Multimedia Tools and Applications (2022) 81:6131–6157
https://doi.org/10.1007/s11042-021-11771-6

A hybrid approach of Weighted Fine-Tuned BERT extraction with deep Siamese Bi – LSTM model for semantic text similarity identification

D. Viji[1] · **S. Revathy**[2]

Received: 27 July 2021 / Revised: 26 October 2021 / Accepted: 25 November 2021 /
Published online: 6 January 2022

Abstract
The conventional semantic text-similarity methods requires high amount of trained labeled data and also human interventions. Generally, it neglects the contextual-information and word-orders information resulted in data sparseness problem and latitudinal-explosion issue. Recently, deep-learning methods are used for determining text-similarity. Hence, this study investigates NLP application tasks usage in detecting text-similarity of question pairs or documents and explores the similarity score predictions. A new hybridized approach using Weighted Fine-Tuned BERT Feature extraction with Siamese Bi-LSTM model is implemented. The technique is employed for determining question pair sets using Semantic-text-similarity from Quora dataset. The text features are extracted using BERT process, followed by words embedding with weights. The features along with weight values, are represented as embedded vectors, are subjected to various layers of Siamese Networks. The embedded vectors of input text features were trained by using Deep Siamese Bi-LSTM model, in various layers. Finally, similarity scores are determined for each sentence, and the semantic text-similarity is learned. The performance evaluation of proposed-framework is established with respect to accuracy rate, precision value, F1 score data and Recall values parameters compared with other existing text-similarity detection methods. The proposed-framework exhibited higher efficiency rate with 91% in accuracy level in determining semantic-text-similarity compared with other existing algorithms.

Keywords BERT · Bi-LSTM · CNN · NLP · Semantic text-similarity · Embedded vectors · Siamese networks

✉ D. Viji
dviji2k@gmail.com

[1] Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

[2] Department of Information Technology, Sathyabama Institute of Science and Technology, Chennai, India

1 Introduction

The measurement on Textual-similarity turns to be the challenging complication as it demands the better semantic understanding of input-sentences. The past models of neural-networks utilized coarse-grained modeling of sentences that has complexity in capturing out the fine-grained word range information for comparisons of semantic-texts. These efforts in exploring the efficient methods for semantic-text similarity detections, has grabbed attentions in NLP-Natural language-processing [4]. These task of text-similarity detections plays the significant role in tasks like text-summarization, minimizing the redundancy of duplicates of documents or the question pairs in portals, question generations, tracking of topics, machine-translation, document clustering, essay scoring etc., The present works upon the determination of text-similarity has been partitioned to three type of approaches such as knowledge-based similarities, corpus-based similarities and string-based similarities [6]. The neural-network model on the basis of Bi-LSTM and ConvNet has been presented for measuring the semantic-text similarity. In this type of model, the similarity focused-layer and the pair-wise interaction model of the words would capture efficiently the fine-grained semantic data [9].

The input data in text-similarity analysis, the pre-processing algorithm which chains out the co-referential entities altogether and then performed the segmentation of words to retain the phrasal verbs meaning and idioms meaning. The similarity analysis is also employed for short-texts as well [24].This classification in this study, algorithm presents the short text representation in the form of two dense type vectors. The first vector is constructed utilizing the word to word similarity on the basis of pre-trained vectors of words and the other dense vector uses the similar word to word-similarity on the basis of external knowledge sources [26]. Along with the sentence pairs, or the question pairs and document comparisons, the para-phrase identification is presented as the new approach [10]. This approach analyses the text-similarity among the sentences pair on the basis of semantic-levels and lexical-levels through integrating keywords usage and neural-networks [39]. Hence focusing on semantic Text-similarity identification, the study employs the hybrid approach of Fine-tuned BERT feature extraction with Siamese network Bi-LSTM classification technique obtained from Quora dataset. The BERT extraction process, extracts the features of the sentence sin question pairs. These embedded vectors were trained by using this Siamese network with Bi-LSTM model. This embedded-vector from BERT process, were connected to input vector that traverses through multi-layer perceptron of Bi-LSTM model. This produces the output-labels of the words. The trained model outputs, their vectors predict the text-similarity of the documents or the question pairs in terms of similarity scores.

The major advantages of the proposed method such as for Bi-LSTM- better features are learned, stores in memory and followed by based on stored input, the output is obtained, for Siamese- the learning process is effective and for BERT feature extraction, it can mapping lot of sentences, vectors are accurate and finally the output weights are passed to Siamese input.

The major contributions of the study are listed below:

- To generate the feature-vectors from preprocessed data using Weighted Fine-Tuned BERT extraction process.

- To train the embedded feature vectors by Deep Siamese network based Bi-LSTM model through multi-layer perceptron. The trained vectors are generated with output-labels.
- To predict the trained model output from the text-similarity between different set of question pairs from quora dataset. Hence the similarity-scores among various sentence pairs were determined by the technique.
- To evaluate the Similarity predictions with respect to various parameters such as accuracy rate, recall value, precision factors and F1-score value. The validation is also employed in different documents comparison, and obtaining text-similarity.

1.1 Paper-organization

The section I describes the basic concepts of Text-similarity and the introductory section for how it is determined by using various NLP approaches for minimizing the duplication and other context issues. The section II enumerates the review of various existing Text-similarity detections models or systems relied in different datasets. The section III explores the proposed hybrid framework approach to implement efficient text-similarity detections by using BERT extraction with Siamese Bi-LSTM model. The outcomes obtained through this proposed-framework are illustrated in section IV. Finally the overall conclusion statements and result inferences were stated in section V.

2 Review of existing work

The modeling task of semantic similarity among the text pair turned out to be the critical Natural language-processing task in question answering applications to plag-detection applications. For this purpose, numerous models were proposed such as traditional feature-engineering techniques and deep learning models.

This type of semantic-text similarity detections is one of the primary roles in natural-language processing. For this initiative, the topic-informed architecture of BERT based model is established for pair-wise similarity in text detections [28]. The BERT based model enables for the feature selection process. This model enhances the performance level upon baselines of strong neural network over the English-language datasets. These enhancements were attained success upon these domain-specific words in qualitative-analysis phase.

The major level of advantage in some pre-trained language-models is the ability that they could efficiently absorb the word context within the sentences. For the analysis and investigation of this efficiency, the Pretrained Bi-Directional Encoder Representations from Transformers model of language processing is implemented and it is fine-tuned upon two QA-Question-answering data-sets and three CQA-community Question-answering data-sets is chosen for this answer-selection process [19]. It has been depicted that this approach attains the maximum enhancement denoting 13.10% in the datasets of Question-answering type and 18.70% in CQA- dataset types in comparison with the other conventional state-of art models. Several NLP applications including information retrieval-engines, dialog medical diagnosis systems for COVID-1 relies in the capability for measuring STS-.

Semantic textual-similarity in some existing data-sets. These models sometimes failed to yield out the performance towards domain-specific environment of COVID-19. For to rectify this gap, The CORD-19STS data-sets is introduced that comprises of around 13,710

sentence-pairs gathered from COVID-19 open-research data-set. As the result, around 1million sentence-pairs were generated by utilizing this various sampling-strategies [7]. The BERT-language model is implemented for scores similarity calculation in accordance with various similarity-levels among different sentence-pairs. This provides of about 32,000 sentence-pairs. The broadly utilized social media entertainment has provided the users with bunches of news depicting about similar events of the society and government. Hence as the attempt for determining the semantic-similarity and paraphrases has turned out as the demand for avoiding the similar news appearing in different channels in different times. This study employs the state of-art techniques for rectifying this issue. The widely used social media have flooded their users with news talking about similar events. Therefore, detecting para-phrases and semantic similarity analysis have become a need to avoid receiving the same news post several times. This research proposes a state-of-the-art approach for PI and STS analysis of news Arabic tweets [25]. The proposed approach employs a set of extracted features based on lexical, syntactic, and semantic computation. Moreover, the approach uses word alignment features to detect the level of similarity between tweets pairs. Likewise, another model which utilizes the knowledge-distillation strategical technique for training out the light-weighted deployment-friendly models of student's information by using this proper weight-initialization and layer-pruning technique. The student-model possesses complete independency of upper teacher-model and results has been generalized through this BERT-like teacher-model [20]. The outcomes of the student model exhibits two-times rapid and attains 96 accuracy rate detections of this fine-tuned BERT-model. This model can be further improvised with data-augmentation and unlabeled information. These BERT topic-models were integrated to bring out the flexible framework. In this scenario, the fake news-detection tools is one of the critical requirement which has to be automated for this detection. In order to address out the mentioned complexity, the framework integrates the potentials of LSTM and CNN-model, and it is utilized with two distinct dimensionality reduction-techniques [36]. The Techniques are Chi-Square and PCA-Principal Component-analysis technique. This implemented is adapted to minimize the feature dimensionality vectors in prior to passing on to classifiers. In this type of qualitative analysis, it is depicted that these enhancements has achieved on instances evolving domain-specific words.

