Attention-Based Convolution Bidirectional Recurrent Neural Network for Sentiment Analysis

Soubraylu Sivakumar, Koneru Lakshmaiah Education Foundation, India*
https://orcid.org/0000-0002-5073-8949

Haritha D., School of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, India

Sree Ram N., School of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, India

Naveen Kumar, School of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, India

Rama Krishna G., School of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, India

Dinesh Kumar A., School of Computer Science and Engineering Koneru Lakshmaiah Education Foundation, India

ABSTRACT

A customer conveys their opinion in natural language about an entity. Applying sentiment analysis to those reviews is a very complex task. The significant terms that influence the polarity of a review are not examined. The terms that have contextual meaning are not recognized and are present across multiple sentences in a review. To address the above two issues, the authors have proposed an attention-based convolution bi-directional recurrent neural network (ACBRNN). In this model, two convolution layer captures phrase-level feature while self-attention in the middle assigns high weight to the significant terms, and bi-directional GRU performs a conceptual scanning of review through forward and backward direction. The authors have conducted four different experiments (i.e., unidirectional, bidirectional, hybrid, and proposed model) on IMDB dataset to show the significance of the proposed model. The proposed model has obtained an F_1 score of 87.94% on IMDB dataset, which is 5.41% higher than CNN. Thus, the proposed architecture performs well compared with all other baseline models.

KEYWORDS

Bidirectional Gated Recurrent Unit (BGRU), Convolutional Bidirectional Recurrent Neural Network (CBRNN), Convolutional Neural Network (CNN), Self-Attention (SA), Sentiment Analysis

1. INTRODUCTION

Sentiment Analysis is the computational study of public opinions, emotions, sentiments, attitudes, and appraisals towards entities. The entity can be an issue, events, services, individuals, products, etc.

DOI: 10.4018/IJDSST.300368

*Corresponding Author

Everyday more reviews are posted on the internet due to advancements in technology. The objective of Sentiment Analysis is to conclude the polarity of the review as positive (good) or negative (bad). So, decision making plays an important role in Sentiment Analysis.

Decision making process helps the customer to get the right product and helps the organization to sell the right product to the customer. It benefits the consumer and corporates based on the decision. Applying a machine learning algorithms to the decision making process on the opinion is a challenging task. The goal behind this is to apply deep learning algorithms to build models that allow automatic extraction of features from the text. When Deep Learning Methods (DLM) is used in feature engineering, automatically the high-level features are learned without any human bias. In deep learning, the features are learned during the training process and no specialized domain knowledge is required by the researchers.

Convolutional Neural Network (CNN) (Yoav Goldberg, 2015) has a local pattern of connection between the neurons of adjacent layers. This connection helps to maintain a special spatially local correlation. This characteristic is helpful in the classification of the sentences in NLP. It finds strong local clues that appear in the different places of inputs regardless of input class membership. The local indicators are nothing but the key phrases that helps to identify the sentiment of a sentence. Consider the movie review from the IMDB dataset

Review 1: *"The movie has a moral story where the actor helps to resolve the social issue from the society."*

A convolutional layer extracts the local features from the movie review includes *"actor", "resolve", "social", "issue"* etc. Self-Attention distinguishes relevant and un-relevant parts of a movie review based on Parts of Speech (POS). It correlates the distinct parts of a longer sequence to compute the weight of a part of a sequence. In the above review, *"actor", "resolve", "social", "issue"* holds noun, verb, adjective, noun tag of POS. These tags are assigned higher weights by the Self-Attention layer and improve (Sindoori et al., 2017) the prediction score of the movie review. GRU (Chen Tao et al., 2017) is divergent of the recurrent network that doesn't have internal memory and has only two gates when compared with LSTM Sepp Hochreiter et al., 1997). The internal design of GRU is simple and takes less training time than LSTM. Bidirectional GRU scans the review in the forward and reverse direction (Li Zhang et al., 2017 and Jianqiao Hu et al., 2017). It has higher learning power to better understand contextual information. It relates the features that are located in distinct parts of the sentence.

The term *"movie"* is located on the left side of the sentence, while *"social issue"* is located on the right side. BGRU understand contextually and correlates the term or phrases located on two extreme ends of the movie review. We have conducted four experiments on the IMDB dataset viz., Unidirectional Neural Network (CNN, LSTM, GRU), Bidirectional Neural Network (BLSTM (Yu Zhao et al., 2017), BGRU), Hybrid Neural Network (CNN+LSTM, CNN+BGRU) and Attention Based Neural Network. The proposed attention based architecture is compared with baseline architecture. It obtained better results than other architecture. The contribution of the research work is listed below:

- We have designed a new architecture by integrating the attention layer with a hybrid convolution bidirectional recurrent neural network (ACBRNN).
- The proposed architecture extracts more relevant terms and assigns high weights to those terms based on a context that influences the polarity of the review.
- We have highlighted the importance of different layers in the proposed architecture with a movie review.

