IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 26, 2019, accepted January 20, 2020, date of publication January 23, 2020, date of current version February 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969041

Duplicate Questions Pair Detection Using Siamese MaLSTM

ZAINAB IMTIAZ[1], MUHAMMAD UMER[1], MUHAMMAD AHMAD[2,3], SALEEM ULLAH[1], GYU SANG CHOI[4], AND ARIF MEHMOOD[1,5]

[1]Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[2]Department of Computer Engineering, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan 64200, Pakistan
[3]Dipartimento di Matematica e Informatica-MIFT, University of Messina, 98122 Messina, Italy
[4]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38542, South Korea
[5]Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan

Corresponding authors: Gyu Sang Choi (castchoi@ynu.ac.kr) and Arif Mehmood (arifnhmp@gmail.com)

This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under the Industrial Technology Innovation Program (No.10063130) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2019R1A2C1006159).

ABSTRACT Quora is a growing platform comprising a user generated collection of questions and answers. The questions and answers are created, edited, and organized by the users. Enormous number of users on the Quora website makes it unavoidable to have multiple questions from different users with similar intent, which raises the issue of duplicate questions. Effectively detecting duplicate questions would make it easier to find high quality answers and help save time, which in turn would result in an improved user experience for writers and readers on Quora. In this paper, Quora Question Pairs dataset is collected from Kaggle for detection of duplicate questions. First, three types of word embeddings involving Google news vector embedding, FastText crawl embedding with 300 dimensions, and FastText crawl sub words embedding with 300 dimensions are implemented individually to vectorize all the questions and train the model. The final features used for prediction are blend of these three types of word embeddings. Then, Siamese MaLSTM (''Ma'' for Manhattan distance) Neural Network model is applied for prediction of duplicate questions in the dataset. Finally, the model is tested on 100000 pairs of questions. The experiments show that the proposed model achieves 91.14% accuracy which is better than the state-of-the-art models.

INDEX TERMS Duplicate question pair detection, text mining, deep learning, MaLSTM, word embedding.

I. INTRODUCTION

Quora is a social media website where questions are posted by users and answered by experts who provide quality insights. Other users can cooperate by editing questions and suggesting more accurate answers to the submitted questions. According to statistics provided by the Director of Product Management at Quora on 17 September 2018 [29], Quora receives 300 million unique visitors every month, which raises the problem of different users asking similar questions with same intent but in different words. Multiple questions with similar wording can cause readers to spend more time to find the best answer, and make writers answer multiple versions of the same question. Therefore, Quora has an important principle for having a single question thread for logically different questions. For example, questions like ''How can I be a good photographer?'' and ''What should I do to be a great photographer?'' are identical because both have the same meaning and should be answered only once. Some questions, like ''How old are you?'' and ''What is your age?'' do not have same wording. However, the context remains the same. Therefore, such questions are also considered duplicate. It can be an overhead to have different pages for such questions. Thus, identifying the duplicate questions at Quora and merging them makes knowledge sharing more efficient and effective in many ways. This way, the seekers can get answers to all the questions on a single thread and writers do not need to write the same answer on different locations for

The associate editor coordinating the review of this manuscript and approving it for publication was Jiahu Qin.

the same question. They can get larger number of readers than if the readers are divided in several threads. Currently, Quora is using Random Forest with many hand-crafted features to merge the duplicate questions into one. This model does not work very efficiently with large amount of data [29]. Inspired by advances in machine learning and deep learning models, Quora organized a competition on Kaggle in 2017. The participants were asked to apply advance techniques of Machine Learning and Deep Learning on the dataset to make the results more reliable and accurate. This work aims to fulfil the same purpose of achieving higher accuracy and saving time, used in complex feature engineering, by applying advance Neural Network Architecture.

Identification of duplicate questions is a crucial task in Natural Language Processing (NLP) with many applications such as Recognizing Textual Entailment (RTE) [1] classifying text, retrieving information and detecting plagiarism and Paraphrase Recognition [2]. It measures the degree of similarity between two interrogative fragments. If the fragments are semantically similar, they can get the same answer and are considered duplicate. The task of identification of duplicate questions can be a great challenge because the true meaning of sentence cannot be known with certainty due to the ambiguous language and synonymous expressions. There are some researches on measuring semantic similarities between sentences. The work [3] suggests to measure the semantic similarity between sentences based on WordNet by using the researcher's own developed tool. The researcher doesn't use any machine learning or deep learning model.

This work proposes a model to identify duplicate pairs of questions. The classification of question pairs is based on the level of similarity between the semantic meaning and similar wording of their text. Quora Question Pairs dataset is downloaded from Kaggle.

To deal with the duplicate questions' detection problem, combination of three different feature engineering is applied with more advanced neural-network based model which is MaLSTM. In this approach, all the questions are vectorized based on Google news vector embedding, FastText crawl embedding and FastText crawl sub words embedding, both individually and as a combination. These word vectors of the questions are used to discover the semantic similarity of words. First, the model is trained on all the features extracted from these three word embedding separately and passed to the MaLSTM Neural Network. After the three techniques have predicted their results, the model takes 33% for both Google news vector embedding, and FastText crawl embedding, and 34% of FastText crawl sub words of the predicted data and combines them by averaging for results. Once the model is trained, it is tested on 100000 records and achieves 91.14% accuracy. For detailed analysis of the results, the model is again tested on 20 unseen records. Out of 20 records, the proposed model correctly predicts 19 records. Finally, the performance is evaluated by calculating the Manhattan distance between the predicted result and the actual result. The range of Manhattan distance is between 0 and 1. Since 0.5 is the center value, that is why we set this center value as threshold. If the Manhattan distance is greater than 0.5, the question pair is predicted as duplicate otherwise it is a non-duplicate.

In Table 1, first and third question pairs have few similar words like 'jealous' in the first and 'web' in the third, but the overall meaning of pair of questions is different. Therefore the questions are not duplicate. Whereas, second and forth question pairs have similar wording as well as similar intent. Hence, they are labelled as duplicate questions.

TABLE 1. Examples of duplicate & non-duplicate questions.

Question 1	Question 2	Label
What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	Non-Duplicate (0)
Why do rockets look white?	Why rockets and boosters are painted white?	Duplicate (1)
What is the web application framework?	What is web application?	Non-Duplicate (0)
How do I make friends?	How to make friends?	Duplicate (1)

The rest of the paper is structured as follows. Section 2 describes a few researches related to this work. Section 3 gives a summary of the proposed model, the dataset, the steps performed on the dataset, and the basic introductions of deep learning models used in this research. In Section 4, results are discussed, and Section 5 gives the conclusion.

