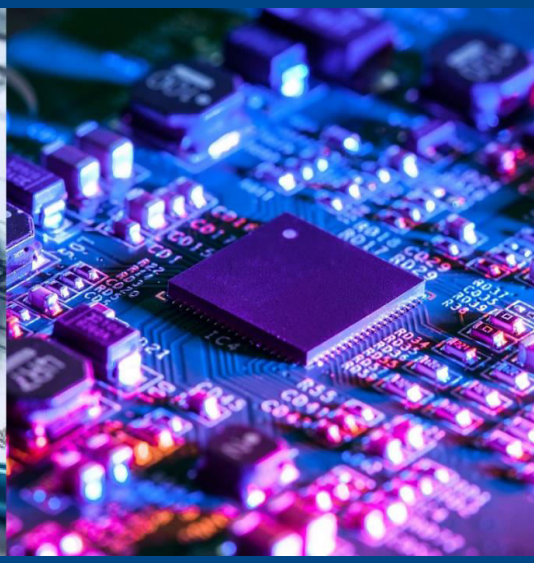
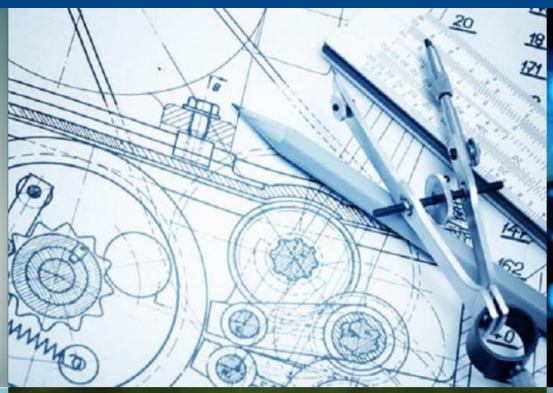
ISSN: 2582-7219

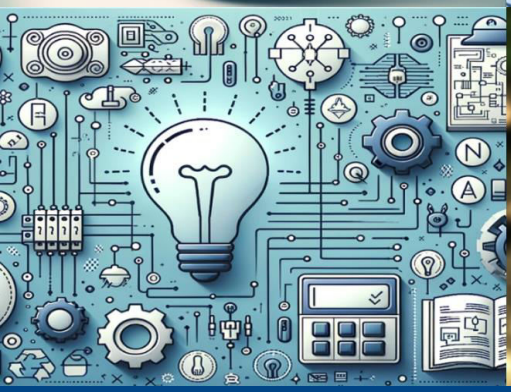




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 6, June 2025**

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806152

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Recognizing Duplicate Questions in QA Platforms with Machine Learning Model

Anil Malav, Mr. Deepak Mahawar

Department of Computer Science, Career Point University, Kota, Rajasthan, India

Assistant Professor, Department of Computer Science, Career Point University, Kota, Rajasthan, India

ABSTRACT: Quora is a widely used question-and-answer platform, it frequently encounters the problem of users posting similar or duplicate questions. This research study investigates how machine learning models can be utilized to effectively handle this issue. Traditional methods such as natural language processing often require significant resources and may fall short in accurately identifying duplicate queries.

Detecting questions that convey the same meaning is vital for providing users with relevant answers tailored to their intent, thereby improving their overall experience. However, the variety in how questions are phrased makes this a challenging task.

In this work, we explore the use of advanced machine learning and deep learning techniques to detect duplicate questions within Quora's question pair dataset. We implement and train multiple models—including Neural Networks, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Bidirectional Long Short-Term Memory networks (BiLSTMs) to distinguish between duplicate and unique questions, aiming to improve the accuracy of duplicate detection.

KEYWORDS: NN, CNN, RNN, BiLSTM, Keras, Tensor Flow.

I. INTRODUCTION

Social media platforms have become an essential part of everyday life, connecting millions of users worldwide with diverse interests and needs. Each platform serves a unique purpose: Facebook focuses on social interaction, LinkedIn caters to professional networking, WhatsApp offers instant messaging and video calls, Stack Overflow supports technical Q&A, and Instagram is centered around photo sharing. Among these, Quora stands out as a question-and-answer platform where users share knowledge, experiences, and opinions across a wide range of topics.

