IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 May 2024, accepted 14 June 2024, date of publication 19 June 2024, date of current version 2 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3416755

RESEARCH ARTICLE

Enhanced Transformer-BLSTM Model for Classifying Sentiment of User Comments on Movies and Books

YUN LIN[1] **AND TUNDONG LIU**[2]
[1]School of Information Science and Technology, Tan Kah Kee College, Xiamen University, Xiamen 363105, China
[2]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen 361005, China

Corresponding author: Yun Lin (linyun@xujc.com)

ABSTRACT Classifying the sentiment of user comments on a website is a crucial task within Natural Language Processing (NLP). Conducting sentiment analysis can aid businesses in gaining a more profound comprehension and examination of users' emotional inclinations towards products or services. This study introduces a sentiment classification model that combines the Transformer and BLSTM architectures to analyze the sentiment of user comments on movie and book websites. By incorporating the strengths of both Transformer and BLSTM, the proposed model mitigates the issue of vanishing gradient by scrutinizing inputs within a long-term context using BLSTM. It employs the multi-head attention mechanism of the Transformer to extract features and capture significant semantic details within the comments. Furthermore, the joint model combines the TF-IDF weights with the vector space, which improves the embedding process. The proposed model's effectiveness was evaluated by categorizing the sentiment of user comments on publicly available datasets containing more than 20,000 movie and book comments. The results indicate that the proposed model is superior to LSTM and CNN in sentiment classification tasks. Moreover, the proposed approach has demonstrated significant improvements, particularly in the training set, achieving an accuracy of 93.81%.

INDEX TERMS NLP, sentiment analysis, transformer, BLSTM, TF-IDF.

I. INTRODUCTION

Our current age is marked by an explosion of data due to internet activities such as social networks, emails, blogs, and news. According to the International Data Corporation (IDC), the world's data is expected to surpass 175 zettabytes by 2025 [1]. As a result, sorting these documents based on sentiments has become a hot topic. Especially classifying the sentiment of user comments on websites is an important task that helps understand users' attitudes and emotional tendencies.

Automatic document categorization (ADC) has gained significant attention in recent years. NLP plays a crucial role in this field. NLP is a sub-field of Artificial Intelligence (AI) that enables computers to understand human languages. The NLP model has been applied in various fields, including topic labeling [2] and sentiment classification [3], [4].

Using NLP models has helped many researchers achieve satisfactory performance in NLP tasks [5], [6], [7], [8]. Previously, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were two common neural-network-based approaches for sentence modeling. CNNs work well in extracting n-gram features at different sentence positions and reducing frequency variations. However, they cannot handle sequential correlation [9]. On the other hand, RNNs are good at sequential modeling, but they fail to efficiently extract features from backward context and have a problem of vanishing gradient [10]. Furthermore, both approaches require high computational costs and a lot of training data.

Bidirectional Recurrent Neural Networks (BRNNs) are a type of neural network that uses recurrent layers to capture the temporal and contextual information of data. They can enhance the model's expressive power and predictive ability by utilizing memory and feedback [11], [12]. However, BRNNs require the entire data sequence to predict a specific

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

location. In building a speech recognition system, the complete sentence must be spoken before the system can process it.

Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) are an effective way to avoid the vanishing gradient problem in RNNs. Previous studies have used LSTMs in language modeling and found that they provide significant speed-ups in training and testing tasks when combined with clustering methods, with only a tiny loss [13], [14]. Although the gradient vanishing/exploding problem in RNNs has been partially addressed by LSTMs, they are still insufficient for handling sequences of 1000 or longer.

Bi-directional LSTM (BLSTM) models are suitable for analyzing sentiment over longer sequences. They can learn long-distance dependencies and utilize future and historical context simultaneously [15], [16]. Combined architectures have also shown promising results in NLP tasks [17], [18]. Aldalbahi et al. [16] proposed a scheme that combines BRNN and LSTM to achieve fast initial access times. This eliminates the need for beam scanning, leading to ultra-low access times and energy efficiencies compared to other existing methods. However, the computation cost of BLSTM can be high, especially for long sequences.

Transformer-based Pretrained Language Models (T-PTLM) are powerful tools that can learn language representations from vast amounts of unlabeled text data and apply this knowledge to different tasks. They have shown remarkable success in NLP and have outperformed traditional machine learning models in various studies, as demonstrated by Koroteev, Nath et al., and Topal et al. [19], [20], [21]. One of the most popular and effective T-PTLM is BERT, which has proven to be more accurate than conventional machine learning models [22], [23].