II. RELATED WORK

Classifying duplicate questions can be a tricky task since the variability of language makes it difficult to know the actual meaning of a sentence with certainty. This task is similar to the paraphrase identification problem, which is a thoroughly researched Natural Language Processing (NLP) task [4]. It uses Natural Language Sentence Matching (NLSM) to decide whether a pair of sentences with same intent is written in different words or not [5]. Feature engineering has been the center of focus for most of the traditional methods developed by different practitioners. The common features used are bag of words (BOW), term frequency and inverse document frequency (TF IDF), unigrams and bigrams. Support Vector Machine (SVM), used with different feature extraction techniques such as BOW or n-gram vectors, is one of the main methods in text categorization [6]. Recently, deep learning approaches have achieved very high performance across several Natural Language Processing (NLP) tasks especially in Semantic Text Similarity [7]–[9]. Deep models, trained with task-specific feature engineering, provided impressive results in semantic analysis and similarity measure. The researcher showed that meaningful semantic symmetries can be captured by using pre-trained word embeddings [10]. Deep models can be combined with word embeddings and used to express the semantic meaning of text chunks with satisfactory accuracy.

LSTM based neural networks have shown great outcomes for tasks such as categorization of text and retrieval of

information [11]. A research [12] proposed supervised and semi-supervised methods based on LSTM that used region embedding method for embedding the text regions of adjustable dimensions. Another work [13], proposed a Neural Network model and studied documents represented in form of vectors in an integrated manner. First, the model used CNN or LSTM to study the vector form of the sentences. Then, the context of sentences and their relations, of a given document, was determined in the distributed vector representation with recurrent neural network (RNN). A novel approach known as the C-LSTM network was used for representation of sentences and classification of text. This architecture combined the capabilities of CNN and LSTM Neural Networks. It used CNN to extract high-level features which were then fed to LSTM [14]. Another research [15], proposed a Tree based LSTM model and used it to predict the similarity between two sentences. Skip-thought based approach was proposed which used skip-gram approach of word2vec from the word to sentence level [16]. First, the sentences were passed through RNN layer to get skip-through vector. Then, it attempted to reconstruct the previous and next sentences.

In spite of aforementioned works, Siamese architecture is one of the most frequently used learning frameworks to project question and answer pairs into a joint space [17]. In another study [7], Siamese LSTM made use of pre-trained word embedding vectors for converting the sentences. For final result, Manhattan distance was calculated to measure the closeness of the pair of sentences. CNNs have achieved great results in classification [18] and in other Natural Language Processing (NLP) tasks [19]. Another research [20] applied Siamese CNN model that used several convolution and pooling processes to produce sentence embeddings. However, using pre-trained word embeddings that are not related to the dataset limits the results of above-mentioned models.

There are only few researches done on Quora dataset [21]. CNN based model used with GloVe embedding, which consists of 100*dimensions* Wikipedia vectors, attained 80.4% accuracy [5]. Another work [22], applied the Siamese GRU using a bi-layer similarity network and achieved 85.0% accuracy. Support vector classifier model trained using the pre-computed features ranging from longest common substring and sub sequences to word similarity based on lexical and semantic resources also attained 85% accuracy. In [23], a bilateral multi-perspective matching (BiMPM) model was applied using the ''matching-aggregation'' framework and 88.17% accuracy was achieved.

Unlike most of the methods mentioned above, this study employs Google news vector embedding, FastText crawl embedding and FastText crawl sub-word embedding for higher level feature engineering. By combining these word embeddings, the size of the training word-vector increases immensely. Since the word embeddings contain word-vectors from various fields, it broadens the range of training domain. This work uses MaLSTM Deep model to read input vectors of each sentence and provides the final hidden state in form of output vector. Afterwards, the similarity between these representations is calculated using Manhattan distance and is used to predict the target label. Overall, the results show that our technique produces more accurate outcomes than other described feature extraction and deep learning strategies. This approach identifies 19 out of 20 pairs of questions successfully.

III. DATASET & PREPROCESSING

A sentence is a set of words which forms phrases and clauses. Meaning of a sentence can be comprehended by inspecting its structure and components. By using Neural Networks, the relationships between words can be examined from several points of view. In this paper, a novel Siamese MaLSTM model is described for discovering the semantic relevance between a pair of questions. The word 'Siamese' refers to the use of two or more sub-identical network structures at the same time. In 'MaLSTM', the first 'Ma' refers to the Manhattan distance estimation technique which is used to measure the similarity between two textual features. While the LSTM is used as a sequence modeling technique which is capable of learning long-term dependencies by processing the input at its three gates. This model proposes an approach in which it combines three feature engineering techniques: GoogleNewsVector, FastText crawl, and FastText crawl subwords.

The blending of these three individually trained word embedding predictions allowed the generation of more accurate predictive results as compared to the traditional deep learning models with single feature engineering technique. First the model is trained on each word embedding individually, then we blend the prediction of each individually trained model by getting first two models by a ratio of 33% and the 3*rd* model by 34% for final prediction.

A. DATASET

Quora released a public dataset that consists of 404, 351 question pairs in January 2017 [24]. The question pairs are from various domains including technology, entertainment, politics, culture, and philosophy. This dataset is downloaded from Kaggle [25] . Each record has a pair of questions and a target class that represents whether the questions are duplicate or not. The dataset is split in 75 and 25 ratio for training and testing respectively. The name of dataset attributes with their description is shown in Table 2.

TABLE 2. Attributes description of dataset.

Name of Attribute	Description
id	ID of each question pair.
qid1	ID of question one.
qid2	ID of question two.
question1	Full content of question one.
question2	Full content of question two.
Is_duplicate	The target label, which is set to 1 if question1 and question2 have similar intent, and 0 if not.

Since the classifier is only concerned with ''question1'', ''question2'' and ''Is_duplicate'', the rest of the attributes of the dataset are ignored. Some examples, from the dataset, of duplicate and non-duplicate questions are shown in Table 3.

TABLE 3. Dataset question pair examples.

Duplicate Questions	Non-Duplicate Questions
What can make Physics easy to learn?	What's causing someone to be jealous?
How can you make physics easy to learn?	What can I do to avoid being jealous of someone?

B. PREPROCESSING

The steps required for organizing the data in understandable format by handling the missing, inconsistent and redundant values is called preprocessing. Various pre-processing steps are performed on experimental dataset. Several NLP techniques are used such as conversion to lower letters of text, stopwords removal, stemming, and tokenization, with the help of freely available libraries such as NLTK and keras's.

After performing the pre-processing steps, the quality of data improves due to the elimination of unnecessary information. The tokenizer function from keras's library is used to tokenize each question into a vector of words, then word embeddings (GoogleNewsVector, FastText crawl, and FastText crawl subwords) are used to extract the quality features. The maximum length of all the questions is set to 20, whereas, the questions with length smaller than 20 are zero padded. Finally, the preprocessed features are fed into the Siamese MaLSTM architecture for label prediction.

