A Deep Learning Model for Classifying Quality of User Replies

Masoumeh Rajabi, Shahla Nemati, Mohammad Ehsan Basiri*
Department of Computer Engineering, Shahrekord University, Shahrekord, Iran.
rajabim91@gmail.com, basiri@sku.ac.ir, s.nemati@sku.ac.ir

Received: 2021/05/28 Revised: 2021/09/01 Accepted: 2021/09/16

***Abstract*— Q&A forums are designed to help users in finding useful information and accessing high-quality content posted by other users in text forums. Automatically identifying high-quality replies posted in response to the initial posts not only provides users with appropriate content, but also saves their time. Existing methods for classifying user replies based on their quality, try to extract quality features from both the textual content and metadata of the replies. This feature engineering step is a time and labor-intensive task. The current study addresses this problem by proposing new model based on deep learning for detecting quality user replies using only raw textual content. Specifically, we propose a long short-term memory (LSTM) model that exploits the embeddings from language models (ELMo) for representing words as contextual numerical vectors. We compared the effectiveness of the proposed model with four traditional machine learning models on the TripAdvisor for New York City (NYC) and the Ubuntu Linux distribution online forums datasets. Experimental results indicated that the proposed model significantly outperformed the four traditional algorithms on both datasets. Moreover, the proposed model achieved about 16% higher accuracy compared to that obtained by the traditional algorithms trained on both textual and quality dimension features.**

Keywords*— *Text Classification; Deep Neural Networks; Social Media Text Processing; Machine Learning.

1. INTRODUCTION

With the dramatic increase in people's access to web services in recent years, users' use of the web to find answers to their questions about buying and selling, renting cars, and finding hotels has increased [1]. This increase has led to the production of a large amount of multimedia data, most of which is textual data [2]. Analyzing this textual data and extracting the information required by users from them is one of the main applications of data mining (DM) and natural language processing (NLP). With the help of methods in DM and NLP, the answers to many questions can be found in text forums, such as finding the best item that meets the needs of users [1], how to repair a device [3], students' questions [4], overall sentiment of sentences and documents [5, 6], and even the initial treatment of many common diseases [7].

Text forum threads (TFT) are pairs of original posts and their responses posted by forum users who are similar in terms of needs, knowledge, or location [8]. TFTs provide convenient means of accessing, sharing and exchanging information between people. Typically, each discussion in forums is started by sending an initial post for asking help from other users and continues by sending responses to this initial post. The overall structure of a typical Q&A text forum is shown in Fig. 1.

Because Q&A forums have many users and provide ease of sending comments, usually, many replies are sent in response to each post on a particular topic [8]. Reading and manually filtering this huge amount of textual data is almost impossible for users who need fast and accurate responses to their questions. Therefore, automatically identifying high-quality replies is an important utility in Q&A forums [8]. This needs an accurate classification method for categorizing the replies according to their quality [8]. Several studies tried to design such classifying methods using traditional machine learning (ML) algorithms.

Existing studies tried to improve the performance of the classification methods by improving the feature extraction and

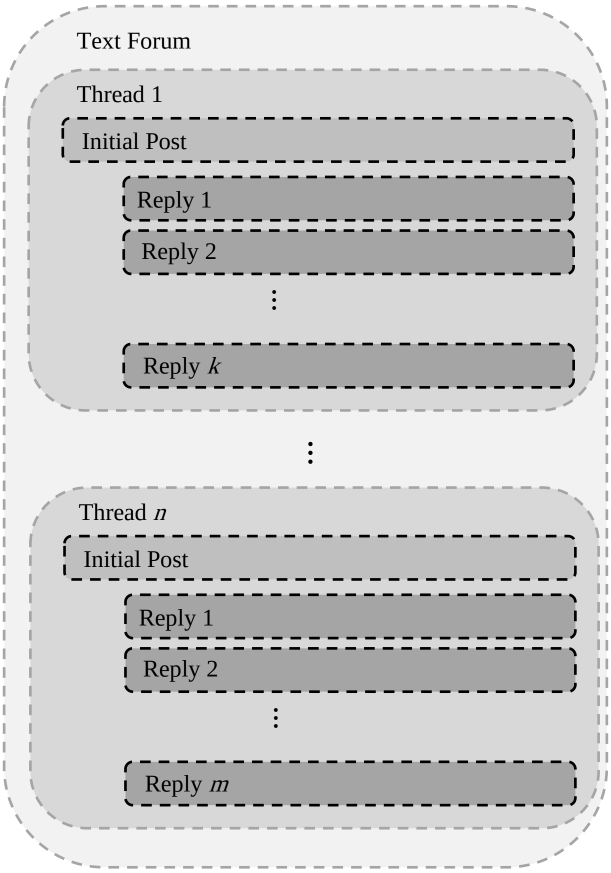


Fig. 1. Structure of a typical text forum and its threads containing initial posts and replies.

10.22133/ijwr.2021.288231.1095 20.1001.1.26454335.2021.4.1.3.9

feature selection steps [8-10]. Extracting meaningful features, Quality Dimensions (QDs), from both the textual content and meta data of replies [8]. QDs are used as measures of the quality of post responses in a typical conversation. Several QDs have been applied in the literature. For example, reply time and the user contribution used in [10] to measure quality support in a stack overflow discussion forum. Number of views and replies are another QD features that were used in [11] for developing a more accurate forum crawler. Lexical, syntactical, external, forum specific, and similarity features have also been considered for evaluating the quality of posts [12]. Relatedness and the acceptance dimension feature have also been used for measuring the relatedness of a post to the topic of discussion [13]. Relevancy, user activity, suitability, ease-of-understanding, respect, and the amount-of-data were six categories of QDs used in [8] for the classification of user replies in three classes of non-quality, low-quality, and high-quality.

