model

GRU

The Gated Recurrent Unit (GRU) is a recurrent neural network (RNN) architecture that addresses the vanishing gradient problem common in regular RNNs.[30] It introduces gating methods to regulate information flow through the network, allowing it to better capture long-term dependencies.

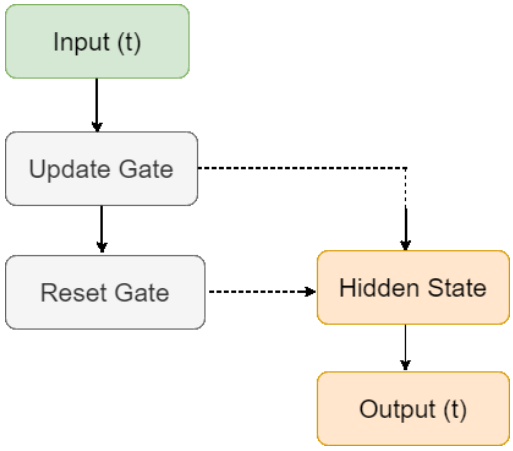


Figure 4.2.4 GRU model

CNN

Convolutional Neural Networks (CNNs) are deep neural networks that process and evaluate visual data, such as photographs. But they can also work on text data. They are distinguished by their ability to autonomously learn spatial hierarchies of features using convolutional layers. Vanishing gradient problem, allowing them to store knowledge over longer time periods.

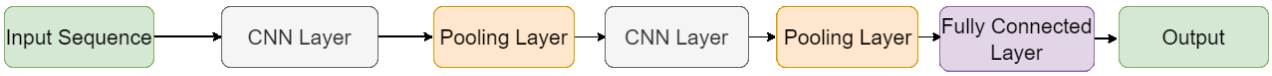


Figure 4.2.5 CNN model

Hybrid Model

The hybrid BiLSTM-CNN model uses BiLSTM layers to detect sequential dependencies and CNN layers to extract spatial properties from input sequences. This combination enables the model to learn both long-term dependencies and local patterns concurrently, making it useful for tasks such as text categorization, sentiment analysis, and sequence labeling. We built hybrid model using BILSTM model and CNN model and additional layers such as pooling, dropout.

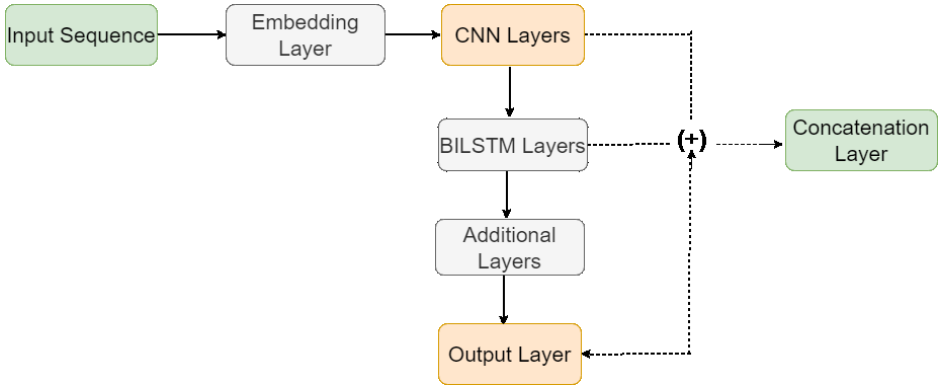


Figure 4.2.6 Hybrid model

We have used these 6 models to train our dataset using 2-k fold cross validation. We used stratified k fold cross validation as out dataset was imbalanced in labels. Also every model was run over 10 epochs per fold.

4.3 Model Evaluation

For model evaluation, we've used evaluation metrics like accuracy, precision, recall, f1 score, hamming loss, roc-auc curve & confusion matrix.

Accuracy

Accuracy reflects how frequently a model correctly classifies input data. It's calculated as the number of correct estimates divided by the total number of predictions.[31] Here True Positives (TP) are instances in which the model accurately predicts the positive class.

True Negatives (TN) are instances in which the model accurately predicts the negative class.