IV. PROPOSED METHODOLOGY

1) WORD EMBEDDING

The Deep models do not understand input in the form of text or speech. In order to make the input understandable for such models, every question must be vectorized. In our proposed model, first layer is embedding layer that accepts the question pairs as input and converts each word into a vector. The embedding dimension is set to 300 and the maximum sequence length is 20. In this work, three different word embeddings of GoogleNewsVector, FastText crawl, and FastText crawl subword are used.

2) GoogleNewsVector

Google provides pre-trained word embedding based on news corpus. This word embedding contains 3 million English words with $300-$ *dimensions*, providing 3 billion word-vectors [26].

3) FastText

FastText is an efficient word representation learning library provided by the Facebook research team. It contains 2 million common crawl words with $300-$ *dimensions*, providing 600 billion word-vectors. It is different from Google word embedding because it provides the n-gram character-level representation of words [27].

4) FastText SUBWORD

FastText Subword contains 2 million word vectors trained with subword information on Common Crawl ($600B$ tokens). Subword embedding provides us more details by converting each word into its sub words. If we want to get the subwords of word 'where' with $n = 3$ the resulting subwords will be, 'whe', 'her', and 'ere'. At the end, it provides the dictionary of union of these subwords [28].

5) SIAMESE DEEP LEARNING NETWORK

Siamese is an artificial neural network that processes two or more input vectors side by side and combines the output vector after sub-identical neural network computation [29]. The weights must also be shared among all the inputs because it reduces training parameters and chances of overfitting. This idea was first proposed in 1994 [29]. The input given to siamese network can be in any form such as numeric , image or text data. Siamese network is useful for several tasks that requires discovering relationship between two patterns such as sentence semantic similarity identification, forged signatures recognition, pattern recognition, and paraphrase identification [31]. Similar inputs are processed with sub-identical network models. The sub-networks extract features from inputs that are similar and comparable. Siamese network applies binary classification at the output, which indicates if the inputs are of the same class or not. If the inputs belong to same class, then it means that those are somehow identical to each other and considered as duplicate. While joining the output of processed inputs, the neuron measures the distance between two feature vectors. Based on calculated distance, questions are considered as duplicate or non-duplicate.

6) MaLSTM

As we know, LSTM is a sequence modeling technique which generates long term sequences by using its multiple inside layers. It consists of four components, o_t output gate, c_t cell memory block (current state determines which information will be fed to next neuron), i_t input gate and the forget gate f_t. The input I_t feeds the LSTM layer in the form of real-valued vectors. The hidden state representations h_t are updated sequentially between the gates. The update steps purely rely on cell memory block c_t. These four components decide which information is used and which information is omitted from the model for final prediction. There are multiple variants of LSTM used to solve different types of problems [32]. The first variant(1) and second variant-(2) that we used in this experiment, are used to generate long term sequences on textual data. It uses sigmoid layer for deciding which information is used for final prediction. It generates output between 0 and 1, where 0 means omit the information and 1 means use the information for final prediction. W_i refers to the weights assigned to input vectors, h_t refers to the current input to neuron and b_i refers to the bias value added to the inputs. The LSTM variants shown in Equations 3 - 6 in

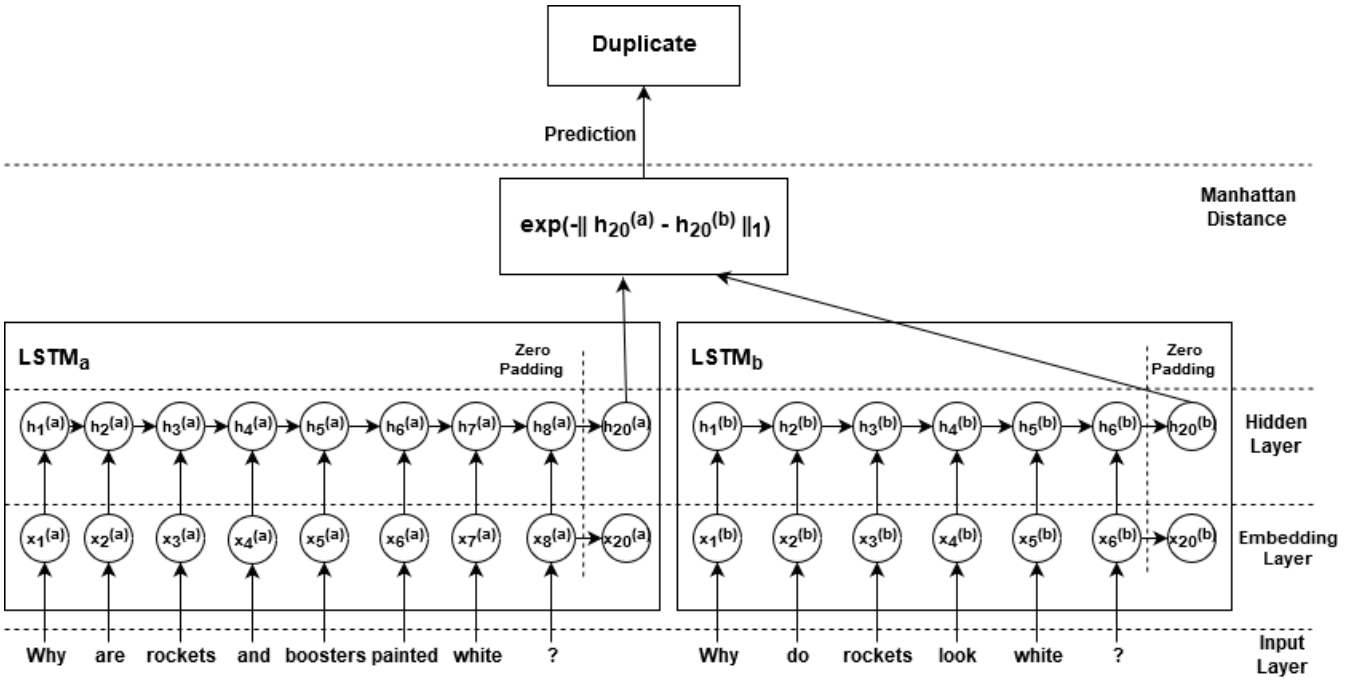


FIGURE 1. Siamese Manhattan LSTM netowrk with zero padding for max sequence length 20.

