International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Context-Preserving Sentiment Classification Using Bi-TCN And BI-GRU with Multi-Head Self-Attention

[*1] K. R. Srinath
Department of Informatics, Osmania University,
Hyderabad, Telangana, India
e-mail: srinath.kr1022@gmail.com

[2] B. Indira
Department of MCA, Chaitanya Bharathi Institute of Technology (CBIT),
Hyderabad, Telangana, India
e-mail: bindira_mca@cbit.ac.in

Abstract— In natural language processing, sentiment classification is the recently used topic. Specifically, the objective of the sentiment analysis is to categorise the polarity expressed on the sentence's target. However, there are some researches for classifying the polarity of the target which outperforms well in their way. Yet, there are some limitations, such as apparent and in-apparent issues, gradient problems, etc., to overcome these issues the context-preserving sentiment classification using BI-TCN (Bidirectional Temporal Convolutional network) and BI-GRU (Bidirectional Gated Recurrent Unit) with Multi-head self-attention is proposed to extracts both the local dependent and global dependent information from the sentence, then it will incrementally extract the supervision information of the target to train the model. Formerly, the model is tested and trained using four datasets and the performance is compared with four existing methods, its accuracy is evaluated using the F1-score, precision, recall, specificity, and MCC (Matthews Correlation Coefficient). Consequently, the proposed approach provides the best accuracy level of 98%..

Keywords- Sentiment classification; BI-TCN; BIGRU; Multi-head self-attention.

I. INTRODUCTION

Nowadays, people express their feelings about products, movies, hotels, etc., on internet platforms such as social media, e-commerce websites, etc., which leads to a large amount of user-generated data on the web but it provides benefits for governments, business organizations, and decision-makers. Sentiment analysis is an extensively used natural language processing method that mines the opinions from the unstructured data which are the contents shared through the internet about the reviews and provides the sentiment polarity as positive, negative or neutral. Sentiment analysis has been used in a wide range of applications, such as information storage, web gathering and retrieval techniques, and many more. Conversely, the challenges in sentiment analysis are inconsistency, emojis, informal grammar, etc., and some of the words should be wisely combined for the best potential performance of the sentiment analysis model. Meanwhile, the machine learning method is trained and tested using the supervised method to analyse and provide the polarity of the sentence. Moreover, many researchers use machine learning algorithms because of their simplicity and high accuracy.

Recently, many researchers provide better results in sentiment classification. [1] used the attention-based LSTM with aspect embedding, [2] provided target-dependent sentiment classification, [3] used the recurrent attention memory network and some methods like long-short term memory (LSTM) and Bi-GRU to solve the exploding and vanishing problem but it does not frequently capture the interdependence characteristics between words. Moreover, finding the difference between the opinion words in multiple targets is difficult because the ABSA model only focuses on the high-frequency word with high sentiment prediction and pays low attention to the low-frequency words which causes the unacceptable performance of the models [4]. Consequently, the word order of the sentences is sensitive in its described target-sensitive sentiment [5]. This issue is also observed and modified by creating a specific word representation related to the target but does not focus on improving attention [4]. The best structure of the complex sentence should gain the dependent information between each word and the other words in the sentence which means global dependent information that can be solved by self-attention [6] but it does not order the words of the sentence.

To overcome the above-mentioned drawbacks, sentiment classification using BI-TCN and BI-GRU with Multi-head self-attention is proposed. TNet-att creates the target representation based on each word by achieving the dependent information of words in the sentences and the target word which is called local dependent information. To gain the best structure of the complex sentence the multi-head self-attention mechanism is used over generating the exact word representation related to the target. Then the inapparent and apparent pattern issues are fixed by training the model using supervised learning. But still, training the attention mechanism with high performance is difficult and time-consuming. Therefore, the proposed method uses the automatically mined supervision information from the training instance to provide the best context information for sentiment classification.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

The main contribution of the proposed method is summarized below:

- The proposed approach ABSA with TNet-att uses the TNet which is the integration of BI-TCN and BIGRU attention layer, CPT layer, and the model is trained using the supervision information and the final layer is the classification layer.
- Initially, the given input sentence and the target are given to the word-level attention layer to create the word representation.
- To capture the contextualized information of the input corpus, the word representation is applied on the Bi-TCN &BIGRU layer and generates the context and aspect word representation.
- The CPT layer extracts the context information and learns more features from the word. Then the multi-head self-attention is introduced in the proposed model to provide the aspect-related sentiment representation to capture the global dependence
- The model is trained using the automatically mined supervision information of the input sentence by extracting the context of words. Finally, the classification layer provides the sentiment polarity of the sentence using the softmax function.

The rest of the paper is organised into a section, section 2 explains the background of the proposed novels, section 3 explains the proposed methods, section 4 explains the experiment and result part, and section 5 explains the conclusion of the proposed paper.

II. BACKGROUND

A. Aspect Based sentiment analysis (ABSA)

The aspect-level sentiment analysis is fine-grained opinion mining towards specific entities, also called targets. The goal of the ABSA is to find the polarity of the target expressed in the reviews by the user and it has a high ability to learn the aspect-related semantic representation of the given sentence compared with other models [7]. [8] Used sentiment analysis in the recognition and classification task. They divide the process into four steps, they are sentiment classification, sentiment polarity, product property selection and sentiment recognition for the product reviews. [9] Proposed the model for fine-grained sentiment analysis, which deals with two tasks they are Aspect target sentiment analysis and Aspect category sentiment analysis. [10] Proposed a sentiment classification using an adaptive recursive neural network. [11] Introduced CNN for ABSA to capture the information from multi-layered sentiment analysis. [12] Proposed ASEGC related to graph convolutional networks to gain efficient information on the ABSA task. [13] Proposed the ABSA model using CNN and GRU. GRU collect the local features generated by the CNN. Although, most of the research works only focus on the local features of the training instances and it does not take responsibility for the global information of the corpus.