This paper is organized in the following way. The related work is presented in Section 2. The proposed architecture and the function of each layer are discussed in Section 3. Section 4 includes the comparative analysis of the various Deep Learning Models with the proposed architecture. Finally in Section 5, concluding remarks along with the future enhancement are presented.

2. RELATED WORK

2.1 Conventional Machine Learning and Deep Learning Models

Detecting emotions from tweets due to different language, slang and the limitation of the number of characters is very challenging. Rose et al., 2018 used the NRC emotion lexicon, online Thesaurus and WordNet-Affect to create a feature vector using lexicon based scheme. The weight assignments are done based on negation and punctuations. Conventional classifier like Naïve Bayes, Random Forest and SVM are employed for classification. An effective way of extracting lexicon features and novel weighting scheme have achieved a maximum accuracy with Random Forest is 73% for tweets. Chronic disease is common in adults, affects individual health and leads to death. Khamparia et al., 2020 have proposed a combined method for data reduction and supervised classification of kidney diseases. The samples are taken from UCI with 400 instances and 25 features. They have employed a Gaussian radial basis function with SVM and Principal Component Analysis (PCA) to differentiate the normal and sick individuals. This proposed architecture obtained better results than other baseline methods.

J. D. Prusa et al., 2017 used a character level approach instead of word text representation. Their approach needs a complex model for classification and also needs more training data. To overcome this limitation, they have introduced a new embedding named log-m approach by replacing 1-of-m approach. This approach obtained better performance result with less training time. The convolution layer padding results in an improvement in performance with an increase in time. M. Hughes et al., 2017 have employed CNN based approach for text categorization of sentence level on a corpus of medical text. They have compared their approach on three other methods like Mean Word Embeddings, Sentence Embeddings and Word Embeddings with BOW. Their results show that the CNN-based approach is performing better than another approach in terms of accuracy by 15%.

The English article errors are corrected automatically by a CNN based model introduced by Sun Chengjie et al., 2015. Instead of employing prior NLP knowledge or humans for extracting the features, they simply took the surrounding words of the articles as features. They have used CMU pronouncing dictionary, for revising the output of their CNN module for when to use ‘a’ or ‘an’ articles. The training of this model is conducted on non-annotated error corpus and annotated error corpus. The A. F. Agarap, 2017 has replaced the Softmax function in the output layer of a GRU with a Support Vector Machine. They have performed analysis on network traffic data from Kyoto university honeypot systems (S. G. Kaveeya et al., 2017 and P. N. Saranu et al., 2018). The cross entropy is replaced in place of the margin based function. The GRU-SVM based approach obtained an accuracy of 84.15% than the traditional Softmax approach with 70.75% accuracy.

J. Y. Lee et al., 2016 have introduced a model based on RNN and CNN for sequential short text classification. This classified short text helps further in the dialog act classification task. A dialog act differentiates a spoken statement in a dialog based on semantic, pragmatic and syntactic criteria. This ANN based approach achieves better results on three datasets when compared to state of art methods. Biosensor plays a major role in the military to the health sector. Fault identification and analysis are the primary concern of sensor data. S. A. Meti et al., 2019 proposed a neural network to identify and classify the faults with auto associative integrated with cascade feed forward propagation method. There is an increase in correlation coefficient with a reduction in mean square error of the proposed architecture compared with the conventional auto associative neural network.

Entity recognition and extraction of relevant information from unstructured documents are difficult. The existing coarse-grained entity recognition methods find fewer predefined entity categories. Fine-grained methods provide better recognition of the entity. Recent works of recognition of the entity are done with bidirectional RNN. Due to the complex structure of BLSTM it takes more amount of time in the training process. K. J. Dhrisya et al., 2020 has proposed Fine-Grained Entity type Classification for Gated Recurrent Unit (FGEC-GRU). This model is applied to the two openly available datasets OntoNotes and FIGER. They have obtained an F1 score of 51.04% and 64.00% for OntoNotes and FIGER respectively. Customers post their reviews about a product or a brand in online forums. The reviews are validated with different classifiers like SVM, Maximum Entropy and Naïve Bayes (N. V. Patel & H. Chhinkaniwala, (2019)). The above machine learning algorithms are applied to movie reviews and positive-negative datasets to compare the performance result using the single fold and five fold validation. They have achieved an 88.39% and 87.60% for movie-review and positive-negative datasets respectively with SVM. The SVM has obtained better performance than Naïve Bayes and Maximum Entropy.