Unlike earlier Q&A sites such as Yahoo Answers and Google Responses, which struggled to maintain content quality due to the influx of irrelevant or low-quality posts, Quora has successfully built a thriving community since its launch in 2009. By 2025, Quora has grown to over 400 million monthly active users, making it one of the largest Q&A platforms globally. The platform supports more than 300,000 topics across 24 languages and attracts a diverse audience, with the United States accounting for about 148 million users and India contributing around 100 million more. This vast and varied user base includes thousands of experts who help ensure the quality and reliability of the information shared. On Quora, anyone can post a question, and knowledgeable users contribute detailed answers, fostering a collaborative environment for learning and discussion. Beyond simply providing answers, Quora encourages users to engage with content by suggesting edits and improvements to existing answers, enhancing the overall quality of information. This collaborative approach, combined with the platform's vast user base, makes Quora a popular destination for people seeking reliable solutions and insights.

The platform's success can be attributed to its ability to connect seekers of knowledge with experts and enthusiasts who provide thoughtful, well-researched responses. This dynamic interaction not only helps users find answers but also promotes continuous learning and intellectual growth. Additionally, Quora's algorithm prioritizes high-quality content, ensuring that the most relevant and accurate answers are presented to users, which further strengthens the platform's credibility. However, with millions of users from around the globe, Quora faces a common challenge: many questions are repeated or phrased differently but essentially ask the same thing. This duplication creates confusion for users

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

trying to find the best answers and poses difficulties for the platform in managing content efficiently. Duplicate questions take up valuable storage space, clutter search results, and reduce the overall user experience. Moreover, the presence of multiple versions of the same question can dilute the quality of responses, as contributors may unknowingly provide similar answers across different threads.

Detecting these duplicate or semantically similar questions is crucial for maintaining Quora's content quality and user satisfaction. This task is complicated by the fact that people often express the same idea in various ways using different words or sentence structures. Natural language is inherently diverse and flexible, which makes it challenging to automatically identify questions that share the same intent but differ in wording or style.

As Quora's knowledge base continues to grow, it becomes increasingly important to filter out redundant or low-quality content to keep users engaged and ensure the platform remains a trusted source of information. To address this, Quora released a dataset on Kaggle aimed at developing models that can identify question pairs with the same intent. By leveraging advanced data processing and machine learning techniques to detect and manage duplicate questions, Quora can streamline content organization, save users' time, and reduce the burden on contributors who otherwise might answer multiple versions of the same question. Ultimately, effective duplicate detection enhances the overall user experience by providing quicker access to the best answers and maintaining a clean, well-organized knowledge repository.

II. RELATED WORK

The problem of detecting duplicate questions is closely related to the well-studied natural language processing (NLP) task known as paraphrase identification, which involves determining whether two sentences convey the same meaning. This task is often approached through natural language sentence matching (NLSM). With the resurgence of neural networks, several neural-based models have been introduced to tackle paraphrase detection effectively. One of the earliest and popular models is the Siamese neural network, which consists of two identical subnetworks that share weights and independently extract features from each input sentence. The outputs of these subnetworks are then compared, often using cosine similarity, to decide if the sentences are paraphrases. While this architecture is efficient and straightforward to train due to shared parameters, it processes the two sentences separately, which can limit the model's ability to capture interactions between them, potentially leading to loss of important contextual information.

To overcome these limitations, a two-dimensional framework called "attend-and-aggregate" was developed. This model captures interactions between sentences by performing phrase-level matching and then aggregating these results into a vector for final classification. To further improve on these methods, Wang et al. proposed the Bilateral Multi-Perspective Matching (BiMPM) model. BiMPM enhances sentence matching by encoding input sentences with Bidirectional Long Short-Term Memory networks (BiLSTMs) and performing matching in both directions from multiple perspectives. The matching results are then aggregated using another BiLSTM layer to produce a fixed-length vector for classification. This bidirectional and multi-perspective approach allows the model to capture richer semantic relationships between sentence pairs.