This paper presents a combined model of Transformer and BLSTM, designed to analyze the sentiment of user comments on movies and books. The model uses the Transformer's attention mechanism to extract character features and incorporates a BLSTM layer in each Transformer block to capture sentence dependencies. Additionally, the model utilizes TF-IDF to generate soft probabilistic weights, which can better illustrate relations in different dimensions. The model was evaluated through sentiment classification tasks using over 20,000 training and testing documents, achieving a maximum accuracy of 93.81%.

The main contributions of our paper are as follows:

- We proposed a combined model (Transformer-BLSTM) for analyzing the sentiment of user comments on websites. It can help businesses understand users' attitudes and emotional tendencies towards products or services, thereby adjusting and improving marketing strategies. It enables companies to make informed decisions and improvements accordingly.
- By combining the Transformer and BLSTM, the proposed method analyzes the inputs in a long-term context and avoids the problem of vanishing gradient. It uses the multi-head attention mechanism for feature extraction and captures important semantic information within the text, which improves the model's ability to analyze semantics. In testing tasks, the model achieves a maximum accuracy of 93.81%.
- TF-IDF weights are added into the combined Transfomer and BLSTM system. It will precisely evaluate the importance of a word in the corpus, which enhances the accuracy and efficiency of document classification. Experimental results show that the performance significantly differs with or without TF-IDF weights.
- The proposed scheme is implemented and evaluated on public datasets that contain more than 20,000 movie and book comments. Results are compared to conventional models, which show the effectiveness of the proposed method.

The structure of the paper is as follows: First, we introduce our proposed model in Section I. Then, we discuss previous works that inspired us to undertake this work in Section II. Next, we describe the system architecture and evaluate the proposed model in Section III and Section IV. Following the assessment of the proposed model's performance, we provide a summary and discuss future work in Section V.

II. RELATED WORK

Sentiment analysis has been a vibrant field within NLP, with numerous approaches proposed to understand and categorize the emotional tone of textual data. This section reviews the literature most pertinent to our study on enhancing the Transformer-BLSTM model for sentiment classification.

Early sentiment analysis research relied heavily on traditional machine-learning techniques. Features such as bag-of-words, TF-IDF, and part-of-speech tags were commonly extracted to train classifiers like Naive Bayes, Support Vector Machines (SVM), and Random Forests [24], [25]. However, these methods often require handcrafted features and do not capture the context and semantic relationships within the text as effectively as modern deep-learning approaches.

The advent of deep learning has revolutionized sentiment analysis. CNNs and RNNs, particularly LSTM, have been widely used due to their ability to capture local and sequential dependencies within text [5], [26], [27]. For example, Patel et al. [28] applied the deep learning-based classification algorithm RNN, measured the classifier's performance based on data pre-processing, and obtained 94.61% accuracy. In the work of [29], they applied LSTM to perform Twitter sentiment analysis. Experiment results show that the combined LSTM-CNN model performs well with the highest accuracy of 87%. Despite their successes, these models struggle with vanishing gradients and are not adept at handling long-range dependencies.

BRNNs and attention mechanisms have been introduced to address some of these limitations. BRNNs process text in forward and reverse orders, providing a more comprehensive understanding of context [11]. Attention mechanisms, however, allow models to focus on different parts of the input sequence, thus improving the capture of contextual information [30]. In the work of [31], they proposed Attention-Based Bidirectional Long Short-Term Memory Networks(AttBLSTM) to capture the most essential semantic information in a sentence. The experimental results show that the method outperforms most of the existing methods.

The Transformer model, introduced by Roy et al. [32], has become a cornerstone in NLP tasks due to its self-attention mechanism, which efficiently processes sequences of varying lengths and captures global dependencies. BERT, a pre-trained Transformer model, has set new standards in language understanding by pretraining on large corpora and fine-tuning for specific tasks [33].

Recent research has explored hybrid architectures, recognizing the complementary strengths of different models. For instance, some studies have combined CNNs and RNNs to leverage CNNs' ability to capture local features and RNNs' capacity for sequential data [34]. Similarly, integrating Transformer models with RNNs, particularly LSTM, has shown promising results in tasks such as text classification and sentiment analysis [35].

