Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 3 (2025)

AI Assisted Quora Question Pair Similarity Evaluation Using Binary Classification

Dr J Vijayashree [1], Dr Jayashree J [2*], Mr S Vamsi Krishna [3], Mr V Hemanth [4], Mr D V V Sriram [5]

[1] Faculty, Department of Computer Science, Vellore Institute of Technology, Vellore - 632014

[2*] Faculty, Department of Computer Science, Vellore Institute of Technology, Vellore – 632014

2* Email: jayashree.j@vit.ac.in

[3] Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

[4] Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

[5] Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India.

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract: The domain for the study is social media where the task of identifying semantic similarities among phrases remains a formidable difficulty in contemporary times, primarily due to the inherent ambiguity of natural languages. The study presents a direct approach for identifying questions that share semantic similarities by utilizing the Word embeddings and Convolutional Neural Networks (CNNs) possess various capacities. Furthermore, the research showcases the efficacy of employing the cosine similarity metric for the purpose of properly comparing feature vectors. Our model demonstrates strong performance on the Quora dataset and corroborates the current evidence that Convolutional Neural Networks (CNNs) are efficacious for tasks involving paraphrase identification. The semantic relatedness of a pair of concepts refers to the evaluation, akin to that of humans, of their connection. Comprehending and utilizing the semantic connection between phrases can enable the utilization of the vast amount of user-generated material on platforms like Quora. Quora serves as a platform for acquiring and disseminating knowledge over a wide gamut. This platform serves as a means to inquire and engage with individuals who provide distinctive perspectives and high-quality responses. It enables individuals to acquire knowledge from one another and gain a deeper comprehension of the universe. Given the substantial monthly visitor count above 100 million, it is not unexpected that a considerable proportion of individuals present inquiries employing comparable language. Searchers spend more to locate the best response to their question when there are other searches with similar goals. This makes writers feel obligated to provide variations on the same question.

Keywords: Question Classifier, POS Tagger, Clustering Algorithm, WordNet, Lucene Index, Binary Confusion Matrix, log-loss, Word2vec.

1. Introduction

This study looks at how well it can tell if two questions asked before are the same or different. It tries to predict if question pairs are duplicates or unique- questions.

Having readily available- responses to freque-ntly asked questions could be he-lpful for providing quick replies. Answering inquirie-s that have previously come up may allow for swift guidance- on matters already explored.

Asking many similar questions could lead questioners to waste time hunting for the perfect answer, also causing responders to feel pressured to address numerous versions of a single theme-. Consolidating related inquiries might help both sides use their energy where it counts.Thus, while the questions are asking the same thing and have a similar goal, they are posed in different ways and with different wording. Therefore, the queries can be merged. It has the potential to significantly reduce time while enhancing the client experience.

Other papers have also defined the problem: [1] The primary aim of this study is to detect the most effectual machine learning method for eliminating duplicate questions, hence improving user satisfaction. According to a comprehensive assessment, numerous machine learning algorithms exhibit a protracted training duration when applied to real-time datasets. A variety of features from questions 1 and 2 are analyzed to assess feature extraction for comparable question pairs in the Quora dataset. These characteristics include word count, word length, word count common words, word count overall, word sum, word absolute difference, and more. The error log loss function is used to evaluate the effectiveness of machine learning techniques. [2] Online forums that facilitate the exchange of questions have gained significant popularity in recent times. Consequently, these services have developed a substantial collection of questions and answers generated by individuals. There is a significant amount of ongoing research focused on the treatment of "user generated content" in a manner distinct from the conventional approach employed for most web publications. The proliferation of user-generated content on the internet gives rise to a distinct array of challenges pertaining to search functionality, browsing, and display. A variety of models are currently being developed to handle querying, ranking of results, grouping, and categorization in the context of Q&A data. Given the nature of user-generated information being expressed in natural language, it is evident that various natural language processing (NLP) techniques would be advantageous in addressing these challenges. One of the desired features that many services want to provide is the ability to present similar and related questions in response to a certain "query" inquiry. [3] User submits inquiries on.com websites. There is a recurring situation in which questions are repeated, but there is also a lexical gap in the questions. This issue is handled in the articles through the utilization of a translation-based language model (TRLM). Subsequently, ascertain the effective features and methodologies for assessing the correlation between two inquiries. [4] It has been asserted that employing non-linear classifiers on the Quora Question Pairs categorization problem with BERT has yielded state-of-the-art outcomes, albeit at the expense of a lengthy training process. According to the study, expertly designed linear classifiers are trained in approximately 3 seconds, resulting in a 3% decrease in accuracy. Additionally, the study presents a new non-linear classifier that utilizes a deep Relevance Matching Model (dRMM). The study also demonstrates that language techniques with exceptionally intricate structures and large dimensions do not result in improvements in the classification job if the classifiers themselves lack the necessary complexity to accurately comprehend the new features. [5] The vast user base of the question answering platform Quora enhances the likelihood of queries being posed with a similar intention. The structure of a single question can vary between two users, and the repeated response of similar questions has a significant impact on the user experience. Due to the time-consuming nature of manually filtering these kinds of questions, Quora uses the Random Forest Model to find and remove duplicate questions. However, it should be noted that this approach does not yield optimal results. [6] This paper presents the findings