In addition to the Siamese and attend-and-aggregate frameworks, attention-based models have proven effective for paraphrase identification. These models explicitly model dependencies between words or phrases in sentence pairs. A notable example is the Attention-Based Convolutional Neural Network (ABCNN), which applies multiple convolutional layers to capture information at different levels of granularity—word, phrase, and sentence—allowing the model to analyse sentences at various levels of abstraction. More recently, transformer-based architectures such as BERT and its variants have revolutionized the field by leveraging self-attention mechanisms to capture deep contextual relationships within sentence pairs. Fine-tuning these pre-trained models on datasets like Quora's question pairs has led to state-of-the-art results in duplicate question detection. Using the Quora question pair dataset, we extracted distinctive features and applied a range of machine learning techniques. Initially, we developed a baseline using traditional machine learning methods after performing feature engineering on the raw data. Our baseline models performed well and established a strong foundation compared to previous studies.

Overall, the literature indicates a clear progression from simpler models that treat sentences independently toward more complex architectures that explicitly model interactions and dependencies between sentences. This evolution highlights the importance of capturing semantic nuances and contextual information to accurately identify duplicates. Our work

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

builds on these insights by combining feature engineering with advanced neural architectures to improve detection accuracy and efficiency

III. DATASET AND PREPROCESSING

Considering the extensive data available, our approach combines neural network modeling with simple word-level and sentence-level similarity measures to improve duplicate question detection. Neural networks, particularly Recurrent Neural Networks (RNNs), excel at handling sequential data because they maintain an internal memory that captures information from previous inputs and computations over time. This ability allows them to process sequences by passing information through cycles within the network. However, RNNs face challenges when it comes to learning long-term dependencies, and their effectiveness often diminishes as the length of the input sequence increases. To overcome these limitations, Long Short-Term Memory (LSTM) networks are commonly used for sequence modeling tasks. LSTMs are particularly effective because they not only handle long-range dependencies better but also provide mechanisms to control how much information is retained or forgotten at each time step. For our task, we design a Siamese network architecture consisting of two identical LSTM subnetworks - one processing each question in the pair. In this setup, each question is passed through a separate but identical LSTM tower that shares the same weights and parameters. Our Siamese LSTM model is trained using pre-trained word embeddings from GloVe, which provide rich semantic vector representations for each word. The model is trained on a large dataset of 283,000 question pairs and validated on 121,000 pairs to assess its performance.

By combining these neural network capabilities with effective preprocessing and feature extraction, our approach aims to accurately detect duplicate questions, enhancing the overall quality and usability of the platform.

A. Dataset

Our research relies on the Quora Question Pairs dataset, made available as part of a Kaggle challenge, featuring a training collection of over 4 lacs question pairs alongside a much larger test dataset with around 2 million question pairs. Since the test dataset no longer contains labels indicating whether the questions are duplicates, the only available evaluation metric for it is accuracy, which can be obtained by submitting predictions to Kaggle's online platform. To enable a more detailed evaluation of our models, including metrics beyond simple accuracy and to perform thorough error analysis, we decided to create our own test dataset by partitioning the original training data. Consequently, our study focuses exclusively on the 4 lacs question pairs from the training set. We split this data into three subsets: training, validation, and testing. This careful division allows us to train our models effectively, fine-tune parameters, and evaluate results in a controlled and meaningful way.

The following fields are present in each sample point:

- id: Individual ID
- qid1: First question's ID
- qid2: Second question's ID
- Is there any overlap between the questions? (A score of 0 indicates no overlap and a score of 1 suggests there is some overlap.)

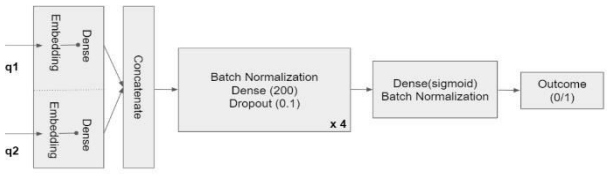


Figure1. Block diagram demonstrating activation normalization in a neural network layer

The dataset we are working with is imbalanced: out of 4 lacs question pairs, 60% pairs are labelled as non-duplicates (0), while 36% are labelled as duplicates (1). Although each pair is unique, individual questions are not necessarily distinct across the dataset. In fact, some questions appear multiple times in different pairs. We also observed that the dataset is not limited to ASCII text. There are 6,228 questions that contain non-ASCII characters, which are present in 8,744 different question pairs. Additionally, we found two pairs where one of the questions is missing, represented as an empty string.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Duplicate_Questions	NonDuplicate_Questions	Is_Duplicate
Which business is good start up in India?	Which question should I ask on Quora?	1
Which business is better to start in India?	What are the questions should not ask on Quora?	0

TABLE1. Examples of Dataset question pair.