To keep the house, office and organization premises away from the fire hazard a prior warning system and emergency services are required to have a safe and smart environment. A fire alarm detection simulator (Muhammad Asif et al., 2020) with various sensors like gas, temperature and humidity are established to collect real-time data to yield a synthetic dataset. This data is evaluated and analyzed with a different classifier to choose an optimal classifier for fire exposure. The WEKA mining tool is employed for the comparative analysis of different classifiers. J48 decision tree classifier obtains a good result with an F1 value of 0.98%.

2.2 Hybrid Deep Learning Models

Nianwen Si et al., 2018 have designed a combined model of the BLSTM and Segment-based CNN to capture rich textual information. The output from the feature extraction layer is fed into the feed-forward network. To predict dependency labels, the feed-forward network is trained with max-margin criteria. Finally, dynamic programming is used to find the best dependency structure from the graph for the sentence. Classification of a sentence based on targets or objects present in a sentence was done by T. T. H. Le et al., 2016. Based on opinionated words a sequence model classifies the sentences into different groups. Later, these grouped sentences are fed into CNN for sentiment classification. They have used four datasets in the evaluation and compared their performance with many models. The sentences are separated into groups using BiLSTM-CRF. Grouping sentences into multiple divisions improves the result of sentiment analysis (V. L. Sarvani et al., 2020).

D. Liang et al., 2016 have proposed AC-BLSTM. It is a combination of Asymmetric CNN with BLSTM. The asymmetric layer is used to learn higher level representations of phrase features. This feature (E. Sreedevi et al., 2019) is fed into the BLSTM to learn the long-term dependency in the sentence. They also proposed a generative model with AC-BLSTM to act as a semi-supervised learning framework.

2.3 Attention Based Deep Learning Models

A new attention based relation extraction is introduced by X Lio et al., 2017. This Att-BGRU-HN extracts the relation from the New york Times Corpus. The addition of highway layers in the network achieved an improvement in performance and outperforms the other relation extraction models. Z. Yang et al., 2016 have used the attention layer to map the knowledge of document organization into the model structure. Traditional architecture doesn't consider the context of a word in the classification of the document. The model structure is designed in a way to consider the sentence and later aggregated to the overall document. Two attention layers are included one for word and another for sentence level.

To perform semantic role labeling Tan Z. et al., 2017 introduced a neural network model with Self-Attention. This network connected and relates two tokens that are separated arbitrarily in a sentence. Self-Attention is better than RNN in processing, analyzing, choosing and processing the information. In deep learning, the recurrent network along with Self-Attention achieved the best performance in classification and semantic labeling tasks. Paulus R. et al., 2017 combined word prediction using the supervised and Reinforcement Learning (RL) method for abstractive summarization. They used intra and inter attention for the summarization of text. Intra attentions are used to record the earlier term weights while calculating the current term in the sequence. In summarization, intra attention recognizes (V. Soniya et al., 2017) only the words that are generated by the decoder. Table 1. lists the survey of existing work methodology and their limitations.

In aspect-based sentiment classification, the attention mechanism contained with RNN primarily focuses only on semantic information rather than important syntactical constraints. To focus on syntactical constraints a Multi-head self-attention based Gated Graph Convolution Network (MGGCN) is introduced by L. Xiao et al., 2020. This network constructs a definition tree by filtering the useless syntactic information with a gate at each node of Graph Network. A syntax-aware context dynamic weighted (SGDW) layer fuses the semantic-rich contextual word with syntactic knowledge based on a certain threshold. These proposed methods are applied on Twitter, rest14, laptop 14, rest10 and rest16 datasets. It has obtained an F1 score of 72.97% (Twitter), 75.03% (rest14), 81.48% (laptop14), 80.68% (rest10) and 88.41% (rest16). A new model combines a biLSTM or a biGRU and an Enhanced Multi-Head Self-Attention mechanism (EMHSA) is proposed by XL. Leng et al., 2021. The EMHSA is a two-layer modified Transformer model. This modified Transformer is that it's masking operation and the last feed-forward layers are removed. Besides, the loss function of this new model is the sum of the weighted RMSE and the cross-entropy loss. It improves the performance of the auto-encoder. EMHSA is used to encode the inter-sentence information as the middle hidden layer. The word embedding word2vec is replaced with BERT to provide better performance. This proposed Hybrid Recurrent Neural Network and EMHSA obtained an F1 score of 75% and 87.4% for SST-1 and IMDB datasets respectively.

3. MATERIALS AND METHODS

Sentiment analysis is a process of deriving opinion about a sentence to determine whether it is positive or negative. A few list of movie reviews are given below:

Review 2: *"Busy is so amazing! I just loved every word she has ever done- freaks and geeks, Dawson' s creek, white chicks, the smokers. after the first time i saw home room i went and got it the next day. i am a big fan of her and she has a lot of fans here in Israel. if someone hasn' t saw is excellent movie than don' t waist more time and go see it now."*

Review 3: *"Terrible movie. Nuff Said. These Lines are Just Filler. The movie was bad. Why I have to expand on that I don' t know. This is already a waste of my time. I just wanted to warn others. Avoid this movie. The acting sucks and the writing is just moronic. Bad in every way."*

The above reviews are examples of positive and negative reviews from the IMDB dataset. Review 2 is positive because of the word *'excellent'*, while review 3 is negative because of the word *'bad'*.