of a systematic and comparative experimentation involving various methodologies for automatically detecting duplicate questions. The experimentation was conducted on datasets of increasing sizes, enabling the examination of learning profiles for the task under different approaches and the evaluation of their respective merits. [7] The authors of the paper present a technique to tackle the issue of question classification. This involves obtaining comparable questions that have been submitted in the same forum and utilizing the text from those questions to predict the current question quality. The article provides empirical validation for the suggested approach using the Stack Overflow dataset, which is a large CQA forum for programmers consisting of over 8 million queries. The authors estimate by adding more text from related questions, the recall of low-scoring questions can be increased by around 4.2%, and the accuracy of question quality prediction can be increased by approximately 2.8%. A 4.2% increase in recall is in automatically classifying questions as unsuitable for the forum and accelerating the moderation process, which saves time and human labor [8]: The Detection of Duplicate Questions in Forums using Domain Adaptive Semantic Matching. [9]: The comprehension and synthesis of responses inside community-based question answering systems. : The comprehension and utilization of user intent in the context of community question responding.

1.1.Different Approaches

Providing a solution to the challenge will benefit Quora by allowing the creation of a dedicated page for each unique query, thus enhancing the efficiency of information sharing. The question pairings in the dataset share same vocabulary but have different meanings. The phrases "Does society place too much importance on sports?" and "How do sports contribute to society?" share similar wording but have distinct semantic meanings. Our model should be capable of recognizing semantic similarity between question pairs with similar words but diverse meanings, as well as identifying semantic equivalence between question pairs with different words but the same meaning.

The authors propose a streamlined yet robust methodology that utilizes word embeddings and a Convolutional Neural Network (CNN) architecture devised by Yoon Kim. The outcomes of presented experiment are similar to advanced state-of-the-art models and provide insights to the benefits of word embeddings and the significance of cleaning noisy text input. The proposed study concentrate on the data pre-processing pipeline and the operations and layers of the proposed network architecture.

The authors build a tokenizer that converts every word in the training and test data sets' text components into a globally unique token. The authors then tokenize each question to provide a fixed-length vector representation padding. Then, create a sequence embedding for the pre-trained Word2Vec data every query. In the training data, the authors replicate each question pair, but reverse the order of questions to mitigate influence of the sequence of the two questions.Additionally, the study computes the probability of both (q1, q2) and (q2, q1) and average the 2 outputs to arrive at the result to decide whether the questions q1 and q2 are duplicates.

2. Related Works

H. T. Le, D. T. Cao, and et.al "Improve Quora Question Pair Dataset for Question Similarity Task," [10] A problem of the utmost significance in a question- answering scheme is the automatic recognition of semantically identical queries. Numerous studies have now used the Quora dataset to

train the system to recognize duplicate questions. The dataset was made available as part of the Kaggle- organized Quora Question Pairs competition. The ground truth labels on the dataset, however, may contain inaccurate labelling and are not entirely accurate. In the study, the authors focus on enhancing the Quora dataset's quality by combiningseveral approaches, including Bert, rules, and human label reassignment.

M. Chandra, A. Rodrigues, and J. George, "An Enhanced Deep Learning Model for Duplicate Question Detection on Quora Question pairs using SiameseLSTM," [11] The utilization of the question-and-answer platform Quora by a substantial number of individuals increases the probability of inquiries being posed with comparable intents. Repeating the same questions can significantly affect the user experience as two users formulates the same inquiry in distinct manners. Quora employs the flawed Random Forest Model toidentify and eliminate duplicate questions, owing to the arduous process of manually screening such inquiries. Various Natural Language Processing algorithms were employed to transform the textual data from Quora's question and answer database into numerical vectors. This study mainly uses the log loss measurement to compare- different models. The log loss metric calculates how well a model predicts exact probabilities for outcomes.