B. Preprocessing

Our dataset, being a substantial collection of human-generated text, inherently contained various anomalies, particularly the presence of non-ASCII characters. For instance, a preliminary tokenization process, without further cleaning, yielded a vocabulary of approximately 1.75 lacs distinct terms. However, by systematically removing numerical digits and punctuation, we were able to significantly reduce this vocabulary to about 1 lac unique words and phrases. To thoroughly understand how different preprocessing techniques influenced our model's test accuracy, we conducted a series of experiments. Given the exponential increase in experimental permutations with each additional preprocessing function, we adopted a methodical approach, starting with trials involving individual or small, logically grouped sets of functions.

Each experimental configuration was rigorously evaluated through three separate runs. Following the training of a linear classifier, which employed hinge loss and Stochastic Gradient Descent (SGD) with 50 iterations and a learning rate of 0.00005, we calculated the average accuracy on our validation set. Non-ASCII characters were specifically removed when working with linear models, while for tree-based models, punctuation was handled with a distinct strategy, adapted to their particular structural requirements.

IV. PROPOSED METHODOLOGY

1. Linear Models

In the field of Natural Language Processing (NLP), N-grams are a fundamental concept, frequently employed in statistical models to enhance their capacity for capturing sequential patterns and contextual information within text. This approach allows models to discern relationships between words that extend beyond individual terms, thereby broadening their overall analytical capabilities. For our baseline analysis, we adopted a similar n-gram-based strategy, constructing three distinct linear models, each leveraging unique sets of n-gram features. The process of extracting these features began with meticulous preprocessing of the raw question text. This involved essential steps such as tokenization, which segments sentences into individual words or sub word units, and the systematic elimination of non-ASCII characters to standardize the input. Following preprocessing and tokenization, we extracted n-grams by counting their occurrences within each question. Each unique n-gram, along with its frequency, was treated as a distinct feature. To ensure compatibility and computational efficiency with scikit-learn's implementations of linear and Support Vector Machine (SVM) models, these features were then transformed into SciPy's sparse CSR (Compressed Sparse Row) matrix format. This model was optimized using Stochastic Gradient Descent (SGD) with L2 regularization, a technique that helps prevent overfitting by penalizing large coefficients. The implementation was carried out using the scikit-learn library. The learning rate for the SGD optimizer was dynamically adjusted using a schedule defined as 1.0 / (t + t0).

Given that the unigram model exhibited the fastest training times, we used it for an initial approximate tuning of the regularization strength and the number of iterations. Through this preliminary optimization, we determined that a regularization parameter of 0.00001 and 20 iterations yielded the most favourable outcomes. As anticipated, the model trained with trigram features, which capture richer contextual information, ultimately delivered the best performance.

2. Neural Network (NN) Model

Neural networks draw their fundamental inspiration from the biological neural systems of the human brain. These computational models consist of interconnected processing units, often referred to as 'neurons,' which work in concert. Through algorithms, they organize and classify raw data, uncover intricate patterns, and continuously refine their understanding. Networks that incorporate multiple intermediate layers are commonly termed deep neural networks. Initially, artificial neural networks (ANNs) were conceived as an endeavor to replicate the human mind's cognitive architecture, specifically to tackle complex problems that proved challenging for conventional algorithmic approaches.

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806152

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

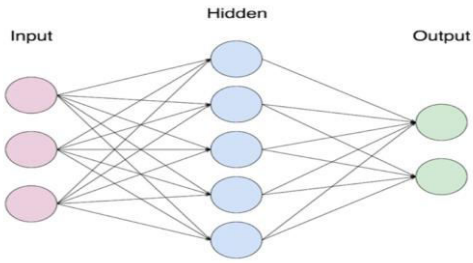


Figure2. Diagram of Classical Neural Network

These examples are typically provided with pre-assigned labels. For instance, an object recognition system, trained on a vast collection of images meticulously tagged with categories like automobiles, houses, or coffee cups, learns to identify consistent visual features associated with each label.