Although the above-mentioned studies tried to achieve high classification performance by designing different types of QDs, they focused on the feature engineering process and hence suffer from same drawbacks; being data-dependent and needing metadata of threads which may be unavailable to APIs and users in many TFTs. These two drawbacks motivate the use of models which relies only on the available textual data of TFTs. In order to resolve these problems, a new deep neural network model is presented in the present study. Deep neural networks are similar to traditional machine learning methods in learning from existing similar data and generalizing to unseen data and needing no explicit feature engineering. The proposed dep model in the current study neither requires a feature engineering phase, nor needs metadata as its input. Specifically, the proposed model only utilizes the textual content of the threads to classify the users' replies according to their quality. An example of such a textual content is shown in Table I.

The proposed model employs pre-trained word embedding trained on large-scale textual datasets. This improves the ability of the proposed model in representing the input data into more meaningful word vectors. To this aim, Word2Vec [15] and the embeddings from language models (ELMo) [16] are tested in the proposed model for converting words to vectors. The main differences between Word2Vec and Elmo are that the former cannot easily handle words it has not seen before and it cannot capture the context while the later addressed these two drawbacks by considering the entire sentence before producing embedding vector for each word in the sentence [17]. Moreover, in order to improve the generalization of the proposed model an augmentation step is applied in which WordNet [18] is used to extend each sentence by adding the synonyms in the sentence. Finally, to exploit orders and context in text, long short-term memory LSTM is used in the model [14].

To assess the performance of the proposed model, two different types of TFTs are used; a TripAdvisor online forum data and an Ubuntu Linux distribution online forum data [8]. The technical words used in the discussion of these forum are different making the classification problem more challenging [8].

The main contributions of this study are four folds:

TABLE I. A TYPICAL EXAMPLE OF INITIAL POST WITH THREE DIFFERENT REPLIES.

Thread Title: Hotel in New York City.	
Initial-Post	We are interested in a cheap hotel in New York City. Any ideas?
High-Quality Reply	My stay was absolutely amazing, check in was flawless, I even got my room upgraded. We were close to a bunch of bars and restaurants. The room and the bathroom was very clean. I'd definitely stay here again!
Low-Quality Reply	My stay was absolutely horrendous.
Non-Quality Reply	Thank you very much. Your opinion about the costs are correct.

- We propose a deep learning model exploiting only the raw text of users' posts.
- We compare the effect of using two different word embeddings for capturing contextual information of the posts.
- We improve the proposed model by augmenting the input using WordNet.

The remainder of the paper continues as follows: Section 2 briefly discusses existing research on TFT categorization and deep neural networks. In Section 3, we introduce the proposed method for TFT quality replies. Sections 4 presents the evaluation results on two datasets of TFT texts. Finally, conclusion and future work are presented in Section 5.

2. LITERATURE REVIEW

Existing research on TFT data classification is presented in the following subsection and a brief overview of the related research on deep learning models used in text classification is presented the next subsection.

2-1. Text forum data classification

In order to improve the classification of users' replies in TFTs, researchers proposed some semantic features, recently. For example, in [18], structural features were extracted from community topics and used in inference networks. They compared three retrieval models including mean cross-ranking and mean average accuracy, and normalized discount accumulated gain on these features. They showed that recovery performance is improved when topic-based structural features are used [18]. Most studies in text classification used traditional ML algorithm for classification [9]. For example, in [19], support vector machine (SVM) is used for classifying topics of online discussion websites into high-quality and low quality. They extracted features using NLP methods with no content analysis to predict the quality of topics. Also, they tried different kinds of rating SVM models with unique features in their study [20].

Some studies proposed the use of lexical features. For example, in [21], prevalent lexical features were used to evaluate user replies in Web communities. Also, they investigated the effect of noise to improve features. To this aim, they first normalized the data and then, trained supervised support machine (SSM) on five lexical features. This study also used traditional ML algorithms such as SVM, minimum sequential optimization (MSO), multinomial naïve Bayes (MNB), and decision trees for classification. They showed that SVM outperforms other ML algorithms on all datasets [21]. In

[22], to improve topic retrieval, subjectivity prior was proposed. They showed that, for both subjective and non-subjective responses, subjective information can improve the performance of the model [22].

Some studies used user-centered features for classification. For example, in [11], to specify projects whose developer support was granted the stack overflow discussion community, users' response time and programmer participation were used. The authors claim that they were the first who assessed the quality of developer support provided by question answering systems. In a similar study, to classify specific online topic-based user posts, inference networks were proposed [23]. Inference network is an advanced probabilistic model which is used to classify posts using several features including content-based, structural, and user-centric features as well as features obtained using emotion analysis of posts. In another study [24], to assess the quality of topics, a function of several QD features was proposed. This study aimed at using quality features to retrieve topics based on a voting mechanism. They showed that quality optimization can be performed to enhance topic retrieval [24].

To expand the ability of topic retrieval systems, quality features and crowdsourcing platforms were assessed on the topic structure in [25]. They proposed several quality features including response emotion, authors activity, response structure, response type, response relevance, for selecting response quality. In order to assess the quality of replies sent to Q&A sites, stack exchange response was used [26]. Specifically, user-centric criteria and data-driven features were used in four topics of entertainment, art, science, and technology. Recently, several quality dimension features were proposed for classifying initial post replies [8]. They showed that traditional ML models including decision tree and SVM classifiers outperformed other ML classification models. An overview of the above-mentioned text forum data classification models is shown in Table II.

2-2. Deep learning methods

Deep neural networks are designed as an alternative for classic machine learning methods to learn expressive and meaningful features from large datasets for several applications including opinion mining [2], analyzing medical reviews [26], online doctor review (ODR) categorization [27, 28], predicting the helpfulness score of user reviews [28], and analyzing tweets [29].