The Deep-Siamese Bi-LSTM-model is implemented for feeding out the embedded vectors from BERT model and predicted the similarity of the text pairs [33]. Hence to demonstrate this work, the Siamese network architecture application has been presented for investigation to larger scale stylistic author-attribution. The system provides the general authorship notion, and it overtakes the key-similarity based technique on 1-shot N-way evaluation and also performs the well-known author-context. From this study, it is depicted that Convolution-neural networks structure of this Siamese sub-network and applicable for higher author numbers, wherein this LSTM model is infeasible approach for training the model. The non-supervised Cross-lingual STS-Semantic textual-Similarity on the basis of Contextual embeddings vectors obtained from BERT-Bi-directional Encoder-Representation is presented in this study. The main objective of this STS-cross lingual approach is to analyze the two text segment's degree within several languages which has the same meaning. The outcomes of the study illustrated that this non-supervised cross-lingual STS metrics utilizing this BERT-extraction model without the process of fine-tuning attains successful outcomes [22]. The achieved performance upon supervised cross-lingual STS approaches or in supervised weakly approached as well. The larger pre-trained language-models including BERT model is significant enhancement upon NLP-tasks. But also this BERT model trains the model for missing words predictions in next sentence or behind the

masks without the Semantic information knowledge, Lexical data knowledge or syntactic information knowledge. This is achieved through this non-supervise pre-training models [27]. This new method is presented for in injecting explicitly the linguistic-knowledge in embedding of word-forms to BERT pre-trained layer. The enhancements on the performance upon semantic similarity multiple data-sets, implies that the information is useful and for missing out content from original-model while in injecting out the counter-fitted embeddings and dependency-based embeddings. Another work which analysis the similarity among the sentences in Korean language. This analysis uses by integrating the deep-learning technique and another approach which assumes lexical-relationships [43]. In this deep-learning technique, five neural-network in associated with RNN, BERT and CNN were utilized. This technique considers lexical-relationships and also employs the cosine-similarity for vectors embedding by using the model of word-representation. For this type of establishing text similarity, text generation is very essential in the approach. Hence for this purpose, automation metric for evaluation in text-generation is applied referred as BERTSCORE. In this method, BERTSCORE performs the computation of similarity score for every token [44]. This token present in sentences of the candidate with every token present in reference-sentences. The BERTSCORE method, does efficient correlation than the levels of human-judgments and also yields out the stronger perfect selection performance compared to other related metrics performance. In addition to this some of the language representation-model is introduced for promoting the better understanding of language [46]. This representation-model incorporates the contextual semantics explicitly obtained from labeling of pre-trained semantic-roles. This representation model is addressed as SemBERT-Semantic aware-BERT model, that is potential for absorb the contextual-sentences explicitly upon the BERT method. Some of this BERT models utilizes the Siamese network structures, in analyzing this one of the study that presents SBERT-Sentence BERT model, is employed for deriving out the meaningful semantical sentences and the embedding. This SBERT model is the pre-trained modified BERT-network structure. This analysis outcomes were made in comparison with Cosine-similarities [32]. This model minimizes the efforts in detecting out the most predominant sentence pairs obtained in pan of sixty five hours with RoBERT method or BERT method to about time range of five seconds with this efficient SBERT method. This model also retains the efficient accuracy rate from that of BERT model. Most of STS-Semantic Text-similarity systems utilizes either this one-hot model of representation or uses this distributed-representation for modeling out sentence-pairs and this has been considered as the regression complexity. Hence a novel model framework is implemented in some studies for integrates the one-hot model representation and distributed-representation for clinical STS-systems [42]. This model utilizes the gated-network. The experimental analysis is relied upon this benchmark data-set and it depicted that this both representations seems to be the efficient fusion for model representation. Similarly another hybridized framework on the basis of Siamese network is proposed in such studies. This hybrid-network combines the Bi-GRU and G-CNN model to learn out the sentences representations [21]. This model also evaluates the similarity calculation between the sentences. This integrated model assumes the global features within sentences and also takes the sentence local features as well and as a result it provides the higher qualified representations of sentences and also evaluates the similarity between these sentences.

The significant task of ASAG-Automatic short-answers grading-systems is in assigning the ordinal appropriate scores to answers of students. For this analysis, the new architecture of neural-networks is implemented for it integrates the Siamese bi-LSTM pooling-layer as well [14]. This mechanism is relied on the basis of Sinkhorn

distance evaluation among the state sequences of LSTM and between the output layers of support-vectors. The learning of similarity between the Chinese sentences is employed by using Siamese architecture of CNN model. This technique utilizes the two type of similarity evaluation metrics such as Manhattan similarity and Co-sine similarity metric [34].The outcomes of the experiments depicted the higher accuracy rate of CNN-Siamese model than other models. In Similar to this, another similarity evaluation method is established as MSE-based knowledge referred as multi-granular semantic-embedding model on the basis of Knowledge-enhancement [30].This model resolves the sentence similarity and the association between the semantic matching of long-text. The accurate rate in representation of documents by categorizing the texts in documents is accomplished by implementing Siamese LSTM basis framework model [35].This enables the learning of accurate representation of documents. This model leverages the sub-network of Siamese LSTM-model to measure out the semantic-distances among the two various documents. The Deep-learning method integrates the Bi-LSTM architecture to attain the greater accuracy in semantic-similarity in question pairs section for CCKS2018-QIM tasks [47]. This detection is relied on the basis of Siamese-network structure. The table similarity is also analyzed by presenting TabSim method and it utilizes the Siamese neural-network. This network achieves the greater range of recall value, higher precisionF1-score values and accuracy rates upon these three various corpora of table-pairs [8].In Similar to this, the matched sentences and non-matched sentence pairs obtain from the sentence-pairs were extracted by using Hungarian-algorithm [41].The Hungarian layer is designed in these embedded vectors. The dynamic-computation graph is described for the optimization of Hungarian-layer, and it embeds other algorithms as well to neural-network. This Semantic textual-similarity is performed by utilizing by LIPN-team. This framework utilizes the support-vector regression model integrating distinct measures of text-similarity [2]. This type of semantic textual-similarity techniques is applied for identification of paraphrases in Arabic-languages on the basis of various combinations of NLP-techniques. One of such technique TF-IDF method for enhancing the words identification is used. This method employs word-2cvec algorithm and it minimizes the complexity of computation and also optimizes the word predictions probability within the context [23].The distributed representation of vectors are also utilized along with this algorithm. Likewise, UNITOR ,a semantic text-similarity model which facilitates the Sem-Evaluation [3]. The task involved in this model and depicted as the SV-regression model complication, wherein this scoring function of this text similarity among the pairs of texts is attained from provided examples.

The Text-similarity implemented in cloud- platform requires a dynamic algorithm for task scheduling to process STS-Semantic textual similarity application. This would minimizes the makesspan time, by increasing the cloud-resources utilisation and satisfying the parameters usage [29].Such resource allocation-model for application processing is accomplished through PSO-Particle Swarm Optimization algorithm defines as PSO-COGENT, for optimizing the cost of execution and computational time of text similarity detection [15, 16]. The Scheduling algorithm ought to optimizes the performance of parameters indicators such as makespan time, energy-consumption, resource utilisation, reliability etc. [17], In another platform as in fog-computing ,utilised for real-time application processing neared to data-sources and enhance the scalability, latency work and aids in resisting the congestion of data load [18].

2.1 Research gaps

The STS-Semantic Text-similarity is also utilised in clinical domain handling the clinical records, such that method utilised for eliminating the redundant data to improvise the clinical decision-making process. However, the size of those clinical semantic-text similarity resources handled by those STS is relatively small, since the data were developed though clinical-notes obtained from single institutional data [40]. The annotation schema uses conventional guidelines of STS, having limited clinical properties. Another limitation is that the dataset used were manually annotated by clinical experts. In some cases, SemEval STS shared tasks are employed through crowd-sourcing on Amazon Mechanical-Turk that is not suitable for such sensitive patient dataset. The extracted features of input text data were subjected to plagiarism that leads to further complexity in some cases. In case of unrealistic plagiarism data-set, PAN corpus has its own drawbacks, such that it brings out larger plagiarism cases. The evaluation outcomes were performed on PAN-Plagiarism [37]. In such cases, the test-beds requirements ought to be emphasized, with more manual and realistic plagiarism cases. BERT Network involved in feature extraction process undergoes major limitation, is that it does not consist of any independent computation of sentence embeddings. The scenario makes the model a difficult task to derive out the Sentence-embedding obtained from BERT model. The researchers ought to pass out single-sentences by using BERT framework, to derive fixed-size embedded vector through averaging the values of outputs [12]. The approach also addresses out the long-term dependency issues and data losses that impacts the existing research models specifically if the size of input data is higher. Since the model in such research, requires more training time and more trained data, out of the baselines [32].

3 Proposed framework

The proposed-framework is employed for efficient identification of Semantic text-similarity by using Weighted Fine-tuned BERT model with Deep Siamese Bi-LSTM model. The Data is obtained from Quora dataset, consisting of various question-pairs sentences. The sentences from this question pairs were taken as the input data and it is loaded. In pre-processing phases, the questions sentences were reduced by removing of Non-ASCII characters, punctuations, and special characters by using python.