TF-IDF is known for its simplicity and effectiveness in highlighting the significance of terms within documents; however, it has been a topic of debate among researchers. Critics argue that TF-IDF may not fully capture the semantic richness of the text, particularly in sentiment analysis, where the subtle nuances of words are crucial. This limitation is due to TF-IDF's inability to reflect the contextual and relational aspects of language, which are often important in determining sentiment. Nevertheless, when combined with deep learning models, TF-IDF has been shown to enhance feature representation and improve classification performance [36].

Our proposed model builds upon these foundations by integrating the Transformer's multi-head attention mechanism with the BLSTM's ability to handle long-term dependencies. Additionally, we introduce TF-IDF weights into our model to better capture the semantic significance of terms within the context of sentiment analysis. This hybrid approach aims to overcome the limitations of individual models and provide a more nuanced understanding of user comments on movies and books. Experiments involved sentiment classification using over 20,000 documents, achieving a maximum accuracy of 93.81%.

III. METHODOLOGY

A. SYSTEM ARCHITECTURE

The system architecture is depicted in Fig. 1. To preprocess the text data, the model first uses TF-IDF before converting it into continuous vector representations. These vectors are then input into a BLSTM network. In addition, character-level features are extracted using a Transformer and a max-pooling layer.

In our study, the Transformer-BLSTM model we propose efficiently analyzes the sentiment of user comments by using specific technical parameters and structural optimizations. We start by preprocessing the data using the TF-IDF method, with a vocabulary size of 5,000, to cover the main words used in the comments. The TF-IDF vectors are then normalized using the $L2$ method to ensure consistency across different documents.

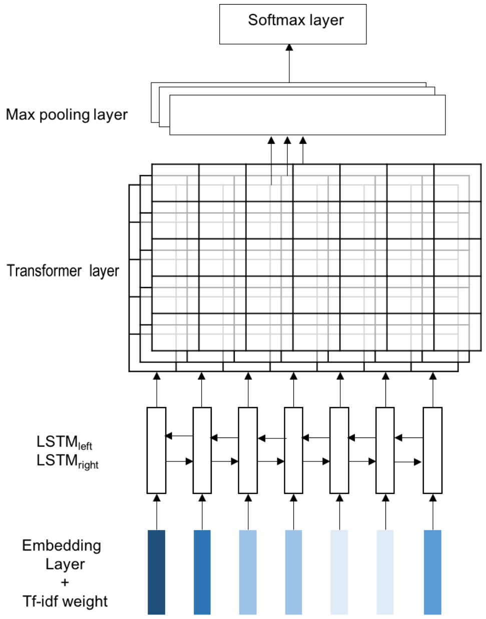

FIGURE 1. The system architecture of the proposed model.

The BLSTM network includes a hidden layer with 512 units, consisting of two stacked BLSTM layers to capture long-term dependencies within the reviews. The initial learning rate is set to 0.001, with a decay rate of 0.1 to adjust the learning rate during training. The batch size is set to 64, considering our computational resources to balance memory usage and training efficiency.

The Transformer layer is configured with eight attention heads to process different aspects of the input text in parallel, with each head operating over a dimensionality of 64. A feed-forward neural network (FFNN) with 2,048 hidden units further extracts deep-level features. Each Transformer block incorporates residual connections and layer normalization to enhance the stability of the training process.

We merge the outputs from both the BLSTM and Transformer models by concatenating them along the feature dimension, where the output sequences from the BLSTM and the pooled features from the Transformer's attention layers are combined into a single comprehensive vector. This unified vector captures the temporal dynamics provided by BLSTM and the contextual information extracted by the Transformer, enhancing the model's ability to process and classify sentiment in user comments. After this, a fully connected layer with 256 units and a Softmax layer is employed to normalize the output probabilities across different classes, which is essential for sentiment classification. The Softmax function creates a probabilistic distribution over the possible sentiment categories by exponentiating its input and dividing it by the sum of exponentiated inputs for each class. During training, the performance of the classification is measured using the cross-entropy loss function. To enhance the model's generalization, we use a dropout rate of 0.5 as a regularization technique and implement early stopping to prevent overfitting. Additionally, hyperparameter tuning is performed using random search along with 5-fold cross-validation to ensure the model's performance on unseen data. This approach helps to optimize the model and improve its accuracy in classifying sentiments.

B. BLSTM

In order to address the vanishing gradient problem in RNNs, our model incorporates a BLSTM layer after the embedding layer. The BLSTM uses memory cells with input, output, and forget gates to manage information flow and maintain sequence dependencies effectively.