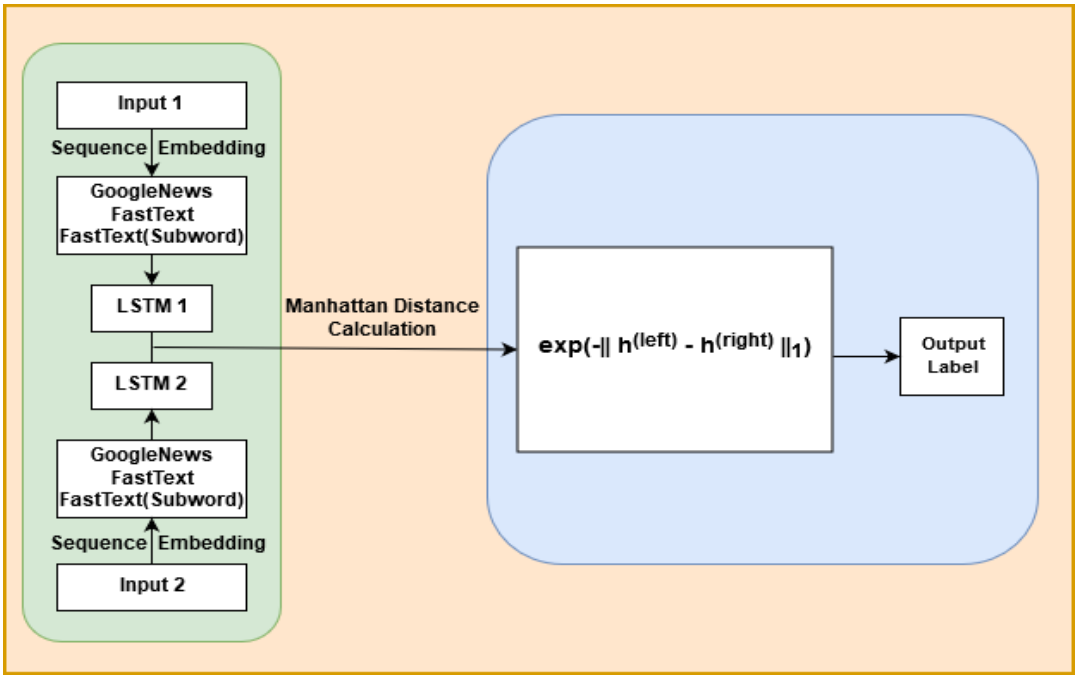
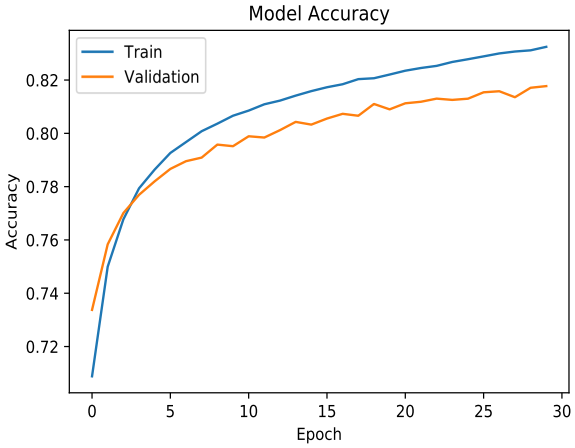


FIGURE 2. Proposed model architecture diagram.

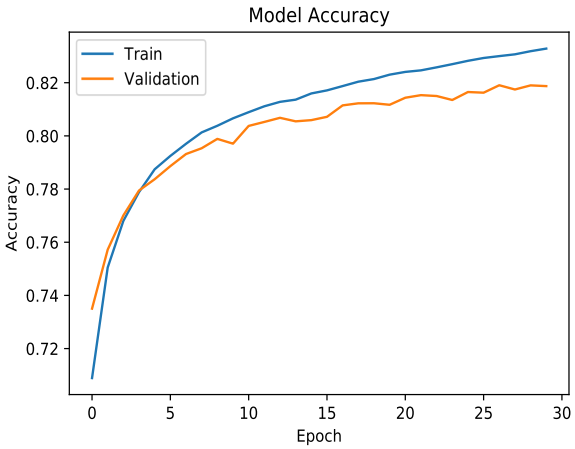
which *tanh* activation function is used to generate sequence based on the subject of the text. If the subject of the sequence processing sentence changes then all previous information stored in the cell memory block also gets erased.

$$I_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$
$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$
$$c_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3)$$
$$c_t = i_t \odot c_t + f_t \odot c_{t-1} \quad (4)$$
$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$
$$h_t = o_t \odot tanh(c_t) \quad (6)$$

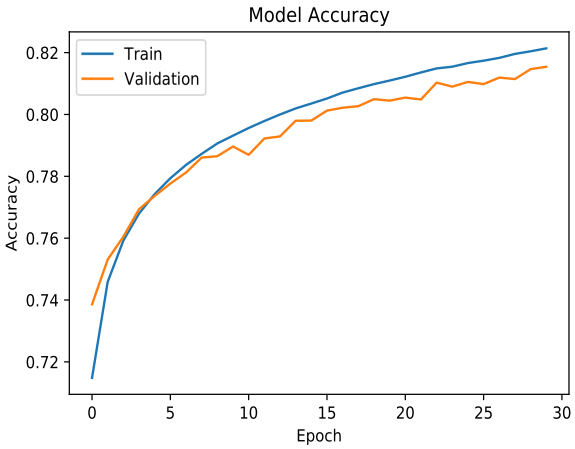
LSTM generate sequences on variable-length space vectors sequences of d_{in}-dimensional vectors. In this experiment,



