Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Siamese Deep Learning Framework for Duplicate Question Detection using Contextual and Lexical Feature Fusion

Meherzadi Muntaha
University of Central Punjab
Muhammad Junaid Asif
junaid.asif@ncp.edu.pk

Artificial Intelligence Technology Centre (AITeC), National Centre for Physics (NCP)
https://orcid.org/0000-0002-3875-0777
Naveed Hussain
University of Central Punjab
Rana Fayyaz Ahmad
Artificial Intelligence Technology Centre (AITeC), National Centre for Physics (NCP)

Research Article

Keywords: Duplicate Question Detection, QUORA, Siamese Neural Network (SNN), BERT Embeddings, GloVe Embeddings, 1D-CNN, Hybrid Deep Learning Model.

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10507899/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Hybrid Siamese Deep Learning Framework for Duplicate Question Detection using Contextual and Lexical Feature Fusion

Muntaha Meherzadi[1], Muhammad Junaid Asif[1,2], Naveed Hussain[1], and Rana Fayyaz Ahmad[2]

[1] Faculty of IT and Computer Sciences (FoIT&CS), University of Central Punjab (UCP), Lahore 54000, Pakistan
[2] Artificial Intelligence Technology Centre (AITeC), National Centre for Physics (NCP), Islamabad 44000, Pakistan

***Abstract*—The Community Question Answering (CQA) platforms such as QUORA, Reddit and Stack Overflow have become important knowledge sharing sites, where users asks questions and are answered by a community of users worldwide. With the expansion of platforms, the frequency of identical and semantically similar questions increases tremendously. Similar questions can vary in words, structure or lexical form, although their meaning is usually the same. Such redundancy results in poor information access, moderation labor, and poor user experience. Duplicate question detection has enabled automatic processing has become a valuable research problem in Natural Language Processing (NLP) and Artificial Intelligence (AI), especially in large and dynamic systems such as QUORA. This research paper presents a hybrid deep learning framework, using Siamese Neural Networks (SNN) to detect duplicate questions. The proposed architecture comprises two parallel operating SNN sub-models; SNN_A and SNN_B. SNN_A leverages BERT embeddings to learn semantic relationships and paraphrasing patterns, and the other sub-model SNN_B captures the lexical semantics and sequential dependencies by incorporating the pre-trained Glove Embeddings, 1D-CNN and Bidirectional Long-short Term Memory (BiLSTM). Moreover, linguistic similarity characteristics are also handcrafted: word overlap, longest common subsequence (LCS), and fuzzy similarity scores are included. A feature fusion system combines both contextual, lexical and handcrafted features into one single representation, which is subsequently sent to a fully connected layer where the duplicate questions are again classified. The experiment has been carried out on the Quora Question Pairs dataset of more than 400,000 labeled question pairs, with the 80:20 stratified split forming the training and validation sets. Accuracy, precision, recall, F1-score, and confusion matrix analysis were used to evaluate model performance using AdamW optimization. The proposed architecture achieves 85.03% accuracy and 0.81 F1-Score, which is better than the baseline models. Results demonstrate that hybrid systems containing contextual and handcrafted elements enhance semantical similarity learning and facilitate duplication detection scalably.**

***Index Terms*—Duplicate Question Detection, QUORA, Siamese Neural Network (SNN), BERT Embeddings, GloVe Embeddings, 1D-CNN, Hybrid Deep Learning Model.**

I. Introduction

QUORA, an American Community-based Question Answer (SQA) platform, founded in June 2009 and publicly available in June 2010, is a popular platform that span on a vide range of social topics and has attracted an increasingly large and diverse users in last years. However, duplicate queries (with different words and having the same semantic sense) tend to occur frequently on Quora. These duplicate occurring questions require manual identification and marking by experienced and high-reputation users. Quora is not a single platform using for question-answering, there are also a variety of different platforms used for this purpose such as Stack Overflow, Yahoo Answers, Reddit and many others.

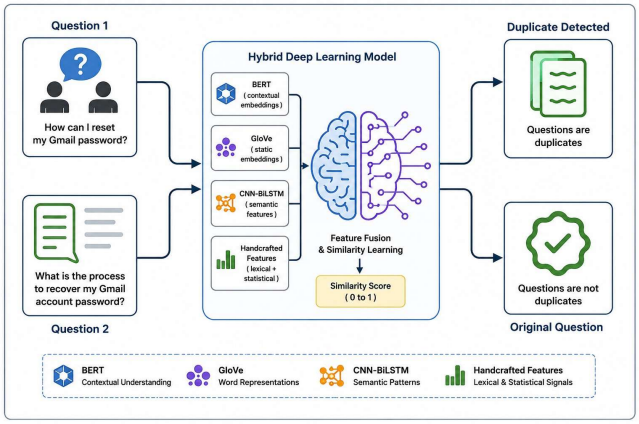


Fig. 1. Overview of Duplicate Question Detection Task

Among various Q&A platform, Quora is one of the most trendy platform, and it had over 300 million Monthly Active Users (MAU) in 2020. This has increased quite diverse to a level exceeding 400 million MAUs by 2025 of which 148 million United States users and over 100 million MAUs India users. The platform can be described as popular especially amongst younger audiences with most of the active users being in the 243˜5 age bracket. [1], [2]. QUORA has much wider scope as compared to Stack Overflow which strictly limits its services to technical and programming related questions and answers [3]. In addition to answering factual questions, the platform also promotes general knowledge exchange, open-ended discussions, and sharing of opinions in many different fields, including science, technology, business, health, lifestyle, and social problems. This broad range of aspects of QUORA will attract more users and encourage them to stick to the free world and not confine themselves to more technical worlds, making the site a global center of informational and conversational knowledge sharing [4].

Due to the broader scope of QUORA, this platform is facing a major challenge of duplicate Q&A, as very often, a query can be posted by another user that has already been asked and answered [5], [6]. This duplication makes retrieval of information more difficult and also reduces the overall quality of posted content. Stack Overflow uses a reputation-based model with reputation points gained by participating in different activities, including answers to questions, up-votes, and edits [7]. Higher-rep-rated users receive more privileges,

such as the ability to flag and label questions as duplicates, and thus actively participate in ensuring the quality and organization of platform [8].

Unlike Stack Overflow, QUORA does not have a system to mark duplicates. It uses, instead, a feature named ”Merge Questions” [9], [10]. Edits are suggested or questions tagged as possible duplicates by regular users, the ultimate decision made by the moderators of Quora or its automated alert system. After merging the duplicate questions, all answers will be combined under one master question, and their duplicate questions will be redirected.

There is a pressing need to suggest an automatic or intelligent methodology of marking identical questions of Quora since the platform is huge and dynamic. Each day, millions of users share questions, which sometimes lead to submitting more queries on the same topic that slightly differ in terms of expression. In its current form, Quora uses both intellectual moderation and automated identification and combination to integrate replicas, yet it is slow, skewed, and resource-intensive. An intelligent system based on deep learning and natural language processing techniques may be able to highlight the semantic similarities between questions, even under different phrasing. Such a system would not only remove redundancy and supplementary management of material, but improve user experience, in that the responses to them will be comprehensive and pulled together within seconds. Moreover, it would conserve a lot of moderation, ensure faster duplicates resolution, and maintain the value and reliability of information in the platform. Lastly, Quora should be scaled by introducing an automated system of duplicates detection, rendering it customer-friendly and knowledge-effective.