B. BIGRU

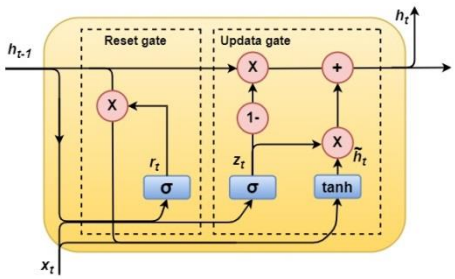

Figure 1. Architecture of Gated Recurrent Unit

Architecture of Gated Recurrent Unit is shown in Figure 1. The variant of the recurrent neural network is a Gated Recurrent Unit (GRU), it has a recursive structure and also has a memory function of processing time series data. It can reduce the gradient explosion and disappearance during the training. GRU has two inputs: the output of the previous time h_{t-1} and the sequence value of the existing time x_t then it has only one output state of the existing timeh_t. It also has two gates update gate and a reset gate which are represented as z_t and r_t the past information's controlled by the reset gate from the existing state then the update gate controls the loss of historical state information. The process of GRU is expressed in Equations (1)-(4)

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (1)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (2)$$

$$\tilde{h}_t = tanh(W_{\tilde{h}} x_t + U_{\tilde{h}}(r_t \odot h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (4)$$

Where W_r, W_z, $W_{\tilde{h}}$, U_r, U_z, $U_{\tilde{h}}$ are the weight of the coefficient matrix, h_{t-1} represent the output state with time $t-1$, h_t represent the output state with time t, x_t is the input sequence with time t, $\tilde{h}_t$ output state with time t, σ is denoted as the sigmoid function which is used to change the intermediate state to the range [0,1], $tanh$ is the hyperbolic tangent function, $\odot$ represent the Hadamard product of the matrix which means a binary operation using two same dimensional matrices and produce the same dimensional matrix as an output. GRU moves only in one direction so it may lose the old information after the current time. But BiGRU moves in both forward and backward directions to capture both the information from the old and present times. Architecture of BIGRU is shown in Figure 2.

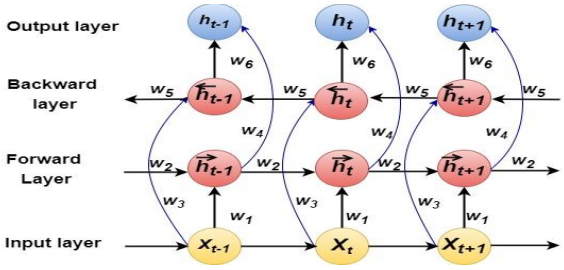

Figure 2. Architecture of BIGRU

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

BiGRU includes an input layer, hidden layer and output layer. The hidden layer h_t has two partitions: forward layer $\vec{h}_t$ and backward layer $\overleftarrow{h}_t$ with the current input. $\vec{h}_{t-1}$ is the forward hidden layer with the $t-1$ time. The backward hidden layer $\overleftarrow{h}_{t+1}$ with the time $t+1$. The process of the hidden layer is expressed in Equation (5)-(7), where $w_i(i = 1,2,\ldots.6)$

$$\vec{h}_t = f(w_1 x_t + w_2 \vec{h}_{t-1}) \quad (5)$$

$$\overleftarrow{h}_t = f(w_3 x_t + w_5 \overleftarrow{h}_{t-1}) \quad (6)$$

$$h_t = g(w_4 \vec{h}_t + w_6 \overleftarrow{h}_t) \quad (7)$$

C. TCN

The traditional convolution cannot capture long sequence-dependent information. So the novel temporal convolutional network (TCN) is proposed by [14] which uses the casual convolution from the residual blocks rather than the convolution block. This block uses the Batch Norm and dropout layer to regularise the network. Although, its prediction in a unidirectional structure does not capture the aspect information of the sentence for classification. Similarly [15] analysed the sentiment through the LSTM and TCN. Thus, [15] modified the TCN by training it with forward and reverse information of the sequence of sentences as input to a model and produced a bidirectional TCN (BiTCN) and then the convolution neural network (CNN) is integrated to predict the protein secondary structure. Similarly [16] proposed TCN-BIGRU shown in Figure 3, which integrates the Bidirectional GRU and TCN because TCN extracts the high-frequency and low-frequency information from the sequence. At the same time, the GRU captures the long-term dependence in a sentence sequence. But, BiGRU is the advanced method of GRU which can learn the current data's long-term information and short-term information together. Therefore, in the proposed approach the Bi-TCN is introduced with BiGRU in the TNet to handle the major issues during the classification and also it changes the size of the receptive field to control and compute the length of the memory sequence in parallel.

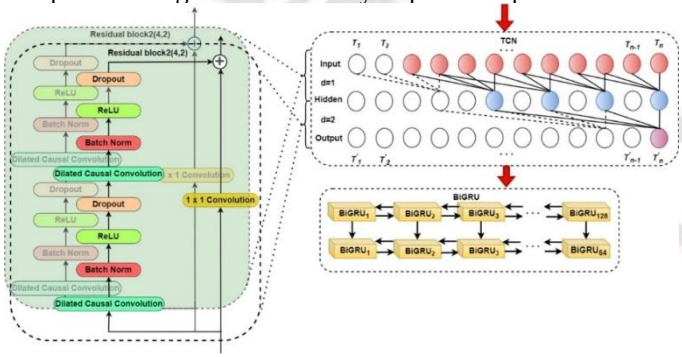

Figure 3. Architecture of TCN-BIGRU layer

D. Multi-head self-attention

The word-level interface between context and aspect is captured using the attention mechanism. [18] Introduced the multi-grained attention for the context embedding in sentiment analysis it captures the context and aspect information using the fine attention mechanism and coarse-grained attention which only use the context vector to find the attention weights. [19] Proposed the multi-head attention with the point-wise-feed-forward network to capture the context and aspect from the hidden information of the sentence. [20] Proposed the multi-attention network which captures the long dependencies in the sentence by self-attention. [21] proved that multi-head attention is not only for machine translation it is also for text classification by combining the multi-head attention with the BiLSTM in sentiment analysis tasks. [22] Proposed a multi-head self-attention transformation network with BiLSTM to create the contextualized word representation to capture the global dependencies.

III. PROPOSED METHODOLOGY

A. Problem Formulation

In the Aspect Based Sentiment Analysis (ABSA), $A = \{A_1, A_2, \ldots., A_L\}$ is a predefined type, $p = \{positive, Negetive, Neutral\}$ are constrained as sentiment polarity.

Word embedding is a method that used the word encoder to convert discrete words to high-dimensional vectors. The word encoder expands the feature extraction by understanding the context of the sentence. It also allocates the same vector to the words with similar meanings in the same context, which is necessary for the classifier. The glove is the latest methodology for word encoding. The input of the word embedding is the sentences with *n* number of words to transfer the word into a dimensional vector. The task of the embedding layer is to encode the sentences as a matrix, $z = [w_1, \ldots. w_i, \ldots. w_n] \in r^{n \times d}$, where $w_i = [x_{i1}, \ldots. x_{ij}, \ldots x_{id}]$ related to the word vector of the given word in the sentence. The pre-trained embedding method Glove is used for word representation that trains the word representation using the co-occurrence of the matrix by the use of an unsupervised method. Both the global and local information of input words are widely counted and the benefits of the neural network model are absorbed by the Glove model.