The BLSTM consists of two sub-layers: a forward layer and a reverse layer, as shown in Fig. 2. The forward layer processes input vectors to generate the hidden vector sequence $\overrightarrow{h}$, while the reverse layer computes the sequence $\overleftarrow{h}$ in the opposite direction. The final output y is a combination of both sequences, allowing the model to leverage both historical and prospective contexts for sequence prediction.

The equations are as follows:

$$\overrightarrow{h} = \boldsymbol{H}(W_{x,\overrightarrow{h}} x_i + W_{\overrightarrow{h},\overrightarrow{h}} \overrightarrow{h}_{i+1} + b_{\overrightarrow{h}}) \tag{1}$$

$$\overleftarrow{h} = \boldsymbol{H}(W_{x,\overleftarrow{h}} x_i + W_{\overleftarrow{h},\overleftarrow{h}} \overleftarrow{h}_{i+1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_i = W_{\overrightarrow{h},y} \overrightarrow{h}_i + W_{\overleftarrow{h},y} \overleftarrow{h}_i + b_y \tag{3}$$

Here, $\boldsymbol{H}$ denotes the sigmoid function, and W and b represent the weight matrices and bias vectors, respectively, which are integral to the gate operation within each memory cell. This bidirectional processing enhances the model's ability to capture long-distance dependencies and contributes to its superior performance in sentiment analysis tasks. The BLSTM's architecture ensures that the network can make informed predictions by considering the entire sequence rather than just a unidirectional view of the data.

By focusing on these critical components, the revised paragraph provides a concise yet comprehensive overview of how the BLSTM layer contributes to the sentiment analysis model's architecture and functionality.

C. FEATURE EXTRACTION

We utilize a Transformer layer to extract character features from the input sentence. The layer comprises multiple encoder blocks, each with a multi-head attention layer and an FNN. Character embeddings are randomly initialized and transformed into a 25-dimensional feature. The Transformer layer offers parallel computing advantages and does not require a deep network for analyzing long sentences.

The results then pass through a max-pooling layer, which reduces the dimension of the extracted feature vector, preserving essential information for prediction. Finally, the vectors are input into a Softmax layer to calculate expected losses during the training process.

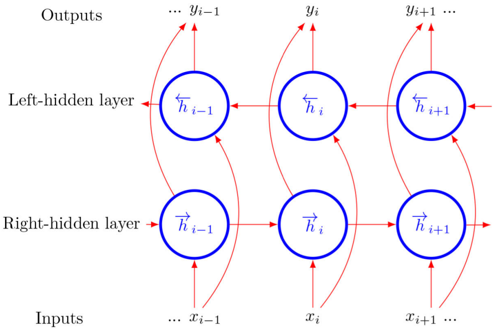

FIGURE 2. The architecture of the BLSTM model.

D. TF-IDF

TF-IDF is a pivotal technique for assessing term significance in text data, combining Term Frequency (TF) and Inverse Document Frequency (IDF) to highlight the importance of words within documents relative to a corpus.

TF quantifies the word frequency in a document, normalized by the document's length to account for varying document sizes, as shown in equation (4):

$$tf_{i,j} = \frac{w_{i,j}}{\sum_k w_{k,j}} \tag{4}$$

IDF adjusts for the overall document frequency of a word, favoring those that are rare across the corpus, as defined in equation (5):

$$idf_{i,j} = \frac{|D|}{1 + |\{j : t_i \in d_j\}|} \tag{5}$$

The TF-IDF weight is computed by multiplying the TF and IDF values, which emphasizes terms that are frequent in a document but rare across documents. This approach ensures that common words are down-weighted while meaningful terms are highlighted, thus enhancing the feature representation for our model.

IV. IMPLEMENTATION AND EXPERIMENTS

In this section, we evaluate the performance of our proposed approach. We sourced datasets from OpenDataLab for our experiment, specifically leveraging subsets of the Book and Movie reviews (BR and MR) datasets. The DBRD dataset provides over 110k book reviews with binary sentiment labels, establishing a benchmark for Dutch sentiment analysis as detailed by Van der Burgh and Verberne in their 2019 paper [37]. Additionally, we utilized a portion of the comprehensive Movie Reviews dataset, which includes 50k polarized reviews for model training and evaluation [38]. It's important to note that our experiments were based solely on selected portions of these datasets to align with the scope of our analysis. The datasets were split into training, development, and testing sets, allocating 80% for training and 10% each for

TABLE 1. Experiment statistics of the BR and MR datasets.