In Tokenization process, each token is assigned to features of sentences. This token is utilized as the symbol for features aggregation from single sentence or in the sentence pairs. After the pre-processing phase, the glove-based word embeddings converts the words of input to vectors. For achieving this, the vocabulary pre-processor is employed, obtained from tensorflow. The lookup functions of tensorflow embedding are utilized for yielding the word-embedding. In Pre-processing phase, vectorization process is performed after Tokenization process. In vectorization process, phrases or the words were changed to corresponding vector-form, consisting of real numbers. The vectors are mapped out in with respective words in pre-trained BERT model. These vectors facilitate to predict the word semantics in question-pairs. The method is also implemented for semantic-text similarity detection between different pdf documents as well. These embedded vectors were trained using Siamese Bi-LSTM network model. For the calculation of Semantic text-similarity in question-pairs, to ensure if the question belongs to label 0 or label 1 turns out to be problem for classification. Hence the hybrid

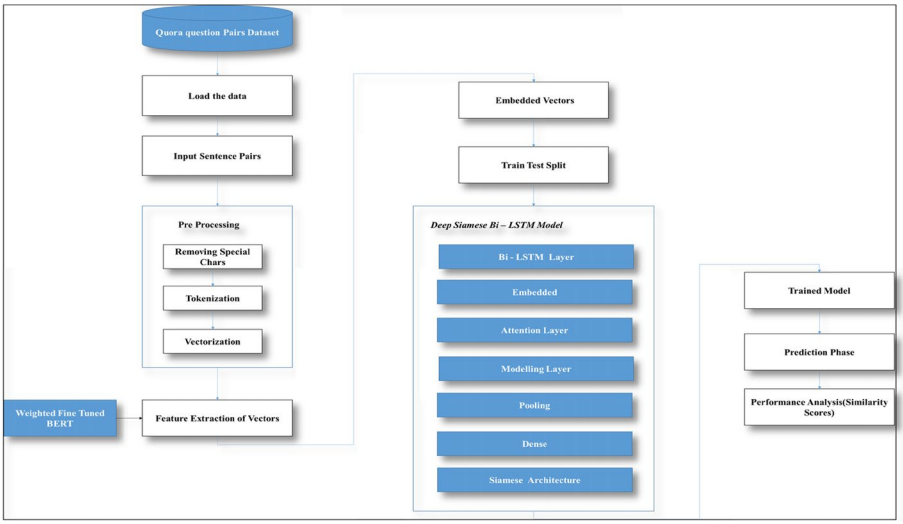

Fig. 1 Overall flow of proposed framework

approach of Siamese network in Bi-LSTM model is implemented. The overall proposed-framework in determining text-similarity is represented in Fig. 1 above. The Siamese network, defined as neural-network with the two similar question pairs or similar documents, evaluates the text-similarity by using shared weight sub-networks. Glove Embedding comprises of 100d vectors of Wikipedia. In Bi-LSTM layer, dependency of the non-continuous and longer distance between the sentence words has been extracted. The two layers of this Bi-LSTM structure consists of encoding of embedded-vector inputs of the words. In attention layer, attention weights were calculated. The key-parts of first-layer's output were taken as inputs for next proceeding layer. The embedded layer outcomes are represented as the weighted sum to vectors, defined as WFT BERT Weights for every sentence. The outcomes are then fed to input for non-linear layers. The distance and angle-information of output alignment of the question pair is obtained using Deep Siamese Bi-LSTM layer. The Similarity score were calculated in each layer for every sentence, with WFH BERT weights output value. Through this similarity score, Efficient Text similarity for input query are evaluated. The trained model outputs, with their vectors predicts the text-similarity of the documents or the question pairs.

3.1 Weighted Fine-Tuned BERT model for feature extraction

The above Fig. 2 represents the framework of Weighted Fine-Tuned BERT Extraction process used in the study. The Text sequence is denoted as $A = \{a_1, \cdots, a_L\}$ where $a_l (1 \leq l \leq L)$ points to sentence and L represents text sequence length. Bidirectional pre-trained BERT model encoded the text sequence $A = \{a_1, \cdots, a_L\}$ to fixed-length sentence vector forms h, represented as input source-element. The vector forms of sentence were represented as stated below:

$$s_l = \text{BERTsent}(a_l) \tag{1}$$

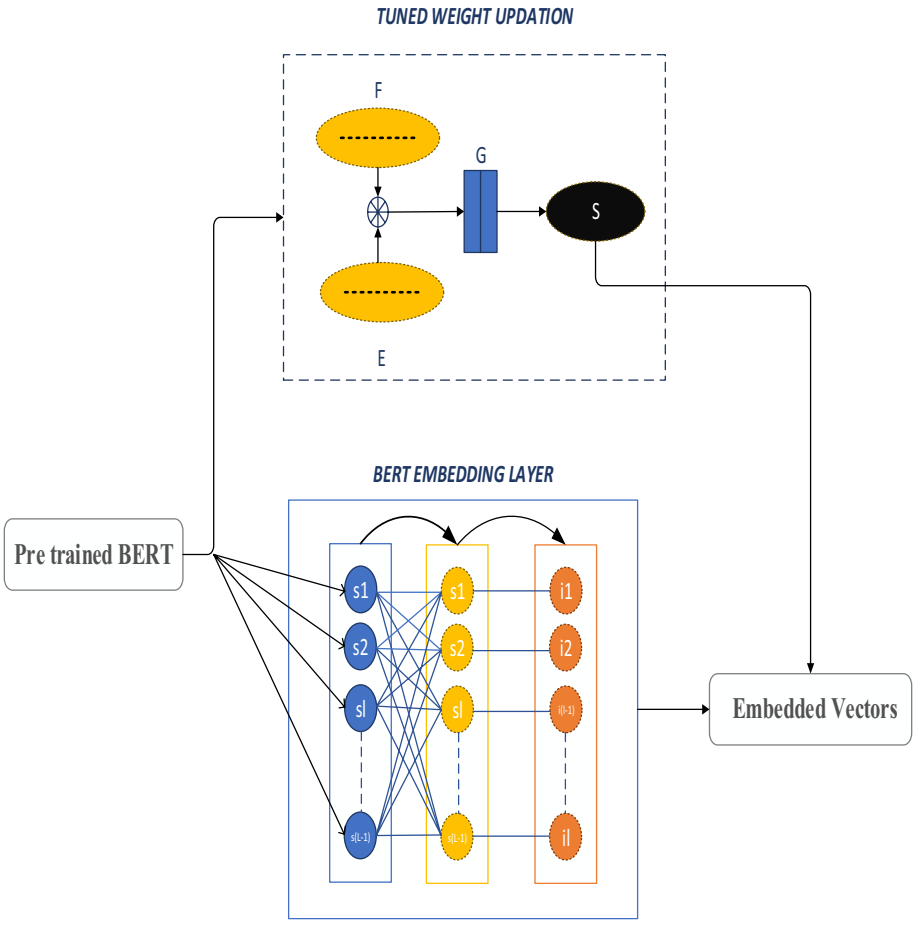

Fig. 2 Weighted Fine-Tuned BERT Extraction Model

Such that $S = \{s_1, \cdots, s_L\}$ are source-elements as depicted in the figure. BERT sent(.) illustrated the encoding sentences into those sentence vectors. The hidden Vector-representation $u_l$of converted sentence Vector-form s_l were obtained by using MLP-Multi-Layer Perception.

$$u_l = \tanh(W_1 s_l + b_1) \tag{2}$$

The Hidden Vector-representation is stated in the above equation wherein b_1 and W_1 denotes to bias and Weight parameter. The general used Text-representation techniques usually neglects the interaction information among all text sentences. This leads to partial Semantics losses. In the study, the entire source-elements were considered as context-information, to get text representation, consisting of more semantics. For instance, interaction information(i_k) between all the source elements$\{s_1, s_2, \cdots, s_L\}$ and one source element (h_k) are captured. The semantics-weight of a_l allocated by a source element a_k represented asα_{kl}.

$$\alpha_{kl} = \frac{\exp\left(ul^{T}u_{k}\right)}{\sum_{l=1}^{L}\exp\left(u_{l}^{T}u_{k}\right)} \tag{3}$$

The i_k were formulated as

$$ik = \sum_{l-1}^{L}\alpha_{kl}S_{1} \tag{4}$$

Similarly to above instance, every single source-element does interaction with entire sources, and obtains the interaction among all source-elements and single source-element shown in Fig. 1 above.

$$I = \left(i_{1}, i_{2}, \cdots, i_{L}\right) \tag{5}$$

The interaction contributed to finalized unequal text-representation, and attention layer is added for getting informative-interaction, similar to classification in Fig. 1. ' s' represents the compatibility score, to weight I. I denotes the interaction-representation. While during process of joint embedding of words, the whole text compatibility score were generated. As a result the final text is represented as below.