To overcome the difficulties of detecting and marking duplicate questions on QUORA, we introduced a hybrid Siamese Neural Network (SNN), which is developed in a manner that it can identify semantic and contextual similarities. The architecture is fed with question pairs in the format of an input and then subjected to various pre-processing functions, including data cleaning, noise removal, tokenization, and embeddings. Then the processed data was subjected to two parallel streams; Stream A consisted of the use of BERT embeddings [11] before a SNN and dense layers, and Stream B consisted of Glove Embeddings into a 1D-CNN-BiLSTM. pipeline and handcrafted features. The extracted features of the two streams are then concatenated to create a combined feature matrix and fed to various machine learning classifiers to determine whether the pair of questions was a duplicate or not.

The research questions that guide this study are as follows:

- RQ1: What effective methods can be used to model semantic similarity between question pairs on large-scale community question-answering platforms to effectively identify duplicate questions even with lexical and static differences?
- RQ2: What is the performance of a hybrid Siamese Neural Network using both contextual embeddings (BERT), static embeddings (GloVe), and hand designed linguistic features on duplicate question detection?
- RQ3: What is the strength of the proposed hybrid structure with real-world, unbalanced data set like Quora Question Pairs dataset?

The key contributions of this research are as follows:

1) We proposed a novel hybrid model based on combination of two parallel streams of Siamese Neural Network (SNN) for effective identification of duplicate questions on the QUORA Platform. The hybrid architecture is much more accurate and robust in detecting a duplicate question as compared to the single-model approach.
2) We presented a context-aware Siamese sub-model SNN_A that uses BERT embeddings to understand deep semantic relations, paraphrasing patterns, and contextual relation between question pairs, which overcomes the short-comings of conventional lexical similarity approaches.
3) The complementary static embedding-based Siamese sub-model SNN_B which uses GloVe embeddings with CNN and BiLSTM layers is created to maintain the global lexical semantics and sequential information which would increase the variety and stability of the representation.
4) The proposed model also includes a collection of handcrafted linguistic similarity methods, such as word overlap, longest common subsequence and fuzzy matching scores, to explicitly represent structural and lexical similarity that might not be entirely reflected by neural embeddings itself.
5) A category of fusion mechanism has been developed to be used effectively to concatenate contextual embeddings, static embeddings and handcrafted features into a common representation space. This combination increases the discriminative ability of the model and allows complementary learning when using heterogeneous sources of features.
6) The results of wide-ranging experimental work on the Quora Question Pairs (QQP) dataset show that the proposed hybrid Siamese architecture performs better in the form of accuracy and F1-score and outperforms classical machine learning baselines and proves the usefulness of hybrid representation learning in duplicate question detection.

The overall structure of this paper is organized as follows to present the research in a clear and systematic manner.

1) **Introduction and Related Work:** The paper start with an introduction, which states the issue of duplicate questions detection on Quora and the interest in creating an intelligent solution. The Related work section talks about the available methods and their limitations in solving semantic similarity challenges.
2) **Materials and Methods:** The methodology section presents the hybrid Siamese Neural Network architecture which is being proposed along with the methods of pre-processing, embedding technique and the dual-stream modeling.
3) **Results and Discussion:** The experimental design and findings section includes the information about the dataset, evaluation metrics, and the performance compared to baseline models.

4) **Conclusions and Future Directions:** At the end of the paper, there is a summary and conclusion that indicates the major findings, practical implications, and possible future research directions.

II. Related Work

Duplicate question detection has become a significant research issue in Community Question Answering (CQA) systems like QUORA where a large amount of user-generated content can result in semantically similar or repeated questions. In the last ten years, numerous studies have investigated different methods such as classic lexical similarity algorithms to state-of-the-art deep learning systems to detect similar pairs of questions. Early research was mainly based on the feature engineering approaches, such as TF-IDF representations [12], bag of words (BoW) [13], word overlap, and similarity measures [14], that offered only computational efficiency but were less effective at representing semantic relations. More recently, contextual embedding models that are based on transformers and Siamese Neural Networks have shown impressive advancements in paraphrase detection and question similarity problems. Even with these developments, there are still difficulties in integrating contextual semantics, lexical information and explicit similarity cues in the same framework. Thus, modern studies are giving more attention to hybrid deep learning models which combine various representation paradigms to improve robustness and performance of duplicate question detection.

One of the major techniques in the text categorization process is support vector machine (SVM) model, which employs several feature extraction methods such as BoW vectors [10]. Recently deep learning methods have shown remarkable results in several NLP tasks, most notably in semantic text similarity [15]–[17]. There have been remarkable results of deep-learning models that are trained on task-specific feature engineering on similarity measure and semantic analysis. Researchers have found that pre-trained word embeddings can capture large semantic symmetries. A combination of word embeddings and deep models is possible to show the semantic meaning of the text [10], [18]. Long short-term memory (LSTM)-based neural networks [19] have shown outstanding performance in activities such as text categorization and information retrieval.

The study [20] proposed a deep learning design of semantic sentence embedding learning based on an encoder-decoder model with a pairwise discriminator. The model which integrates local reconstruction loss with global similarity constraints produces discriminative embeddings and is effective on paraphrase detection tasks showing effectiveness on the QUORA Question Pairs dataset. A Feed forward neural network [21] was trained on TF-IDF features to determine whether a pair of questions is duplicate or not. The pro-posed experimental results indicate high performance with accuracy surpassing numerous baseline methods. Nevertheless, the model is mostly based on lexical characteristics and fails to represent deep contextual semantics and complementary representation fully. This drawback makes the consideration of hybrid architecture that will integrate both contextual embeddings and semantic modeling necessary, which is the rationale behind the offered hybrid Siamese framework in the given study.

Another study presented by Dhakal et al. [22] addresses the issue of duplicate questions in community question-answering systems like QUORA by using deep learning to handle semantic question matching. The authors use the Quora Question Pairs dataset of over 400,000 question pairs with labels and conduct a comprehensive preprocessing operation of tokenization, normalization, stemming, and stop-word elimination. The key part is the feature engineering, in which lexical, fuzzy matching, TF-IDF, and vector distances as well as Word Mover Distance features are obtained and then used to train an artificial neural network (ANN) with several hidden layers. The proposed model has reached the accuracy of about 86.09 which indicates the effectiveness of integrating handcrafted linguistic features and neural models in semantic similarity learning. Nevertheless, the method is also based on a lot of engineered features and shallow ANN architecture that could restrict contextual semantic understanding. This shortcoming underscores the importance of hybrid systems that combine contextual embeddings like BERT with deep Siamese systems, which is in line with the impetus of the current study.

Rishickesh et al. [23] compares the different traditional and machine learning approaches to detecting duplicate questions on the QUORA dataset. Various models including Na¨ıve Bayes, Logistic Regression, Support Vector Machine, Artificial Neural Networks, Recurrent Neural Networks, and XGBoost were experimented with the help of NLP preprocessing and count-vectorizer/TF-IDF as the feature vectors. Among all models, XGBoost achieved an accuracy of around 80.9%. A deep learning approach based on Siamese LSTM was presented by Chandra et al [24]. This study presents a comparison of the classic machine learning algorithms like Logistic Regression, SVM, and XGBoost against deep learning algorithms and concludes that semantic similarity learning is better in Siamese networks. The Siamese LSTM has the lowest log-loss (0.21), and the results of the experiment demonstrate that the model is more effective than classical models. The methodology, however, is predominantly based on fixed embeddings, which implies the necessity of the hybrid architecture with the contextual representations, which is why the current study is motivated.