Existing deep learning methods for textual data classification usually utilize either convolutional neural network (CNN) or recurrent neural network (RNN) [30]. Although there are several studies addressing the text classification problem using deep models [29], there are few studies that proposed a deep model for TFT text classification.

For text forum classification, most of deep learning methods were proposed to classify questions [31]. For example, in [31], a bidirectional LSTM (Bi-LSTM) network was suggested for Indonesian QA for classification of questions based on their type. They divided records into greeting, daily conversation, and meetings categories. In order to classify the questions, they evaluated LSTM, RNN, and Bi-LSTM and reported that Bi-LSTM achieved a 0.90% accuracy.

TABLE II. RELATED RESEARCH ON TFT DATA CLASSIFICATION.

Ref	***Year***	***Type***	***Description***
[21]	2010	Probabilistic	Structural features
[22]	2014	Traditional ML	Ranking
[23]	2015	Traditional ML	Lexical features
[24]	2015	Traditional ML	Topic modeling
[11]	2015	Traditional ML	cooperation features and Response time
[23]	2016	Probabilistic	Inference network
[24]	2016	Retrieval	Voting
[25]	2018	Traditional ML	Quality and service features
[26]	2019	Traditional ML	Data driven features
[8]	2019	Traditional ML	Quality of initial posts
[31]	2019	Deep neural net.	Q&A Categorization
[32]	2020	Deep neural net.	Bingali Q&A
[33]	2020	Deep neural net.	Ranking

In [32], a context-based deep learning Seq2Seq model was presented to classify Bengali questions. They utilized context and question in the encoder and related answer in the decoder model. In order to evaluate the classification system, they used 2,000 Bengali texts. The main reason they used a seq2seq model was that the Bengali language is a resource limited language and using their proposed method they were able to utilize English resources for training. Finally, in [33], a new CNN model was proposed to rank answers listed in a candidate pool of answers and select the most suitable ones. In order to improve the system's knowledge, two public datasets were used, namely, TrecQA [34] and WikiQA [35].

2-3. Research gap

In summary, most of existing methods for text forum data processing exploited traditional ML algorithms. Feature engineering in the traditional ML models make them of little use. In order to address this problem, we propose a deep learning system for classification of text forum data. Although there are some studies in the literature that utilized deep models, they were proposed for Q&A data classification. Therefore, the main novelty of current study is proposing a deep model for classification of text forum data. The proposed deep model exploits only the raw text of users' posts. Also, we improved the proposed model by augmenting the input using WordNet.

3. LONG SHORT-TERM MEMORY

Recurrent neural networks models intend to extract the dependencies of words by considering text data as an array of words [28]. However, simple recurrent neural network models do not perform well. Among many types of recurrent neural networks, LSTM is the most widely used model used to record long-term dependencies [28]. These networks address the problem of vanishing gradients in RNN [28] with the introduction of a specific cell for storing values at different time intervals, and three other gates for regulating information flow through the gate [39].

There are some differences between LSTM concepts and traditional RNNs. Specifically, control flow inside the cell is performed using the forget, update (also known as input), and the output gates. Moreover, a memory cell, c exists in the structure. Also, the network has an input from hidden memory, h, and an input, x, and produces two outputs; one output is c_t and the other is h_t which has two parts: one is passed to the next time step and the other produces output for the current time step [40]. the functions used in the above-mentioned gates are shown in (1) to (6):

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (2)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (3)$$

$$\tilde{s}_t = \tanh\,(w_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$s_t = f_t \odot s_{t-1} + i_t \odot \tilde{s}_t \quad (5)$$

$$h_t = o_t \odot \tanh(s_t) \quad (6)$$

where, σ is sigmoid function, $\odot$ represents the element-wise multiplication of the vectors, w_x is weights for input gate, h_{t-1} is the weight for previous cell's output, x_t is used as the weight for current step's input, and b_x is the bias for gate the input gate [29]. The overall view of the structure of an LSTM cell is shown in Fig. 2 [38].

4. Proposed Model

The framework of the proposed model for classifying TFT users' responses based on their quality is shown in Fig. 3. This figure has four main parts; pre-processing, data augmentation, vectorization, and deep learning-based classification. Each part of the proposed model will be described in more details in the following sections.

4-1. Pre-processing

The objective of data pre-processing is to clear the text of users' replies. To this aim, punctuations, URLs, and irrelevant characters are removed from the texts of each user reply which previously converted to lowercase. After that, answers are tokenized and stop words are removed. Finally, stemming using Snowball Stemmer which is also known as the Porter2 stemming algorithm is used for stemming. This decreases the size of the vocabulary and the resulting vector in the next step.

4-2. Data Augmentation

In the current study, we used WordNet for augmenting user replies. The reason for using the WordNet is to amplify the replies by adding synonyms of each word. This step can expand the model's performance because the datasets used in the current study have small number of training data. Also, the generality of the proposed deep model is increased.

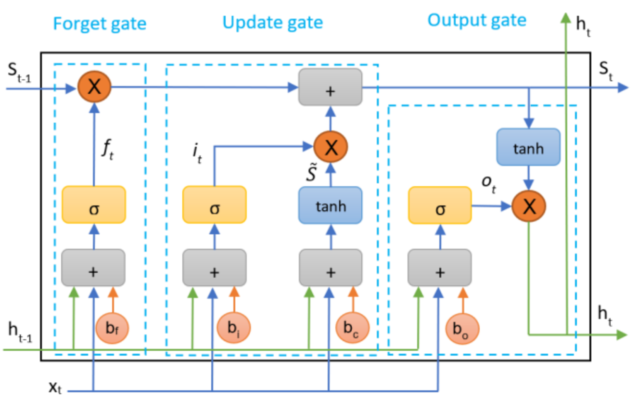


Fig. 2. The overall structure of an LSTM cell reproduced from [40].