False Positives (FP) occur when the model wrongly forecasts the positive class. False Negatives (FN) are occasions where the model incorrectly predicts the negative class.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} - - - - - - - (1)$$

Precision

Precision is defined as the ratio of accurately predicted positive observations to all expected positives.[32] It reflects how often of the model's favorable predictions are accurate.

$$Precision = \frac{TP}{TP + FP} - - - - - - - (2)$$

Recall

Recall, also known as sensitivity or true positive rate, is the proportion of accurately anticipated

positive observations to actual positives. It evaluates the capacity of the model to detect every positive occurrence.

$$Recall = \frac{TP}{TP + FN} - - - - - - - - - (3)$$

F1-Score

The F1 Score is a harmonious average of recall and precision. It gives a single statistic that balances a model's precision and recall, which is very valuable when dealing with imbalanced datasets.

$$F1 - Score = 2\frac{(Precision * Recall)}{(Precision + Recall)} - - - - - - - - - (4)$$

Hamming Loss

Hamming loss is a metric used to evaluate the performance of multi-label classification models.[33] It measures the fraction of labels that are incorrectly predicted, considering both false positives (incorrectly predicted positive labels) and false negatives (missed true positive labels). Essentially, it provides an average measure of the errors per label.

Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model by comparing the predicted labels against the actual labels. It provides a detailed breakdown of the types of errors made by the model and helps in understanding the performance in a more granular way.

4.4 Summary

Implementing a deep learning model involves several key steps: data preprocessing to prepare and clean datasets, selecting appropriate model architectures like RNNs or CNNs tailored to specific tasks such as Quranic verse classification or semantic search, training the model using labeled data to adjust parameters and optimize performance metrics, and evaluating its effectiveness through metrics like accuracy and F1 score. This process ensures robust and efficient models capable of handling complex text analysis tasks with accuracy and reliability.

CHAPTER 5
Result & Analysis

5.1 Overview

Result analysis involves evaluating deep learning models through metrics like accuracy, precision, recall, and F1 score to gauge performance. It includes interpreting confusion matrices to understand prediction strengths and weaknesses, identifying errors for improvement, and using visualizations like ROC curves to optimize model thresholds. Qualitative assessments ensure outputs meet expectations, guiding enhancements for better applicability in tasks such as Quranic verse classification and semantic search. This iterative process refines models, enhances accuracy, and fosters innovation in text analysis and interpretation.

5.2 Experimental Result

Total 6 different models were used on the dataset up to 10 epochs per fold. The following table shows the evaluation metrics of these models.

Table 5.2.1 Performance Analysis of models

Models	Hamming Loss	Accuracy	Precision	Recall	F1-Score
Simple RNN	0.094	87.68%	0.85	0.37	0.27
LSTM	0.094	87.68%	0.85	0.37	0.27
BILSTM	0.093	89.89%	0.75	0.26	0.27
GRU	0.095	89.29%	0.83	0.24	0.23
CNN	0.094	87.91%	0.85	0.37	0.27
Hybrid	0.03	99.91%	0.86	0.84	0.85

Simple RNN model gives the accuracy about 87.68% and hamming loss is 0.094 that means the prediction of wrong level is not so much. In confusion matrix, we can see that almost for all labels the prediction was pretty correct. But for label Allah it predicts allah for 63 verse whereas the actual prediction was the other label and for label disbelief it predicts other label for 43 verses whereas it

was disbelief label. And in roc curve, the auc is near to 0.77 which shows the clear difference between positive and negative instances.

The Lstm model got 87.68% accuracy same as simple rnn. Their hamming loss and other values were quite similar. Also in confusion matrix their predictions were quite similar. Also the auc value is same 0.77. The BILSTM model got accuracy about 89.29% which shows the progess compared to previous models. The hamming loss was also small 0.095. For label allah, 33 actual prediction was predicted as other label and 31 verses were labeled as allah but the actual result was other label. The auc value is 0.81 which shows clear difference between positive and negative values.

The GRU model also got accuracy about 89.29% same as bilstm and hamming loss was 0.095. For label allah, 33 actual prediction was predicted as other label and 36 verses were labeled as allah but the actual result was other label. The auc value is 0.79 which shows clear difference between positive and negative values.

For cnn model, the accuracy was 87.91% and other values like hamming loss, confusion matrix, auc value were quite similar as rnn, lstm. But our hybrid model surpassed every model which got 99.91%. and hamming lass was only 0.03 that means the wrong predictions were too low. In confusion matrix, for love/marriage/divorce the prediction was wrong for 9 instances. Also the auc value was 0.97. The following figure shows hybrid's training & validation accuracy and loss and confusion metrics of every label and roc curve.