There may have an *M* target in a sentence represented as T^S and each target in a sentence with m_i term is denoted as T_i^S which is derived in Equations (8) & (9)

$$T^S = T_1^S, T_2^S, \ldots\ldots, T_M^S \quad (8)$$

$$T_i^S = \{w_i, w_{(i+1)}, \ldots, w_{(i+m_i-1)}\} \quad (9)$$

The prediction of sentiment polarities of the *M* target and also the prediction of sentiment polarity for each of the N aspect types are derived from Equations (10)-(12).

$$P^T = \{P_{1,}^T P_2^T, \ldots, P_M^T\} \quad (10)$$

$$A^S = \{A_1^S, A_2^S, \ldots, A_N^S\} \quad (11)$$

$$P^A = \{P_1^A, P_2^A, \ldots, P_N^A\} \quad (12)$$

Where the polarity of T_M^S sentence is denoted as P_M^T. The sentiment polarity of A_N^S is denoted asP_N^A.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

B. Proposed Transformation Network Attention (Tnet-Att) Based on Sentiment Classification

Figure 4 shows the overall framework of the proposed TNet-ATT which is the integration of TNet and the attention mechanism. The framework consists of five components: TCN-BiGRU word-level attention layer, Context-Preserving Transformation (CPT) layer, Multi-head self-attention layer, Model Training with Automatically Mined Attention Supervision Information layer, and Sentiment classifier.

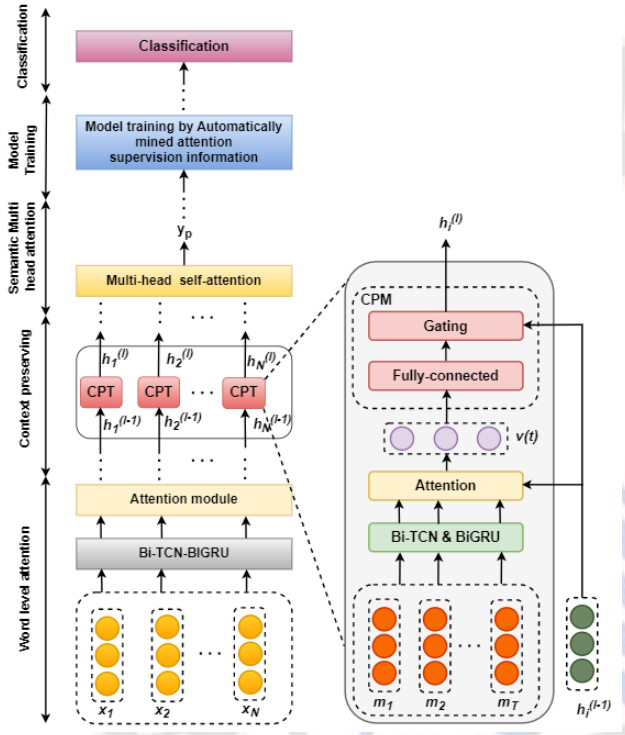

Figure 4. Overall framework of the proposed method

C. Bi-TCN & BIGRU word-level attention layer

In the proposed approach the bottom layer is an attention layer using BI-TCN & BIGRU. The attention mechanism is used to choose the key characteristics, then the selected characteristics are extracted using the Bi-TCN and then the BIGRU captures long dependence and obtains future information. The input of the attention mechanism is the text after the word embedding which transfers the words into a word vector and then the BIGRU provide the output from the hidden layer h_i then, u_i obtained by a linear layer and δ_i is obtained by softmax function for each word. Each word vector has a different weight from the attention mechanism. Then the characteristic extraction over the word is done by the Bi-TCN using two residual blocks both consisting of two convolutional layers with kernel size 4, dilation factor 1 for the first residual block and 2 for 2nd residual block. The input of the BIGRU is the output from the Bi-TCN to extract the long-term correlation between future information and present information. Then both Bi-TCN and BIGRU combined and transforms the input into the contextualized word representation expressed in Equation (13).

$$h^{(l-1)} = \left(h_1^{(l-1)}, h_2^{(l-1)}, \dots, h_N^{(l-1)}\right) \quad (13)$$

D. Context-Preserving Transformation (CPT) layer

The previous Bi-TCN-BIGRU word-level attention layer does not consider the target information because the attention-based approach retains the word-level features fixed and combined as a representation of the sentence using the weights. So, the CPT layer is introduced in the proposed approach shown in Figure 4. In each CPT layer, TST (Target Specific Transformation) is used to combine the word representation and target representation by computing the importance of target words based on each sentence rather than the whole sentence and using another TCN-BIGRU to gain the target word vector representation $v(t)$ with attention mechanism and then the vector representation $v(t)$ joined with the word representation. Moreover, the Context preserving Mechanism is used in the CPT layer to retain the context information and learn more about the features of words. As a final point, the context information is combined in all layers to enable the understanding of the target of word representation then the word representations are updated as expressed in Equation (14).

$$h^l = f\left(h^{(l-1)}\right) = \left(h_1^l, h_2^l, \dots, h_N^l\right) \quad (14)$$

E. Multi-head self-attention layer

The word level attention measures the significance of words, an output of word embedding. The syntactic or semantic features in the same sentence can be captured effectively using Self-attention. If a pair of words is connected directly then gaining the interdependent feature over a long distance is easy.

The multi-head mechanism is proposed by [6] to measure the dot-product multiple times in parallel. The keys, values and queries are projected to dimensions d_k, d_v and d_q. All individual output is concatenated and projected linearly expected dimension. The outcome of the multi-head self-attention is expressed in Equation (15)

$$multiHead(Q,K,V) = [head_1, head_2, \dots . . head_h]w^O \quad (15)$$

Where, $w^o$is the transformation matrix, the attention value of the entire sentence is Multi-head, and $concat$ is the splicing operation.

Each of the $head_i$ is calculated using Equation (16)

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (16)$$

Where the parameter matrices are W_i^Q, W_i^K, W_i^V to learn the model then the output matrix is $MATT = \{matt_1, \dots . matt_i, \dots . matt_h\}$.

Finally, the residual concatenation of $Multi\_head$ and z gets the sentence matrix illustrated in Equation (17)

$$X = residual\_Connect\ (z, Multi\_head) \quad (17)$$

Among them, $X \in R^{L\times D}$ is the output of multi-head attention, and $residual\_Connect$ is the residual operation.

In the proposed TNet-ATT the multi-head attention mechanism provide the aspect-related sentiment representation o is shown in Equation (18)

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

$$o = Attention(h(x), v(t)) \quad (18)$$

Where $v(t)$ is the generated aspect representation and $h(x)$ is the word-level semantic representation.

F. Model Training layer

In the proposed approach the model is trained using the automatically mined supervision information, this process is explained using the algorithm. Initially, the model is trained using the training corpus Q with the parameters $\theta^{(0)}$. Then K number of iteration is taken to train the model which extracts the influential context word of all the instances as attention supervision information, it can be done by introducing $\emptyset$ denoted as initialization of two words set for every training example (x, y, z) and also it secure all the extracted context words. $S_a(x)$ and $S_m(x)$ are the two words set, $S_a(x)$ is an active effect of context word with sentiment prediction of x, it will stay in the refined model training. $S_m(x)$ is a misleading effect on context words which has low attention weight.