3. Convolutional Neural Network (CNN) Model

This section introduces our Convolutional Neural Network (CNN) model, representing the initial deep learning approach employed in this study. It outlines how specialized machine learning and ensemble algorithms are trained within our framework to accurately predict whether question pairs within the dataset are semantically identical or distinct. This framework is meticulously designed to advance the investigation into robust duplicate query pair detection. CNNs are particularly adept at processing textual data by identifying local patterns and salient features, akin to their success in recognizing visual patterns in images. Our methodology leverages these inherent strengths, training the CNN to discern nuanced semantic similarities between questions, thereby forming a crucial component of our overall detection strategy.

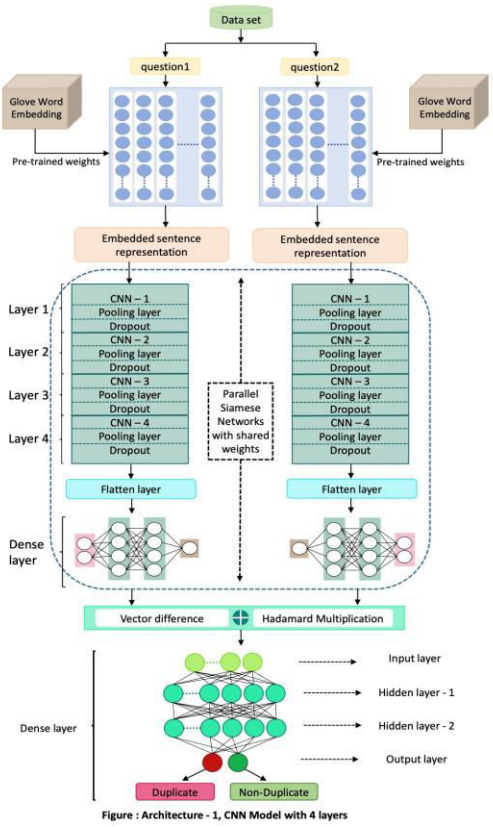


Figure3. Four Layer CNN Model

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806152

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4. Recurrent Neural Networks (RNN) Model

A Recurrent Neural Network (RNN) is a type of neural network specifically designed to handle sequential data, represented as a series of elements x(1), x(2),..., x(τ) indexed by time step t. These networks are particularly advantageous for tasks involving ordered inputs, such as speech recognition and natural language processing. RNNs are termed "recurrent" because they apply the same set of operations to each element within a sequence, with the current output being contingent upon previous computations. This inherent design grants RNNs a form of "memory," allowing them to retain and utilize information processed earlier in the sequence.

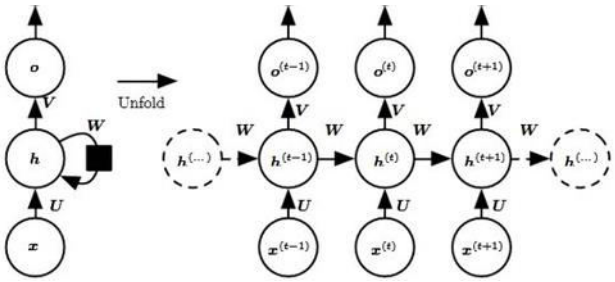


Figure4. Architecture Model of RNN Network

5. Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory networks (BiLSTM), a type of recurrent neural network, are particularly effective in capturing the sequence of words in text, which is vital for understanding contextual meaning. The literature review highlights how models like RNNs and BiLSTMs process input sequentially, maintaining a hidden state at each step to track the relationships between words. Such models are especially useful when analysing queries that exhibit complex sequential patterns, where context from both directions significantly enhances comprehension.