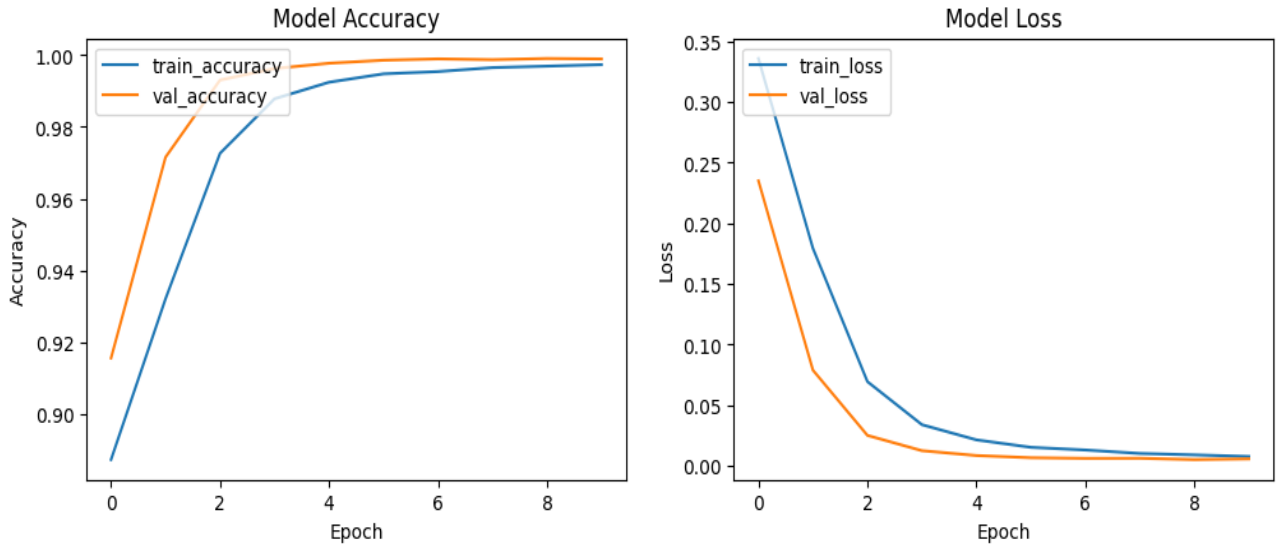


Figure 5.2.1 Accuracy & Loss of Hybrid

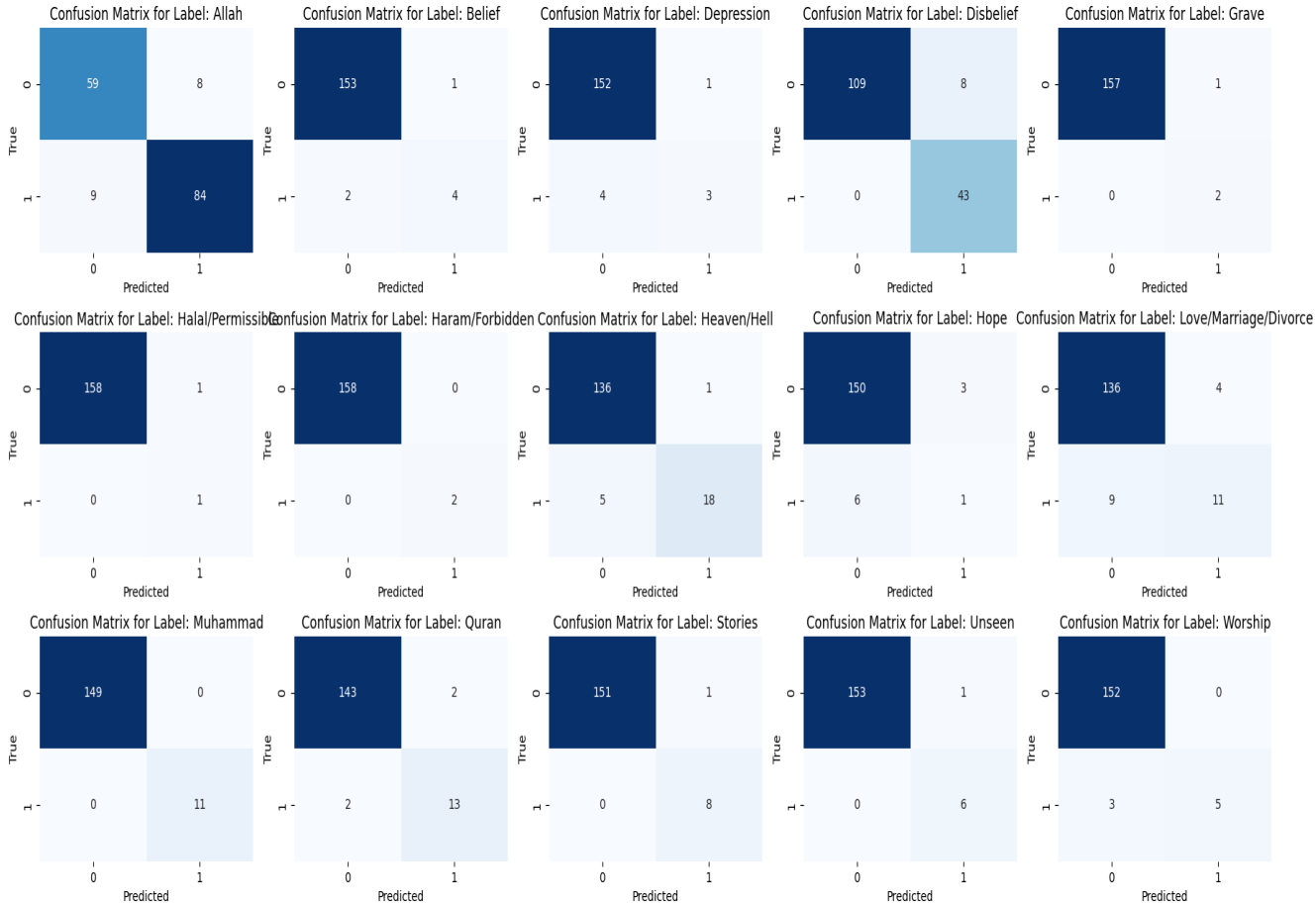


Figure 5.2.2 Conf_matrix of Hybrid

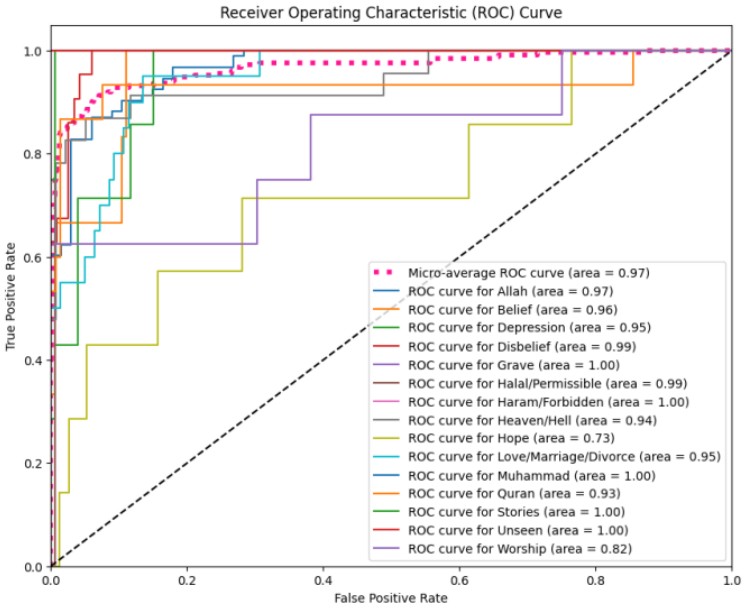


Figure 5.2.3 Roc-Auc of Hybrid

For generating verses based on cosaine similarity between embedding verses using sentence transformer model, the user can search by a topic and the top 5 related verses will be shown srom

our dataset. It was built using streamlit via visual studio.