A hybrid model [25] with weighted fine-tuning of BERT to generate feature extractors and Siamese Bi-LSTM enhances sentence representation by learning contextual, lexical and sequential dependencies. This model produces embedded vectors by using BERT and compares similarity scores between question pairs by using shared Bi-LSTM layers. Evaluations on the Quora dataset through experiment show a strong enhancement in accuracy, precision, recall, and F1-score when using hybrid Siamese architecture-based methods as compared to traditional machine learning and isolated deep learning solutions, making hybrid Siamese architectures effective.

To learn contextual relationships in between sentences, a neural network model that combined Recurrent Neural Networks (RNN) with distributed vector representations was presented [26]. Learning document level embeddings and

sequential dependencies, the method enhanced the semantic understanding of text categorization problems, which implies that contextual representation learning is successful in solving similarity and classification problems. One of the studies presented C-LSTM (Convolutional-LSTM) network [27] that combines CNN and LSTM to enhance sentence representation and text classification. CNN only captures local n-gram features whereas LSTM can capture the long-term dependencies in the sequences and thus something more semantically modeled is richer. The results of experiments indicate that this hybrid architecture is better than standalone CNN and LSTM models in classification problems. Attention-based CNN models have also been used to replicate question detection based on which convolutional layers capture local textual representations and the attention mechanism identifies semantically significant words or phrases [10].

Another study [28] implemented a multi-perspective convolutional neural network sentence similarity where sentences were encoded by CNN layers in a multi-convolution and pooling approach, and similarity between local regions of sentence representations was then organized. This architecture based on the Siamese style enhances the learning of semantic similarity and competes with other architectures developed on paraphrase and textual similarity.

Overall, the literature reviewed shows that there is a definite shift between the traditional methods of lexical similarity to detect duplicates to more sophisticated deep learning frameworks. The initial strategies based on handcrafted features and classical machine learning offered baseline performances but failed to learn deep semantic relationships. Later studies came up with neural network models including CNN, RNN, LSTM and Siamese architecture that have greatly enhanced semantic similarity learning. Recent works used contextual embeddings, such as those of BERT and ELMO, which improved the quality of representations, but tended to rely on single streaming models. Even with these developments, there is still a limitation in the successful integration of contextual, lexical, and explicit similarity information in the same framework. This is why the current study can fill this gap by introducing a hybrid Siamese architecture combining contextual embeddings, static embeddings, and handcrafted features to generate stronger and scalable duplicate question detection.

III. Materials and Methods

This section presents a detailed description of proposed architecture as shown in Fig 2. The figure shows a novel end-to-end hybrid deep learning model, which is a combination of various techniques to detect duplicate questions in the Quora Question Pairs (QQP) dataset. The proposed framework consists of four major sections including data pre-processing, features extraction, two Siamese Neural Networks (SNN) connected in parallel and a feature-fusion based classification module. Building on the base, the core architecture consists of two-parallel streams of Siamese Neural Networks (SNN), referred as Model A and Model B, each responsible to extract various forms of semantic information of the text.

Model A made use of an embedding based model which used BERT-base transformer to create deep contextual embeddings to enable the system to capture paraphrases, synonyms and fine semantic details. On the other hand, Model B was developed based on pretrained GloVe embeddings, 1D-CNN and Bidirectional LSTM layers to learn static and sequential semantic relationships and lexical patterns respectively. Question 1 and Question 2 were independently processed using shared-weight branches in both models and maintain the inherent symmetry of Siamese architecture and provide similar results in terms of feature acquisition across inputs.

The results of Model A and Model B were subsequently concatenated with the handcrafted features in a classification module relying on fusion which integrates deep contextual representations, global static embeddings, local n-gram events, extended sequence dependencies, and explicit signs of similarity in lexicon. This fusion process allowed the model to utilize the strengths of every component in a complementary manner, which allowed assessing the similarity of questions more resiliently and thoroughly. The last classification was done through a deep neural network that included attention and dropout regularization and finally determined whether two questions were related to the same underlying intent. This structured methodological framework provided a complex and detailed representation of textual semantics which led to better performance in duplicate question detection than traditional machine-learning benchmarks.

A. Data Acquisition:

We utilized Quora Question Pairs (QQP) [29] dataset for analysis, training and testing purposes. QQP composed of more than 400,000 question pairs, where each pair of questions was annotated by human experts as duplicates or non-duplicates. The recorded instances are indicative of the inherent diversity of user-generated content collected from the internet such as paraphrasing, grammar mistakes and textual size. QQP, unlike the controlled laboratory data, represents a real-world language diversity and therefore is a powerful indicator of semantic similarity and paraphrase detection studies.

The data presented in the dataset is in the form of free-form texts and considerably different in structure, as some can be short and include a few keywords, whereas others can be long and multi-sentence in the form of a query. All samples are stored in tabular format with four classes such as question1, question2, isDuplicate, and unique identifiers. That dataset does not involve any specialized signal processing but needs a lot of preprocessing of the raw textual data to normalize and standardize it. To ensure the consistency of the experiment, the dataset is typically purged using text normalization measures like lowercasing, deleting punctuation, spaces, and tokenization. Table 1 provides a summary of statistics of QQP dataset.

B. Data Preprocessing:

The data preprocessing stage initiated with text normalization, where all the characters in both questions were made lower case to remove inconsistencies due to case but not se-mantically different. After normalization, punctuation, special characters and unnecessary characters were eliminated because