Data	Positive			Negative		
	Train	Dev	Test	Train	Dev	Test
BR	9594	1199	1199	9578	1197	1197
MR	800	100	100	800	100	100
Total	10394	1299	1299	10378	1297	1297

TABLE 2. Experiments on BR dataset with TF-IDF.

Dataset	Confusion Matrix			Evaluations(%)		
		Positive	**Negative**	**P**	**R**	**FM**
Train	**Positive**	8800	968	90.09	91.72	90.89
	Negative	794	8610			
Dev	**Positive**	1162	191	85.88	96.91	91.06
	Negative	37	1006			
Test	**Positive**	1144	185	86.08	95.41	90.50
	Negative	55	1012			

TABLE 3. Experiments on BR dataset with no TF-IDF.

Dataset	Confusion Matrix			Evaluations(%)		
		Positive	**Negative**	**P**	**R**	**FM**
Train	**Positive**	8709	1356	86.53	90.77	88.60
	Negative	885	8222			
Dev	**Positive**	1115	199	84.86	92.99	88.74
	Negative	84	998			
Test	**Positive**	1109	203	84.53	92.49	88.33
	Negative	90	994			

TABLE 4. Experiments on MR dataset with TF-IDF.

Dataset	Confusion Matrix			Evaluations(%)		
		Positive	**Negative**	**P**	**R**	**FM**
Train	**Positive**	731	70	91.26	91.37	91.32
	Negative	69	730			
Dev	**Positive**	90	9	90.91	90	90.45
	Negative	10	91			
Test	Positive	88	13	87.13	88	87.56
	Negative	12	87			

development and testing. Detailed information regarding the data used is meticulously outlined in Table 1. Our study aims to evaluate the proposed method's performance by classifying positive and negative reviews.

From a classification perspective, the terms 'True Positive (TP),' 'False Positive (FP),' 'True Negative (TN),' and 'False Negative (FN)' are used to evaluate the effectiveness of the results. Furthermore, diverse performance metrics are employed to evaluate the classifier's effectiveness.

- Precision(P): P is the ratio of TP to the sum of TP and FP. It measures the accuracy of the positive predictions.
- Recall(R): R is the ratio of TP to the sum of TP and FN, which measures the ability of a classifier to find all the positive instances.
- F-Measure (FM): FM is a weighted harmonic mean of precision and recall, balancing their trade-offs.
- Accuracy (A): A is a measure of the overall correctness of a classifier, indicating the proportion of correct predictions out of the total number of predictions.

A. EFFECT OF TF-IDF

We first compare the architectures with and without TF-IDF weights on the BR dataset. The results presented in Table 2 and Table 3 demonstrate the positive impact of incorporating TF-IDF weights on model performance. Specifically, the precision of our model, which quantifies the accuracy of positive predictions, reaches an optimal value of 90.09% in the training set when TF-IDF is utilized. This represents a marked improvement over the model that does not employ TF-IDF weights, underscoring the benefits of TF-IDF in enhancing the model's predictive capabilities.

Furthermore, the FM score, a single metric that combines both precision and recall into a weighted harmonic mean, allows for a more balanced evaluation of the model's performance. It provides a comprehensive assessment by considering both precision and recall metrics. An FM score of 90.50% on the test set, which is 2.17% higher than the model without TF-IDF weights, indicates that our model has achieved an effective balance between identifying true positives and minimizing false positives. This result not only validates the effectiveness of our approach but also emphasizes the importance of TF-IDF in sentiment analysis tasks where the context and specific word choices are pivotal.

The results from the MR dataset, as detailed in Table 4 and Table 5, corroborate the findings from the BR dataset. The inclusion of TF-IDF weights consistently leads to better performance across all evaluated metrics, reinforcing the notion that TF-IDF is instrumental in capturing the semantic richness of text data. By assigning higher weights to terms that are frequent in a document but rare across the corpus, TF-IDF allows our model to prioritize the most contextually relevant words, which is crucial for accurately determining the sentiment of user comments.

In summary, the integration of TF-IDF weights into our model's architecture has proven to be a pivotal factor in improving the precision and overall performance of our sentiment classification tasks. The detailed examination of our results not only showcases the significance of TF-IDF in our model but also contributes to the broader understanding of how TF-IDF can enhance the predictive power of sentiment analysis models.

B. PERFORMANCE COMPARISON

To evaluate the effectiveness of our proposed method, we compared it with several baseline models, including LSTM, CNN, and our model without TF-IDF. The results are presented in Table 6. The experimental results demonstrate the effectiveness of our model. We observed that the joint architecture of BLSTM and Transformer outperforms single LSTM and CNN models in sentiment classification tasks. For example, on the training set, the accuracy of the joint model without TF-IDF is 91. 31%, while it is 85. 79% and 86. 03% for CNN and LSTM, respectively.