The algorithm [1] explains the training of the model with the extraction of context words of all instances. The initial step is to create an aspect representation using the parameter $\theta^{(k-1)}$ from the previous iteration. Then, form a new sentence X to replace the previous extracted words of sentence x based on $S_a(x)$ and $S_m(x)$. Similarly, the context word extracted from sentence x is isolated during the prediction of sentiment in sentence X and therefore the essential context words from sentence X are extracted efficiently. Finally, the word representation is shown in Equation (19)

$$h(X) = \{h(X_i)\}_{i=1}^{N} \quad (19)$$

Based on the $v(t)$ and $h(X)$ the parameter θ^{k-1} forced to find the sentiment polarity of X as y_p. Then the word saliency score vector $\alpha(X) = \{\alpha(X_1, \alpha(X_2), \dots \dots \alpha(X_N)\}$ are continuously introduced in this process which is expressed as $\sum_{i=1}^{N} \alpha(X_i)$=1, where $\alpha(X_i)$ is denoted as the measure of X_i on the sentiment prediction of X.

In the third step, the variance of $\alpha(X_i)$ is measured using the entropy $E(\alpha(X_i))$ derived in Equation (20)

$$E\alpha\big((X_i)\big) = -\sum_{i=1}^{N} \alpha(X_i) \log\left(\alpha(X_i)\right) \quad (20)$$

Entropy is used to find any context words in X during the sentiment prediction. If there is any influential context word and extract the context word X_m along with the maximum influence weight as attention supervision information then the entropy must be less than the threshold ϵ_α. Thus, it will produce different prediction results if the prediction is correct then the X_m is added in the $S_a(x)$ otherwise added in the $S_m(x)$.

Then, in the fourth step, the new training corpus Q^k is created by combining the X, y, and z as triples and merging with the collected ones. To update the $\theta^{(k-1)}$, Q^k is forced for the next iteration. Therefore the model will find more influential context words so, it will take K iterations to extract the influential context words. At last, these extracted words of training examples will be comprised into the Q and form a final training corpus Q_s along with attention supervision. This extraction is used to train the model

Algorithm:1 Training model

Q: training corpus;
$\boldsymbol{\theta^i}$: model parameters;
$\boldsymbol{\epsilon_\alpha}$: the entropy threshold of attention weight distribution;
K: the maximum number of training iteration
$\boldsymbol{\theta^{(0)} \leftarrow Train(Q, \theta^i)}$
for $\boldsymbol{(x, y, z) \in Q}$ do
  $\boldsymbol{s_a(x) \leftarrow \emptyset}$
  $\boldsymbol{s_m(x) \leftarrow \emptyset}$
end for
for k=1, 2,…., k do
  $\boldsymbol{Q^{(k)} \leftarrow \emptyset}$
  for $\boldsymbol{(x, y, z) \in Q}$ do
    $\boldsymbol{v(t) \leftarrow GunAspectRep(y, \theta^{(k-1)})}$
    $\boldsymbol{X \leftarrow MaskWord(x, s_a(x), s_m(x))}$
    $\boldsymbol{h(X) \leftarrow GenWordRep(X, v(t), \theta^{(k-1)})}$
    $\boldsymbol{y_p, \alpha(X) \leftarrow SentiPred(h(X), v(t), \theta^{(k-1)})}$
    $\boldsymbol{E(\alpha(X)) \leftarrow CalcEntropy(\alpha(X))}$
    if $\boldsymbol{E(\alpha(X)) < \in_\alpha}$ **then**
      $\boldsymbol{m \leftarrow argmax_{1 \le i < N} \alpha(X_i)}$
      if $\boldsymbol{y_p == z}$ then
        $\boldsymbol{s_a(x) \leftarrow s_a(x) \cup \{X_m\}}$
      else
        $\boldsymbol{s_m(x) \leftarrow s_m(x) \cup \{X_m\}}$
      end if
    end if
    $\boldsymbol{Q^{(k)} \leftarrow Q^{(k)} \cup (X, y, z)}$
  end for
  $\boldsymbol{Q^{(k)} \leftarrow TrainQ^{(k)}; \theta^{(k-1)}}$
end for
$\boldsymbol{Q_s \leftarrow \emptyset}$
for $\boldsymbol{(x, y, z) \in Q}$ do
  $\boldsymbol{Q_s \leftarrow Q_s \cup \big(x, y, z, s_a(x), s_m(x)\big)}$
end for
$\boldsymbol{\theta \leftarrow Train(Q_s)}$
Return: $\boldsymbol{\theta}$

The extracted context word using the algorithm is used to expand the training of the proposed model and a soft attention normalizer is also introduced to optimize the objective of training which is expressed in Equation (21).

$$\triangle\left(\alpha\big(s_a(x) \cup s_m(x)\big), \hat{\alpha}\big(s_m(x) \cup s_m(x)\big); \theta\right) \quad (21)$$

Where $\alpha(*)$ and $\hat{\alpha}(*)$ are denoted as the model-induced weight distribution and expected influence weight distribution of the words in the $s_a(x) \cup s_m(x)$ and $\triangle\left(\alpha(*), \hat{\alpha}(*); \theta\right)$ is a Euclidean Distance loss. As per the analysis, the context word of $s_a(x)$ is equally focused during the training of the model with the expected influence weight with the same value $\frac{1}{s_a(x)}$. By the way, the first extracted influence of words will be enhanced and the later extracted words are reduced. The overfitting of high-frequency words and underfitting of low-frequency context words with sentiment polarity then the misleading effect $s_m(x)$ in the sentiment polarity of X is directly set to the weight 0. Finally, the training objective of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

the proposed model is based on the log-likelihood of the gold truth which is expressed in Equations (22) & (23)

$$j(Q;\theta) = -\sum_{(x,y,z)\in Q} j(x,yz;\theta) \quad (22)$$

$$= \sum_{(x,y,z)\in Q} d(y)\,.\,logd(x,y;\theta) \quad (23)$$

Then the training with attention to supervision information is shown in Equation (24)

$$j(Q;\theta) = -\sum_{(x,y,z)\in Q_s} j(x,y,z;\theta) + \gamma \triangle \left(\alpha\left(s_a(x) \cup s_m(x)\right), \hat{\alpha}\left(s_m(x) \cup s_m(x)\right);\theta\right) \quad (24)$$

Where $j(x,y,z;\theta)$ is the convolutional training objective, $d(y)$ is the one-hot vector of y, $d(x,y;\theta)$ is represented as sentiment distributed pair (x,y) predicted by the model, "." Represents the dot product. γ denoted as hyper-parameter, $\gamma > 0$ balances the preference between the loss function regularization.

G. *Classification layer*

The input of the classification layer is the objective of the training corpus with automatically mined supervision information Q_s. Then the softmax function is implemented to calculate the predictive probabilities distribution for all the instances. Finally, the result of the classification layer is expressed in Equations (25) & (26)

$$z = w_s Q_s + b_s \quad (25)$$

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{3} \exp(z_j)} \quad (26)$$

Where w_s and b_s are the weight and bias terms and j are the instances.