(a) Google News Vector Embedding

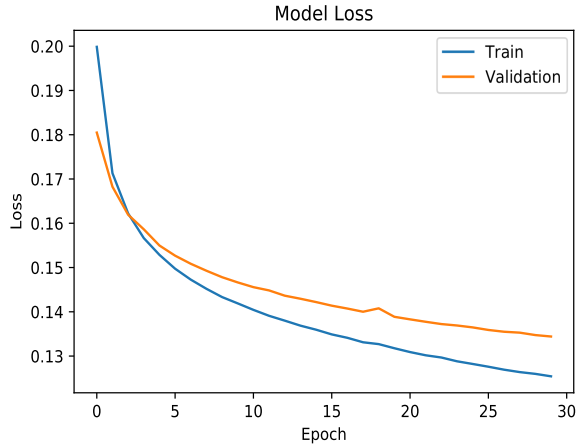


(b) FastText Embedding

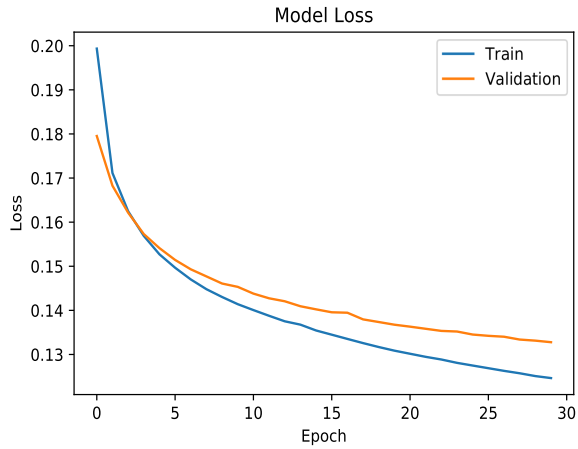


(c) FastText Subword Embedding

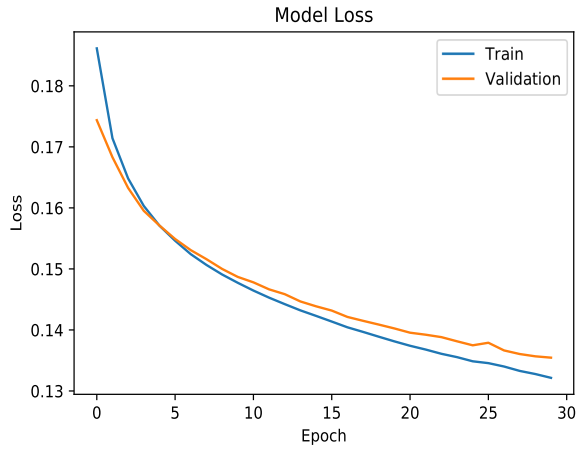
FIGURE 3. Training and validation *accuracy* of all models.



(a) Google News Vector Embedding



(b) FastText Embedding



(c) FastText Subword Embedding

FIGURE 4. Training and validation *loss* of all models.

the input dimensional vectors size is 300 ($dim = 300$). Each question is represented as word sequence of word vectors $(a_1, a_2, a_3, \ldots, a_N)$. The maximum sequence length is set to 20. No question can be greater than 20 in length. The questions having length lesser than 20 are zero-padded.

The Manhattan LSTM model (MaLSTM) uses Siamese architecture, where two identical LSTM sub-networks (LSTM_a and LSTM_b) processes a sentence in the input sentence pair. These input sentences are converted into

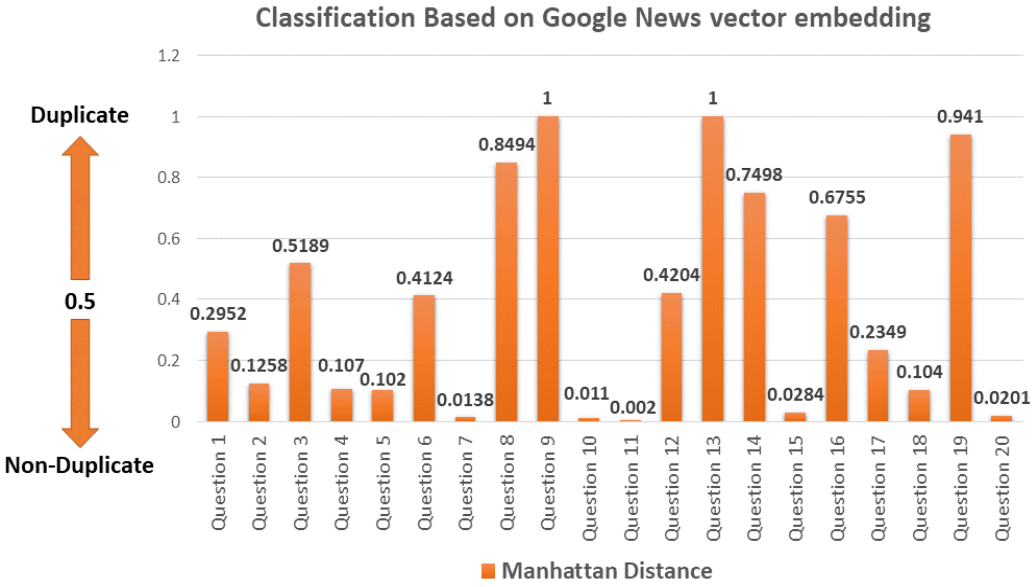


FIGURE 5. GoogleNewsVector embedding Manhattan distance score.

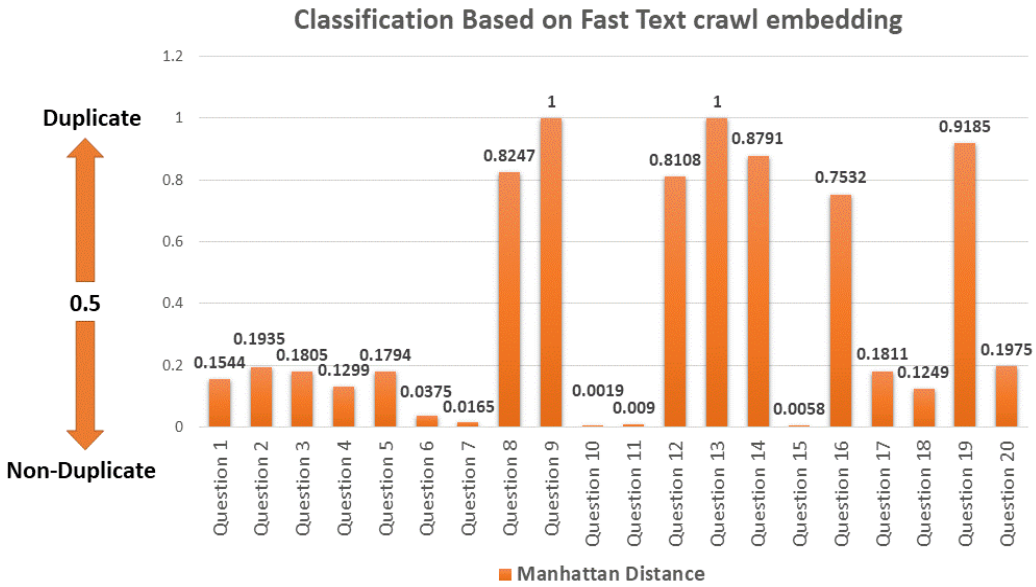


FIGURE 6. FastText crawl embedding Manhattan distance score.

real-valued word vector form and are assigned equal weights which converts the variable length input sequences to the fixed length vector form [7]. Each assigned weight processes one question. MaLSTM calculates Manhattan distance for the final prediction. Manhattan distance outperforms a little than the other substitutes such as cosine similarity [7]. The proposed Siamese MaLSTM architecture diagram is shown in Figure 1. The reason for choosing Manhattan Distance among other similarity measure is that we are working on large set of word embedding consisting of multiple dimensions. It has been observed by many researchers that Manhattan distance similarity measure not only performs well on very high dimensional data but also takes less time for computation since Manhattan distance finds the similarity between textual features by calculating the absolute distance between two points that lies at axes of right angle [4], [6], [23]. Manhattan equation for two points x and y is shown in Equation 7

$$M_a = |x_1 - x_2| + |y_1 - y_2| \tag{7}$$

In Equation 7, the x_1 and y_1 refers to output of first model and x_2 and y_2 to second. The absolute difference between them shows the similarity measure between the two inputs given to the model. In this experiment, we set a threshold of 0.5 to classify the questions as duplicate or non-duplicate. If the final distance measure value of Manhattan distance is