TABLE 5. Experiments on MR dataset with no TF-IDF.

Dataset	Confusion Matrix			Evaluations(%)		
		Positive	Negative	P	R	FM
Train	Positive	726	85	89.52	90.75	90.13
	Negative	74	715			
Dev	Positive	92	11	89.32	92	90.64
	Negative	8	89			
Test	Positive	87	14	86.13	87	86.57
	Negative	13	86			

TABLE 6. Accuracy of the classification tasks on Train, Dev, and Test sets.

Models	Train	Dev	Test
LSTM	85.79%	85.5 %	85.01 %
CNN	86.03 %	85.89 %	85.77 %
Our model(no TF-IDF)	91.31 %	88.2 %	87.77 %
Our model(with TF-IDF)	93.81 %	90.51 %	89.98 %

Furthermore, the performance of our model on the test dataset reinforces its effectiveness. The model without TF-IDF achieved an accuracy of 87.77%, which is notably higher than the 85.01% and 85.77% accuracies of the LSTM and CNN models, respectively. This indicates that our model is not only robust during training but also generalizes well to unseen data. Incorporating TF-IDF further enhanced the model's performance, with an accuracy of 89.98% on the test set. The TF-IDF component emphasizes the importance of term frequency and inverse document frequency, allowing the model to capture the semantic nuances and weight terms better accordingly, which is particularly beneficial for sentiment classification tasks.

We have observed that our model's accuracy on the test dataset is 87.77% without TF-IDF and 89.98% with TF-IDF. These figures slightly differ from the training set accuracies of 91.31% and 93.81%, respectively. The slight decrease in accuracy shows that our model needs to capture more nuances from the test set despite its ability to generalize from the training data. The variance may be due to overfitting to the training data's specific patterns, which can hinder the model's ability to generalize to the test data. However, the small gap between training and test accuracies also indicates that our model has not significantly overfit, suggesting it has developed a robust framework for sentiment classification.

The consistency of our model's superior performance across training and test datasets suggests that the joint architecture of BLSTM and Transformer and the optional TF-IDF weighting provide a more nuanced and comprehensive approach to sentiment analysis.

C. LOSS ANALYSIS

A loss versus epoch graph provides a clear visualization of our progress during neural network training. Since the MR dataset is smaller in size compared to the BR dataset, there may be more random fluctuations, which could lead to noisy loss value curves, making the optimization dynamics of the model less apparent. We choose the BR dataset to present the most comprehensive and representative results that accurately reflect the performance and training dynamics of our proposed Transformer-BLSTM model.

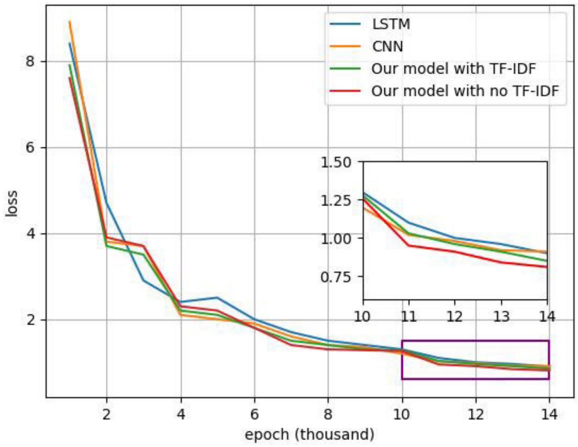

FIGURE 3. Loss vs Epochs on BR dataset.

In our experiments, we use the softmax layer to measure losses. Fig. 3 illustrates a comparison of frame loss on the BR set for various LSTM, CNN, and our model. The results of this assessment show a significant increase in loss rates when using models with a limited number of epochs, indicating notable discrepancies across all models. It is observed that the CNN and LSTM models outperformed other models between 2000 and 5000 epochs. This is attributed to their superior capability of effectively extracting features in tasks related to image processing. In sentiment analysis, short text data can be likened to image structures, allowing them to capture local features and patterns in the initial epochs accurately. The shared parameter mechanism enables the learning of general feature representations even with small datasets, thereby improving performance in the early stages. However, as the number of epochs increases, our model displays lower training loss values compared to other models, indicating a better fit to the training data. Furthermore, our model exhibits a more consistent decrease in training loss, suggesting smoother learning and potentially enhanced generalization capability.