$$T = sI \tag{6}$$

Every sentence is represented as $a_l = \{wrd_1, wrd_2, \cdots, wrd_n\}$, such that wrd_1 denotes each word within a sentence. Pre-trained BERT model encodes every sentence$a_l = \{wrd_1, wrd_2, \cdots, wrd_n\}$ to their respective word embedding-forms, $E_l = \{E_1, \cdots, E_n\}$, This is stated as below

$$V_l = \text{BERTtoken}\left(a_l\right) \tag{7}$$

In the notation, $\text{BERTtoken}\left(a_l\right)$ represents word encoding to their word vectors. The word-embedding representation of entire text sequence $A = \{a_1, \cdots, a_L\}$ were represented as$E = \{E_1, E_2, \cdots, E_l\} = \{\{e_1, \cdots, e_n\}, \{e_1, \cdots, e_n\}, \cdots, \{e_1, \cdots, e_n\}\}$, such that n describes the total count of words. Additionally, b points to corresponding text sequence label A, that could be encoded to their label embedding forms $F = \{f_1, f_2, \cdots, f_K\}$ determined through BERT, wherein K represents count of classes. $F = \{f_1, f_2, \cdots, f_K\}$ would be represented as

$$f_k = \text{BERTtoken(b)} \tag{8}$$

The words and labels were embedded to one joint-space. It seems quite tedious to learn within Joint-space. The simple approach for calculating compatibility among pairs of label words is the cosine similarity. In the below notation, G denotes the compatibility among the pairs of label words, which is stated below.

$$G = \left(F^{T}E\right) \oslash \hat{G} \tag{9}$$

The element-wise division, participated in vectors operation or matrix operation is denoted by operator $\oslash$. The normalized matrix were represented by $\hat{G}$, consisting of $K \times L$ size. Every normalized matrix element is determined as $\hat{g} = \| ck \| \| el \|$, such that $\| \cdot \|$ denotes the norms 2. e_l and $f_k$represents l^{th} word embedding and k^{th} label embedding. The relative spatial-information between consecutive

words, were calculated using non-linear function, during obtaining the label word-pairs compatibility. For example, text-sequence were centered at position q, such that the length of text-sequence is 2i + 1.In provided pairs of label-phrase, the label to-token compatibility is represented as $G_{q-i:q+i}$.

In below equation, e_q describes the high level compatibility stigmatization among whole lables and q^{th} phrase. The function is evaluated as below.

$$e_q = \mathrm{ReLU}\left(G_{q-i:q+i} WRD_2 + b_2\right) \tag{10}$$

Such that b_2 denotes bias and WRD_2 weight.The maxpooling operation obtains the highest compatibility value among q-th phrase in accordance with whole labels. The operation is represented as below.

$$m_q = \max\left(e_q\right) \tag{11}$$

The entire text sequence compatibility score are represented as stated below.

$$s = \mathrm{SoftMax}(m) \tag{12}$$

In the notation, m and L denotes the vector and length. S_q represents q^{th} Softmax element.

$$S_q = \frac{\exp\left(m_q\right)}{\sum_{q^{1}=1}^{L} \exp\left(m_{\acute{q}}\right)} \tag{13}$$

Figure 1 illustrates the Dual-label embedding process. The compatibility-score(s) of whole text-sequence were evaluated by using words learning embedding and labels embedding. The compatibility score is used for capturing high interactive-information, and to weight finalized interactive text representation $I = \left(i_1, i_2, \cdots, i_L\right)$. Further to this, labels could learn more count of textual-content, the classifier could effectively leverages those weighted labels, for text classification. And also the compatibility score(s) were employed, to weight finalized label-vectors F_k.

$$T = \sum q s_Q i_q \tag{14}$$

$$\acute{F} = \sum q s_Q F_k \tag{15}$$

T denotes final representation of text and the final label-representation represented as $\acute{F}$ (Table 1).

At the first stage, Feature extraction process is represented in the steps included in the above algorithm 1. The features set of the question from the Quora website where been extracted from the question section or input statement. Similarly the similarity-scores of the different statements has also been extracted in the analysis phase. The proposed-framework does the extraction of features subset from the questions. The dataset provided is $(S,Y) = \left\{\left(S_1, y_1\right)\ldots\left(S_m, y_m\right)\right\}$. The trained classification -model represents C: S◇Y. The soft-label setting is assumed such that attacker could query out the classifier, for obtaining output-probabilities, upon provided input. The model parameters, training data is not provided for access. For example, weight example is denoted by S_weight. The weighted instance S_weight need to get generated for provided input-pair, such the condition

Table 1 Weighted Fine-Tuned BERT feature extraction

Algorithm-1: Weight Fine Tuned BERT Feature-Extraction
1. **Input:** Sentence $= [a_1, \ldots, a_n]$, ground truth label y
2. **Output:** S_weight Embedded Matrix for the weight updation in Deep Siamese Bi – LSTM Model
3. Initialization: S_weight ← A
4. Compute token importance $I_i \forall a \in S$
5. for i in descending order of I_i do
6. $S_M \leftarrow$ S_weight$_{[1:i-1]}$ [M] S_weight$_{[i+1:n]}$
7. Predict top − K tokens A for sim Sim $\in S_{sim}$
8. A ← FILTER$_{(T)}$ * $\acute{F}$
9. L = {} // python − style dict
10. for t ∈ A do
11. L[a] = S_weight$_{[1:i-1]}$[a]S_weight$_{[i+1:n]}$
12. End
13. if ∃ a ∈ A s.a (L[a]) ≠ y then
14. **Return**: S_weight ← L[a`] where (L[a`]) ≠ y
15. L[a`] has maximum similarity with S
16. else
17. S_weight ← L[a` ] where L[a`] causes maximum
18. reduction in probability of y in (L[a`])
19. end if
20. end
21. Return: S_weight ← None

$C(S_{weight}) \neq y$ ought to be satisfied. Simultaneously, S_weight ought to be grammatically corrected with semantically similar as S. In order to generate, Weighted example S_weight, two categories of token level perturbations such as token replacement and new token insertion is introduced.(i) Token replacement with condition a ϵ S with other token. (ii) New token insertion a} within S.Few token within the input, contributed high for Final Prediction through C in compared to others. The Token replacement or new token insertion could have stronger impacts in modifying the classifier's predictions. The token-importance I_i were estimated for every 'a' through deleting 'a' from S and also through computation of probability decrease in prediction of correct label(y). The insertion and replacement operations, were performed on token (a), through determining the similarity, as a result similarity token insertion occurs. The pre-trained BERT-model utilized for prediction of similarity token. The similarity tokens would fits well to text context and grammar of text. If there is occurrence of multiple-tokens causes C for S misclassification, during token replacement, token is chosen that makes S_weight more similar to Original S on the basis of similarity scores. If misclassification does not occur, another token is chosen that reduces probability of prediction. The token perturbations are applied iteratively unless either C(S_weight) ≠ S or entire S tokens are perturbed.

3.2 Siamese Bi-LSTM-model for prediction

3.2.1 Long-Short-Term Memory (LSTM)

This Siamese LSTM-Model is described in the study, that it explored for determining the semantic-relatedness between the documents. The document pair denoted as Dc_i and $Dc_j$is the input of the model. The variables wt_t^i (wt_t^j)and em_t^i (em_t^j) represents the t-th word for Dc_i (Dc_j) document and em_t^i represents the word-embedding. The LSTM model is treated as the encoder. In the document, the words within the documents are represented as (wt_1^i….wt_t^i ….wt_{Ti}^i). The T denoted as subscript represents the total count of words within the document. The document may consist of arbitrary words count. At every step of this model, The LSTM-framework generates the hidden-state hi_t^i which interpreted as sequence representation. (wt_1^i….wt_t^i ….wt_{Ti}^i). The LSTM-encoder transforms all words within document to distributed vector. The following equations are represented for document encoding. The following Eqs. (16), (17), (18), (19), (20), (21) states the memory-state detections and forget determinations obtained through sigmoid functions in Eqs. (16) and (17)

$$i_t = \text{sigmoid}\,(wt_i em_t + U_i hi_{t-1} + c_i) \tag{16}$$

$$fr_t = \text{sigmoid}\,(wt_f em_t + U_f hi_{t-1} + c_f) \tag{17}$$

$$\widetilde{cs}_t = \tanh(wt_{cs} em_t + U_{cs} hi_{t-1} + c_c) \tag{18}$$

$$cs_t = i_t \odot \widetilde{cs_t} + fi_t \odot cs_{t-1} \tag{19}$$

$$ou_t = \text{sigmoid}\,(wt_o em_t + U_o hi_{t-1} + c_o) \tag{20}$$

$$hi_t = ou_t \odot \tanh \odot cs_t \tag{21}$$

The above equation determines the memory-state and forget-state by using this LSTM-model .The fi_t denotes the forget-state of the model [11].

In this equation, the variable cs_trepresents the memory-state and the forget state is represented as fi_t. ou_tDenotes the output-gate and $i_t$represents the input-gate. Θ represents the component-wise-product. The weights matrix for hidden-layer ht and embeddings are defined as U and M. These representation are noticed in Eqs. (18) and (19).