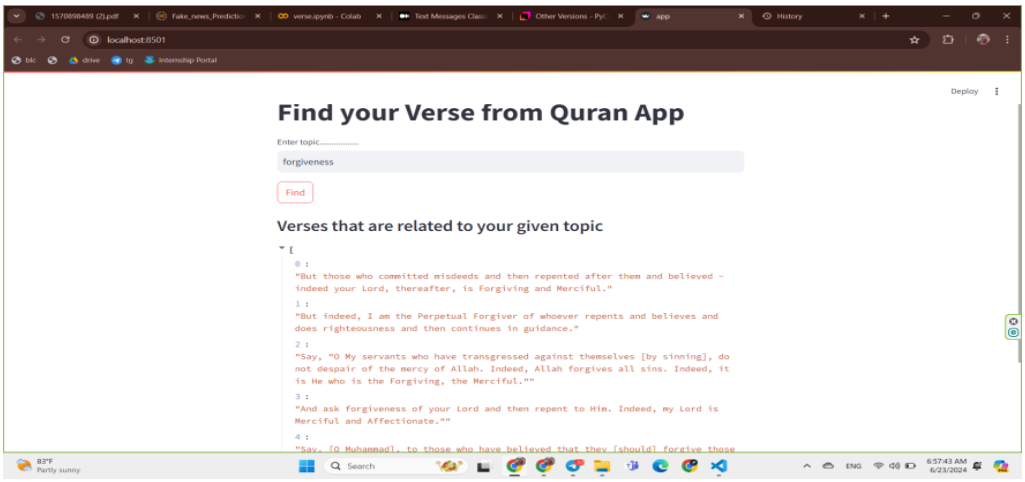


Figure 5.2.4 Generating verses

5.3 Comparative Analysis

In a comparative analysis of deep learning models applied to tasks like Quranic verse classification and semantic search, several key models such as RNNs, CNNs, LSTM, BiLSTM, and hybrid architectures are evaluated based on their effectiveness and suitability. Each model offers distinct advantages: RNNs excel in sequential data processing, CNNs in spatial pattern recognition, LSTM and BiLSTM in capturing long-term dependencies, and hybrid models combining these strengths for enhanced performance. The analysis involves metrics like accuracy, precision, recall, and F1 score, alongside qualitative assessments to determine which model best aligns with specific task requirements. This comparative approach informs decision-making for deploying optimal models in text analysis and interpretation, ensuring robust performance and applicability in diverse contexts.

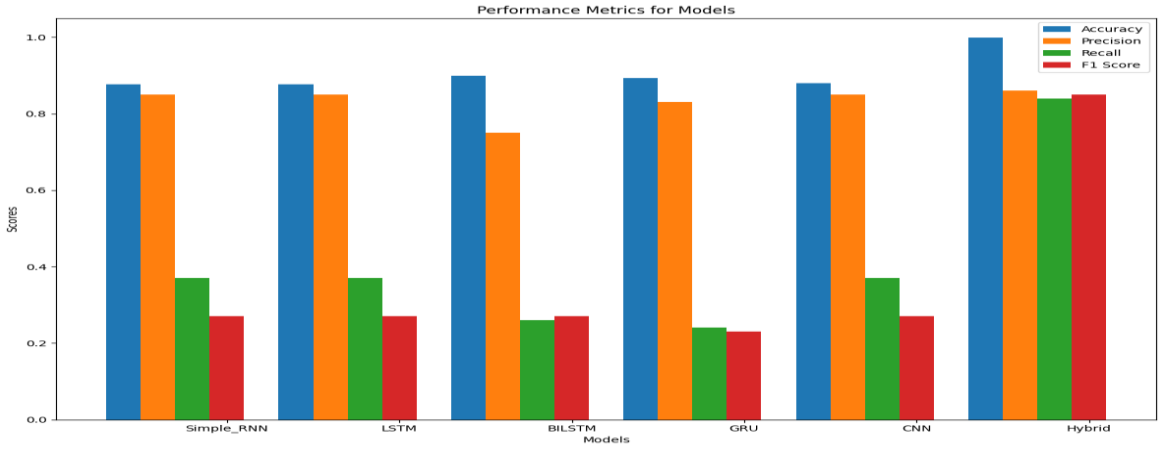


Figure 5.3.1 Comparative Analysis

5.4 Summary

Result analysis in the context of deep learning involves evaluating model performance through metrics like accuracy, precision, recall, and F1 score. It includes interpreting confusion matrices to understand prediction strengths and weaknesses, identifying errors for improvement, and using visualizations such as ROC curves to optimize model thresholds. Qualitative assessments ensure that model outputs meet expectations and are aligned with task requirements, guiding refinements and enhancements for better performance in applications like Quranic verse classification and semantic search. This iterative process aims to validate model effectiveness, optimize outcomes, and drive advancements in text analysis and interpretation.

CHAPTER 6

Impact on Society, Environment and Sustainability

6.1 Impact on Life

The impact of the project on life focuses on how the outcomes and applications of the deep learning models for Quranic verse classification and semantic search can benefit individuals. This includes facilitating easier access to religious texts, enhancing understanding of complex concepts in the Quran, and potentially aiding in personal spiritual growth and reflection. By improving access to and interpretation of Quranic verses, the project aims to positively influence the spiritual and intellectual lives of users.