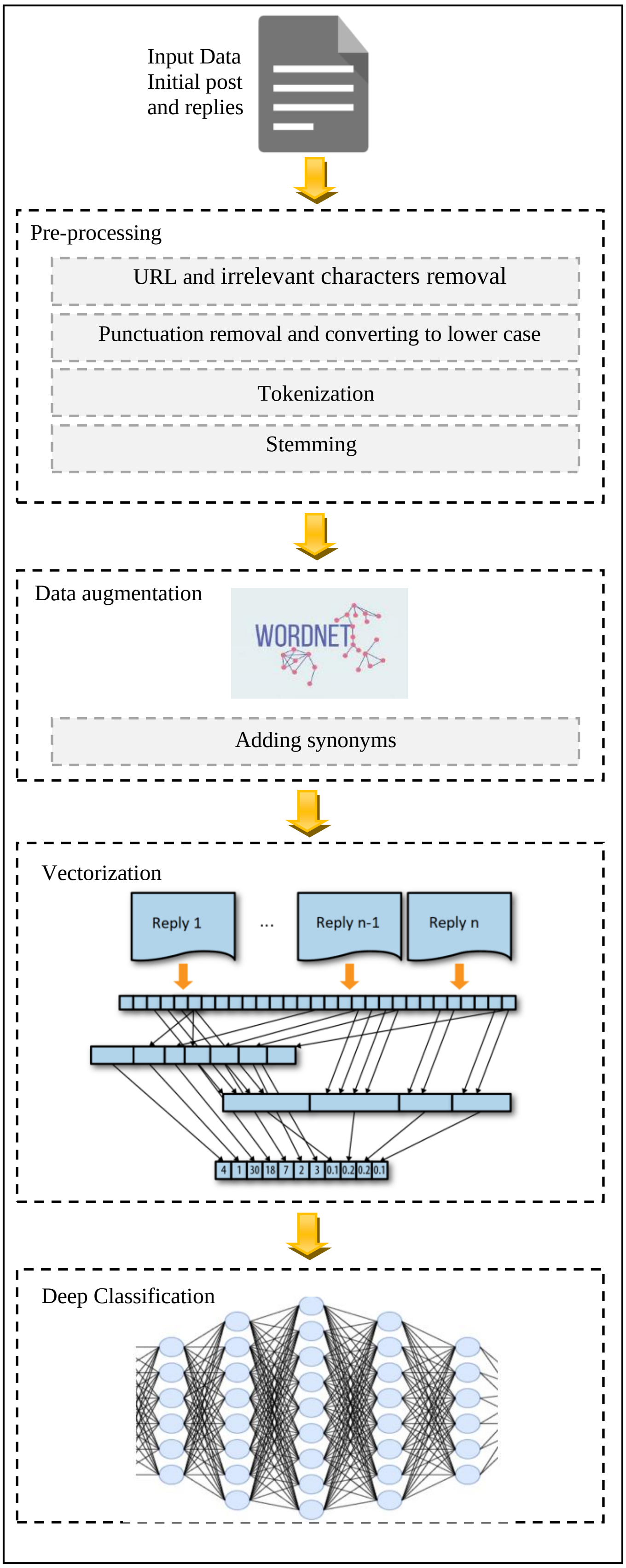


Fig. 3. Overall structure of a the proposed model.

4-3. Vectorization

Vectorization is the process of converting text into numerical representation [36]. In the current study, each user reply is converted to a numerical vector using ELMo [15]. ELMo is a character-based contextual embedding with application in many textx mining and deep learning applications [15]. Despite Word2vec [16], in ELMo, every word has a different meaning based on the context. This makes ELMo more data-sensitive than Word2Vec.

4-4. Deep Learning for Classification

An LSTM deep neural network is used in the proposed model for the classification of the replies (See Fig. 4). In the LSTM layer of the proposed model, 100 LSTM cells are used to process the text sequences. The outputs of these cells are sent to a fully connected layer which extracts meaningful features from the output of the LSTM layer.

5. EXPERIMENTS AND RESULTS

5-1. Dataset

Two datasets of the TripAdvisor online forum for New York City (NYC) and the Ubuntu online Linux distribution forum are used in the current study [8]. These datasets were manually labeled into three classes of non-quality or irrelevant, low-quality, and high-quality answers. The first dataset contains discussions related to travel planning to New York city, and the Ubuntu dataset contains discussions about different aspects of Ubuntu Linux.

The Ubuntu dataset was collected from a technical discussion forum and the NYC dataset was gathered from a general discussion forum. The Ubuntu dataset contains 726 user responses and the NYC dataset has 312 replies classified into the above-mentioned three classes. Each class in the dataset is specified with one of 1, 2 and 3 digits. For the Ubuntu dataset, there are 92, 217, and 417 samples in classes 1, 2, and 3, respectively. In the NYC dataset, 85, 102, and 125 samples belong to classes 1, 2, and 3, respectively.

5-2. Experimental setup

In order to carry out the experiments we implemented the proposed system and other methods using Keras library of Python language in Google Colab environment [40]. The machine on which the experiments were conducted used a Tesla K80 GPU with 12GB GDDR5 VRAM and two 2.00GHz Intel(R) Xeon(R) CPUs with 6MB cache, and 13GB RAM.

In the proposed model and the GRU model, the output layer consists of three SoftMax cells and a spatial dropout with rate 0.2 was used. In the compile time, Adam optimizer with accuracy and categorical cross-entropy was used. The parameters used in the deep models are shown in Table III.

Four traditional ML classifiers including support vector classifier (SVC), Naïve Bayes (NB), logistic regression (LR), artificial neural network (ANN), and a similar deep neural network method (i.e., GRU) were compared with the proposed model, in the experiments.