IV. EXPERIMENT AND RESULT

A. *Dataset*

The proposed approach use four datasets: 1) IMDB movie reviews dataset which is taken from (https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews) it contains 50,000 data 25,000 for training and 25,000 for testing. 2) Twitter entity sentiment analysis dataset taken from (https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis), 3) Airline reviews from Twitter Airline Sentiment dataset (https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment), 4) Cell phone reviews dataset from Amazon taken from Kaggle (https://www.kaggle.com/code/mamunalbd4/amazon-cell-phones-reviews/data).

B. *Performance metrics*

The performance metrics used in this paper are accuracy, precision, recall, and F1-score are expressed in Equations (27)-(32)

$$Accuracy = \frac{TP+TN}{FN+FP} \quad (27)$$

$$Precision = \frac{TP}{FP+TP} \quad (28)$$

$$Recall = \frac{TP}{FN+TP} \quad (29)$$

$$F1\ Score = \frac{2\times Precision\times Recall}{Recall+Precision} \quad (30)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(FN+TN)(FP+TN)(TP+FN)}} \quad (31)$$

$$specificity = \frac{TN}{TN+FP} \quad (32)$$

Where MCC refers to Matthews Correlation Coefficient which is utilized to evaluate the performance of binary classification between the ranges 1 to -1. TP refers to a True positive, TN refers to a True Negative, FP refers to a false positive and FN refers to a false negative.

C. *Parameter settings*

The experiment of the proposed model is executed in python software using the windows 10 operating system, the parameters used in the experiment are charted in the table 1

TABLE I. DESCRIPTION OF PARAMETERS

	Parameter	values
Bi-TCN BIGRU	Dropout rate	0.4
	Normalization	Batch normalization
	Activation function	RLU
	Kernel Size	4
Multi-Head self-attention	kernel	50
	Learning rate	0.001
	No. of self-attention head	6
Dense	Activation function	Softmax
	Size of Hidden layer	7

D. *Analysis of the result*

The proposed model is tested and trained using four different datasets they are: IMDB movie reviews, Twitter entity sentiment analysis, Twitter airline reviews and Amazon mobile reviews. These datasets are partitioned into two, one half for testing and another half for training. The performance of the proposed approach using this dataset is evaluated using F1-score, recall, precision, MCC, and specificity, and then to know the efficiency of the proposed approach, it is compared with some of the existing methods like TD-LSTM [2], TSMN-ASC [5], BGRU-Capsule [23], and ABSC-MAN [20].

The proposed model is trained with super attention learning with the automatically mined supervision information. The training of the attention mechanism is guided by the automatically and incrementally extracted information from the training instance. From the heat map shown in Figure 5, the bolded words are target aspects and the different highlighted words depend on the weight of the attention. However, Ans. /pred = ground-truth/predicted label. Finally, the result shows that the training is done without changing any grammatical function using the attention mechanism.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

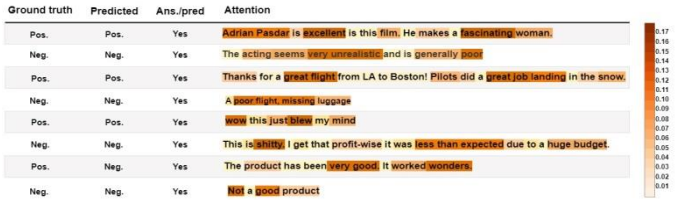

Figure 5. Visualization of attention for the sentences from the dataset

Figure 6 to Figure 9 shows the relationship between the epochs and F1 for the four datasets. The epochs are the iterations for the training set of the model. Conversely, if the epoch increases then the overfitting problem will be created which will reduce the ability of the model. Therefore the correct epochs should be selected for the classification. In the proposed model the growth of epochs will increase the classification performance F1 score and maintain stability when the epoch is 70.

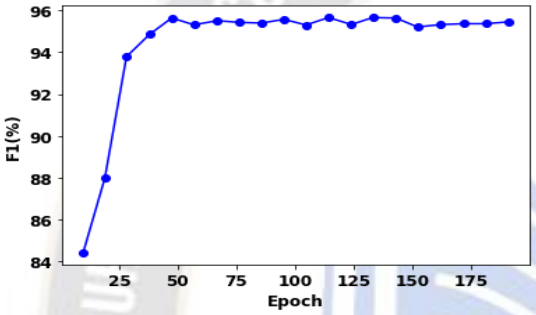

Figure 6. Relation between F1 and epochs for IMDB movie reviews dataset

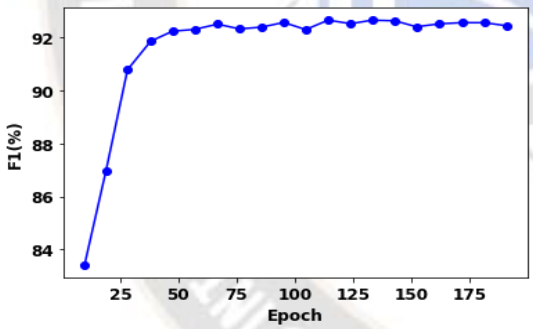

Figure 7. Relation between F1 and epochs for Twitter entity sentiment analysis dataset

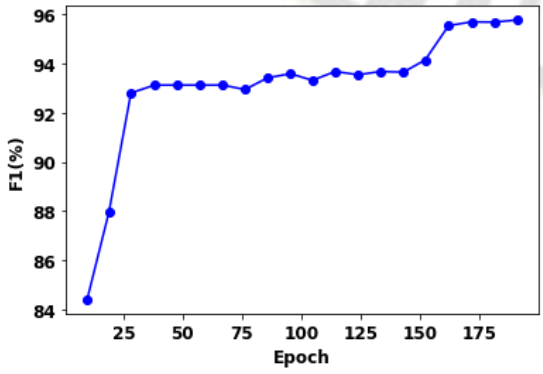

Figure 8. Relation between F1 and epochs for Twitter Airline reviews dataset

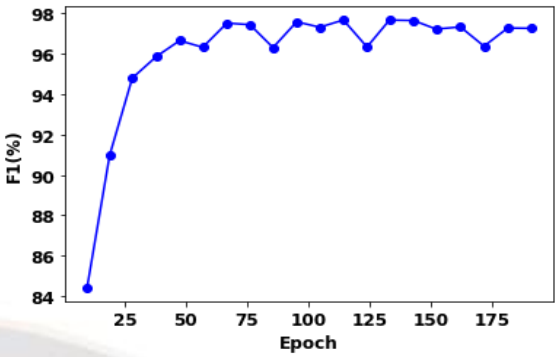

Figure 9. Relation between F1 and epochs for Amazon mobile reviews dataset

In the experiment, the model is tested using different iterations. The dissimilar iteration will affect the model. If the iteration of the model increased then initially the performance of the model will rise and then it will fall. From Figure 10, when the iteration number is reduced below 8 the performance will get increased. Conversely, if the iteration number is raised above 8 then the precision and accuracy of the model gradually fall and it will reduce the performance of the model.