3.2.2 Siamese-Networks

The LSTM-network model hidden vector is defined as the vector-representation for every document denoted in equations in (1) to (6). The weights of the network in two Siamese-networks were shared. The Euclidian-distance depicted as the negative similarity-function and it expresses the relatedness degree among the documents pair upon network top-layer. The Euclidian distance among the vectors of the documents were taken as the finalized scores of textual-semantic distances. Let Z is considered as the binary-label allocated to pair of documents. Z=0 meant that the documents pairs were similarly deemed. If the value of Z is 1, then this indicates that pairs are negative, such that the pairs are dissimilarly

deemed. This Siamese-LSTM network model thus estimates the semantic-relatedness for document pairs with the aid of Euclidean distance-measure. In this Siamese LSTM, the model is trained utilizing the BPTT-back propagation through-time algorithm on the basis of contrastive-loss function. This network architecture compels the sub-networks of LSTM for capturing out the differences of textual semantics in the training phases. After the training phases is completed, the strategy is applied for mapping out the document to semantic-vector of fixed type sizes through one of two LSTM trained sub-networks. Then the resulted semantic-vectors is fed in 3layers DNN-deep neural-network. These vectors estimate the probability distributions as well across various classes.

3.2.3 Architecture of Siamese network based Bi-LSTM model

The efficient capability in sequence modelling were proved through LSTM model specifically long sequences modelling with inherent sequential non-linear patterns.

The typical LSTM model would be formulated as stated below.

$$fg_t = \sigma(W_f.[h_{t-1}, x_t] + b_{fg}) \tag{22}$$

$$a_t = \sigma(W_a.[h_{t-1}, x_t] + b_a) \tag{23}$$

$$S'_t = \tanh(W_s.[h_{t-1}, x_t] + b_s) \tag{24}$$

$$S_t = fg_t * S_{t-1} + a_t * S'_t \tag{25}$$

$$b_t = \sigma(W_b[h_{t-1}, x_t] + b_b) \tag{26}$$

$$h_t = b_t * \tanh(S_t) \tag{27}$$

Such that b_t, fg_t and a_t represents output, forget-gate and input activation-vectors at specific time-step (t). $h_t$and x_tDenotes the output vector and input-vector of LSTM unit. The LSTM cell-state is stored in S_t. Further W and b that is subscribed through fg,a,s described the parameter-matrices, which is learnt in training process. The dot product-operation is represented by. and the element-wise multiplication-operation is represented by *.

Even though, the LSTM model does has the memory technique comprising of gating-function, to protect longer term information, the recurrent-structure, faces still the issues of long dependencies learning. This is due to the long-path backward signals and forward signals, which ought to be traversed within network. Particularly, the trajectories head information ought to traverse total recurrent network length, in order to attain embedding output vector. The Bi-directional structure is applied to reduce the traversing path of information. (depicted in Fig. 3).

$$t'r = f(tr) \tag{28}$$

$$\vec{h} = \overrightarrow{LSTM}(t'r) \tag{29}$$

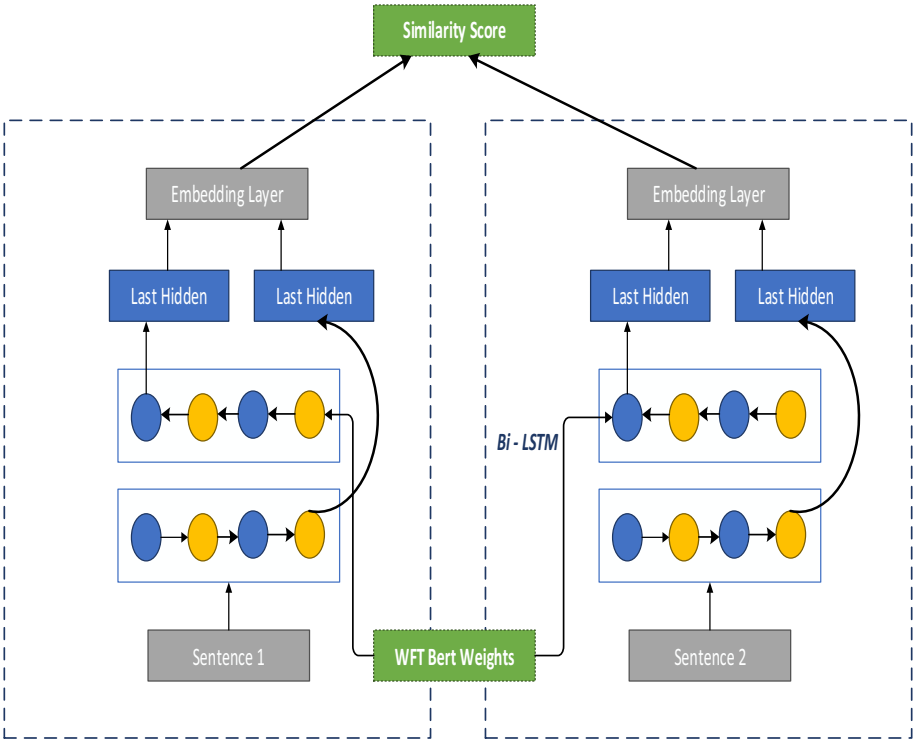

Fig. 3 Siamese network with Bi-LSTM based trajectory encoder

$$\overleftarrow{h} = \overleftarrow{\text{LSTM}}(t'r) \tag{30}$$

$$w = g([\vec{h}_T, \overleftarrow{h}_1]) \tag{31}$$

In the above equations, the document-trajectory ($t'r$) or query trajectory is determined. fg represents the linear time-distributed transformation, consisting of tanh activation method. The function does encoding of geo-coordinates. The encoded-trajectory is denoted by $t'r$. The linear-transformation that provides the embedding vector is g. The embedding vector points to w. The two BiLSTM directions consists of $\overleftarrow{\text{LSTM}}$ and $\overrightarrow{\text{LSTM}}$. The directions are reverses factors of input-trajectory .The outcomes of $\overleftarrow{h}$ and$\vec{h}$ are obtained (Table 2).

In this proposed-framework, the Siamese network approach is integrated with Bi-LSTM model for generating the trained model for similarity evaluation in this algorithm 2. The Siamese-network is trained first on the embedded vectors or the labeled instances by utilizing the triplet loss function. The technique is utilized for the predictions of labels of the sentence pairs for those non-labeled instances. Proceeding to this, the embedding vectors of labeled instances and non-labeled examples were pass on to LLGC-Local learning with Global-consistency. After few iterations count, an appropriate unlabelled examples percentage are selected on the basis of LLGC scores and it is added with labeled-instances for next following iteration. The model is used for classification of question-answer pair sentences obtained from Quora data-set. In this first, the training of the labels were performed

Table 2 Siamese Bi-LSTM model

Algorithm-2 : Siamese Network with Bi-LSTM Model
1. **Input:** Labeled-examples as (aL, bL), number of iterations j and selection percentage c 2. **Output :** w – Similarity Score 3. **Given Parameters:** $A1_1$,$A1_2$,$A2_1$,$A2_2$,$A3_1$,$A3_2$ 4. **Initialize**: e_0,$f_0 = 0$ 5. **for** 1 to j do a. train_siamese_neural_network(aL, b b. $implant_L$ = siamese_neural_network (aL) c. $labels_{e0}$,$dist_{f0}$ = $Bi - LSTM($ $implant_L$, aL) d. Calculate the $A1_1$,$A1_2$,$A1_3$; e. Update the cell state of $A2_1$,$A2_2$, $A2_3$ with $labels_{e0}$ $A3_1$,$A3_2$ f. Calculate theof $A1_1$,$A1_2$ g. (aL, bL) = concat((aL, bL), ($A3_1$, $A3_2$)) **6. end for** **7. Output** $w_0 = [w_1 ... w_j]$

by Siamese-BiLSTM model. This framework, Siamese training-phase training phase algorithms executed for twenty-five iterations. For this Siamese- Bi-LSTM integrated algorithm-2, the small labeled subset instances or examples were selected in accordance with semi-supervised learning-practice, acquiring balanced count of examples for every class. The rest of the labels were assumed as non-labeled ones. The outcomes of the algorithm-2 were determined in three random-executions by utilizing random-initialization of the parameters ($A1_1$, $A1_2$,$A2_1$,$A2_2$,$A3_1$,$A3_2$) of Siamese-network in every run, given the initially labeled-examples were selected randomly. The parameters of the trained model is then concatenated with the provided-parameters in the final step.