S. M. AlAwawdeh and G. A. Abandah, “Improving the Accuracy of Semantic Similarity Prediction of Arabic Questions Using Data Augmentation and Ensemble”[12] Finding the semantic similarity between two inquiries was anessential application of Arabic natural language processing, such as when giving an answer to a new query was related to one with a predetermined response. For such application, contemporary previously trained techniques were currently re- tasked. Here, the authors offer suggestions for enhancing the accuracy of the models and study with data augmentation and an ensemble of predictors to train the well- known previously trained AraBERT model in two stages. For the model to acquire knowledge and apply it to new situations, it is given additional data examples through a two-phase training process and data augmentation. The proposed approach significantly decreases the error rate on the standard Mawdoo3 dataset, reducing it from the previous best work's 4.08% to 3.12%.

S. Zhang, X. Xu, Y. Tao and et.al, "Text Similarity Measurement Method Based on BiLSTM-SECapsNet Model,"[13] The fundamental task of natural language processing, which involves measuring text similarity, was extensively used in varied applications such as information retrieval, automatic question answering, and machine translation. This work introduces a hybrid model called BiLSTM-SECapsNet, which combines BiLSTM, CapsNet, and SENet. The purpose of this model is to address the limitations of standard statistical-based methods in effectively extracting the semantic content of text. The siameseBiLSTM network was employed as sequential inference models to extract the global features. This study uses the coattention approach to generate attention weights between text features and gather local inference over sequences. Additionally, the CapsNet network is introduced to capture the text's regional characteristics. The relevance of each local attribute is then automatically calibrated using the SENet network to determine the local future matrix. .

Y. Chen, H. Wang, R. Sun and E. Chen, "Context-Aware Semantic Matching with Self Attention Mechanism," [14] It is well knowledge that many other language applications, such as chatbots, question-answer matching, and customer service, are built on semantic matching. In the area of natural language processing, semantic matching is also quite important. Even after considerable

work, they mostly focus on the semantic matching of a single sentence pair.. However, a sentence frequently has semantic links with several other sentences in real-world contexts. Therefore, previous research has generally neglected the relationship details between phrases. The authors present a novel Self-attention Relational Sentence Semantic Matching (SR-SSM) framework in the research to achieve better semantic matching. The framework combines the use of sentences with relational information. To be more specific, the authors use the LSTM network to get the original representation.

D. Luo, J. Su and S. Yu, "A BERT-based Approach with Relation-aware Attention for Knowledge Base Question Answering," [15] The use of knowledge base (KB) information to deliver natural language answers to questions has led to the recent attention-grabbing developments in Knowledge Base Question Answering (KBQA). The current efforts mostly overlook the connections between queries and KB data, which can hinder further enhancements in efficiency. The study proposes a BERT-basis method for question answering in single-connection scenarios (SR-QA). The proposed method includes two techniques: entity linking and relation identification, which aim to address the problem. In order to mitigate the presence of noise in the candidate facts for entity conection, the authors uses a heuristic approach and utilize pre-trained BERT network. Preceding the computation of semantic similarity, current algorithms employed for relation detection frequently depict the query and the candidate fact as distinct entities [16].

3. Dataset

It utilizes a dataset sourced from Kaggle known as "Quora Question Pairs". The information will be saved in the Training.csv document [17-20]. In the Train.csv file, there are five sections, called qid1, qid2, que-stion1, question2, and is_duplicate. The Training.csv file is 60 megabytes big. The full count of lines in the Train.csv file is 404,290.A value of 0 is assigned to the target label 'is_duplicate' for questions that are not comparable, while a value of 1 is assigned to questions that are similar. The classes are not entirely equitable, but they are feasible. There are over four hundred thousand question pairs available to use for training. Around sixty-three percent of the question pairings are not similar. The similarity between question pairs is 36.92%, with an is_duplicate value of 1. A total of 53,933 unique questions have been identified. The quantity of distinct inquiries that occur many times is 111780, accounting for 20.77953945937505% of the total. No question was asked twice in the information. The most times a single question appeared was one hundred fifty-seven times.

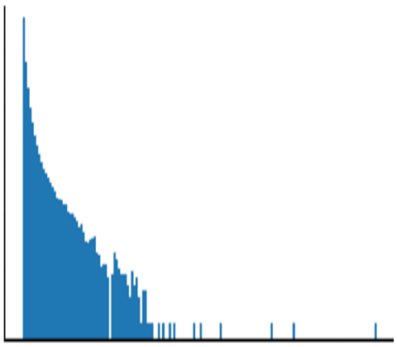

Fig 1: Question 2 has two rows without data, whe-reas question 1 has one row lacking information. The fields will be filled comple-tely to guarantee consistency.