In the implementation of the traditional machine learning methods, default Scikit-learn parameters were used. Specially, for ANN, three hidden layers, each with two neurons using

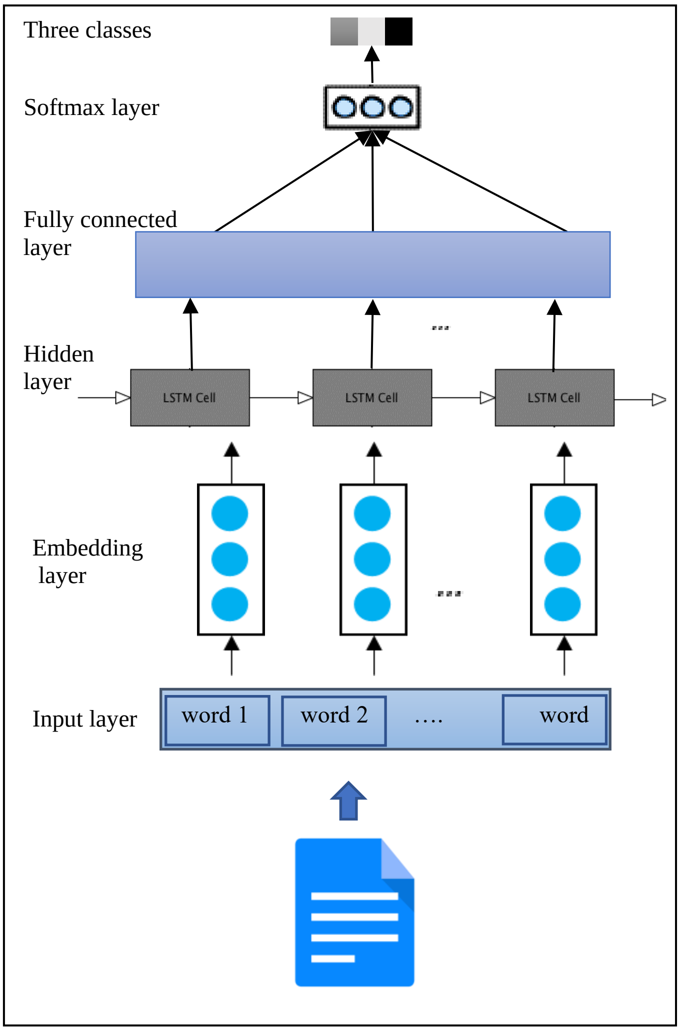


Fig. 4. Block-box diagram of the proposed model.

TABLE III. PARAMETER SETTINGS OF THE PROPOSED MODEL.

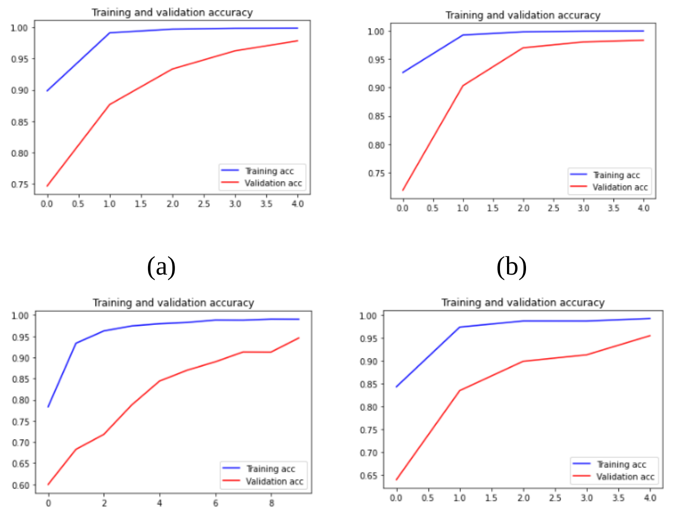
Deep Model	Parameters		
	Number of cells	***dropout rate***	***recurrent dropout***
LSTM	100	0.2	0.2
GRU	100	0.25	0.2

Adam solver and Relu activation function, for SVC, RBF kernel with gamma set to 10, for NB, multinomial function, and for LR, multinomial with LBFGS solver are used.

5-3. Results

The comparison of the accuracy of the proposed method with GRU on both datasets using the Elmo and Word2Vec embeddings is depicted in Fig. 5 and Fig. 6.

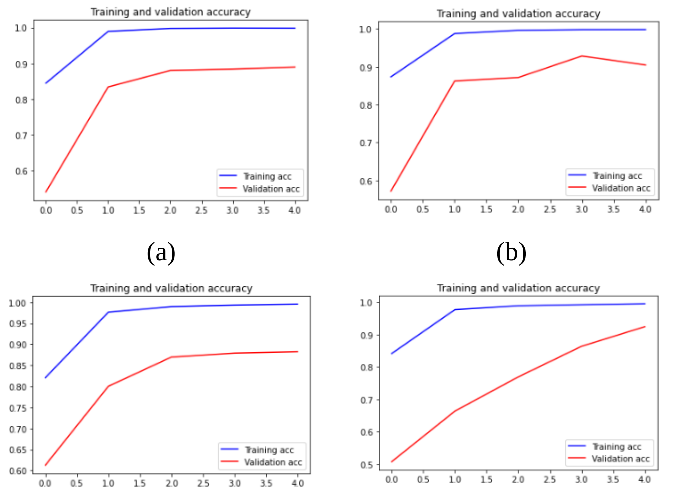
As shown in the figures, both embeddings perform well with the proposed deep model but, for the Elmo diagram, there is more convergence between training and test data lines from a point onwards. In order to better compare the methods and to assess the effect of data augmentation, Tables IV to VII compare the proposed method and the GRU model in terms of accuracy, loss, precision, recall, and F1-measure on both datasets using both embeddings described in the previous subsection.



(a) (b)

(c) (d)

Fig. 5. Comparison of the accuracy of the ELMo embedding in the proposed method (a) with the GRU method (b), and the Word2Wec embeddings in the proposed method (c) and the GRU method (d) on the Ubuntu dataset.