4 Results and discussions

4.1 Dataset description

The Quora Question pair's dataset has been used for analysis phase. This dataset nearly consists of 400 thousand of question-pairs that are labeled as non-duplicates and duplicates labels. In the dataset, the vocabulary consists of around 85,000 of words that mapped to unique identification numbers and 100d glove-embedding were utilized for vector conversion of those sentence features. The experimental analysis is employed utilizing the

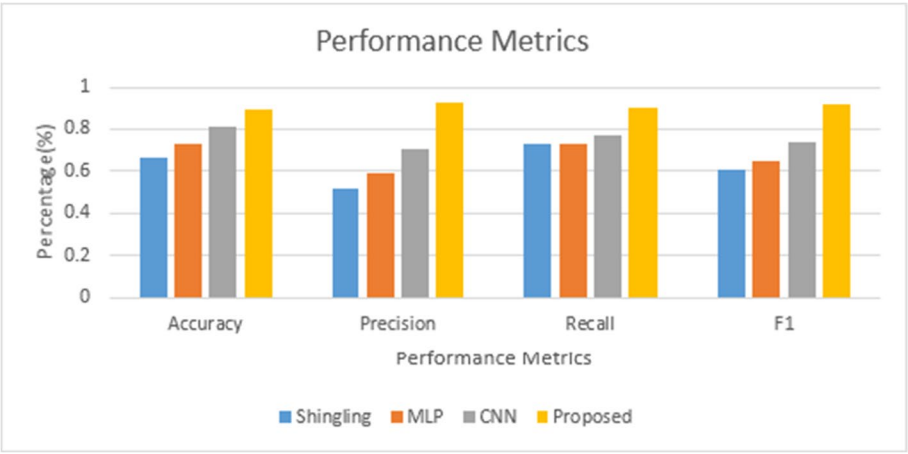

Fig. 4 Performance Comparison evaluation

Quora-dataset for thirty epochs and utilizes the batch-size of sixteen. The training-set consists of around 361,745 question-pairs and the testing phase question-pairs consist of around 36,174 count. The accuracy rate of the model is enhanced by increasing the training steps count. The gain in the performance is attained across the simpler multi-layer perceptron method by using this Siamese Network-model architecture. The enhanced increase to 14% performance gain is achieved over shingling-technique. Further more than 80% rate of accuracy is attained by using this Siamese CNN-architecture model. For implementing the non-linearity, the sigmoid or ReLU activation-function is utilized and the loss in the features can be evaluated by Adam-optimizer. The functions are implemented by using Tensorflow-GPU. The testing phase is also carried out by using PDF files as inputs and calculations of similarity scores for those pdf documents. The single pdf files is made in comparison with other files at each iteration.

4.2 Performance-evaluation results

The assessment in the performance analysis of the proposed-framework is established by making the comparisons of various performance parameters such as accuracy rate, precision value in detecting the text-similarity scores, Recall values and the F1 score evaluation. These metrics value were determined for existing deep-learning methods such as Shingling-method, CNN technique and MLP technique with proposed-framework.

The above Fig. 4 enumerates the performance comparisons of different techniques of text-similarity detection methods. From the graphical representation, it is well depicted that the proposed-framework exhibited higher range of accuracy rate in predictions, recall values, F1-Score values and also in precision factors than the other existing techniques in Duplicate detection of text-similarity between the documents or in question pairs in this Fig. 4.

The representation of the performance of various semantic text-similarity detection methods were illustrated in the Table 3 above. The proposed-framework have 0.90% of accuracy rate value, 0.9 precision rate values, 0.9% of recall data and 0.92% of F1-score

Table 3 Analysis of proposed and existing system with respect to various parameters

Methods	Accuracy	Precision	Recall	F1
Shingling	0.6657	0.5151	0.7297	0.6039
MLP	0.7263	0.5878	0.7245	0.649
CNN	0.8044	0.7003	0.7688	0.733
Proposed	0.9	0.93	0.91	0.92

values [1]. These values are higher than the equivalent metrics values of shingling technique [13], CNN-technique [45] and MLP method [31]. This Table 3 above shows that proposed framework evolving Siamese Bi-LSTM model networks with BERT extraction is efficient in text similarity prediction than other existing methods.

Similarly, specific evaluation metrics such as recall factors, F1-score values and the precision rate of the proposed-framework is made compared with other algorithms such logistic-regression model, decision-tree algorithm, KNN-algorithm, RF-Random-forest algorithm, and SVM technique is represented in Fig. 5 above. The validation of the model provided the output that this framework outperforms the other algorithms in identifying the Text-similarity within the sentences and evaluating the similarity-scores [38]. from the Fig. 5 above it is found that the metrics values in like the precision in determining the similarity scores, recall values and f1-score values is higher for proposed-framework in compared to the other algorithms.

The Table 4 above demonstrates that the proposed-framework of Deep Siamese Bi-LSTM model found to generate higher rate as 0.9% of precision rate, 0.910 of Recall value, with 90% of accuracy rate and 0.92% of F1-score values in text-similarity and duplicate detection through this BERT with Deep Siamese Bi-LSTM network model. These rate values are higher than the other algorithms. This evaluation analysis depicted that proposed-framework exhibited higher efficiency rate in identification of text-similarity and in predictions of similarity scores.

The above Fig. 6 illustrates the graphical representation of accuracy rate evaluations of proposed-framework with other algorithms as well. This framework, exhibited higher accuracy detection rate in determining the semantic-text similarity within the questions pairs originated from quora dataset or between different pdf files. This Siamese network model process each sentence and attained accuracy rate detections percentage in sentence similarity.

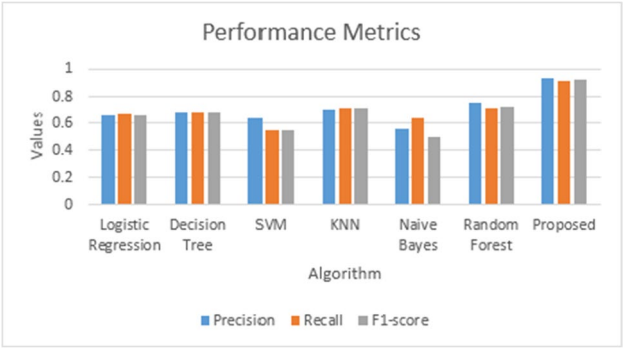

Fig. 5 Comparison evaluation of proposed framework with existing algorithms

Table 4 Comparison analysis of proposed framework with different algorithms of various parameters

Methods	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.66	0.67	0.66	67.1
Decision Tree	0.69	0.69	0.69	69.3
SVM	0.64	0.55	0.55	60
KNN	0.71	0.72	0.72	71.9
Naive Bayes	0.56	0.64	0.5	63.7
Random Forest	0.76	0.72	0.73	72.3
Proposed	0.93	0.91	0.92	90

The accuracy rate values were incorporated in a single Table 5 and made in comparison with the other existing algorithms, involved in duplicate-detections of sentences within question pairs or the duplicates of the sentences stating the same questions. Among the rate of accurate detection the Bi-LSTM Siamese model overtakes the performance of other algorithms. This attains the efficient detection rate of 90%. This fetches the duplicate sentence pairs within the different type of questions obtained from the larger number of quora question-pair samples.

The present proposed-framework model is employed in Quora Dataset where the portal consists of various set of question pair sentences. The similarity between these semantic-texts are analyzed and predictions of the similarity scores were calculates which depicts the performance level of the model. The above Fig. 7 represents the implementation of the model relying in another MRPC-Microsoft Paraphrase corpus dataset [26], and the performance in this dataset is evaluated. This figure demonstrates this assessment. This MRPC-dataset comprises of sentence-pairs that is extracted automatically through online sources of news with along human-annotations, to ensure if the sentence are equivalent semantically.

The accuracy rate of existing algorithms were compared with similarity accuracy rate proposed-framework from quora dataset. From the Fig. 8, it is clearly seen that Weighted Fine-tuned BERT LSTM model exhibits higher accuracy rate of 90%.

The above Table 6 illustrates the accuracy rate value representation and F1 score comparison of proposed model with other methods such as Multi-perspective LSTM, Siamese LSTM model, L.D.C, ESIM, DINN, Enhanced RCNN, Siamese CNN model and multi-perspective CNN model. The proposed model to determine text-similarity

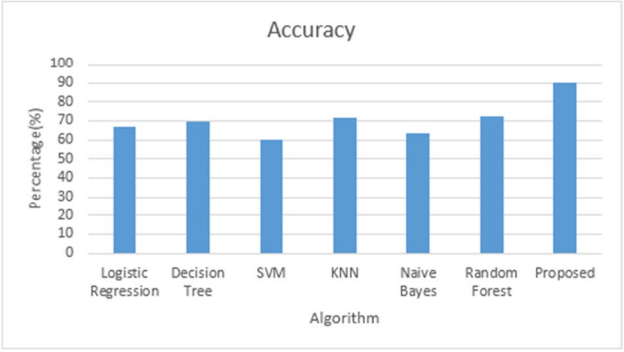

Fig. 6 Performance analysis of proposed-framework in terms of accuracy parameter

Table 5 Comparison analysis of proposed system with respect to accuracy rate

Methods	Accuracy
Logistic Regression	67.1
Decision Tree	69.3
SVM	60
KNN	71.9
Naïve-Bayes	63.7
Random Forest	72.3
Proposed	90

using Weighted Fine-Tuned BERT extraction with Bi-LSTM method, determined the

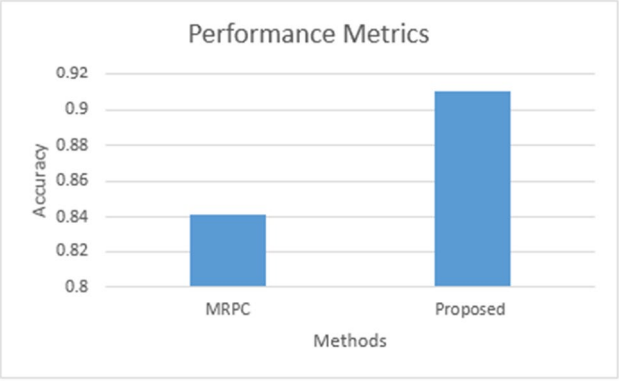

Fig. 7 Accuracy rate evaluation of proposed framework upon MRPC dataset [26]

text-similarity measure in 90% accuracy and F1 score with 92% [29] (Fig. 9).