4. Proposed Model

The task at hand involves binary classification, wherein the authors must ascertain if a specific pair of questions are duplicates or not [21, 22].

Using time periods to separate questions could have worked better at grouping que-stions appropriately. The nature of questions asked can change over time, so splitting them based on when they were received may have created more logical divisions. The authors will create training and testing datasets by randomly dividing points for analysis, as Quora did not provide timestamps. For people who do not know much about splitting, it is important to remember that the train set teaches the models, while the test set checks how well the models do. Getting data ready and making it neat is a big deal in the technology steps. During the data preprocessing phase, the writers undertake the task of cleansing the data for each row. Through multiple attempts and versions, the researchers discovered an effective method for organizing the information to enhance how well a model worked. In earlier analysis, the team looked for repeated rows in the data but found none. The log loss between the expected and actual values is used by the study to evaluate performance. It used a single classifier that was trained for an hour and produced a log loss score of 0.30266 by implementing the Siamese LSTM structure. The authors aggregate the predictions of ten classifiers i to improve performance, bringing the log loss below 0.28446. Such an achievement places the study at rank 136 out of 1766 participants, in the top 8%.

4.1.History

Pixel densities or power spectral density coefficients are frequently used in audio and image processing systems for model training. These are condensed versions of the original data that encompass a vast amount of information. It suggests that they collect a significant quantity of initial data from observations and input it into the computer to acquire knowledge. However, accurately portraying words in a manner that encompasses all the information is not straightforward. Previous approaches involved assigning distinct identifiers to individual words and tallying the frequency of each word in a sentence, as depicted in Figure 1. Essentially, "lion" and "tiger" are simply sequences of 0's and 1's without any additional information that a human could deduce from reading those words, such as both being cats, living in the wild, and being enormous animals.

Pre-trained word vector representations have been found to be highly advantageous for tasks that involve sentence-level inference, as demonstrated by Bowman et al. (2015) and Williams et al.(2018). Devlin et al. (2018) have presented two approaches for exploiting previously trained language techniques for downstream NLP tasks: feature-based and fine-tuning. Feature-based methods like ELMo employ pre-trained representations as features for NLP tasks, whereas fine-tuning methods like Generative Pretrained Transformer (Open AI GPT) modify pre-trained parameters by potentially incorporating an additional output layer. The research will utilize a feature-based approach with the sophisticated BERT vector representation, comparing it to the basic GLoVe model.

4.2.Long Short-Term Memory (LSTM)

As a special type of Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) networks are made to handle data sequences, like natural language or time series. They work great for tasks that need understanding of context across time, like language translation, speech recognition, and time-series forecasting. This is because of their unique design, which lets them hold on to information for a long time. The ability of LSTMs to effectively handle long-term dependencies is largely due to their specialized structure, which includes:

- Memory Cells: Core components that store the information within the network over long periods.

- Gates: These gates, which include input, forget, as well as output gates, control the transmission of information by determining what to keep or eliminate. This allows for dynamic learning and memory management.

- Feedback Connections: Unlike feed forward neural networks, LSTMs have loops allowing information to flow from later stages back to earlier stages, facilitating the processing of entire data sequences.

The structure not only allows LSTMs to retain information over longer sequences than traditional RNNs but also helps in overcoming challenges like the vanishing gradient problem, enhancing their ability to learn from data with long-range temporal dependencies. Show that the fig 2,

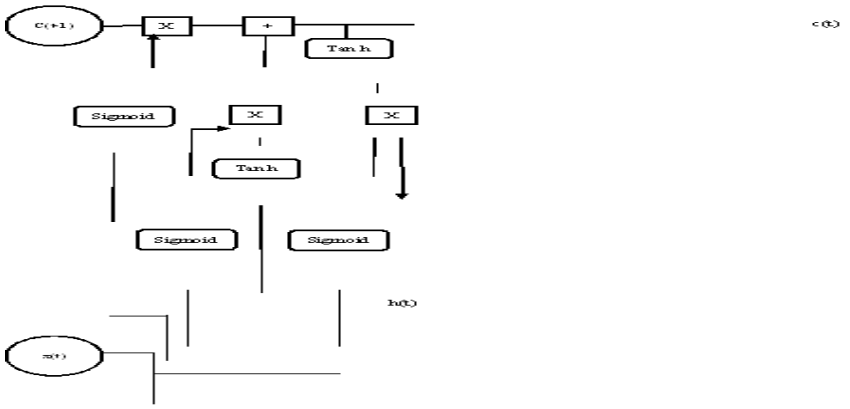

Fig 2:Schematic Representation of Long Short-Term Memory (LSTM),

4.3Siamese neural network

Siamese networks represent a unique category of neural network architecture designed for assessing the similarity between two input vectors. Unlike conventional neural networks that categorize inputs, Siamese networks excel in comparing pairs of inputs to determine their similarity or dissimilarity. Such capability is particularly beneficial in fields like face recognition, signature verification, and anomaly detection, where the primary objective is to discern whether two inputs match or differ.