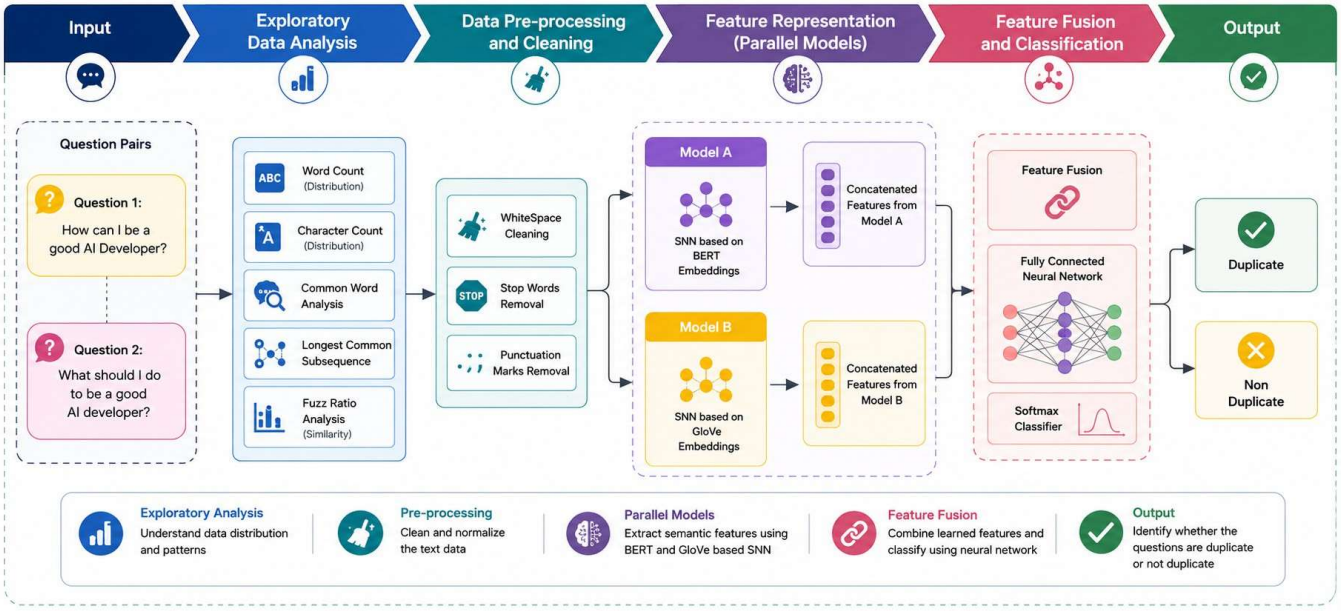


Fig. 2. Overall Proposed Architecture for QUORA Duplicate Question Detection

TABLE I
SUMMARY OF THE QQP DATASET

Aspect	Summary
Number of Classes	05
Names of Classes	qid1, qid2, question1, question2, isduplicate
Total Question Pairs	404,351 pairs
Duplicate Question Pairs (Label = 1)	149,263 pairs
Non-duplicate Question Pairs (Label = 0)	255,088 pairs (~63.1%)
Unique Questions	Approximately 500,000+
Average Question Length	12–15 words
Maximum Question Length	200+ words
Minimum Question Length	1~2 words

they usually create noise unlike adding semantic representation of the text. Also, any additional whitespace caused by punctuation deletion or uneven formatting by users was changed to single space to be structurally consistent.

Since user-generated content often contains symbols, e.g. and use of symbolic representations, e.g. a dollar, euro and rupee, these were translated into their textual counterparts (dollar, euro, and rupee), enhancing semantic readability and preventing misrepresentations during embedding lookup.

Missing or null entries were also handled in the pre-processing pipeline. Questions that were empty or had an invalid structure were filled with empty strings in a systematic manner so that the pipelines would not break and the tensors could keep their dimensionality when being trained. Once the cleaning and normalization process was done, a separate clean version of every question (clean_q1 and clean_q2) was saved. These purified versions were made consistent across handcrafted feature computation, GloVe-based embedding extraction, and BERT tokenization, so that all latter steps of the proposed hybrid model have the same methodology.

In general, the noise reduction, textual consistency, and linguistic regularities in this preprocessing strategy worked well in training the deep learning modules to learn more expressive semantics and enhance the accuracy of the hybrid Siamese architecture and its performance.

C. Model A: SNN based on BERT Embeddings

Model A referred as the primary contextual semantic encoder in the proposed hybrid architecture with the aim of capturing rich and context aware representations of the deep linguistic connections between the question pairs. As illustrated in Fig 3, both question 1 and question 2 were independently processed on two parallel but equally parametrized branches, which is in line with the symmetrical configurations of Siamese Neural Network. This weight-sharing approach ensured that the two questions were mapped by the same semantic space so that similarity can be meaningfully compared.

The process started with tokenization by using the Word-Piece BERT tokenizer, which divides each question into sub-word units and incorporated special tokens [CLS] at the start and [SEP] at the end. The representation enabled BERT to support identifying sentence boundaries and provided the support to extract the text in a holistic interpretation. Three complementary types of embedding of the tokens were then created: token embeddings containing the single constituent sub-words, positional embeddings encoding the linear position

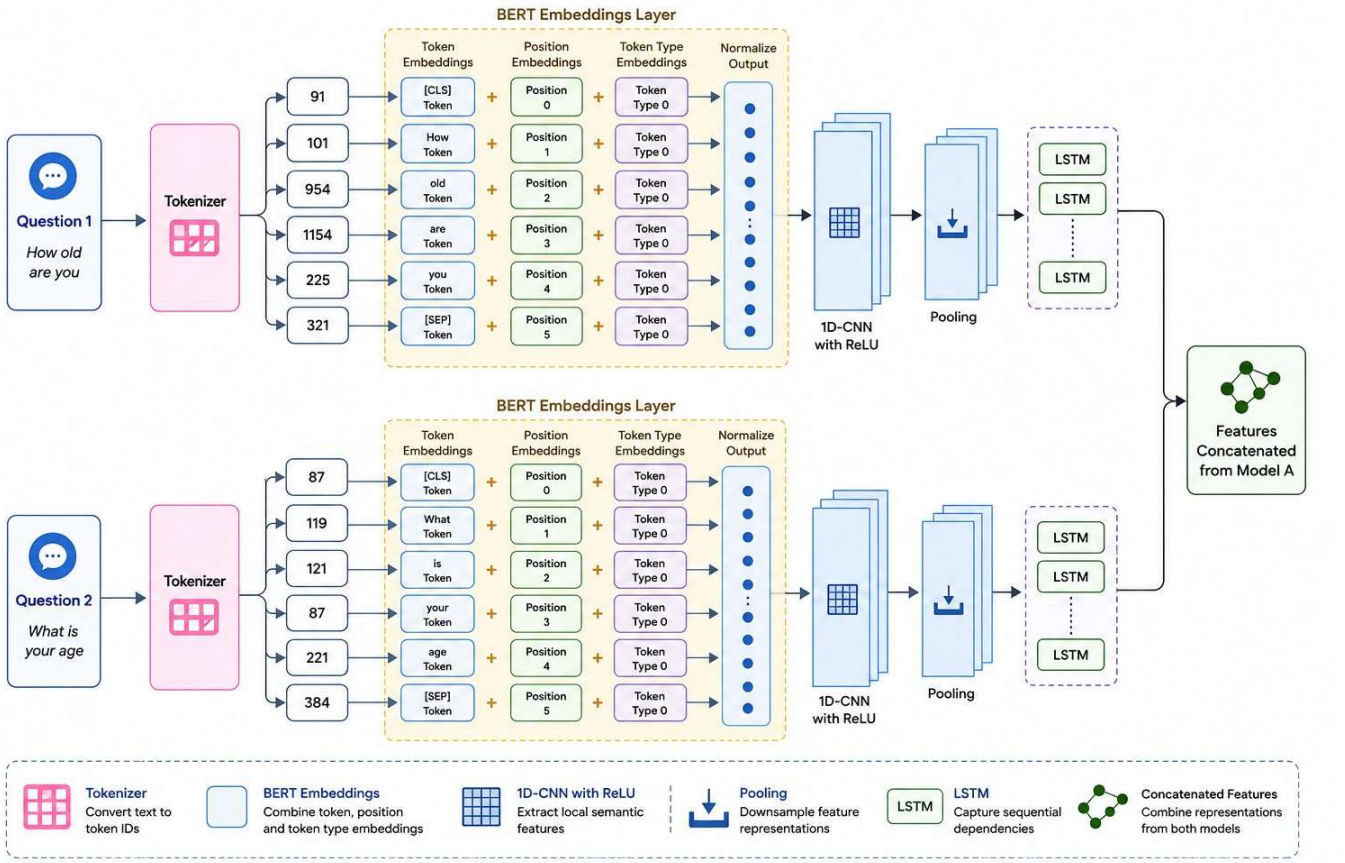


Fig. 3. Architecture of Model A -Contextual Siamese Neural Network with BERT embeddings, which encodes pairs of questions with shared weights before applying CNN and BiLSTM layers to extract deep semantic features before feature fusion.