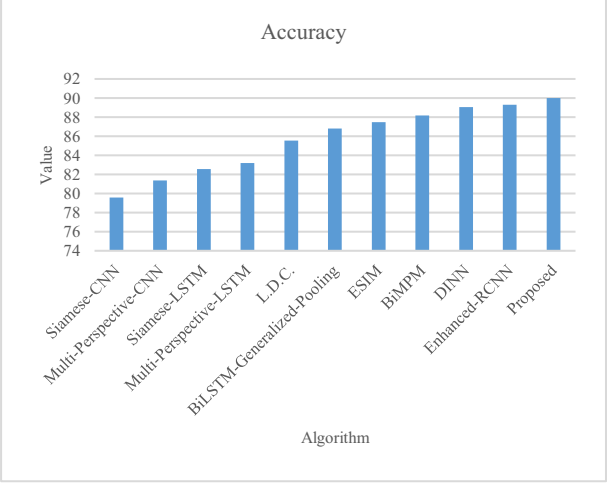

Fig. 8 Comparison of proposed framework with existing algorithms [29]

Table 6 Performance evaluation of proposed framework

Model	Accuracy	F1 -Score
Siamese-CNN	79.6	Nil
Multi-Perspective-CNN	81.38	Nil
Siamese-LSTM	82.58	Nil
Multi-Perspective-LSTM	83.21	Nil
L.D.C.	85.55	85.23
BiLSTM-Generalized-Pooling	86.82	85.9
ESIM	87.5	87.44
BiMPM	88.17	87.96
DINN	89.06	89.01
Enhanced-RCNN	89.3	89.47
Proposed	90	92

Similarly, the accuracy value of proposed Bi-LSTM model were made compared with other methods by taking quora question pairs dataset. The accuracy of the framework is higher for proposed model, compared to other methods such as BERT-Base, Enhanced RCNN algorithm and Enhanced RCNN model.

The Table 7 above enumerates the accuracy rate comparisons and F1-score evaluation of various methods [29], compared with proposed model. The results of the graphical representation of proposed Weighted Fine-Tuned BERT extraction with Bi-LSTM model exhibited higher accuracy rate of 90.89% and F1-Score value of 92.64%.

The F1 score evaluation of proposed model is analyzed with F1-score values of existing methods. The result outcomes in Fig. 10 depicted that F1-score of proposed framework, seems to be higher than other algorithms. The outcomes proved the efficiency of propose text-similarity method [29] (Table 8).

The above Table 8 denotes the accuracy rate value representations of model relying in MRPC dataset and the accuracy rate comparisons of proposed-framework upon quora

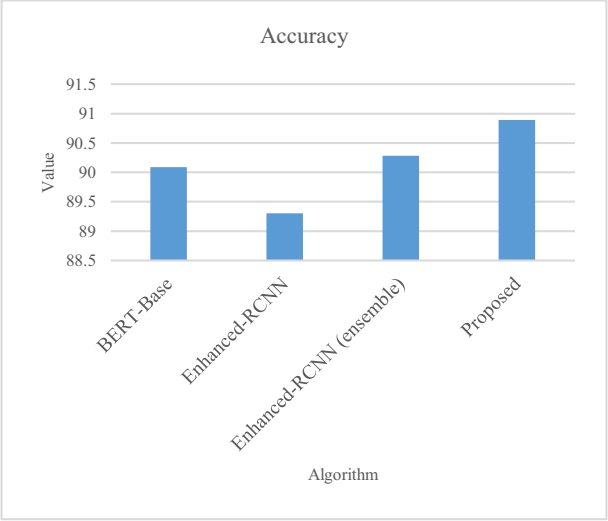

Fig. 9 Comparison evaluation of accuracy rate of proposed model algorithm [29]

Table 7 Performance evaluation with existing method

Model	Accuracy	F1 -Score
BERT-Base	90.09	90.2
Enhanced-RCNN	89.3	89.47
Enhanced-RCNN (ensemble)	90.28	90.35
Proposed	90.89	92.64

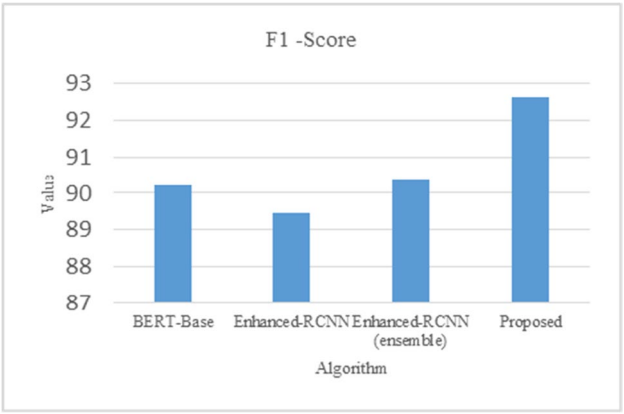

Fig. 10 F1-Score evaluation of proposed framework

dataset. From the results analysis, it is found that proposed-framework shown higher rate of accurate degree predictions in determining the semantic-text similarity by using embedded vectors representation. The outcomes of the model exhibited 0.91% of accuracy rate, higher than framework model upon MRPC dataset. This accuracy rate upon MRPC dataset found to 0.84% lesser than Siamese-Bi-LSTM network model (Table 8).

The proposed framework relying in quora dataset is made compared with the accuracy rate of other pre-trained unsupervised BERT model of Deep-Bidirectional transformers in this Fig. 11 above. This comparisons evaluated that the proposed-model exhibits higher rate of accuracy ranges in determining the Text-similarity and performing the NLP-tasks than this BERT pre-trained model [5] (Table 9).

The features contributing to the Semantic text-similarity is extracted using BERT model upon MRPC dataset, and this accurate prediction detections in this text-similarity is evaluates and it exhibited 89% of accuracy in this Table 9 above [5]. In this analysis, Deep BERT-Bidirectional Transformers method is utilized for understanding the language similarity. This method is a pre-trained analysis design. But the proposed-framework provided 91% of accurate similarity prediction determinations upon this Quora dataset. These well-defined accuracy rate predictions showed the precise approach and efficient technique to determine the semantic-text similarity between the question pairs obtained from quora

Table 8 Performance evaluation of proposed-system with model upon different dataset

Methods	Accuracy
MRPC	0.8417
Proposed	0.91

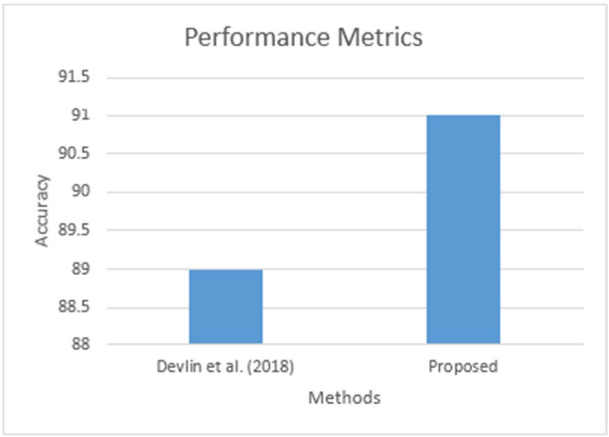

Fig. 11 Accuracy rate analysis of proposed-framework with other unsupervised BERT model in different dataset

Table 9 Performance evaluation of proposed-framework with different dataset model

Methods	Accuracy
Devlin et al. (2018) (BERT)	89
Proposed	91

Fig. 12 Execution time performance analysis

dataset and all establishing the similarity scores between different semantic-text documents, that utilizes the embedded-vector represents for shorter and longer texts .

In this Semantic-Text Similarity detections, another performance analysis parameter in assessing the efficiency of proposed-framework is execution time analysis. The Execution time for various file sizes has been considered were been fetched through our framework. The execution time of different file-sizes within our framework is represented in the Table These values were plotted in graph. This graphical-representation of execution time variation for different file-sizes is determined. This is depicted in the above Fig. 12. The Fig. 12 illustrates the variation of execution time increases with respect to File sizes within our system. But the execution time of this Semantic-text similarity detection were executed or performed in few seconds itself. Hence this depicted the efficient performance of system.

Table 10 Performance evaluation of proposed-model (Execution-Time)

File size (KB)	Time (Seconds)
5	1.23
10	2.56
15	3.15
20	4.56

The above table represents the representation of execution time values in detecting the Semantic-Text similarity scores by using our proposed-framework relying quora-dataset. The File Size taken for this similarity detection were 5 Kb, 10 kb, 15 Kb and 20 Kb. The Similarity detection for Semantic-Text, in how many seconds this is determined for these above mentioned various file sizes is represented in the Table 10. The execution time for those sizes are 1, 23 s, 2.5 s, 3.15 s, and 4.56 s as such taken by executing proposed-model algorithms. From the Table 10 , it is specified that our proposed-framework determines the Text-Similarity for every file size in few seconds. This in turn states the efficiency of the framework in terms of execution time performance analysis. However the major limitation of the study exhibited as increased computational time.