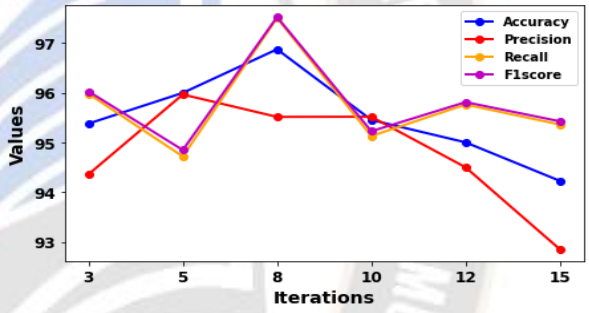

Figure 10. Number of iterations of the model

Dropout is used in the model to improve the overview of the proposed model, which is illustrated from Figure 11 to Figure 14. It shows the different dropout values selected for the experiment using four datasets separately. However, when the dropout value is 0.4 then the performance of the model is optimal. Moreover, Figure 15 to Figure 18 shows the equipotential plot for the f-measures for the four datasets. The proposed approach reaches the highest precision, recall and f1-score in all four datasets.

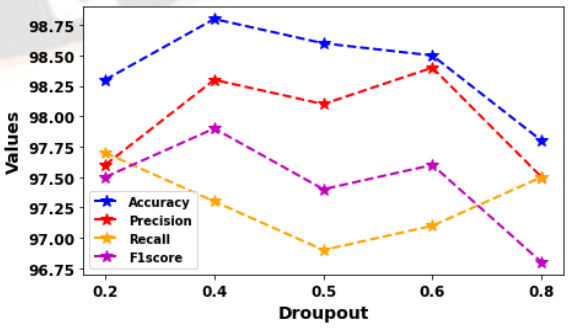

Figure 11. Dropout value of the model using the IMDB movie reviews dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

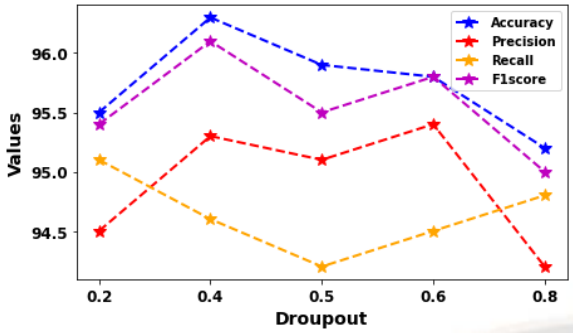

Figure 12. Dropout value of the model using Twitter entity sentiment analysis dataset