of the tokens to maintain the syntax structure, and segment embeddings distinguishing between the sentences where necessary, but in our research each question was handled separately.

$$Q = [w_1, w_2, w_3, \ldots, w_n] \tag{1}$$

$$X = E_{token} + E_{position} + E_{seg} \tag{2}$$

All these three embeddings were added and applied to the BERT-base transformer encoder, which has 12 transformer blocks, 12 self-attention heads, and hidden state dimensionality of 768.

$$H = BERT\,(H) = [h_1, h_2, h_3, \ldots, h_T] \tag{3}$$

The encoder used successive rounds of multi-head attention and feed forward transformations, producing deep contextual representations which semantically reflected the characteristics of each question, paraphrasing, syntactic and long-range relationship. The contextual embedding that was attributed to the [CLS] token happened to be especially informative, since it was the sum of semantic meaning of the question. To get more representational depth out of BERT and use the embeddings to fit the duplicate-question detection task, the task-dependent neural layers drew upon the contextual outputs with additional processing. These layers included:

1) **1D Convolutional Neural Network (1D-CNN):** The output embeddings obtained from BERT were then fed to 1D-CNN with ReLU as activation function. Main purpose of this layer is to detect the salient local contextual patterns, such as common co-occurring n-grams, phrase patterns and semantics micro-patterns where there may be duplicate intent. The convolutional filter took on locally localized semantic signatures that may not be stressed out by BERT itself because of its global attention mechanism.

$$C = ReLU\,(W_c * H + b_c) \tag{4}$$

2) **Max Pooling:** After the CNN layer, max-pooling operation was performed to minimize the dimensionality of the feature sequence but preserving most prominent activations.

$$p = max\,(C) \tag{5}$$

3) **Bidirectional Long Short-Term Memory (BiLSTM):** Following the max pooling layers, the features were then fed into BiLSTM layer, processing the sequences both in forward and backward directions. This helps to learn to respond to the question with regards to how past and future answers to a question impact the meaning of it, a vital attribute when encountering paraphrased questions that differ in form but have similar intent.

$$u = \overrightarrow{LSTM}(p) + \overleftarrow{LSTM}(p) \quad (6)$$

The output of question 1 and question 2 were then concatenated to form the final contextual embedding vector for model A, integrating transformer-based knowledge from BERT, local feature representation from CNN, and sequential representation from BiLSTM. This resulting embedding served as a pivotal element in the hybrid fusion mechanism, enriching the duplicate question classification process with robust contextual semantics.

$$h_A(1) = f_A(Q_1) \quad (7)$$

$$h_A(2) = f_A(Q_2) \quad (8)$$

$$z_A = |h_A(1)| \oplus |h_A(2)| \quad (9)$$

D. *Model B: SNN based on GloVe, CNN, and BiLSTM*

Model B as illustrated in Fig 4 is referred to as the static embedding component of the proposed hybrid architecture with the main aim to obtain the lexical semantics, local phrases, and sequential patterns with the help of pretrained GloVe embeddings along with convolutional and recurrent neural layers.

$$Q = [w_1, w_2, w_3, \ldots, w_n] \quad (10)$$

$$x_t = E_{Glove}(w_t) \quad (11)$$

$$X = [x_1, x_2, x_3, \ldots, x_n] \quad (12)$$

The process started by feeding the question pairs to rule based tokenizer for tokenization, dividing each question into small units represented by individual words. As compared to tokenization process carried out in model A using BERT tokenizer, this tokenizer preserves complete words rather than giving sub-words. This allowed it to be used directly in compatibility with 300-dimensional pretrained GloVe embeddings, which were created using large token co-occurrence statistics. These crossword global semantic representations enabled Model B to encode lexical similarity even in a context of different contextual shades across questions.

After the tokenization, every word was assigned to its corresponding GloVe vectors. In the case of out-of-vocabulary words, the zero or randomly weighted embedding was used to maintain dimensional similarity. The output of the resulting sequence of embeddings was passed forward to the next components of the model B including 1D-Convolutional Neural Network, max pooling and Bi-LSTM. 1D-CNN with ReLU activation function was used to detect the local lexical and se-mantic patterns such as bi-grams and trigrams. These patterns
frequency serves as robust indicators of question similarity. The question was scanned by the convolutional filters using sliding windows to detect recurring phrase structures and fixed expressions typical of a paraphrased or reworded question.

$$C = ReLU(W_g * X + b_g) \quad (13)$$

After the convolution process, the temporal dimensions of feature vector were reduced by applying the pooling operation while retaining the most significant features. These pooled features were then passed by the bidirectional LSTM to identify long range dependencies and add sequential context to the representation in both forward and backward directions. It makes the model to generalize the connections between distant words as well as the ability to understand grammatical designations like subject predicate frameworks, question formats (e.g. What is...?,), and other consecutive prompts creating semantically comparable associations.

$$p = max(C) \quad (14)$$

The same parameterized BiLSTM layers were used for both questions, producing context-aware embeddings for each question. It ensures that the two streams of embeddings were similar and comparable to each other with respect to similarity. The final embeddings produced for both question 1 and question 2 were then concatenated, produced the output feature vector of SNN with Glove Embeddings. Which is later concatenated with the output of handcrafted features.

$$u = \overrightarrow{LSTM}(p) + \overleftarrow{LSTM}(p) \quad (15)$$

$$h_B(1) = f_B(Q_1) \quad (16)$$

$$h_B(2) = f_B(Q_2) \quad (17)$$

$$z_B = |h_B(1)| \oplus |h_B(2)| \quad (18)$$

Another component of model B is handcrafted feature extractions, designed for capturing the explicit lexical and structural similarities between the question pairs which deep neural encoding may fail to encode completely. Character-level and word-level statistics were calculated, such as the level of character length and word counts of Question 1 and Question