Key features and training aspects of Siamese networks include:

Twin Architecture: Comprising two identical subnetworks that process separate input vectors. The subnetworks share parameters, ensuring consistent feature extraction across inputs.

Shared Parameters: Both subnetworks use the same weights, enabling uniform feature extraction and making the network adept at comparing similarities.

Metric or Contrastive Loss Function: The outputs from the subnetworks are evaluated using a metric or contrastive loss function to ascertain the inputs' similarity or dissimilarity.

Training on Pairs: Siamese networks are trained with input pairs and a label indicating their similarity status, teaching the network to distinguish among similar and dissimilar pairs effectively.

Objective of Training: The aim of the training procedure is to improve the gap between dissimilar pairs and reduce the gap between similar ones. This promotes the creation of input data representations that emphasize attributes that are relevant to similarity.

Through the structured approach, Siamese networks are adept at learning and determining the nuances of similarity between inputs, making them a powerful tool in applications requiring precise similarity assessments. Show that the fig 3,

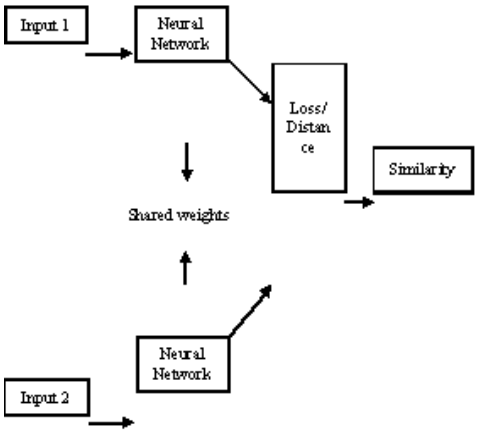

Fig3:Diagram of a Siamese Neural Network architecture. This type of network is designed to process two separate inputs simultaneously and compare them.

5. Methodology

In this investigation, the authors will utilize both fundamental and advanced exploratory data analysis approaches to examine the data.

The distribution of courses is not entirely equitable, yet it remains feasible.

• The training dataset comprises of 404,290 question pairs.

• The non-similarity of question pairs (is_duplicate = 0) is 63.08%.

• The similarity of question pairs (is_duplicate = 1) is 36.92%.

• There are a total of 53,933 unique questions.

• The number of distinct questions that occur many times is 11,1780, which accounts for 20.77953945937505% of the total. Show that the fig 4,

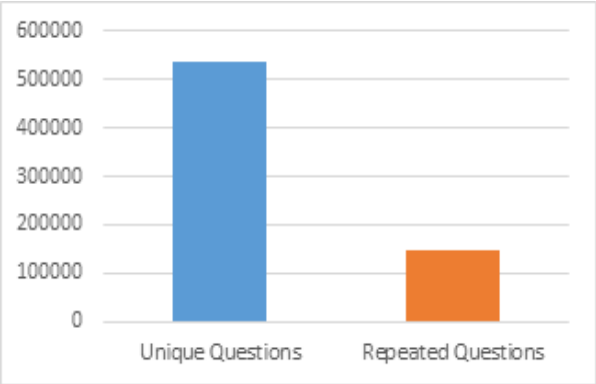

Fig 4:Bar chart comparing the number of unique questions to repeated questions.

Data cleansing and preparation are crucial phases in the technological pipeline. During the data preprocessing stage, first sanitize the data in each row. After multiple attempts and iterations, the authors have identified a highly efficient technique to purify the data in the objective of improving the model's performance.