5 Conclusion

The Deep learning techniques outbreaks the baselines in text-similarity in eliminating the duplication of data in sentences, Documents wise comparisons and in question-pairs analysis. The present study implements the hybrid approach of Weighted Fine-tuned BERT extraction process with Deep Siamese networks Bi-LSTM model in quora question pair dataset. Hence in preprocessing method, special character removal and used vectorization method for the conversion of words to appropriate vector representations. The BERT extraction process, extracts the features of the sentence from question pairs. Those embedded vectors were trained by using Siamese network with Bi-LSTM model. The layers of Bi-LSTM structure does the encoding in feature-vectors. This embedded-vector were connected to input vector that traverses through multi-layer perceptron of Bi-LSTM model, with the addition of shared Weighted Fine-Tuned Weight values. The trained model outputs, with their embedded vectors predict the text-similarity of the documents or the question pairs. The study were validated among different pdf documents comparisons for text-similarity detections by using the model. The proposed-framework showed the better efficiency in determining text-similarity and the prediction score evaluation than the other existing algorithms. The model exhibited higher accuracy rate of 91% in text-similarity identifications compared with state of art approaches.

Authors' contributions I Am D. Viji Hereby State That The Manuscript Title Entitled "A Hybrid Approach of Weighted Fine-Tuned BERT Extraction with Deep Siamese Bi – LSTM Model for Semantic Text Similarity Identification" Submitted To The Multimedia tools and applications, I Confirm That This Work Is Original And Has Not Been Published Elsewhere, Nor Is It Currently Under Consideration For Publication Elsewhere. And I Am Research Scholar In the Department of CSE, Sathyabama Institute of Science and Technology, and Chennai.

I'm the corresponding author of our paper, my contribution work on this paper is to Writing, developing, and reviewing the content of the manuscript. And my co-author Dr. S. Revathy works were to cite the figure,

table and references. Equally I have done 50% and my Co-author have done 50%. We are the entire contributors of our paper. And no other third party people are not involved in this paper.

Funding This research work was not funded by any organization/institute/agency.

Code Availability N\A.

Declarations

Conflict of interest I confirm that this work is original and has either not been published elsewhere, or is currently under consideration for publication elsewhere. None of the authors have any competing interests in the manuscript.

Ethics approval No animals or human participants are involved in this research work.

Informed consent I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

References

1. Abishek K, Hariharan BR, Valliyammai C (2019) An enhanced deep learning model for duplicate question pairs recognition. Soft Computing in Data Analytics (ed). Springer, Berlin, pp 769–777
2. Buscaldi D, Roux JL, Flores JJG, Popescu A (2013) Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features. In: Second Joint Conference on Lexical and Computational Semantics, p 63
3. Croce D, Annesi P, Storch V, Basili R (2012) Unitor: Combining semantic text similarity functions through sv regression. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp 597-602
4. Deudon M (2018) Learning semantic similarity in a continuous space. In: Advances in neural information processing systems, pp 986-997
5. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
6. Gomaa WH, Fahmy AA (2013) A survey of text similarity approaches. Int J Comput Appl 68:13–18
7. Guo X, Mirzaalian H, Sabir E, Jaiswal A, Abd-Almageed W (2007) Cord19sts: Covid-19 semantic textual similarity dataset. arXiv preprint arXiv:02461, 2020
8. Habibi M, Starlinger J, Leser U (2008) TabSim: A Siamese Neural network for accurate estimation of table similarity. arXiv preprint arXiv:10856, 2020
9. He H, Lin J (2016) Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In: Proceedings of the conference of the north American chapter of the Association for Computational Linguistics: human language technologies, 2016, pp 937-948
10. He H, Gimpel K, Lin J (2015) Multi-perspective sentence similarity modeling with convolutional neural networks. In: Proceedings of the conference on empirical methods in natural language processing, 2015, pp 1576-1586
11. Shih C-H, Yan B-C, Liu S-H, Chen B (2017) Investigating siamese lstm networks for text categorization. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) IEEE, pp 641–646
12. Jang B, Kim M, Harerimana G, Kang S-u, Kim JW (2020) Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. Appl Sci 10:5841
13. Kashefi O, Mohseni N, Minaei B (2010) Optimizing document similarity detection in Persian information retrieval. J Convergence Inf Technol 5:101–106
14. Kumar S, Chakrabarti S, Roy S (2017) Earth mover's distance pooling over Siamese LSTMs for automatic short answer grading. In: IJCAI, pp 2046–2052
15. Kumar M, Sharma SC (2018) Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment. Comput Electr Eng 69:395–411

16. Kumar M, Sharma SC (2018) Cost and energy efficient scheduling in cloud environment with deadline constraint. Sustain Comput Inform Syst 19:147–164
17. Kumar M, Sharma SC, Goel A, Singh SP (2019) A comprehensive survey for scheduling techniques in cloud computing. J Netw Comput Appl 143:1–33
18. Kumar M, Dubey K, Pandey R (2021) Evolution of emerging computing paradigm cloud to fog: applications, limitations and research challenges. In: 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp 257-261
19. Laskar MTR, Hoque E, Xiangji Huang J (2019) Utilizing bidirectional encoder representations from transformers for answer selection. In: International Conference on Applied Mathematics, Modeling and Computational Science Springer, pp 693–703
20. Li JY, Zhang J. From massive pre-trained models to small low-latency deployment:distilling BERT for sentence similarity identification
21. Li Y, Zhou D, Zhao W (2020) Combining local and global features into a Siamese network for sentence similarity. IEEE Access 8:75437–75447
22. Lo C-k, Simard M (2019) Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp 206-215
23. Mahmoud A, Zrigui M (2017) Semantic similarity analysis for paraphrase identification in Arabic texts. In: Proceedings of the 31st Pacific Asia conference on language, information and computation, pp 274-281
24. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Aaai, pp 775-780
25. Mohammad A-S, Jaradat Z, Mahmoud A-A, Jararweh Y (2017) Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. Inf Process Manag 53:640–652
26. Nguyen HT, Duong PH, Cambria E (2019) Learning short-text semantic similarity with word embeddings and external knowledge sources,. Knowl Based Syst 182:104842
27. Peinelt N, Rei M, Liakata M (2010) GiBERT: Introducing Linguistic Knowledge into BERT through a Lightweight Gated Injection Method. arXiv preprint arXiv:12532, 2020
28. Peinelt N, Nguyen D, Liakata M (2020) tBERT: Topic models and BERT joining forces for semantic similarity detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 7047-7055
29. Peng S, Cui H, Xie N, Li S, Zhang J, Li X (2020) Enhanced-RCNN: an efficient method for learning sentence similarity. In: Proceedings of The Web Conference 2020, pp 2500-2506
30. Peng D, Hao B, Tang X, Chen Y, Sun J, Wang R (2020) Learning long-text semantic similarity with multi-granularity semantic embedding based on knowledge enhancement. In: International Conference on Control, Robotics and Intelligent System, 2020, pp 19-25
31. Rao J, Liu L, Tay Y, Yang W, Shi P, Lin J (2019) Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp 5373-5384
32. Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084
33. Saedi C, Dras M (2021) Siamese networks for large-scale author identification. Comput Speech Lang 70:101241
34. Shi H, Wang C, Sakai T (2020) "A Siamese CNN Architecture for Learning Chinese Sentence Similarity," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 24-29
35. Shih C-H, Yan B-C, Liu S-H, Chen B (2017) Investigating siamese lstm networks for text categorization. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp 641-646
36. Umer M, Imtiaz Z, Ullah S, Mehmood A, Choi GS, On B-W (2020) Fake news stance detection using deep learning architecture (cnn-lstm). IEEE Access 8:156695–156706
37. Vani K, Gupta D (2018) Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges. Inf Process Manag 54:408–432
38. Viswanathan S, Damodaran N, Simon A, George A, Kumar MA, Soman K (2019) Detection of duplicates in Quora and Twitter corpus. Advances in big data and cloud computing. ed: Springer, Berlin, pp 519–528

39. Wang X, Li C, Zheng Z, Xu B (2018) Paraphrase recognition via combination of neural classifier and keywords. In: International Joint Conference on Neural Networks (IJCNN), 2018, pp 1-8
40. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M et al (2020) MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval 54:57–72
41. Xiao H (2020) Hungarian layer: A novel interpretable neural layer for paraphrase identification. Neural Netw 131:172–184
42. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X et al (2020) Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inf Decis Mak 20:1–7
43. Yoo Y, Heo T-S, Park Y (2021) A novel hybrid methodology of measuring sentence similarity. arXiv preprint arXiv:2105.00648
44. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (1904) Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:09675, 2019
45. Zhang X, Rong W, Liu J, Tian C, Xiong Z (2017) Convolution neural network based syntactic and semantic aware paraphrase identification. In: International Joint Conference on Neural Networks (IJCNN), 2017, pp 2158-2163
46. Zhang Z, Wu Y, Zhao H, Li Z, Zhang S, Zhou X, et al (2020) Semantics-aware BERT for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9628-9635
47. Zhu Z, He Z, Tang Z, Wang B, Chen W (2018) A semantic similarity computing model based on Siamese network for duplicate questions identification. In CCKS Tasks, pp 44-51