6.2 Impact on Society and Environment

The societal impact of the project revolves around its contributions to education, research, and cultural preservation. By enabling efficient classification and retrieval of Quranic verses, the project supports academic research in Islamic studies and enhances access to religious teachings for diverse communities. Moreover, by leveraging sustainable computing practices and ethical data handling, the project minimizes environmental impact and promotes responsible technological development within society.

6.3 Ethical Aspects

Ethical considerations in the project encompass data privacy, fairness, and transparency. Ensuring user data confidentiality and informed consent in data collection and usage protects user privacy. Fairness involves mitigating biases in model training and evaluation, particularly in culturally sensitive contexts such as religious text analysis. Transparency in model decisions and interpretations fosters trust among users and stakeholders, promoting ethical standards throughout the project lifecycle.

6.4 Sustainability Plan

The sustainability plan outlines strategies for long-term project viability and impact. This includes

maintaining and updating the deep learning models to adapt to evolving user needs and technological advancements. Additionally, incorporating sustainable practices in model training, deployment, and maintenance ensures minimal environmental impact. Collaboration with stakeholders, continuous monitoring of performance metrics, and resource optimization contribute to the project's sustainability and scalability over time.

6.5 Summary

The summary encapsulates the project's achievements, challenges, and future directions. It highlights key findings from result analyses, discusses implications for further research or applications, and reinforces the project's contributions to Quranic studies and text analysis. By summarizing the impact on life, society, and ethical considerations, as well as outlining a sustainable plan, the summary provides a comprehensive overview of the project's outcomes and its potential for continued growth and relevance in the field of deep learning and Islamic studies.

CHAPTER 7
Conclusion & Future Work

7.1 Conclusions

The conclusions section outlines the major results and outcomes of the study. It highlights the effectiveness of deep learning models in Quranic verse classification and semantic search, emphasizing key metrics such as accuracy, precision, recall, and F1 score. The section also discusses how the project addresses its objectives, validates its hypotheses, and contributes to advancements in text analysis and interpretation within the context of Islamic studies. Conclusions may also touch upon practical implications for users and stakeholders, reinforcing the project's significance and potential impact.

7.2 Further Suggested Works

Additional proposed works present options for future research and development based on the project's findings and limits. This includes exploring enhancements to existing models, such as integrating newer deep learning architectures or refining algorithms for better performance. Additionally, suggestions may involve expanding the dataset to improve model robustness, conducting comparative studies with other religious texts, or exploring applications in related fields like natural language processing or cultural heritage preservation. These recommendations aim to advance knowledge and applications in Quranic studies and deepen understanding of computational methods in religious text analysis.

7.3 Limitations/ Conflict of Interests

The limitations section acknowledges constraints or challenges encountered during the project, such as dataset limitations, computational resources, or inherent biases in model predictions. It transparently discusses potential areas where the model may not perform optimally or where further improvements are needed. Conflict of interests, if applicable, addresses any biases or influences that may have impacted the project's execution or interpretation of results. By identifying these limitations and conflicts, the section ensures transparency and credibility in research outcomes, guiding future investigations and mitigating potential biases. These sections collectively provide a comprehensive overview of the project's outcomes, implications for future research, and considerations for ethical and transparent research practices within the domain of deep learning applied to Quranic text analysis and interpretation.

REFERENCES

[1] Ahsan, Z. (2013). The Qur'an, basic scientific research and technology (perspective). Revelation and Science, 3(01).

[2] Fatma, G., Pirzada, N., & Begum, S. (2022). Problems, Illusions and Challenges Faced by a non-Arabic Speaker in Understanding Quran: A Sub-Continental Study. Journal of Positive School Psychology, 6(2), 5422-5426.

[3] Kadhim, B. J., Merzah, Z., & Ali, M. M. (2023). Translatability of the Islamic terms with reference to selected Quranic Verses. International Journal of Linguistics, Literature and Translation, 6(5), 19-30.

[4] Nasir, W. Deep Learning for Natural Language Processing: Current Trends and Future Directions.

[5] Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., & Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200.

[6] Bashir, M. H., Azmi, A. M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2023). Arabic natural language processing for Qur'anic research: a systematic review. Artificial Intelligence Review, 56(7), 6801-6854.

[7] Mohamed, E. H., & Shokry, E. M. (2022). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*, *34*(3), 934-945.

[8] Zouaoui, S., & Rezeg, K. (2021). A novel quranic search engine using an ontology-based semantic indexing. *Arabian Journal for Science and Engineering*, *46*, 3653-3674.