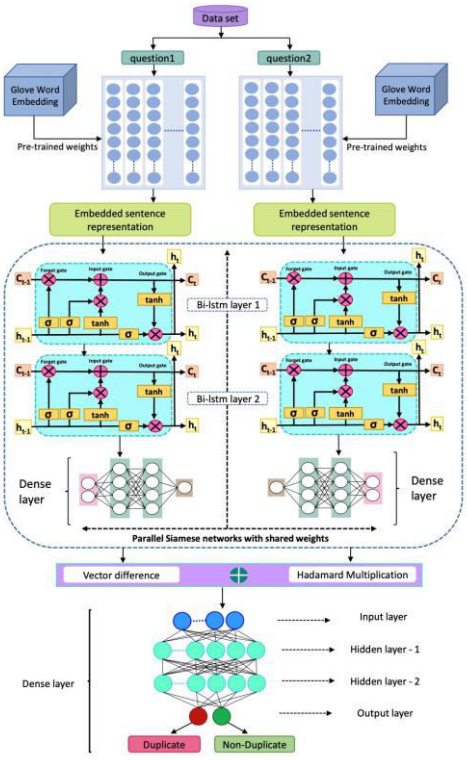


Figure5. Two-Layer Long Short-Term Memory Model

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Our proposed approach for duplicate question pair detection utilizes a BiLSTM-based model, incorporating a slightly distinct additional feature. In this neural network architecture, we feed each question of a pair as separate inputs.
The initial hidden layer of this network is an Embedding Layer. This layer transforms individual words into dense numerical vectors, with specified dimensions: an input size derived from the vocabulary (plus one for padding), an output size of 300, and an input sequence length of 100.

The architecture employs a parallel Siamese Network structure, where an Embedding Layer and an LSTM Layer are applied independently to each question input. Additionally, we extract a "Common Words Feature" between the question pair.

Subsequently, the output from each LSTM layer and the "Common Words Feature" are individually passed through separate Dense Layers in a parallel fashion. All three outputs from these Dense Layers are then combined using a Concatenation layer. The merged output then proceeds through a sequence of layers: Batch Normalization (to accelerate and optimize training), Dropout (to prevent model overfitting), and a final Dense Layer. We specifically applied a dropout weight of 0.17 within the LSTM layers.

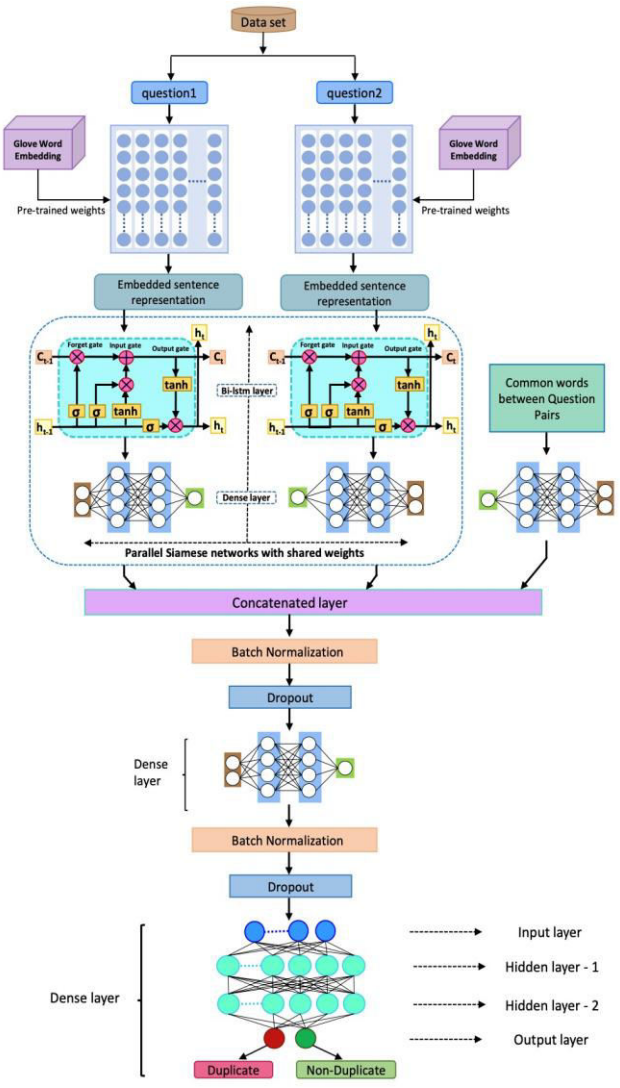


Figure6. Common Word Metrics in LSTM Model

© 2025 IJMRSET | Volume 8, Issue 6, June 2025| DOI:10.15680/IJMRSET.2025.0806152

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. RESULTS AND DISCUSSION

The performance summary of the top models from each category is presented below using classification reports. These reports provide detailed insights into key evaluation metrics such as precision, recall, and F1-score, helping to compare the effectiveness of each model in identifying duplicate question pairs accurately.