(a) (b)

(c) (d)

Fig. 6. Comparison of the accuracy of the Elmo embedding in the proposed method (a) with the GRU method (b), and the Word2Wec embeddings in the proposed method (c) and the GRU method (d) on the NYC dataset.

As shown in Tables IV to VII, in most settings, the LSTM model outperforms the GRU model. The reason for this can be the difference in internal mechanism of the models. Moreover, ELMo performs better than Word2vec in Ubuntu dataset but on NYC dataset, Word2vec performs better. This can be the result of structural differences of sentences in datasets. Finally, we can observe that the data augmentation (Tables V and VII) can improve the model significantly. This may be due to the fact that WordNet synonyms improve the generality as the model saw more words. In other words, data this augmentation step reduces the chance of finding new words in the test phase.

To show the utility of using deep learning models, the best results obtained by the algorithms are compared in Fig. 7 and Fig. 8. It should be noted that common features including

TABLE IV. COMPARISON OF ACCURACY AND LOSS FOR THE PROPOSED METHOD AND THE GRU MODEL WITHOUT DATA AUGMENTATION.

Embedding / *dataset*		*Accuracy*		*Loss*	
		ELMo	***Word2vec***	***ELMo***	***Word2vec***
Ubuntu	***LSTM***	0.60	0.61	0.93	**0.80**
	GRU	**0.62**	0.60	0.87	0.84
NYC	***LSTM***	0.50	**0.54**	0.96	**0.87**
	GRU	0.32	0.37	1.10	1.03

TABLE V. COMPARISON OF ACCURACY AND LOSS FOR THE PROPOSED METHOD AND THE GRU MODEL WITH DATA AUGMENTATION.

Embedding / *dataset*		*Accuracy*		*Loss*	
		ELMo	***Word2vec***	***ELMo***	***Word2vec***
Ubuntu	***LSTM***	**0.96**	0.96	0.02	**0.01**
	GRU	0.95	0.91	0.06	0.1
NYC	***LSTM***	0.92	**0.93**	**0.02**	0.05
	GRU	0.91	0.93	0.1	0.03

TABLE VI. COMPARISON OF PRECISION, RECALL, AND F1-MEASURE OF THE PROPOSED METHOD WITH THE GRU MODEL WITHOUT DATA AUGMENTATION.

Embedding / *dataset*		*Precision*		*Recall*		*F1*	
		ELMo	***Word2vec***	***ELMo***	***Word2vec***	***ELMo***	***Word2vec***
Ubuntu	***LSTM***	**0.54**	0.48	**0.50**	0.49	**0.50**	0.43
	GRU	0.54	0.45	0.49	0.49	0.50	0.42
NYC	***LSTM***	0.48	**0.54**	0.38	**0.54**	0.32	**0.53**
	GRU	0.41	0.42	0.43	0.46	0.36	0.42

TABLE VII. COMPARISON OF PRECISION, RECALL, AND F1-MEASURE OF THE PROPOSED METHOD WITH THE GRU MODEL WITH DATA AUGMENTATION.

Embedding / *dataset*		*Precision*		*Recall*		*F1*	
		ELMo	***Word2vec***	***ELMo***	***Word2vec***	***ELMo***	***Word2vec***
Ubuntu	***LSTM***	**0.95**	0.94	0.94	0.94	**0.95**	0.94
	GRU	0.95	0.94	0.94	**0.95**	0.95	0.94
NYC	***LSTM***	**0.93**	0.92	**0.93**	0.91	**0.93**	0.91
	GRU	0.93	0.92	0.92	0.91	0.92	0.91

unigram, bigram, and trigram are tested and the best results were reported in the figures for traditional ML algorithms. Considering Fig. 7, an important observation is that traditional ML algorithms, in almost all cases, perform than the proposed model in the absence of data augmentation. This confirms that traditional ML algorithms can obtain higher test accuracy than

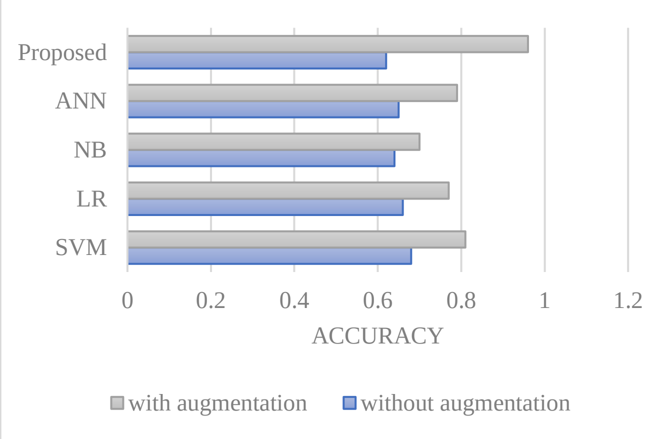


Fig. 7. Comparison of the best accuracy obtained using the proposed model and four traditional ML algorithms on Ubuntu dataset.

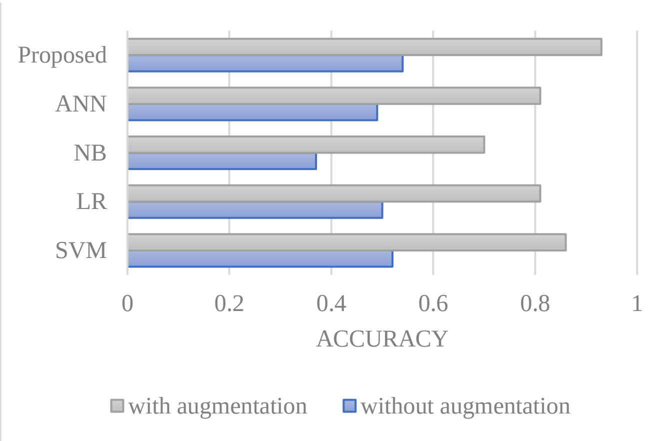


Fig. 8. Comparison of the best accuracy obtained using the proposed model and four traditional ML algorithms on NYC dataset.

the proposed method when less data is provided. However, the proposed deep model outperformed all traditional models using augmentation.