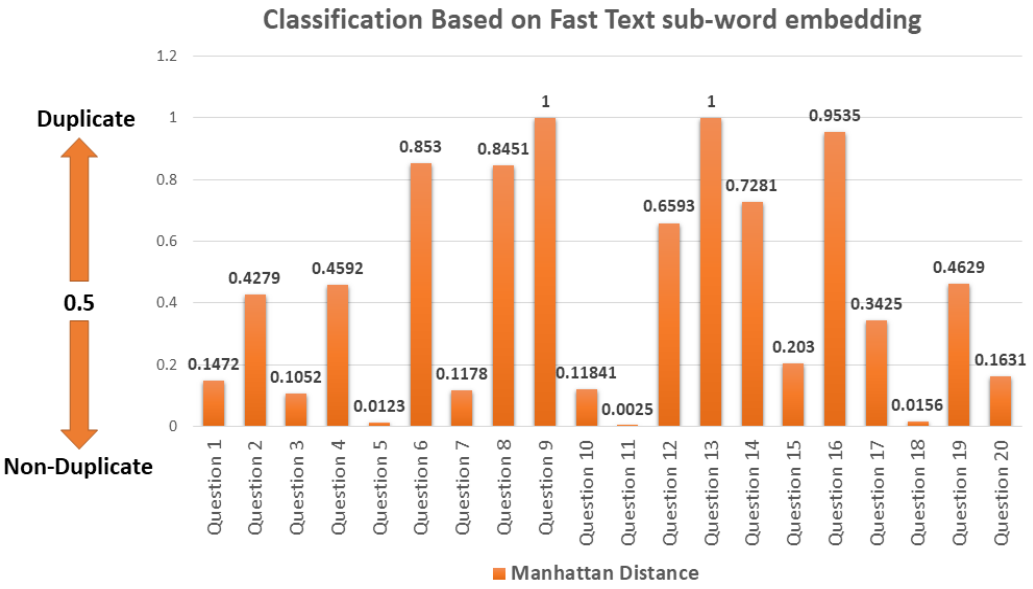


FIGURE 7. FastText crawl subword embedding Manhattan distance score.

greater than 0.5, the question pair is classified as duplicate - otherwise a non-duplicate. The complete flow of experiment is shown in Figure 2.

V. RESULTS AND DISCUSSION

In the final set of experiments, the Siamese LSTM model is first trained on each word embedding (GoogleNewsVector, FastText and FastText subword) individually and later we use the blend of these trained models prediction for final prediction. The models are trained on 303K number of samples and are tested on 100k instances.

The training is performed using a 2GB Dell PowerEdge T430 graphical processing unit on 2x Intel Xeon 8 Cores 2.4Ghz machine which is equipped with 32 GB DDR4 Random Access Memory(RAM). The training takes 3.5 hours to run epochs on 'Quora Question Pair Dataset' on each embedding and to show the classification results. The results are shown in Table 4 where it is evident that by using Siamese-MaLSTM with Google News Vector, FastText Crawl and FastText Crawl Subword, the model achieved 81.77%, 82.77% and 82.57% accuracy, respectively. It can be observed that upon combining the predicted results of all three of these approaches with 33% of Google News Vector, 33% of FastText Crawl and 34% of FastText Crawl subwords, the obtained accuracy is 91.14% which is much higher than other state of the arts models. The training and validation accuracy of models is shown in figure 3 and loss in figure 4

For better understanding of how our models predicting classes, we have extracted 20 test samples from the test data and models on that. The results accuracy is shown in table 5.

Predicted results were observed in each case mentioned above. When Google News Vector is used with Siamese LSTM, 16 out of 20 records are predicted successfully as shown in Table 6. Google news vector has around 3 billion word-vector tokens which are relatively lesser than the other two embedding used in this paper. Moreover, the words-vectors are only related to news domain. Whereas, the Quora question pair dataset used in this work contains records of several domains. That is why, the model trained on this word embedding is unable to identify the questions that are from other domains.

When FastText crawl with Siamese MaLSTM is tested, 18 out of 20 records are predicted correctly as shown in Table 6. FastText crawl has 600 billion word-vector tokens, which are a lot more than the Google news vector embedding. It contains word-vectors from various domains and that is why it provided better results by predicting two more records accurately. Remaining two records that are predicted wrong are question pair 6*th* and 9*th*. In 6*th* question, after removal of stop words, remaining features has very less similarity. For example, 1*st* question has ''astrology'', ''rise'' and ''cap''

TABLE 4. Model results using different word embeddings.

Model	Word Embedding	Accuracy	Precision	Recall	F-Score
Siamese LSTM	Google News Vector	81.77%	78.77%	69.93%	74.09%
Siamese LSTM	FastText Crawl ($600B$)	82.77%	79.20%	70.58%	74.64%
Siamese LSTM	FastText Crawl Subword ($600B$)	82.57%	78.26%	70.29%	74.04%
Siamese LSTM	Blend of word embeddings	91.14%	83.68%	81.17%	82.41%

TABLE 5. Accuracy results using different word embeddings.

Model	Word Embedding	Accuracy
Siamese LSTM	Google News Vector	80.0%
Siamese LSTM	FastText Crawl ($600B$)	90.0%
Siamese LSTM	FastText Crawl Subword ($600B$)	90.0%
Siamese LSTM	Blend of word embeddings	95.0%

TABLE 6. Question pair classification using google news vector embedding and Manhattan distance.

Sr#	Question 1	Question 2	Actual Label	Predicted Label Google News	Predicted Label FastText	Predicted Label FastText Subword	Predicted Label Blending
1	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0	0	0	0	0
2	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0	0	0	0	0
3	How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0	1	0	0	0
4	Why am I mentally very lonely? How can I solve it?	Find the remainder when 23^{24} is divided by 24,23?	0	0	0	0	0
5	Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0	0	0	0	0
6	Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1	0	0	1	1
7	Should I buy tiago?	What keeps childern active and far from phone and video games?	0	0	0	0	0
8	How can I be a good geologist?	What should I do to be a great geologist?	1	1	1	1	1
9	When do you use ã instead of ä?	When do you use "&" instead of "and"?	0	1	1	1	1
10	Motorola (company): Can I hack my Charter Motorolla DCX3400?	How do I hack Motorola DCX3400 for free internet?	0	0	0	0	0
11	Method to find separation of slits using fresnel biprism?	What are some of the things technicians can tell about the durability and reliability of Laptops and its components?	0	0	0	0	0
12	How do I read and find my YouTube comments?	How can I see all my Youtube comments?	1	0	1	1	1
13	What can make Physics easy to learn?	How can you make physics easy to learn?	1	1	1	1	1
14	What was your first sexual experience like?	What was your first sexual experience?	1	1	1	1	1
15	What are the laws to change your status from a student visa to a green card in the US, how do they compare to the immigration laws in Canada?	What are the laws to change your status from a student visa to a green card in the US? How do they compare to the immigration laws in Japan?	0	0	0	0	0
16	What would a Trump presidency mean for current international master's students on an F1 visa?	How will a Trump presidency affect the students presently in US or planning to study in US?	1	1	1	1	1
17	What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0	0	0	0	0
18	Why do girls want to be friends with the guy they reject?	How do guys feel after rejecting a girl?	0	0	0	0	0
19	Why are so many Quora users posting questions that are readily answered on Google?	Why do people ask Quora questions which can be answered easily by Google?	1	1	1	0	1
20	Which is the best digital marketing institution in banglore?	Which is the best digital marketing institute in Pune?	0	0	0	0	0