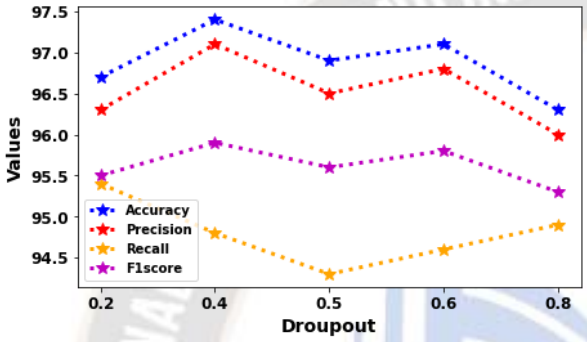

Figure 13. Dropout value of the model using Twitter Airline reviews dataset

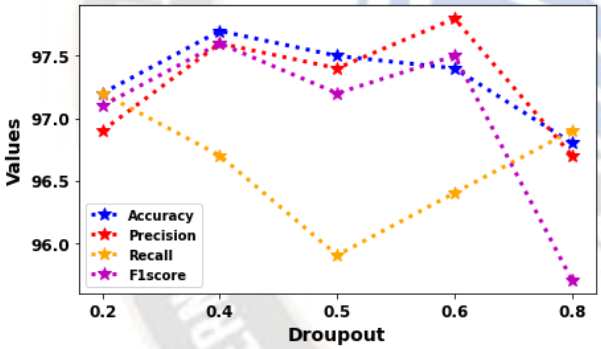

Figure 14. Dropout value of the model using the Amazon mobile reviews dataset

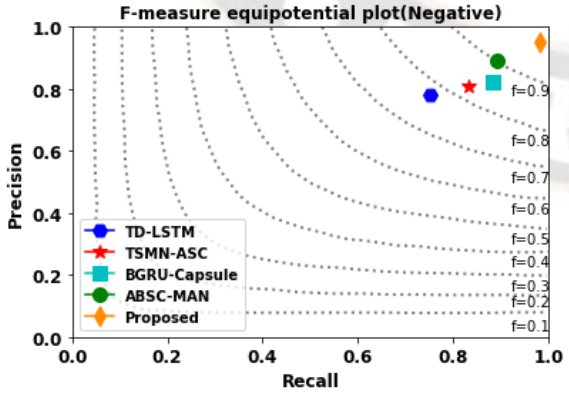

Figure 15. Equipotential plot for the IMDB movie reviews dataset

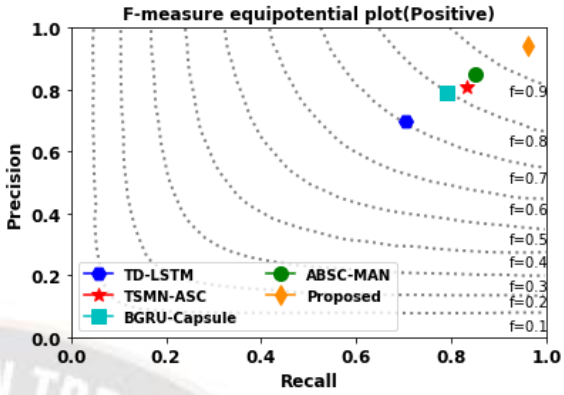

Figure 16. Equipotential plot for the Twitter entity sentiment analysis dataset

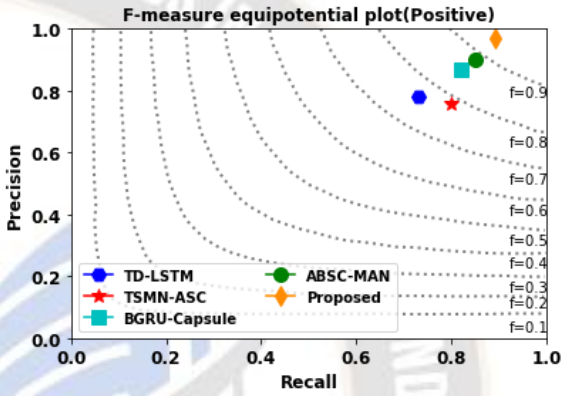

Figure 17. Equipotential plot for the Twitter Airline reviews dataset

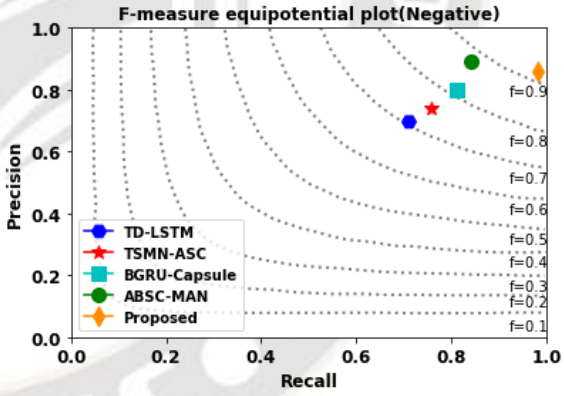

Figure 18. Equipotential plot for the Amazon mobile reviews dataset

Figure 19 shows the accuracy during the iteration of training and validation. It shows that when the number of iterations is low then the accuracy is increasing rapidly therefore by increasing the iteration the accuracy remains constant after so many iterations. The execution time of the proposed model is low compared to the existing method in all datasets represented in Figure 20.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

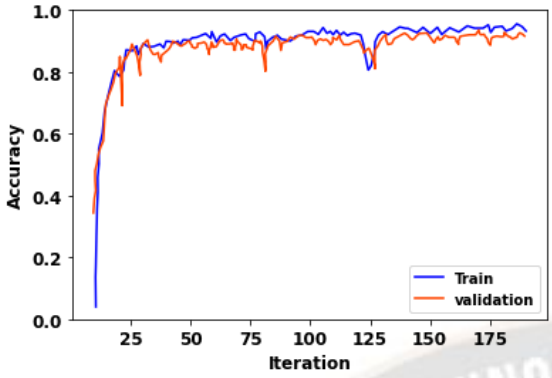

Figure 19. Accuracy of the model during the iteration

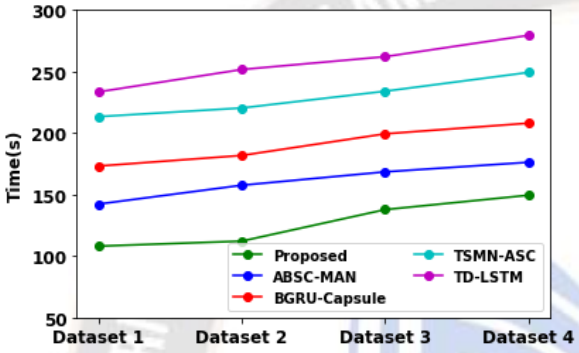

Figure 20. Execution time of the model using the datasets

The confusion matrix is used to define the sentiment classification of the model. The evaluation in the confusion matrix will be compared with similar existing models. The confusion matrix also provides the values for the performance metrics. The confusion matrix for the proposed model with four datasets is displayed in Figures 21–24.

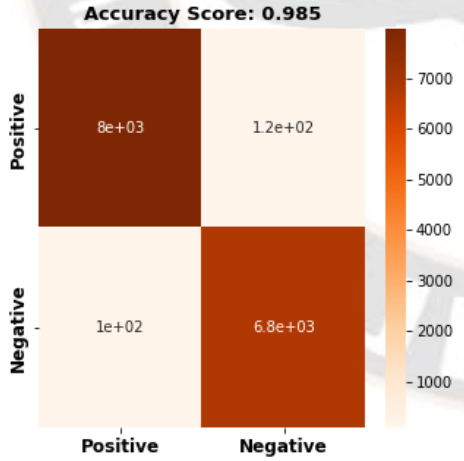

Figure 21. Confusion matrix for the IMDB movie reviews dataset

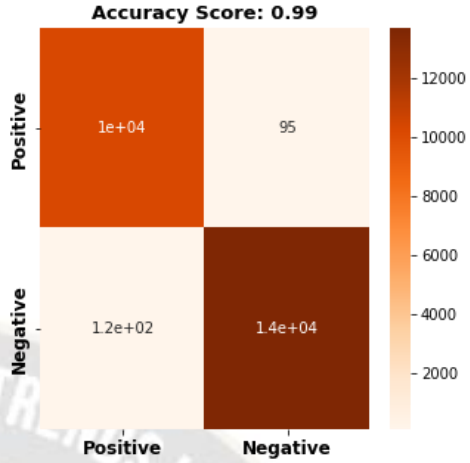

Figure 22. Confusion matrix for the Twitter entity sentiment analysis dataset

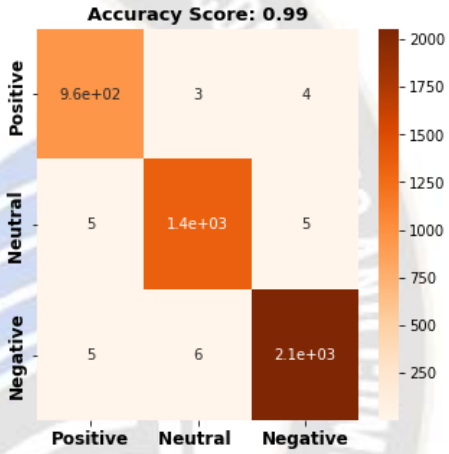

Figure 23. Confusion matrix for the Twitter Airline reviews dataset

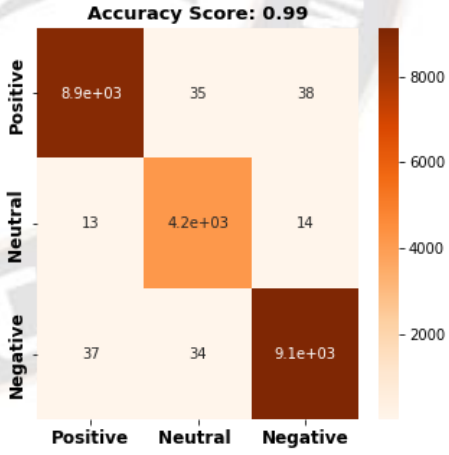

Figure 24. Confusion matrix for the Amazon mobile reviews dataset

E. Comparison of performance metrics

The bar graph from Figures 25 to 28 illustrates a comparison of performance metrics between the existing method and the proposed method using four datasets. The proposed approach performs better in all metrics when using four datasets. However, from the evaluation, all the existing methods provide above 80% accuracy, precision, recall, and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

F1-score in all datasets except the TD-LSTM model. The IMDB movie reviews dataset provides 98% accuracy, which is the highest accuracy in the proposed approach. The lowest accuracy value is 95%, which is produced by the Twitter entity sentiment analysis.