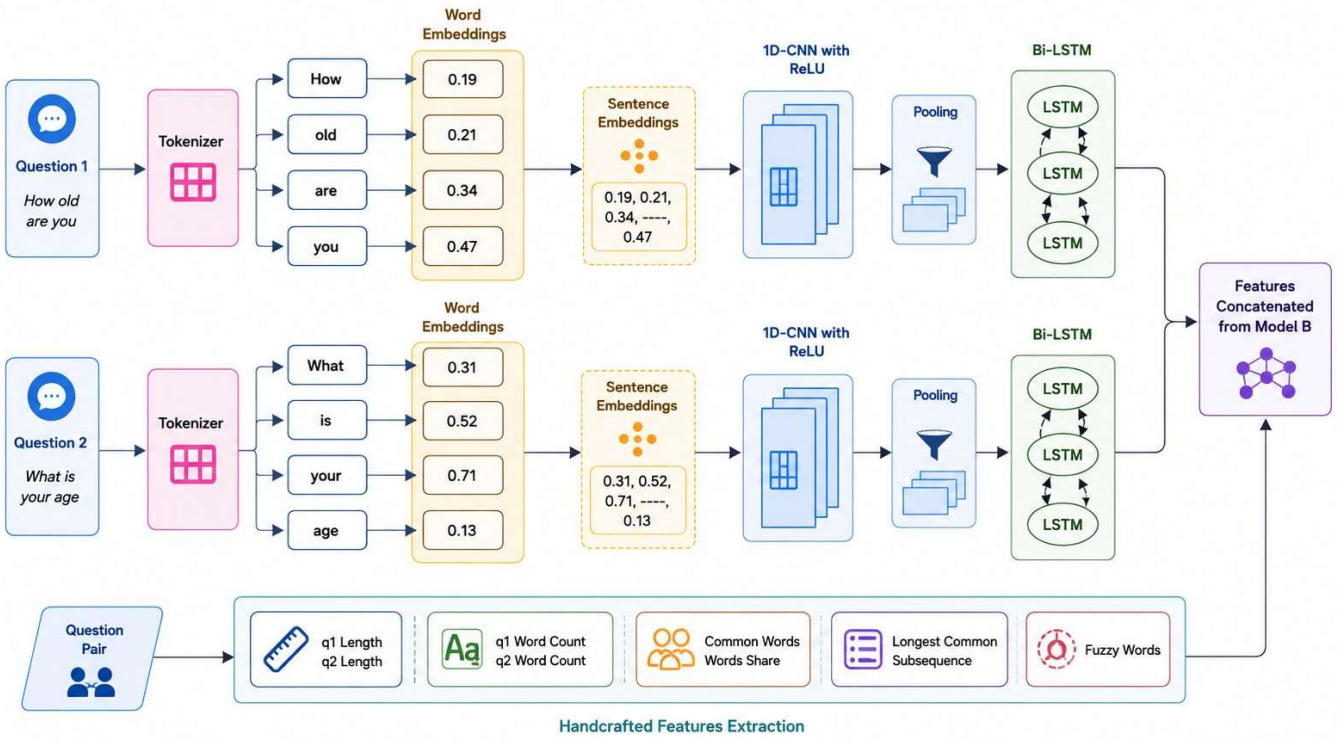


Fig. 4. Model B architecture, Static Siamese Neural Network using Glove embeddings with CNN and BiLSTM layers which shows how lexical feature extraction and sequential representation learning are used together with handcrafted similarity features before fusion.

2, of each question pair, which produced the indicators of the textual complexity and structural similarity. Lexical overlap was quantified by counting the number of shared unique words in questions and a ratio of word-share that is the proportion of shared words to use of all unique words in the 2 questions. The Longest Common Subsequence (LCS) score was also calculated to obtain sequential overlapping patterns that can perhaps be paraphrased. To improve further on similarity modeling, some fuzzy metrics of string-matching were introduced, like Fuzzy Q-ratio and partial-ratio, where the system can identify soft lexical matches, even in cases where the wording is not the same. These handcrafted features were then combined to one single numerical input, which was further concatenated with the output feature vector obtained through SNN with GloVe embeddings. Let the handcrafted feature vector be:

$$f \in R^K$$

Overall, the proposed hybrid deep learning architecture enhances the ability of the model to identify duplicate and non-duplicate pairs of questions. The feature vector obtained from model A focused on dynamics contextual semantics, whereas model B emphasizes static global semantics and phrase-level patterns based on large-scale co-occurrence statistics. The combination of these elements allowed representing question similarity in a more comprehensive way.

E. Feature Fusion and Classification:

The fusion layer is the essential core of the proposed hybrid architecture, which incorporated the heterogeneous representations produced through model A and model B into one unified semantic space, which can be used for final classification. Particularly, the contextual embeddings obtained by Model A, the one capturing deep bidirectional semantic dependencies by processing with BERT, CNN, and BiLSTM, were added to the fixed lexical-semantic embeddings generated by Model B, the ones obtained via the use of GloVe vectors, convolutional feature extraction, and sequential modeling. This combination created a holistic representation which combined high-level contextual knowledge and stable global semantic data.

$$z = |z_A| \oplus |z_B| \oplus |f| \tag{19}$$

Besides using these neural embeddings, the feature array of linguistic labels that encode the explicit lexical overlaps, the ratio word-share, length-based statistics, fuzzy string similarity and the longest common subsequence was tacked onto the fused representation. Having this element of handcrafted aspect enabled the model to take advantage of the superficial similarity aspects, which can be serious predictors of copying yet not adequately captured in the spaces of deep embedding alone.

After concatenation, the resultant feature vector was subjected to an attention mechanism that aims at highlighting the most discriminative semantic dimensions and notching obsolete or noisy elements. This attention-refined vector was then passed through a dense feed-forward network with dropout regularization so that it would learn optimally and reduce overfitting. Lastly, a softmax output layer was employed which resulted in binary classification accurately defining whether duplicate or non-duplicate.

$$\alpha = Softmax(W_a z + b_a) \tag{20}$$

$$z' = \alpha . z \tag{21}$$

$$s = \sigma (W_d z' + bd) \tag{22}$$

The multi-source fusion approach allowed the model to collaboratively harness contextual semantics, lexical structure, and explicit similarity measures, leading to a more holistic and accurate duplicate question detection model.

$$o = Softmax(W_c s + b_c) \tag{23}$$

$$o = [P(duplicate), P(nonduplicate)] \tag{24}$$

IV. Experimental Setup

A. *Training Parameters and Hyperparameter Configuration:*

To facilitate effective model learning and achieve a representative class distribution, the Quora Question Pairs dataset was split into an 80:20 training and validation sample by stratified and sampling, keeping the balance of duplicate and non-duplicate pairs in the training and validation sample. The suggested hybrid Siamese Neural Network was trained with the help of an optimized component of hyperparameters, which were chosen via empirical tuning and optimal practice of transformer-based actions and sequence-modeling designs. They were trained using an NVIDIA GeForce RTX 4070 Ti, which has enough processing ability to fine-tune the BERT effectively, run BERT on batches of queries, and train the dual-branch hybrid architecture both end-to-end.

Table II presents the hyperparameters of optimizer configuration, learning schedule, dimensions of neural networks, and training configuration. These values have been ascertained through preliminary experimentation, previous literature advice and stability reasons of the model. This design took into consideration balance between convergence speed, overall generalization performance and computer efficiency.

TABLE II
Training Parameters and Hyperparameters

Parameter	Value
Hardware	RTX 4070 Ti GPU
Batch Size	32
Epochs	5–10
Optimizer	AdamW
Learning Rate	1×10^{-5} to 2×10^{-5}
Loss	Cross-entropy
Weight Decay	0.01
Max Seq Len	128 tokens
GloVe Dim	300
Pooling	Max-pooling
BiLSTM Units	256 (bi-dir)
Dropout	0.3–0.5
Classifier	Dense + Softmax
Train/Val Split	80:20

B. *Evaluation Metrics:*

An extensive predefined measure of evaluation metrics was used to evaluate strictly the predictive performance and the generalization ability of the proposed hybrid architecture. Accuracy, precision, recall, and the F1-score [30]–[34] constituted the main measures that can provide one with a different perspective of the model behavior.