[9] Beirade, F., Azzoune, H., & Zegour, D. E. (2021). Semantic query for Quranic ontology. *Journal of King Saud University-Computer and Information Sciences*, *33*(6), 753-760.

[10] Bsoul, Q., Abdul Salam, R., Atwan, J., & Jawarneh, M. (2021). Arabic Text Clustering Methods and Suggested Solutions for Theme-Based Quran Clustering: Analysis of Literature. *Journal of Information Science Theory and Practice*, *9*(4), 15-34.

[11] Wahdan, A., Hantoobi, S., Salloum, S. A., & Shaalan, K. (2020). A systematic review of text classification research based on deep learning models in Arabic language. *Int. J. Electr. Comput. Eng*, *10*(6), 6629-6643.

[12] Elmitwally, N. S., & Alsayat, A. (2020). The multi-class classification for the first six surats of the Holy Quran. *International Journal of Advanced Computer Science and Applications*, *11*(1), 327-332.

[13] Utomo, F. S., Suryana, N., & Azmi, M. S. (2020). Stemming Impact Analysis On Indonesian Quran Translation And Their Tafsir Classification For Ontology Instances. *IIUM Engineering Journal*, *21*(1), 33-50.

[14] Arkok, B., & Zeki, A. (2020). Classification of Holy Quran Verses based on Imbalanced Learning. *International Journal on Islamic Applications in Computer in Science and Technology*, *8*, 11-24.

[15] Choirulfikri, M. R., Lhaksamana, K. M., & Al Faraby, S. (2022). A Multi-Label Classification of Al-Quran Verses Using Ensemble Method and Naïve Bayes. *Building of Informatics, Technology and Science (BITS)*, *3*(4), 473-479.

[16] Aiman, A. S., & Lhaksmana, K. M. (2022). Topic Classification of Quranic Verses in English Translation Using Word Centrality Measurement. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(5), 803-809.

[17] Abdullahi, A., Samsudin, N. A., Rahim, M. H. A., Khalid, S. K. A., & Efendi, R. (2021). Multi-label classification approach for Quranic verses labelling. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, *24*(1), 484-490.

[18] Rostam, N. A. P., & Malim, N. H. A. H. (2021). Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University-Computer and Information Sciences*, *33*(6), 658-667.

[19] Alashqar, A. M. (2023). A Classification of Quran Verses Using Deep Learning. *International Journal of Computing and Digital Systems*, *14*(1), 1-xx.

[20] Borhani, M. (2020). Multi-label Log-Loss function using L-BFGS for document categorization. *Engineering Applications of Artificial Intelligence*, *91*, 103623.

[21] Yulianto, F., Lhaksmana, K. M., & Murdiansyah, D. T. (2021). Classifying Quranic Verse Topics using Word Centrality Measure. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(3), 594-601.

[22] quran_verse.csv, last accessed at <<https://drive.google.com/file/d/1aStt0EqibTKbI52u4ZU1WUBweXU5gHqK/view?usp=sharing >>, last accessed on 13-7-2024 at 12.00 AM.

[23] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

[24] Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3), 509-553.

[25] Sivakumar, S., Videla, L. S., Kumar, T. R., Nagaraj, J., Itnal, S., & Haritha, D. (2020, September). Review on word2vec word embedding neural net. In 2020 international conference on smart electronics and communication (ICOSEC) (pp. 282-290). IEEE.

[26] Agarwal, S., & Arora, S. (2007, May). Context Based Word Prediction for Texting Language. In RIAO (pp. 360-368).

[27] Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

[28] Staudemeyer, R. C. (2015). Applying long short-term memory recurrent neural networks to intrusion detection. South African Computer Journal, 56(1), 136-154.

[29] Elfaik, H., & Nfaoui, E. H. (2020). Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text. Journal of Intelligent Systems, 30(1), 395-412.

[30] Zargar, S. (2021). Introduction to sequence learning models: RNN, LSTM, GRU. Department of Mechanical and Aerospace Engineering, North Carolina State University.

[31] Bernadó-Mansilla, E., & Garrell-Guiu, J. M. (2003). Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. Evolutionary computation, 11(3), 209-238.

[32] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).

[33] Norouzi, M., Fleet, D. J., & Salakhutdinov, R. R. (2012). Hamming distance metric learning. Advances in neural information processing systems, 25.

[34] Alshammeri, M., Atwell, E., & ammar Alsalka, M. (2021). Detecting semantic-based similarity between verses of the Quran with Doc2vec. *Procedia Computer Science*, *189*, 351-358.