	precision	recall	f1-score	support
0	0.78	0.91	0.84	76350
1	0.79	0.55	0.65	44937
micro avg	0.78	0.78	0.78	121287
macro avg	0.78	0.73	0.74	121287
weighted avg	0.78	0.78	0.77	121287
samples avg	0.78	0.78	0.78	121287

Figure7. CNN Model: Classification Report

	precision	recall	f1-score	support
0	0.78	0.91	0.84	76350
1	0.78	0.58	0.66	44937
micro avg	0.78	0.78	0.78	121287
macro avg	0.78	0.74	0.75	121287
weighted avg	0.78	0.78	0.78	121287
samples avg	0.78	0.78	0.78	121287

Figure8. BiLSTM Model: Classification Report

	precision	recall	f1-score	support
0	0.89	0.83	0.86	76350
1	0.74	0.83	0.78	44937
micro avg	0.83	0.83	0.83	121287
macro avg	0.82	0.83	0.82	121287
weighted avg	0.83	0.83	0.83	121287
samples avg	0.83	0.83	0.83	121287

Figure9. BiLSTM with Common Words Feature: Classification Report

We have further compared the f1 score (f1-0 and f1-1) of all the three best-performing models:

S.No.	Model	f1—0	F1-1
1	CNN	0.84	0.65
2	BiLSTM	0.84	0.66
3	BiLSTM with Common Words	0.86	0.78

Table2. f1 Score Comparison of 3 Proposed Models

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

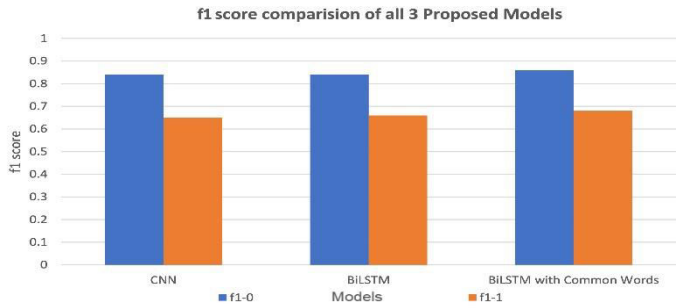


Figure10. Graph of f1 Score Comparison of 3 models

S.No.	Model	Accuracy (%)	Log loss
1	CNN	80.41	0.463
2	BiLSTM	80.59	0.463
3	BiLSTM with Common Words	84.84	0.368