Based on the analysis, we have found that the rate of loss reduction is the fastest among all the architectures, especially within the first 2500 epochs. Our proposed model shows a convergence of the loss rate at around 0.8, which is quicker than alternative models. This contributes to improved accuracy in classification tasks. The empirical evidence suggests that our model outperforms others in training loss, indicating that our tested model could be better suited for sentiment analysis in movie and book reviews.

V. CONCLUSION

This paper introduces new contributions to the field of document sentiment analysis. Firstly, we propose a Transformer-BLSTM model that automatically analyzes user comments on a website for sentiment analysis. The model helps the website better understand users' emotional tendencies, which can aid in making critical decisions. Secondly, the model uses both Transformer and BLSTM to analyze inputs in a long-term context, overcoming the problem of vanishing gradient. We employ the multi-head attention mechanism for feature extraction, which allows the model to capture essential semantic information in the text, improving its ability to analyze semantics. Moreover, we added TF-IDF weights to the combined Transformer and BLSTM system. This enhances the accuracy and efficiency of document classification by precisely evaluating the importance of a word in the corpus.

The model was implemented and evaluated using a public dataset of more than 20,000 comments for sentiment classification tasks and was compared to traditional models. In testing, the proposed model achieved a maximum accuracy of 93.81%. Furthermore, performance with or without TF-IDF weight showed significant differences, with the highest FM improvement reaching 2.17% in the experiments. These experiments have effectively demonstrated the effectiveness of the work.

In the future, we aim to evaluate various features of the model to improve its accuracy. Additionally, we hope to expand our work to different NLP tasks.

REFERENCES

[1] E. Burgener and J. Rydning, "High data growth and modern applications drive new storage requirements in digitally transformed enterprises," A White Paper, Dell Technol., NVIDIA, Santa Clara, CA, USA, IDC Report US49359722, 2022.

[2] A. Ghourabi, "A BERT-based system for multi-topic labeling of Arabic content," in *Proc. 12th Int. Conf. Inf. Commun. Syst. (ICICS)*, May 2021, pp. 486–489.

[3] L. Kong, C. Li, J. Ge, F. Zhang, Y. Feng, Z. Li, and B. Luo, "Leveraging multiple features for document sentiment classification," *Inf. Sci.*, vol. 518, pp. 39–55, May 2020.

[4] G. Choi, S. Oh, and H. Kim, "Improving document-level sentiment classification using importance of sentences," *Entropy*, vol. 22, no. 12, p. 1336, Nov. 2020.

[5] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020.

[6] T. Vu, T. Wang, T. Munkhdalai, A. Sordoni, A. Trischler, A. Mattarella-Micke, S. Maji, and M. Iyyer, "Exploring and predicting transferability across NLP tasks," 2020, *arXiv:2005.00770*.

[7] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for NLP tasks," in *Proc. 15th Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, 2020, pp. 179–183.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, and T. Rocktäschel, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[9] J. Kong, L. Zhang, M. Jiang, and T. Liu, "Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition," *J. Biomed. Informat.*, vol. 116, Apr. 2021, Art. no. 103737.

[10] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in neural NLP: Modeling, learning, and reasoning," *Engineering*, vol. 6, no. 3, pp. 275–290, Mar. 2020.

[11] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, and Q. Li, "Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 105856.

[12] Y. Qiao, D. Wiechmann, and E. Kerz, "A language-based approach to fake news detection through interpretable features and BRNN," in *Proc. 3rd Int. Workshop Rumours Deception Social Media (RDSM)*, 2020, pp. 14–31.

[13] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2733–2742, Oct. 2020.

[14] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU neural network performance comparison study: Taking yelp review dataset as an example," in *Proc. Int. Workshop Electron. Commun. Artif. Intell. (IWECAI)*, Jun. 2020, pp. 98–101.

[15] M. Khan, H. Wang, A. Riaz, A. Elfatyany, and S. Karim, "Bidirectional LSTM-RNN-based hybrid deep learning frameworks for univariate time series classification," *J. Supercomput.*, vol. 77, no. 7, pp. 7021–7045, Jul. 2021.

[16] A. Aldalbahi, F. Shahabi, and M. Jasim, "BRNN-LSTM for initial access in millimeter wave communications," *Electronics*, vol. 10, no. 13, p. 1505, Jun. 2021.