[35] Mostafa, A., & Mohamed, O. (2022, June). GOF at Qur'an QA 2022: Towards an Efficient Question Answering For The Holy Qu'ran In The Arabic Language Using Deep Learning-Based Approach. In Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection (pp. 104-111)

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:
CLASSIFICATION OF ENGLISH TRANSLATED HOLY QURAN & GENERATING VERSES BASED ON SERCH INDEX

Student ID: [193-15-13522]

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a **pneumonia detection** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding. 18	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**

CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project requires data collection of English translated Quranic verses and labeling them by separating into different classes. For classification of this verses, I've to know the Deep learning models like RNN, LSTM, BILSTM, CNN, etc which covers K3 and K4.	**Page no:** [15-19] **Section:** [3.2] **Page no:** [24-27] **Section:** [4.2]
				This project requires a systematic approach or methodology which involves data collection, pre -processing, training deep learning based models and testing & validating the dataset to ensure accuracy which needs to be designed before implementation. Also for pre -processing & training models, I've to use Python	**Page no:** [15-19] **Section:** [3.2]

				programming language that covers K5 & K6.	
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to advance multilabel text classification using deep learning, showcasing a comprehensive understanding of current methodologies.	**Pa-ge no:** [9-13] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing challenges such as scraping and manually labeling some amount of quranic verses based on user index topics or themes. Also needed to be extra conscious for maintain authenticity as I 'm using English translated Holy quranic verses. So that no misconception can arise. Also faced limitations of traditional methods and the complexities of multilabel classification using DL.	**Page no:** [15-19] **Section:** [3.2]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	Finding the appropriate model seems challenging as most of works are on only classification of quranic verse whereas I've dealt with classification as well as generation of quranic verses based on search index. And also different studies used different types of datasets & algorithms that's why I've conducted in -depth analysis for choosing the suitable algorithm.	**Page no:** [9-13],[24-27] **Section:** [2.2,2.3],[4.2]

4.	**EP4: Familiarity of Issues**	Yes	CO8	For understanding the English translated quranic verse and specifically how to label or classify them based on users need & connection with the verses which covers EP4.	**Page no:** [9-14] **Section:** [2.2, 2.3, 2.4]
5.	**EP7: Interdependence**	Yes	CO5	This project comprises several subsystems like quranic verse collection , labeling , preprocessing , training deep/machine learning models, testing/validating, generating verses based on search index which ensures **EP-7**.	**Page no:** [15-23] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep learning frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements in multilabel text classification problem.	**Page no:** [20-23] **Section:** [3.3, 3.4]

2.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by accelerating and promoting Qurans knowledge through multilabel quran verse classification and generating of verses based on search methods, while also promoting environmental sustainability by sharing English translated Islamic knowledge just by searches.	**Page no:** [34] **Section:** [6.2]
3.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a multilabel verse classification using dl techniques like rnn and cnn demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [9-14] **Section:** [2.1, 2.2,2.3,2.4]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [20-23] **Section:** [3.4]
2	CO9	Yes	The project demonstrates adherence to ethical principles by ensuring transparency, fairness, and cultural sensitivity in the application of deep learning models to Quranic text analysis. It prioritizes the accurate representation of religious teachings, maintains user privacy, and transparently documents the research process. This commitment ensures responsible knowledge dissemination and societal well-being, aligning with ethical standards in the application of advanced technologies for religious and cultural studies which comply with **CO9**.	**Page no:** [34] **Section:** [6.3]
3	CO12	Yes	The project's dedication to continuous	**Page no:** [15-32]

			improvement and adaptation within the dynamic field of AI and text analysis is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental evaluation. This commitment ensures that the project stays updated with the latest advancements, refining techniques to effectively address contemporary challenges in Quranic text classification and retrieval.	**Section:** [3.2, 3.3, 3.4, 4.2, 4.3, 5.2, 5.3]

Title

ORIGINALITY REPORT

13%	11%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
3	fastercapital.com Internet Source	1%
4	Submitted to University of North Texas Student Paper	1%
5	Submitted to Chiang Mai University Student Paper	<1%
6	www.ijraset.com Internet Source	<1%
7	medium.com Internet Source	<1%
8	arxiv.org Internet Source	<1%
9	Shikha Verma, Aditi Sharan, Nidhi Malik. "Efficient Classification of Hallmark of Cancer Using Embedding-Based Support Vector	<1%