that are not present in 2*nd* question. And the 2*nd* question has "ascendant" and "triple" that are only present in 2*nd* question. Other wrong predicted question has special symbols in it and the model could not identify the correct result in that case as well.

Finally, when FastText crawl subwords with Siamese LSTM is tested, 18 out of 20 records are predicted correctly as shown in Table 6. Like FastText crawl, Fast Text crawl subwords also has 600 billion word-vectors from several domains. But it also considers the subwords of each word. This word embedding is also unable to differentiate between the special symbols present in the 9*th* question. The other question that is wrongly predicted is 19*th*. As we know that, in fastText subword, we have n-gram of each word which leads to richer word2vec dictionary. Sometimes due to n-gram breaking, the word root forms are changed due to which the semantic similarity between two sentences is not accurately identified.

The MaLSTM accurately scores 95% result when we use it with all three of the feature engineering techniques. It predicts 19 out of 20 records accurately as shown in Table 5. The only label which was predicted wrongly was due to the presence of special symbols in 9*th* pair. It is observed that even after combining the techniques, the model was not able to differentiate between special characters ã and ä in first question, and between "*and*" and "&" in second question of the 9*th* question pair because the model was not trained on special symbols in any of the word embedding.

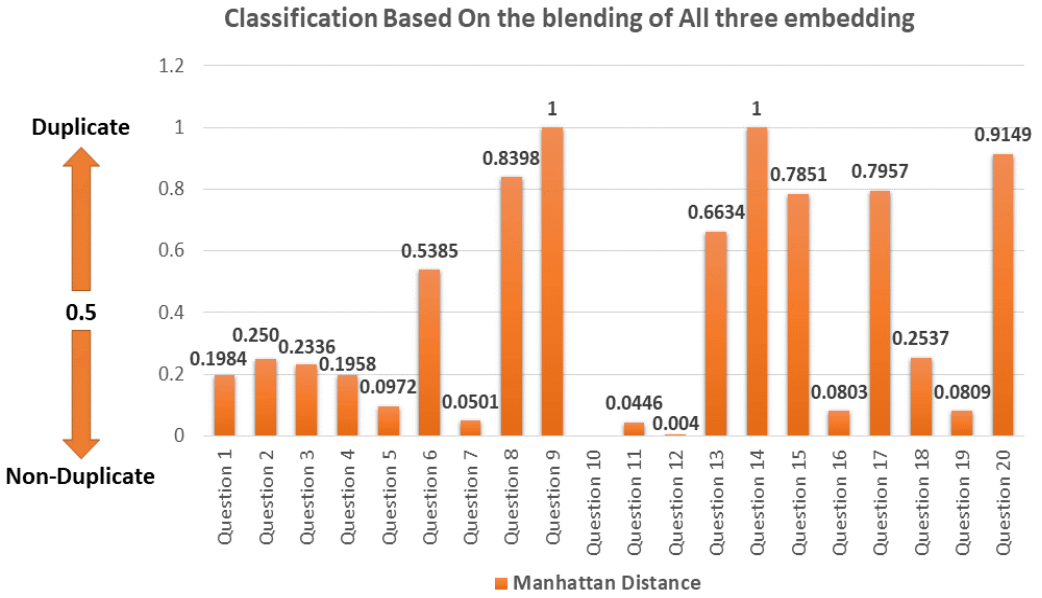


FIGURE 8. Word embedding blending Manhattan distance score.

VI. CONCLUSION AND FUTURE WORK

This work proposed a model that identifies duplicate question pairs by combining the three word embedding (i.e., Google News Vector, FastText Crawl, and FastText Crawl Subword) feature extraction techniques which results in a much better accuracy as compared to these embeddings individually. Furthermore, this work proposed a novel Siamese MaLSTM model which accounts the Manhattan distance to determine the semantic similarity among the questions with 95% accuracy which is way better then state-of-the-art works. Upon closely looking at the manhattan values, in blend of different word embedding predictions, the manhattan score classifies the question pairs in more accurate way than any other embedding. The duplicate question score is very close to 1 while the non-duplicate pair values are much closer to zero. This determines the correctness and exactness of our proposed technique. The future work entails the Hybrid Neural Networks with attention layer with several similarity measuring techniques and experimental analysis on a larger dataset.

REFERENCES

[1] S. AbdelRahman and C. Blake, "A rule-based human interpretation system for semantic textual similarity task," in *Proc. 1st Joint Conf. Lexical Comput. Semantics, 6th Int. Workshop Semantic Eval.*, vol. 2, 2012, pp. 536–542.

[2] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proc. 1st Joint Conf. Lexical Comput. Semantics*, vol. 2, 2012, pp. 385–393.

[3] T. N. Dao and T. Simpson. *Measuring Similarity Between Sentences*. Accessed: Sep. 15, 2019. [Online]. Available: https://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement

[4] W. Zhu, T. Yao, J. Ni, B. Wei, and Z. Lu, "Dependency-based Siamese long short-term memory network for learning sentence representations," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0193919.

[5] Z. Wang, W. Hamza, and R. Florian, "Bilateral multi-perspective matching for natural language sentences," *CoRR*, vol. abs/1702.03814, pp. 1–7, Jul. 2017.

[6] B. N. Patro, V. K. Kurmi, S. Kumar, and V. P. Namboodiri, "Learning semantic sentence embeddings using sequential pair-wise discriminator," *CoRR*, vol. abs/1806.00807, pp. 1–15, Mar. 2018.

[7] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2786–2792.

[8] M. Tsubaki, K. Duh, M. Shimbo, and Y. Matsumoto, "Non-linear similarity learning for compositionality," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2828–2834.