Assuming TP, TN, FP, and FN to represent true positives, true negatives, false positives and false negatives respectively. These values were obtained using the confusion matrix. Accuracy defines the fraction of the correctly classified pairs of questions as a whole and is given as. Precision determines the accuracy of duplicated predictions and the ratio of correct predictions that have been duplicated to all predictions that have been made. The recall is a measure that tests the capability of the model to identify all the real duplicate question pairs. Since the QQP data has an unequal class equilibrium, the F1-score was employed as its main measure. This is defined as the harmonic average of the precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \tag{26}$$

$$Recall = \frac{TP}{TP + FN)} \tag{27}$$

$$Recall = 2x\frac{Precision x Recall}{Precision + Recall)} \tag{28}$$

V. Results and Discussion

As presented in table III, the proposed model indicates its ability to discriminate the duplicate and non-duplicate questions with an impressive accuracy of 85.03%. For the non-duplicate category, the model demonstrates impressively high precision of 91.23% and F1-score of 87.67%. It indicates that the model can detect the non-duplicate questions accurately regardless of lack of semantic equivalence.

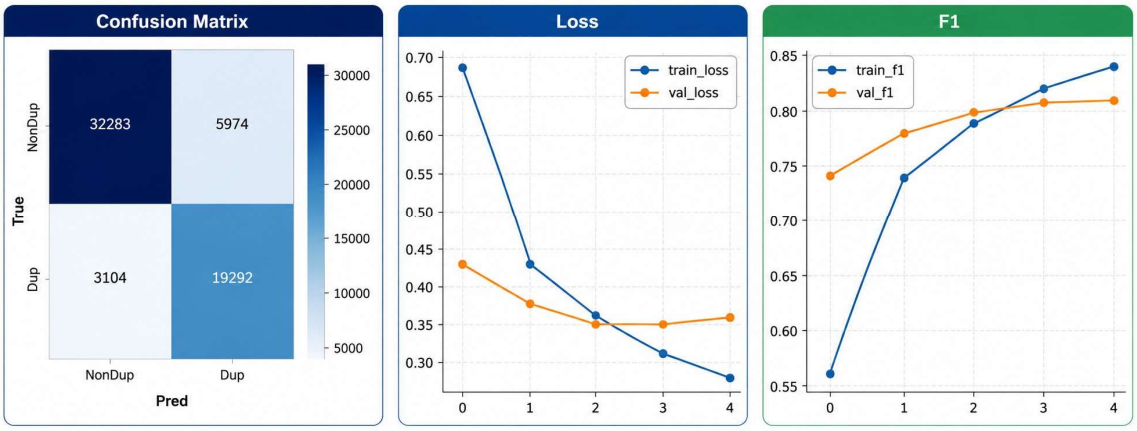


Fig. 5. Confusion matrix and training dynamics of the proposed Hybrid Siamese Neural Network, illustrating classification performance along with training–validation loss convergence and F1-score progression across epochs on the QQP dataset.

On the other hand, the model's competitive F1-score (0.8095) and strong recall (0.8614) for the duplicate category demonstrate its ability to identify paraphrased or semantically similar queries even with little lexical overlap.

The relatively low precision for the duplicate class indicates some non-duplicate pairs were falsely classified as duplicates, which is consistent with identified false-positive instances in semantic similarity tasks. However, the model's high recall on duplicates emphasizes how useful it is in practical situations where missing true duplicates can result in redundant content, a worse user experience, and ineffective information retrieval. Together, these findings show that the model can achieve strong and thorough semantic understanding by combining GloVe-based static embeddings, BERT-based contextual representations, and manually created similarity features.

The Quora Question Pairs dataset of more than 400,000 pairs of questions labelled with answers were tested experimentally. Training of the model was done with AdamW optimizer and was tested based on accuracy, precision, recall, F1-score and confusion matrices. The proposed hybrid architecture realized a 85.03 accuracy and 0.81 F1-score, which is better than the traditional machine learning baselines and single stream deep learning models. The ablation study also established that all the elements of the hybrid architecture contextual embeddings/static embeddings/handcrafted features have a positive influence on the overall result, which might prove the role of multi-representation learning importance.

To conclude, the suggested hybrid Siamese Neural Network structure would give a powerful and scalable solution to the detection of duplicate questions in full-scale community-based question -answering systems. The model brings together different representation paradigms in one architecture and generates better semantic similarity learning, classification, and has greater generalization. Results of this study will be used in the development of intelligent text similarity modeling as well as showing how hybrid deep learning methods can be applied in real-life NLP.

As shown in confusion matrix, the model achieves a high number of true negatives (32,283) and true positives (19,292),

TABLE III
CLASSIFICATION PERFORMANCE OF THE PROPOSED HYBRID SNN MODEL

Class	Precision	Recall	F1-score	Support
Non-Duplicate (0)	0.9123	0.8438	0.8767	38,257
Duplicate (1)	0.7636	0.8614	0.8095	22,396
Accuracy	—	—	**0.8503**	60,653

confirming strong discriminative ability. Misclassifications are approximately 5,974 false positives and 3,104 false negatives. False positives primarily arise in cases of high lexical overlap but differing intent, while false negatives occur in abstract, paraphrased, or indirectly phrased duplicates.

A. Ablation Study:

The model variants that were used were as follows:

1) **Model A (Siamese BERT)**: The implemented architecture kept the contextual embedding stream utilizing BERT and implemented it in a Siamese architecture. It tested the performance of contextual semantics as such, and not with auxiliary features.
2) **Model B (Siamese GloVe + CNN + BiLSTM)**: This variant employed the use of fixed GloVe embeddings along with CNN and BiLSTM layers within a Siamese network without the use of contextual embeddings or hand-defined features.
3) **Model A + Model B (Without Handcrafted Features)**: The feature concatenation of both Siamese streams was carried out, although handcrafted linguistic features were eliminated to determine the effect of deep representations only.
4) **Full Hybrid Model (Proposed)**: This design combined Model A, Model B and handcrafted design using a layer of fusion, and then the final classifier.

VI. CONCLUSION

The current study dealt with the issue of duplicate questions detection on large-scale community question answering

TABLE IV
ABLATION STUDY ANALYZING THE CONTRIBUTION OF INDIVIDUAL COMPONENTS OF THE PROPOSED HYBRID FRAMEWORK, INCLUDING BERT EMBEDDINGS, GLOVE EMBEDDINGS, AND HANDCRAFTED LINGUISTIC FEATURES

Model Configuration	Accuracy (%)	Precision	Recall	F1-score
Siamese BERT (Model A)	83.9	0.79	0.84	0.81
Siamese GloVe + CNN + BiLSTM (Model B)	80.6	0.75	0.82	0.78
Model A + Model B	84.4	0.80	0.85	0.82
Full Hybrid Model (Proposed)	**85.03**	**0.81**	**0.86**	**0.81–0.82**

(CQA) forums, specifically on QUORA, where contextually related questions are usually presented in different languages. Redundant behavior, disjointed responses, ineffective search, and elevated moderation work are caused by the presence of duplicate questions. Conventional methods of semantic relations between question pairs using lexical similarity measures, manually developed features, and classical machine learning methods have had limited ability to model deep semantic relationships. Despite the recent advancements in deep learning contextual embedding-based approaches to semantic understanding, the use of only one representation paradigm is not adequate to achieve high-quality duplicate detection in the real world.