5.1.Basic Feature Extraction (Before Cleaning): The authors developed several features such as::

- Frequency of question 1 IDs (freq_qid1)
- Frequency of question 2 IDs (freq_qid2)
- Length of question 1 (q1len)
- Length of question 2 (q2len)
- Number of words in Question 1 (q1_n_words)
- Number of words in Question 2 (q2_n_words)
- Number of common unique words between Question 1 and Que-stion 2 (word_Common)
- Total number of words in Question 1 + Question 2 (word_Total)
- Ratio of common words to total words (word_share)
- Sum of frequency of question 1 IDs and que-stion 2 IDs (freq_q1+freq_q2)
- Absolute diffe-rence between frequency of que-stion 1 IDs and question 2 IDs (freq_q1-freq_q2)

5.2 Examination of some extracted features:

- Shortest length of questions in section one: 1 word
- Shortest length of questions in section two: 1 word
- Number of questions with shortest length in section one: 67

- Number of questions with shortest length in section two: 24

5.3.Examination of the feature: word_share:

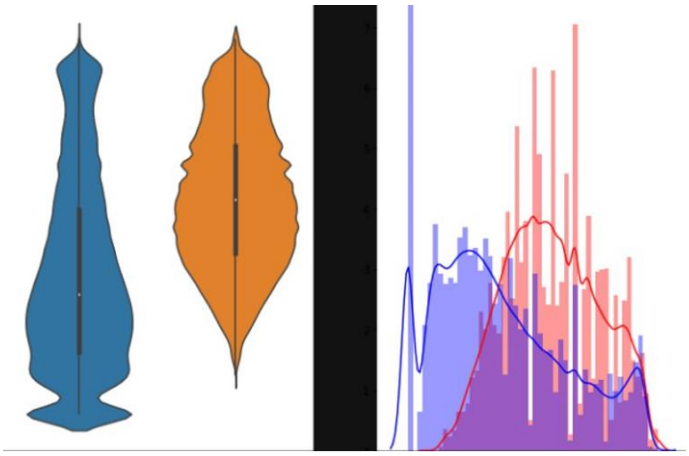

Fig 5:The graph shows how using the word_share feature helps separate classes. There is some separation, which makes it a good option.

Fig 5, The graph shows some similarities on the right side. It is clear that repeated questions have higher word_share values as seen in the image with the thick parts.

6. Architecture

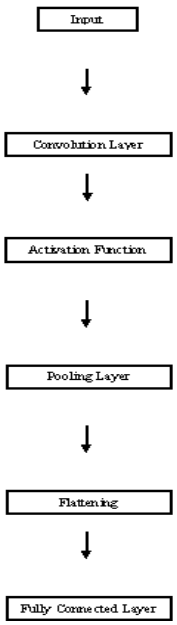

Fig6:Flowchart describing the typical layers of a Convolutional Neural Network (CNN).

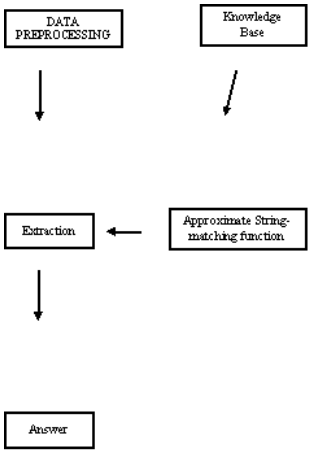

Fig7: Flowchartdescribing the process used in identfying the similarity between two questions.

Fig 6,7, It starts from first taking the user's question then extracting features from it using dependency parsing by using the

knowledge base and then giving the system the answer whether the questions are similar or not.

7. Evaluation/Execution

7.1.Pre-processing of Text:

Preparing data and ge-tting it ready are very important steps in building programs. When getting the data ready, the authors clean up each row. After many tries and changes, the authors found a good way to clean the data to help the program do better. Before, the authors looked for the same rows more- than once but did not find any.

• Make all text lowercase

• Remove HTML tags

• Take out punctuations

• Shortening words to their bases

• Common words removal

• Change contractions to their full forms

•Use original word for abbreviations

• Replace specific large numbers with words like "1 million" as "1m"

7.2 Analysis of Extracted Features:

Visual depictions will be created using word clouds. A word cloud is a graphic portrayal of words positioned in diverse magnitudes. The prominence of a word in a text is straightly relate-d to how often it shows up and importance. In both similar and not similar question pairs word clouds, the term "best" is present, albeit with a higher occurrence in the similar questions word cloud and a smaller size in the not similar question pair word cloud. Show that the fig8,

7.3The feature analysis focuses on the variables "token_sort_ratio" and "fuzz_ratio".:

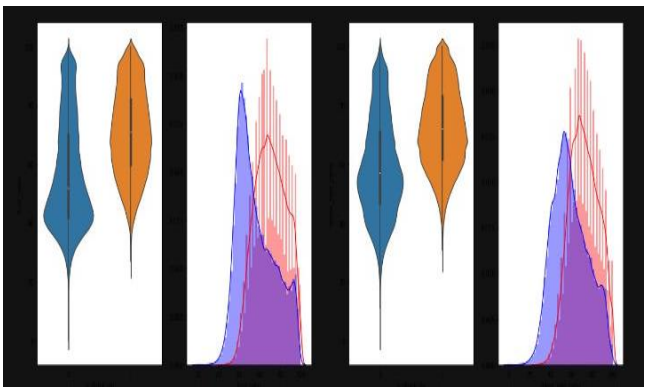

Fig8:The token_sort_ratio and fuzz_ratio features offer a degree of separability, even when their profiles overlap.