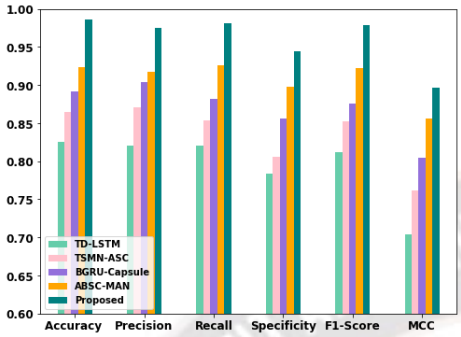

Figure 25. Comparison of the model using the IMDB movie reviews dataset

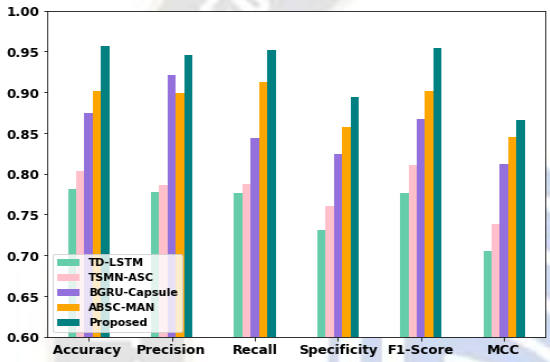

Figure 26. Comparison of the model using Twitter entity sentiment analysis dataset

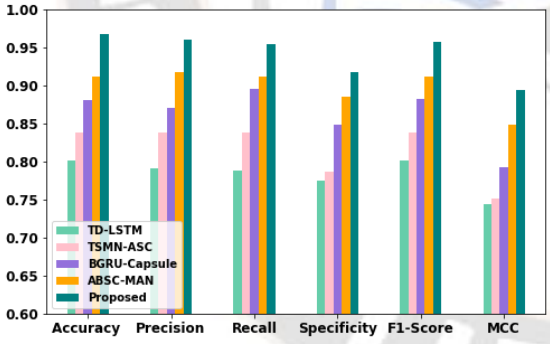

Figure 27. Comparison of the model using the Twitter Airline reviews dataset

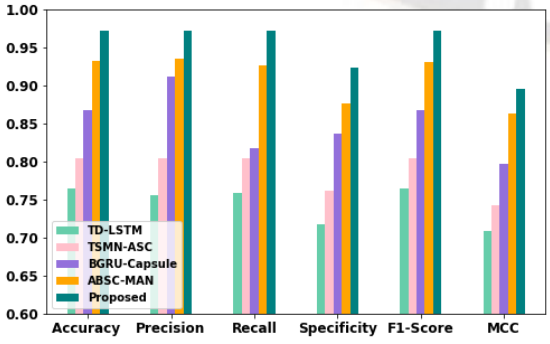

Figure 28. Comparison of the model using the Amazon mobile reviews dataset

V. Conclusion

The proposed context-preserving sentiment classification using BI-TCN and BIGRU with multi-head self-attention used the aspect-related sentiment representation to capture the global dependencies to provide the high sentiment polarity for the reviews from four datasets by minimizing the gradient problem, apparent and in apparent pattern issues during the classification. The performance of the proposed approach is measured using F1-score, recall, precision, MCC, and specificity these are compared with the four existing methods such as TD-LSTM, TSMN-ASC, BGRU-capsule and ABSC-MAN using the IMDB dataset, movie reviews dataset, Twitter airline reviews dataset and Amazon mobile reviews dataset. Finally, the comparison reveals that the proposed approach provides the highest accuracy of sentiment classification with high performance. Accordingly, the highest value of F1- score and precision is 97%, the highest value of recall is 98%, the MCC is 89%, the specificity is 94% and the accuracy is 98%. Future work aims to propose sentiment classification using advanced neural networks and attention mechanisms.

Acknowledgment

I confirm that all authors listed on the title page have contributed significantly to the work, have read the manuscript, attest to the validity and legitimacy of the data and its interpretation, and agree to its submission.

References

[1] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," Proc. 2016 conference on empirical methods in natural language processing, Nov. 2016, pp. 606-615.

[2] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," 2015. *arXiv preprint arXiv:1512.01100*.

[3] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," Proc. 2017 conference on empirical methods in natural language processing, Sep. 2017, pp. 452-461.

[4] X. Li, L. Bing, W. Lam, B. Shi, "Transformation networks for target-oriented sentiment classification," 2018. arXiv preprint arXiv:1805.01086.

[5] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, "Target-sensitive memory networks for aspect sentiment classification," Proc. 56th Annual Meeting of the Association for Computational Linguistics, Jul. 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need. Advances in neural information processing systems," vol. 30, 2017.

[7] A. Zhao, and Y. Yu, "Knowledge-enabled BERT for aspect-based sentiment analysis," Knowledge-Based Systems, vol. 227, no. 107220, 2021.

[8] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[9] W. Xue, W. Zhou, T. Li, and Q. Wang, "MTNA: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews," Proc.

Eighth International Joint Conference on Natural Language Processing, Nov. 2017, pp. 151-156.

[10] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," Proc. 52nd annual meeting of the association for computational linguistics, Jun. 2014, pp. 49-54.

[11] E. F. Ayetiran, "Attention-based aspect sentiment classification using enhanced learning through CNN-BiLSTM networks," Knowledge-Based Systems, vol. 252, no. 109409, 2022.

[12] Y. Xiao, and G. Zhou, "Syntactic edge-enhanced graph convolutional networks for aspect-level sentiment classification with interactive attention," IEEE Access, vol. 8, pp. 157068-157080, 2020.

[13] N. Zhao, H. Gao, X. Wen, and H. Li, "Combination of convolutional neural network and gated recurrent unit for aspect-based sentiment analysis," IEEE Access, vol. 9, pp. 15561-15569, 2021.

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018. arXiv preprint arXiv:1803.01271.

[15] S. Mai, S. Xing, and H. Hu, "Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1424-1437, 2021.

[16] Y. Zhang, Y. Ma, and Y. Liu, "Convolution-Bidirectional Temporal Convolutional Network for Protein Secondary Structure Prediction," IEEE Access, vol. 10, pp. 117469-117476, 2022.

[17] F.Teng, Y. Song, and X. Guo, "Attention-TCN-BiGRU: An Air Target Combat Intention Recognition Model," Mathematics, vol. 9, no. 19, pp. 2412, 2021.

[18] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," Proc. 2018 conference on empirical methods in natural language processing, pp. 3433-3442 , 2018.

[19] Q. Zhang, and R. Lu, "A multi-attention network for aspect-level sentiment analysis," Future Internet, vol. 11, no. 7, pp. 157, 2019.

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

[21] F. Long, K. Zhou, and W. Ou, "Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention," IEEE Access, vol. 7, pp. 141960-141969, 2019.

[22] Y. Lin, C. Wang, H. Song, and Y. Li, "Multi-Head Self-Attention Transformation Networks for Aspect-Based Sentiment Analysis," IEEE Access, vol. 9, pp. 8762-8770, 2021.