To address these restrictions, the current research proposed a hybrid deep learning system grounded on parallel Siamese Neural Networks that includes contextual embeddings provided by BERT, fixed embeddings by GloVe, and handcrafted similarity features provided by linguistics. The suggested framework allows complementary learning of representations through a synthesis of semantic context, lexical structure and explicit similarity cues as a single feature fusion process. The Siamese design is very advantageous to provide a uniform encoding of the pairs of questions and enhance similarity learning in shared space.

The Quora Question Pairs dataset of more than 400,000 pairs of questions labelled with answers were tested experimentally. Training of the model was done with AdamW optimizer and was tested based on accuracy, precision, recall, F1-score and confusion matrices. The proposed hybrid architecture realized a 85.03% accuracy and 0.81 F1-score, which is better than the traditional machine learning baselines and single stream deep learning models. The ablation study also established that all the elements of the hybrid architecture contextual embeddings/static embeddings/handcrafted features have a positive influence on the overall result, which might prove the role of multi-representation learning importance.

REFERENCES

[1] "Top 35 Social Media Platforms (2025) — explodingtopics.com." https://explodingtopics.com/blog/top-social-media-platforms. [Accessed 01-10-2025].

[2] "Quora Statistics 2025: Users, Traffic Source And Growth Data — resourcera.com." https://resourcera.com/data/social/quora-statistics/#:~:text=Quora%20is%20one%20of%20the,previous%20year%2C%20300%20million%20MAU. [Accessed 01-10-2025].

[3] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical software engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[4] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: an analysis of quora," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 1341–1352, 2013.

[5] N. Ansari and R. Sharma, "Identifying semantically duplicate questions using data science approach: A quora case study," *arXiv preprint arXiv:2004.11694*, 2020.

[6] Z. Mundargi, S. Kalal, A. Kashyap, O. Kamble, S. Pande, and S. Mhamane, "Quora question duplication detection: An ml approach for identifying semantically equivalent questions," in *International Conference summit on Artificial Intelligence*, pp. 161–172, Springer, 2024.

[7] S. Wang, D. M. German, T.-H. Chen, Y. Tian, and A. E. Hassan, "Is reputation on stack overflow always a good indicator for users' expertise? no!," in *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 614–618, IEEE, 2021.

[8] M. Ahasanuzzaman, M. Asaduzzaman, C. K. Roy, and K. A. Schneider, "Mining duplicate questions in stack overflow," in *Proceedings of the 13th International Conference on Mining Software Repositories*, pp. 402–412, 2016.

[9] S. I. K. M. S. Pandi and H. Shomee, "Simplifying question-answering on quora with machine learning,"

[10] A. Altamimi, M. Umer, D. Hanif, S. Alsubai, T.-H. Kim, and I. Ashraf, "Employing siamese malstm model and elmo word embedding for quora duplicate questions detection," *IEEE Access*, vol. 12, pp. 29072–29082, 2024.

[11] T. Zubair, M. J. Asif, F. Kamiran, H. H. Saeed, and R. F. Ahmad, "Efficient adaptation of llms for hate speech detection in low-resource languages: A comparative study on roman urdu," in *23rd International Bhurban Conference on Applied Sciences & Technology (IBCAST 2026)*, 2026.

[12] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," *International journal of computer applications*, vol. 181, no. 1, pp. 25–29, 2018.

[13] M. Raj, J. P. Singh, A. K. Singh, A. K. Dey, P. K. Singh, and S. Kesarwani, "Detection of duplicate question pairs by applying proposed bow, tf-idf, and use approach," in *International Conference on Soft Computing: Theories and Applications*, pp. 21–33, Springer, 2023.

[14] M. Ektefa, M. A. Jabar, F. Sidi, S. Memar, H. Ibrahim, and A. Ramli, "A threshold-based similarity measure for duplicate detection," in *2011 IEEE Conference on Open Systems*, pp. 37–41, IEEE, 2011.

[15] G. Majumder, P. Pakray, A. Gelbukh, and D. Pinto, "Semantic textual similarity methods, tools, and applications: A survey," *Computacio´n y Sistemas*, vol. 20, no. 4, pp. 647–665, 2016.

[16] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[17] J. Bjerva and R. O¨ stling, "Cross-lingual learning of semantic textual similarity with multilingual word representations," in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 211–215, 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[19] P. Sravanthi and B. Srinivasu, "Semantic similarity between sentences," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 1, pp. 156–161, 2017.

[20] B. N. Patro, V. K. Kurmi, S. Kumar, and V. Namboodiri, "Learning semantic sentence embeddings using sequential pair-wise discriminator," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2715–2729, 2018.

[21] A. Hashmi, "An nlp approach to efficient duplicate question detection using neural networks and tf-idf," in *2025 4th International Conference on Computing and Information Technology (ICCIT)*, pp. 395–400, IEEE, 2025.

[22] A. Dhakal, A. Poudel, S. Pandey, S. Gaire, and H. P. Baral, "Exploring deep learning in semantic question matching," in *2018 IEEE 3rd international conference on computing, communication and security (ICCCS)*, pp. 86–91, IEEE, 2018.

[23] R. Rishickesh, R. Ram Kumar, A. Shahina, and A. Nayeemullah Khan, "Identification of duplication in questions posed on knowledge sharing platform quora using machine learning techniques," *Int J Innovative Technol Exploring Eng (IJITEE)*, vol. 8, no. 12, pp. 2444–2451, 2019.

[24] M. Chandra, A. Rodrigues, and J. George, "An enhanced deep learning model for duplicate question detection on quora question pairs using siamese lstm," in *2022 IEEE International conference on distributed computing and electrical circuits and electronics (ICDCECE)*, pp. 1–5, IEEE, 2022.

[25] D. Viji and S. Revathy, "A hybrid approach of weighted fine-tuned bert extraction with deep siamese bi–lstm model for semantic text similarity identification," *Multimedia tools and applications*, vol. 81, no. 5, pp. 6131–6157, 2022.

[26] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using lstm for region embeddings," in *International Conference on Machine Learning*, pp. 526–534, PMLR, 2016.

[27] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," *arXiv preprint arXiv:1511.08630*, 2015.

[28] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1576–1586, 2015.

[29] "Question Pairs Dataset — kaggle.com." https://www.kaggle.com/datasets/quora/question-pairs-dataset. [Accessed 25-11-2025].

[30] M. J. Asif, "Crowd scene analysis using deep learning techniques," *arXiv preprint arXiv:2505.08834*, 2025.

[31] M. J. Asif, H. Khan, R. Tehseen, R. F. Ahmad, M. Asad, S. T. Hussain Rizvi, and S. Saqib, "Exploring convolutional neural networks for rice grain classification: An explainable ai approach," in *2025 International Conference on Frontiers of Information Technology (FIT)*, pp. 1–6, 2025.

[32] M. J. Asif, S. Saqib, R. F. Ahmad, M. Asad, and S. T. H. Rizvi, "Conv-lstm for real-time spatio-temporal analysis of crowd behavior in public spaces,"

[33] M. J. Asif, M. S. Rafaqat, U. Nazakat, U. Khan, and R. F. Ahmad, "Towards automated solar panel integrity: Hybrid deep feature extraction for advanced surface defect identification," 2026.

[34] R. Khan, M. J. Asif, and R. F. Ahmad, "Integrating triaxial imu sensors and ensemble learning for effective parkinson disease severity classification," *Preprint Submitted to arXiv*, 2026.