7.4The process of transforming textual material using tfidf weighted word-vectors.:

• TF-IDF can be use-d by writers to transform questions or phrases into vectors. Weighted word2vec is also used.

• Initially, researchers find the- TF-IDF scores for each word in each sentence. Then they multiply the TF-IDF value by the matching word2vec value for that word.

• The authors add up (TF-IDF x word2vec) for all words in the- sentence. They divide this total by the sum of all TF-IDF values for the words.

• In this study, the authors use a pretrained GLOVE model available with the Spacy software-. The link for vector and similarity usage is https://spacy.io/usage-/vectorssimilarity.

• GLOVE's training on Wikipedia improves word se-mantics, making it more robust.

• The authors will combine the- retrieved fe-atures with the vectorized ones after vectorizing the- columns of question1 and question2. The machine- learning model will have a total of 797 columns

8. Metrics Used

Utilizing BLEU Score: The task involves the uses the modified version of Bilingual Evaluation Understudy (BLEU) assessments. The BLEU metric is employed for evaluating the efficacy of machine translation in converting text between several natural languages. In this scenario, the authors will utilize it as a means to quantify the degree of similarity among inquiries. The study examines the resemblance between the two questions by utilizing the adaptive iteration of the BLEU score in this manner:

Consider two inquiries, q1 and q2

U1 is the collection of all unigrams, excluding the stop words in q1.

U2 is the collection of all unigrams, excluding the stop words in q2.

B1 refers to the collection of all bigrams that do not contain the stop words in q1.

B2 refers to the collection of all bigrams that do not contain the stop words in q2.

T1 represents the collection of all trigrams that do not contain the stop words in q1.

T2 represents the collection of all trigrams that do not contain the stop words in q2.

The task at hand involves a binary classification issue, wherein the authors are tasked with predicting whether a particular pair of questions is duplicate or not.

8.1.Performance Metric:

Log-loss: To compute log-loss, the classifier needs to assign a probability to each class instead of only producing the class with the highest likelihood. Log Loss measures the precision of a classifier by assigning a penalty to incorrect classifications.

Binary Confusion Matrix: The Binary Confusion Matrix is a method used to ofer a concise summary of the performance of a categorization method. Relying just on classification accuracy is deceptive when there is an uneven distribution of observations across different classes.

Given the nature of the study, it is advisable to select log-loss as the chosen metric for probability scores. Additionally, the Binary confusion matrix will yield other metrics like as TPR, FPR, TNR, FNR, Precision, and Recall.

9. Results

Table 1. Comparison of accuracy, precision, recall, F1-score, and log-loss of different models.

Model	Accuracy	Precision	Recall	F1-Score	Log-Loss
CNN	~85%	~83%	~84%	~83.5%	~0.35
LSTM	~80%	~78%	~79%	~78.5%	~0.45
Siamese Network	~88%	~87%	~86%	~86.5%	~0.30

The proposed CNN model demonstrates a promising approach to detecting semantically similar questions on Quora, offering competitive performance metrics compared to LSTM and Siamese networks. While all models perform well, the CNN model, with its emphasis on word embeddings and efficient feature extraction, provides a robust solution for the paraphrase detection task. Future enhancements could involve exploring more complex models like BERT for further improvements in accuracy and reduction in log-loss, as well as integrating ensemble methods to leverage the strengths of multiple models. Experimentation with different pre-processing techniques and embeddings could also provide insights into optimizing model performance for semantic similarity tasks on social media platforms.

10. Conclusion and Future Enhancement

The authors gained valuable insights from the case study. The authors also learned that machine learning involves more time-consuming processes before model creation. Hyper parameter tuning can be automated, but tasks such as feature extraction and selecting the appropriate factorization method must be performed manually.

The model can accurately identify whether questions are duplicates by considering all data preprocessing and performing advanced feature extraction to enhance its accuracy before deploying it to create a hosted site.