[17] O. Alharbi, "A deep learning approach combining CNN and bi-LSTM with SVM classifier for Arabic sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 165–172, 2021.

[18] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020.

[19] M. V. Koroteev, "BERT: A review of applications in natural language processing and understanding," 2021, *arXiv:2103.11943*.

[20] S. Nath, A. Marie, S. Ellershaw, E. Korot, and P. A. Keane, "New meaning for NLP: The trials and tribulations of natural language processing with GPT-3 in ophthalmology," *Brit. J. Ophthalmology*, vol. 106, no. 7, pp. 889–892, Jul. 2022.

[21] M. O. Topal, A. Bas, and I. van Heerden, "Exploring transformers in natural language generation: GPT, BERT, and XLNet," 2021, *arXiv:2102.08036*.

[22] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," 2020, *arXiv:2005.13012*.

[23] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," *Frontiers Artif. Intell.*, vol. 6, Mar. 2023, Art. no. 1023281.

[24] K. Dhola and M. Saradva, "A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Jan. 2021, pp. 932–936.

[25] F.-E. Lagrari and Y. Elkettani, "Traditional and deep learning approaches for sentiment analysis: A survey," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 6, no. 4, pp. 1–7, 2021.

[26] S. Bodapati, H. Bandarupally, R. N. Shaw, and A. Ghosh, "Comparison and analysis of RNN-LSTMs and CNNs for social reviews classification," in *Advances in Applications of Data-Driven Computing*. Berlin, Germany: Springer, 2021, pp. 49–59.

[27] L. Li, T.-T. Goh, and D. Jin, "How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4387–4415, May 2020.

[28] P. Patel, D. Patel, and C. Naik, "Sentiment analysis on movie review using deep learning RNN method," in *Intelligent Data Engineering and Analytics: Frontiers in Intelligent Computing: Theory and Applications (FICTA)*, vol. 2. Cham, Switzerland: Springer, 2020, pp. 155–163.

[29] A. C. M. V. Srinivas, C. Satyanarayana, C. Divakar, and K. P. Sirisha, "Sentiment analysis using neural network and LSTM," in *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1074, no. 1, 2021, Art. no. 012007.

[30] Q. Xu, L. Zhu, T. Dai, and C. Yan, "Aspect-based sentiment classification with multi-attention network," *Neurocomputing*, vol. 388, pp. 135–143, May 2020.

[31] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 207–212.

[32] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 53–68, Feb. 2021.

[33] Z. Khan and Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3034–3042.

[34] D. Munandar, A. F. Rozie, and A. Arisal, "A multi domains short message sentiment classification using hybrid neural network architecture," *Bull. Electr. Eng. Informat.*, vol. 10, no. 4, pp. 2181–2191, Aug. 2021.

[35] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[36] H. Liu, X. Chen, and X. Liu, "A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022.

[37] B. van der Burgh and S. Verberne, "The merits of universal language model fine-tuning for small datasets—A case with Dutch book reviews," 2019, *arXiv:1910.00896*.

[38] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 142–150.

YUN LIN received the M.S. degree from Xiamen University, Xiamen, China, in 2017. He has been a Lecturer with the School of Information Science and Technology, Tan Kah Kee College, Xiamen University, since 2017. His current research interests include machine vision and intelligent algorithms.

TUNDONG LIU was born in February 1970. He received the Ph.D. degree in engineering from the University of Science and Technology of China, in June 2003. From August 2015 to August 2016, he was a Visiting Scholar with the University of Calgary, Canada, for one year. From 2010 to 2013, he was the Deputy District Head of Science and Technology in Fengze, Quanzhou. From 2013 to 2015, he was the Vice Dean of the School of Information Science and Technology, Xiamen University. From 2016 to 2019, he was the Vice Dean of Xi'an Jiaotong University and Xiamen University. Since 2021, he has been the Vice Principal of Xiamen University Affiliated Experimental School. He is a member of Xiamen Association for Science and Technology and Fujian Provincial Association for Science and Technology, the Chairperson of Xiamen Automation Society, and the Director of Chinese Automation Society. In 2011, he received the "National Advanced Individual in County-Level Science and Technology Assessment" by the Ministry of Science and Technology. In 2017, he received the Key Talent of Xiamen City. In 2018, he received the Outstanding Science and Technology Worker of Fujian Province. In 2020, he received the First Prize for Teaching Achievements in Fujian Province. In 2021, he was approved as the Class C High-Level Talent in Fujian Province. In 2022, he was approved as the Leading Talent in Xiamen City.