Finally, to compare the obtained results with those reported in [8] using traditional ML methods with quality dimension features and only textual features, Table VIII reports the best results obtained using these three models. The first point in Table VIII is that traditional ML methods with quality dimension features perform better compared to when they use only textual features (i.e., n-gram features). The second observation is that the proposed model performs better than ML methods using both text-based and quality dimension features. This emphasizes the ability of deep models in automatically extracting meaningful features from textual data.

6. Conclusion

With the rapid improvement of Web services, users can find the answers to most of their questions in general and technical text forum threads. These threads are invaluable public data for users who want to exploit the experience of other users to answer to their questions. However, due to the fast growth of users and their comments, finding high-quality replies is turned into a challenging problem. In order to find such high-quality replies, machine learning algorithms have been utilized in previous studies. These algorithms usually extract features from the replies and use them for classification. Recently, semantic and quality of dimension features have also been used for improving the quality of the classification system. However, the main drawback of this approach is that feature engineering is a domain-dependent and labor-intensive task. In the current study, we proposed a deep learning model based on ELMo word embedding and LSTM neural network. The proposed model neither needs explicit feature extraction nor depends on the data for designing the feature space. Moreover, the results of experiments showed that, the proposed method with only text of user replies, achieves better accuracy compared to that of traditional ML methods. This achievement by the proposed model is more important when it significantly outperformed machine learning algorithm that utilized relevancy, author activeness, timeliness, ease-of-understanding, politeness, and amount-of-data features in addition to the text of replies. Improving the proposed model by using other types of word embeddings and deep neural networks may be considered as a future work. Another line for the future work can be fusion of traditional and deep models for enhancing the system.

TABLE VIII. Comparison of best results obtained using the proposed model with text-based and quality dimensions features.

Dataset / *Feature type*	*Ubuntu*	*NYC*
n-gram	0.68	0.52
Quality dimension	0.80	0.77
Proposed deep model	**0.96**	**0.93**

References

[1] Ullah, Fahim, and Samad ME Sepasgozar. "Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data." Sustainability 12, no. 11 (2020): 4382.

[2] Abdar, Moloud, Mohammad Ehsan Basiri, Junjun Yin, Mahmoud Habibnezhad, Guangqing Chi, Shahla Nemati, and Somayeh Asadi. "Energy choices in Alaska: Mining people's perception and attitudes from geotagged tweets." Renewable and Sustainable Energy Reviews 124 (2020): 109781.

[3] Laitala, Kirsi, Ingun Grimstad Klepp, Vilde Haugrønning, Harald Throne-Holst, and Pål Strandbakken. "Increasing repair of household appliances, mobile phones and clothing: Experiences from consumers and the repair industry." Journal of Cleaner Production 282 (2021): 125349.

[4] Aderibigbe, Semiyu Adejare. "Online Discussions as an Intervention for Strengthening Students' Engagement in General Education." Journal of Open Innovation: Technology, Market, and Complexity 6, no. 4 (2020): 98.

[5] Basiri, Mohammad Ehsan, Arman Kabiri, Moloud Abdar, Wali Khan Mashwani, Neil Y. Yen, and Jason C. Hung. "The effect of aggregation methods on sentiment classification in Persian reviews." Enterprise Information Systems 14, no. 9-10 (2020): 1394-1421.

[6] Basiri, Mohammad Ehsan, Moloud Abdar, Arman Kabiri, Shahla Nemati, Xujuan Zhou, Forough Allahbakhshi, and Neil Y. Yen. "Improving sentiment polarity detection through target identification." IEEE Transactions on Computational Social Systems 7, no. 1 (2019): 113-128.

[7] Sarrouti, Mourad, and Said Ouatik El Alaoui. "SemBioNLQA: a semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions." Artificial intelligence in medicine 102 (2020): 101767.

[8] A. Osman, N. Salim, and F. Saeed, "Quality dimensions features for identifying high-quality user replies in text forum threads using classification methods," *PloS one*, vol. 14, no. 5, p. e0215516, 2019.

[9] Nemati, Shahla, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Y. Yen, and Vladimir Makarenkov. "A hybrid latent space data fusion method for multimodal emotion recognition." IEEE Access 7 (2019): 172948-172964.

[10] Nemati, Shahla, and Ahmad Reza Naghsh-Nilchi. "Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video retrieval." In 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 222-228. IEEE, 2017.

[11] Squire, Megan. "Should We Move to Stack Overflow?" Measuring the Utility of Social Media for Developer Support." In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, vol. 2, pp. 219-228. IEEE, 2015.

[12] Sun, Jianling, Hui Gao, and Xiao Yang. "Towards a quality-oriented real-time web crawler." In International Conference on Web Information Systems and Mining, pp. 67-76. Springer, Berlin, Heidelberg, 2010.

[13] Weimer, Markus, and Iryna Gurevych. "Predicting the perceived quality of web forum posts." In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP), pp. 643-648. 2007.

[14] Wanas, Nayer, Motaz El-Saban, Heba Ashour, and Waleed Ammar. "Automatic scoring of online discussion posts." In Proceedings of the 2nd ACM Workshop on Information Credibility on the Web, pp. 19-26. 2008.

[15] Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[16] N. Reimers and I. Gurevych, "Alternative weighting schemes for elmo embeddings," *arXiv preprint arXiv:1904.02954*, 2019.