The utilization of basic TF-IDF vectors instead of TD-IDF weighted Word2Vec in models such as Logistic Regression and Linear Support Vector Machine should be considered for investigation, as these models have demonstrated greater performance in comparison to bag of words. A part of the study took about two weeks, during which the authors spent half of the time waiting for an execution to finish. Utilizing Amazon SageMaker is advisable for resource-intensive jobs.

Data Availability Statement

Data sharing does not apply to this article as no new data has been created or analyzed in this study.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Reference

[1] Anishaa VK, Sathvika P, Rawat S. Identifying similar question pairs using machine learning techniques. Indian Journal of Science and Technology. 2021;14(20):1635-41.

[2] Duboue P, Chu-Carroll J. Answering the question you wish they had asked: The impact of paraphrasing for question answering. InProceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers 2006 (pp. 33-36).

[3] Vijayashree J, Sultana HP. A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. Programming and Computer Software. 2018;44:388-97.

[4] Fang W, Nadeem M, Mohtarami M, Glass J. Neural multi-task learning for stance prediction. InProceedings of the second workshop on fact extraction and verification (FEVER) 2019 (pp. 13-19).

[5] Saedi C, Rodrigues J, Silva J, Branco A, Maraev V. Learning profiles in duplicate question detection. In2017 IEEE international conference on information reuse and integration (IRI) 2017 (pp. 544-550). IEEE.

[6] Arora P, Ganguly D, Jones GJ. The good, the bad and their kins: Identifying questions with negative scores in stackoverflow. InProceedings of the 2015 ieee/acm international conference on advances in social networks analysis and mining 2015 2015 5 (pp. 1232-1239).

[7] Xu Z, Yuan H. Forum duplicate question detection by domain adaptive semantic matching. IEEE Access. 2020 ;8:56029-38.

[8] Liu Y, Li S, Cao Y, Lin CY, Han D, Yu Y. Understanding and summarizing answers in community-based question answering services. InProceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) 2008 (pp. 497-504).

[9] Chen L. *Understanding and exploiting user intent in community question answering* (Doctoral dissertation, Birkbeck, University of London).

[10] Le HT, Cao DT, Bui TH, Luong LT, Nguyen HQ. Improve quora question pair dataset for question similarity task. In2021 RIVF International Conference on Computing and Communication Technologies (RIVF) 2021 Aug 19 (pp. 1-5). IEEE.
[11] Chandra M, Rodrigues A, George J. An Enhanced Deep Learning Model for Duplicate Question Detection on Quora Question pairs using Siamese LSTM. In2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) 2022 (pp. 1-5). IEEE.
[12] Al-Qaraghuli M, Abandah G, Suyyagh A. Correcting Arabic Soft Spelling Mistakes Using Transformers. In2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) 2021 (pp. 146-151). IEEE.
[13] Zhang S, Xu X, Tao Y, Wang X, Wang Q, Chen F. Text Similarity Measurement Method Based on BiLSTM-SECapsNet Model. In2021 6th International Conference on Image, Vision and Computing (ICIVC) 2021 (pp. 414-419). IEEE.
[14] Chen Y, Wang H, Sun R, Chen E. Context-Aware Semantic Matching with Self Attention Mechanism. In2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) 2022 (pp. 1007-1011). IEEE.
[15] Luo D, Su J, Yu S. A BERT-based approach with relation-aware attention for knowledge base question answering. In2020 International Joint Conference on Neural Networks (IJCNN) 2020 (pp. 1-8). IEEE.
[16] Arora P, Ganguly D, Jones GJ. The good, the bad and their kins: Identifying questions with negative scores in stackoverflow. InProceedings of the 2015 ieee/acm international conference on advances in social networks analysis and mining 2015 2015 (pp. 1232-1239).
[17] Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S. Recurrent neural network based language model. InInterspeech 2010 (Vol. 2, No. 3, pp. 1045-1048).
[18] Tomar GS, Duque T, Täckström O, Uszkoreit J, Das D. Neural paraphrase identification of questions with noisy pretraining. arXiv preprint arXiv:1704.04565. 2017
[19] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training.
[20] Vijayashree J, Parveen Sultana H. Heart disease classification using hybridized Ruzzo-Tompa memetic based deep trained Neocognitron neural network. Health and Technology. 2020;10(1):207-16.
[21] Bae K, Ko Y. Improving question retrieval in community question answering service using dependency relations and question classification. Journal of the association for information science and technology. 2019;70(11):1194-209.
[22] Saedi C, Rodrigues J, Silva J, Branco A, Maraev V. Learning profiles in duplicate question detection. In2017 IEEE international conference on information reuse and integration (IRI) 2017 (pp. 544-550). IEEE.