Table3. Comparison of Accuracy and Log Loss of 3 Models

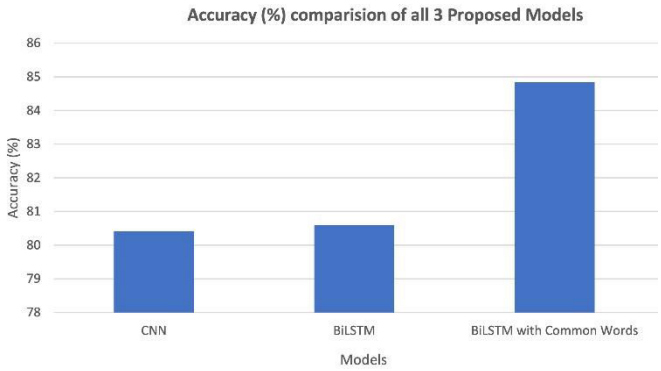


Figure11. Graph of Accuracy Comparison for 3 Proposed Models

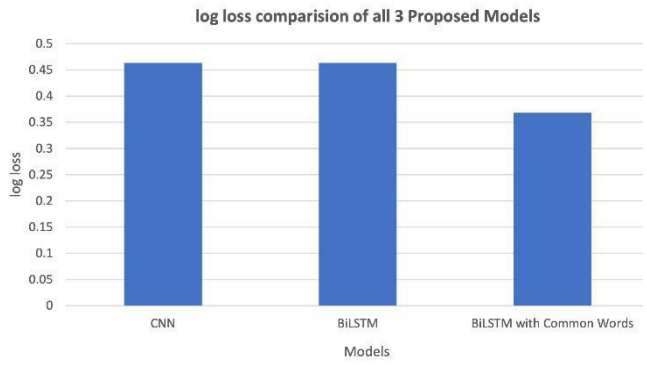


Figure11.2. Graph of Log Loss Comparison for 3 models

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

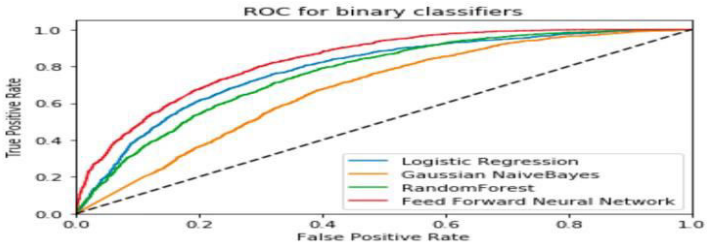


Figure12. Comparison between different the model

Our findings indicate that the BiLSTM model, augmented with the Common Words Feature, achieved the highest accuracy at 85.11%, notably without data augmentation. The two-layer BiLSTM model recorded an accuracy of 80.59%, slightly outperforming the two-layer CNN model, which reached 80.41%.

REFERENCES

[1] Kaggle Quora 4 Lacs Questions Pair CSV Format Dataset: https://www.kaggle.com/datasets/quora/question-pairs-dataset.
[2] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi perspective matching for natural language sentences. arXiv preprint arXiv:1702.03814, 2017.
[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Sa kinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In Advances in neural information processing systems, pages 737–744, 1994.
[4] Shuohang Wang and Jing Jiang. A compare-aggregate model for matching text sequences. arXiv preprint arXiv:1611.01747, 2016.
[5] Neural paraphrase identification of questions with noisy pre-training. arXiv preprint arXiv:1704.04565, 2017.
[6] Wenpeng Yin, Hinrich Schu tze, Bing Xiang, and Bowen Zhou. Abcnn: Attention- based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 4:259–272, 2016.
[7] Machine learning : https://rb.gy/0okprl.2018.
[8] Machine learning working : https://rb.gy/skdpbw. 2020.
[9] Advantages and disadvantages of machine learning : https://rb.gy/s8yuze.2020.
[10] Convolutional Neural Network CNN upgrad blog : https://www.upgrad.com/blog/convolutional-neural-networks/2020.
[11] Structure of cnn : https://rb.gy/nxajwf. 2018.
[12] Tutorials Points. Applications of neural networks : https://rb.gy/dimitb.2020.
[13] Hyper-parameters of cnn : https://rb.gy/fqy8a3.2019.
[14] DENNY BRITZ. Stride of cnn : https://rb.gy/goumbj.2015.
[15] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2. Lille, 2015.
[16] Branislav Holl nder. https://rb.gy/ofpm8g.2018.
[17] Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. Quora question pairs, 2018. [18]
[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 815–823, 2015.
[19] Laura Leal-Taix´e, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese
[20] cnn for robust target association. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 33–40, 2016.
[21] Object co-segmentation using the deep Siamese network. arXiv preprint arXiv:1803.02555, 2018.
[22] Matching resumes to jobs via deep siamese network. In Companion Proceedings of the The Web Conference 2018, pages 87–88, 2018.
[23] Quora question pairs — Kaggle [online] is available: https:// www.kaggle.com/c/quora- question-pairs.
[24] Wenpeng Yin, Hinrich Schu tze, Bing Xiang, and Bowen Zhou. Abcnn: attention-based
[25] Convolutional neural network for modeling sentence pairs. corr abs/ 1512.05193, 2015.
[26] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. Enhancing and combining
[27] Sequential and tree lstm for natural language inference. ar xiv preprint arXiv:1609.06038, 2016.
[28] Kaggle: Your home for data science [online] available: https://www.kaggle.com/.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com