[17] Zaidi, Syed Farhan Alam, Faraz Malik Awan, Minsoo Lee, Honguk Woo, and Chan-Gun Lee. "Applying Convolutional Neural Networks With Different Word Representation Techniques to Recommend Bug Fixers." IEEE Access 8 (2020): 213729-213747.

[18] F. Bond and K. Paik, "A survey of wordnets and their licenses," *Small*, vol. 8, no. 4, p. 5, 2012.

[19] S. 6atia and P. Mitra, "Adopting inference networks for online thread retrieval," in Proceedings of the AAAI Conference on Artificial Intelligence, 2010, vol. 24, no. 1.

[20] J.-T. Lee, M.-C. Yang, and H.-C. Rim, "Discovering high-quality threaded discussions in online forums," Journal of Computer Science and Technology, vol. 29, no. 3, pp. 519–531, 2014.

[21] Obasa, Adekunle Isiaka, Naomie Salim, and Atif Khan. "Enhanced lexicon based model for web forum answer detection." In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC), pp. 237-243. IEEE, 2015.

[22] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra, "Using subjectivity analysis to improve thread retrieval in online forums," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9022, pp. 495–500.

[23] S. Bhatia, P. Biyani, and P. Mitra, "Identifying the role of individual user messages in an online discussion and its use in thread retrieval," *Journal of the Association for Information Science and Technology*, vol. 67, no. 2, pp. 276–288, Feb. 2016.

[24] A. Heydari, M. Tavakoli, Z. Ismail, and N. Salim, "Leveraging quality metrics in voting model based thread retrieval," World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 10, no. 1, pp. 117–123, 2016.

[25] A. Osman, N. Salim, F. Saeed, and I. Abdelhamid, "Quality features for summarizing text forum threads by selecting quality replies," in *Advances in Intelligent Systems and Computing*, 2019, vol. 843, pp. 47–56.

[26] Fu, Hengyi, and Sanghee Oh. "Quality assessment of answers with user-identified criteria and data-driven features in social Q&A." Information Processing & Management 56, no. 1 (2019): 14-28.

[27] S. M. Jiménez-Zafra, M. T. Mart in-Valdivia, M. D. Molina-González, and L. A. Ureña-López, "How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain," *Artificial intelligence in medicine*, vol. 93, pp. 50–57, 2019.

[28] A. Shukla, W. Wang, G. G.- Ritu, C. M. I. Y. Can, and undefined 2019, "Catch me if you can—Detecting fraudulent online reviews of doctors using deep learning," *papers.ssrn.com*.

[29] M. Parimala, R. M. Swarna Priya, M. Praveen Kumar Reddy, C. Lal Chowdhary, R. Kumar Poluru, and S. Khan, "Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach," in *Software - Practice and Experience*, 2020.

[30] Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep Learning--based Text Classification: A Comprehensive Review." ACM Computing Surveys (CSUR) 54, no. 3 (2021): 1-40.

[31] R. Anhar, T. Adji, N. S.-2019 5th International, and undefined 2019, "Question Classification on Question-Answer System using Bidirectional-LSTM," *ieeexplore.ieee.org*.

[32] Keya, Mumenunnessa, Abu Kaisar Mohammad Masum, Bhaskar Majumdar, Syed Akhter Hossain, and Sheikh Abujar. "Bengali Question Answering System Using Seq2Seq Learning Based on General Knowledge Dataset." In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2020.

[33] Wang, Dong, Ying Shen, and Hai-Tao Zheng. "Knowledge Enhanced Latent Relevance Mining for Question Answering." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4282-4286. IEEE, 2020.

[34] S. Gu, X. Luo, H. Wang, J. Huang, Q. Wei, and S. Huang, "Improving answer selection with global features," *Expert Systems*, vol. 38, no. 1, Jan. 2021.

[35] Y. Yang, W.-T. Yih, and C. Meek, "WIKIQA: A Challenge Dataset for Open-Domain Question Answering," Association for Computational Linguistics, 2015.

[36] Basiri, Mohammad Ehsan, and Arman Kabiri. "HOMPer: A new hybrid system for opinion mining in the Persian language." Journal of Information Science 46, no. 1 (2020): 101-117.

[37] G. Rao, W. Huang, Z. Feng, and Q. Cong, "LSTM with sentence representations for document-level sentiment classification," *Neurocomputing*, vol. 308, no. 1, pp. 49–57, 2018.

[38] Jiang, Changhui, Yuwei Chen, Shuai Chen, Yuming Bo, Wei Li, Wenxin Tian, and Jun Guo. "A mixed deep recurrent neural network for MEMS gyroscope noise suppressing." Electronics 8, no. 2 (2019): 181.

[39] Manaswi, Navin Kumar. "Understanding and working with Keras." In Deep Learning with Applications Using Python, pp. 31-43. Apress, Berkeley, CA, 2018.

Masoumeh Rajabi received her B.S. degree in software engineering from Arak university in 2015 and her M.S. from Shahrekord University in 2021. Her research interest includes natural language processing, deep learning, and data mining.

Shahla Nemati was born in Shiraz, Iran in 1982. She received the B.S. degree in hardware engineering from Shiraz University, Shiraz, Iran, in 2005, the M.S. degree from Isfahan University of Technology, Isfahan, Iran, in 2008, and the Ph.D. degree in computer engineering from Isfahan University, Isfahan, Iran, in 2016. Since 2017, she has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. Her research interests include data fusion, affective computing, and data mining.

Mohammad Ehsan Basiri received the B.S. degree in software engineering from Shiraz University, Shiraz, Iran, in 2006 and the M.S. and Ph.D. degrees in Artificial Intelligence from Isfahan University, Isfahan, Iran, in 2009 and 2014. Since 2014, he has been an Assistant Professor with the Computer Engineering Department, Shahrekord University, Shahrekord, Iran. He is the author of three books and more than 60 articles. His research interests include sentiment analysis, natural language processing, deep learning, and data mining.