[9] B. Rychalska, K. Pakulska, K. Chodorowska, W. Walczak, and P. Andruszkiewicz, "Samsung Poland NLP team at SemEval-2016 task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, San Diego, CA, USA, 2016, pp. 602–608.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrasesand their compositionality," *CoRR*, vol. abs/1310.4546, pp. 1–9, Oct. 2013.

[11] P. Sravanthi and B. Srinivasu, "Semantic similarity between sentences," *Int. Res. J. Eng. Technol.*, vol. 4, no. 1, pp. 156–161, 2017.

[12] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using LSTM for region embeddings," in *Proc. ICML*, 2016, pp. 1–9.

[13] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, Sep. 2015, pp. 1422–1432.

[14] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A C-LSTM neural network for text classification," Nov. 2015, *arXiv:1511.08630*. [Online]. Available: https://arxiv.org/abs/1511.08630

[15] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, vol. 1, Jul. 2015, pp. 1556–1566.

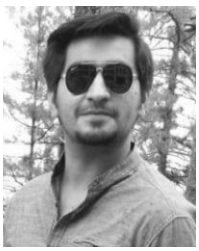
[16] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," *CoRR*, vol. abs/1506.06726, pp. 1–11, Jun. 2015.

[17] L. Yu, K. M. Hermann, P. Blunsom, and G. S. Pulman, "Deep learning for answer sentence selection," 2014, *arXiv:1412.1632*. [Online]. Available: https://arxiv.org/abs/1412.1632

[18] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[19] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.
[20] H. He, K. Gimpel, and J. Lin, "Multi-perspective sentence similarity modeling with convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1576–1586.
[21] C.-H. Shih, B.-C. Yan, S.-H. Liu, and B. Chen, "Investigating Siamese LSTM networks for text categorization," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 641–646.
[22] Y. Homma, S. Sy, and C. Yeh, "Detecting duplicate questions with deep learning," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–8.
[23] K. Abishek, B. R. Hariharan, and C. Valliyammai, "An enhanced deep learning model for duplicate question pairs recognition," in *Soft Computing in Data Analytics*. Singapore: Springer, 2019, pp. 769–777.
[24] N. S. Dandekar Iyer and K. Csernai. (2017). *First Quora Dataset Release: Question Pairs*. Accessed: Sep. 15, 2019. [Online]. Available: https://data.quora.com/FirstQuora-DatasetRelease-Question-Pairs
[25] Quora. (Nov. 2017). *Quora Question Pairs, Version 1*. Accessed: Sep. 20, 2019. [Online]. Available: https://www.kaggle.com/c/quora-questionpairs/data
[26] M. Mihaltz. (May 2016). *word2vec-Googlenews-Vectors*. Accessed: Sep. 20, 2019. [Online]. Available: https://github.com/mmihaltz/word2vecGoogleNews-vectors
[27] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *CoRR*, vol. abs/1612.03651, pp. 1–13, Dec. 2016.
[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, pp. 1–12, Jun. 2016.
[29] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, San Francisco, CA, USA, 1993, pp. 737–744.
[30] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. 7th Annu. Neural Inf. Process. Syst. Conf.*, San Francisco, CA, USA, 1994, pp. 737–744.
[31] P. Premtoon. (2017). *What are Siamese Neural Networks, What Applications are They Good for, and Why?* Accessed: Sep. 15, 2019. [Online]. Available: https://www.quora.com/What-are-Siamese-neural-networks-what-applications-are-they-good-for-and-why
[32] K. Greff, R. K. Srivastava, J. Koutník, R. Bas Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *CoRR*, vol. abs/1503.04069, pp. 1–12, Oct. 2015.

ZAINAB IMTIAZ received the B.S. degree from the Department of Computer Science, University of Central Punjab (UCP), Pakistan, in 2015. She is currently pursuing the master's degree in Computer Science with the Khwaja Fareed University of Engineering and IT (KFUEIT). She worked as Microsoft Dynamic AX Developer, from 2015 to 2017. She is also serving as a Research Assistant with the Fareed Computing and Research Center, KFUEIT, Pakistan, and an Assistant Lecturer with the Computer Science Department. Her recent research interests are related to data mining, machine learning, and deep learning-based text mining.



MUHAMMAD UMER received the B.S. degree from the Department of Computer Science, Khwaja Fareed University of Engineering and IT (KFUEIT), Pakistan, from October 2014 to October 2018, where he is currently pursuing the master's degree in computer science. He is also serving as a Research Assistant with the Fareed Computing and Research Center, KFUEIT. His recent research interests are related to data mining, mainly working machine learning and the deep learning-based IoT, text mining, and computer vision tasks.

MUHAMMAD AHMAD is currently an Assistant Professor with the Department of Computer Engineering, Khwaja Fareed University of Engineering and Information Technology, Pakistan. He is currently an Associated with the Research Group of Advanced Image Processing Research Lab (AIPRL), a first Hyperspectral Imaging Lab, Pakistan. He is also an Associated with the University of Messina, Messina, Italy, as a Research Fellow. He authored a number of research articles in reputed journals and conferences. His current research interests include machine learning, computer vision, remote sensing, hyperspectral imaging, and wearable computing. He is also a Regular Reviewer for Springer Nature journals, NCAA, the IEEE TIE, the IEEE TNNLS, the IEEE TGRS, the IEEE TIP, the IEEE GRSL, the IEEE GRSM, the IEEE JSTARS, the IEEE TMC, the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE ACCESS, the IEEE COMPUTERS, the IEEE SENSORS, the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, MDPI Journals, *Optik*, *Measurement Science and Technology*, *IET Journals*, and the *Transactions on Internet and Information Systems*.

SALEEM ULLAH was born in AhmedPur East, Pakistan, in 1983. He received the B.Sc. degree in computer science from Islamia University Bahawalpur, in 2003, the M.I.T. degree in computer science from Bahauddin Zakariya University, Multan, in 2005, and the Ph.D. degree from Chongqing University, China, in 2012. From 2006 to 2009, he worked as a Network/IT Administrator in different companies. From August 2012 to February 2016, he worked as an Assistant Professor with Islamia University Bahawalpur, Pakistan. He has been working as an Associate Professor with the Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, since February 2016. He has almost 13 years of industry experience in field of IT. His research interests include ad hoc networks, congestion control, and security.

GYU SANG CHOI received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, in 2005. He was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT) for Samsung Electronics, from 2006 to 2009. Since 2009, he has been a Faculty Member with the Department of Information and Communication, Yeungnam University, South Korea. His research areas include non-volatile memory and storage systems.

ARIF MEHMOOD received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, South Korea, in November 2017. Since November 2017, he has been a Faculty Member with the Department of Computer Science, KFUEIT, Pakistan. He is currently serving with The Islamia University of Bahawalpur, as an Assistant Professor. His recent research interests are related to data mining, mainly working on AI and deep learning-based text mining, and data science management technologies.