3.1 Preprocessing

It converts the raw text into valuable information. Without preprocessing it contains noisy, irrelevant and redundant data. Preprocessing steps in the classification task improve the performance by removing the unwanted text. The preprocessing is done before classification viz., punctuation removal, case conversion and stop word removal. A single movie review is taken from the IMDB dataset and the preprocessing steps are applied. It is shown in Figure 1. In punctuation removal, two commas and one full stop are removed from the review. The second step in the preprocessing is case conversion, it converts capital *'A'* and *'S'* to lower case. Stop words *'a'*, *'of'*, *'and all the'*, *'of a'*, and *'or a'* are removed from the movie review.

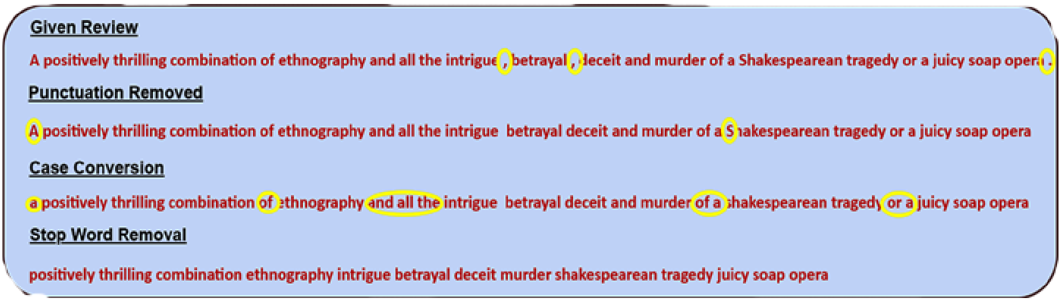
Table 1. Lists the survey of existing work methodology and their limitations

Sl. No.	Author	Methodology Used	Limitations
1.	(Rose S.L. et al., 2018)	Combining lexicon feature extraction and weighting method with Random Forest	Global and Local features are not considered for detecting emotions in tweets.
2.	(Khamparia A. et al., 2020)	Principal Component Analysis (PCA) and Gaussian radial basis function with SVM for classification of Chronic diseases.	Opposite word pairs located very close Representation to each other in the vector space is not addressed.
3.	(J. D. Prusa et al., 2017)	CNN with log-m embedding for classification.	It involves with computational complexity. High end hardware is required for the implementation.
4.	(Mark Hughes et al., 2017)	CNN are used for text categorization of medical text.	More domain specific terms are required for better classification.
5.	(Sun Chengjie et al., 2015)	CNN are used to correct the article errors.	Lacks in domain specific knowledge in correcting the errors.
6.	(A. F. Agarap, 2017)	GRU with SVM for classification of network traffic data.	GRU has low learning efficiency and slow convergence rate.
7.	(Ji Young Lee et al., 2016)	CNN for classification of short text.	They lag performance in modelling long sequences.
8.	(S. A. Meti et al., 2019)	Auto Associative Neural Network for Fault identification and analysis.	Co-occurrence of words is not considered for classification.
9.	(Nianwen Si et al., 2018)	BLSTM and CNN are used for feature extraction. Feed forward network are used for prediction.	Global features are not captured and addressed.
10.	(T. T. H. Le et al., 2016)	BLSTM and CRF for grouping of sentence. CNN for sentence classification.	Words located at distance location in a sentence are not given importance.
11.	(D. Liang et al., 2016)	Asymmetric CNN with BLSTM for classification.	Terms having relative importance are not given higher weights.
12.	(X Lio et al., 2017)	Att-BGRU-HN for Relation Extraction.	Phrases level features are not extracted.
13.	(Z. Yang et al., 2016)	Two attention layers for document classification.	Each sentence is encoded in complete isolation.
14.	(Tan Z. et al., 2017)	Semantic role labelling using Self-Attention.	Polysemy words are not addressed.
15.	(Paulus R. et al., 2017)	Intra and inter attention for abstractive summarization.	Long term dependency is not addressed.

3.2 Proposed Architecture

We have introduced an Attention based Convolution Bidirectional Recurrent Neural Network (ACBRNN). This architecture is composed of two layer convolution neural network at the top, a self-attention network in the middle and a bidirectional GRU at the bottom. Figure. 2. shows the architecture diagram of the proposed ACBRNN. This ACBRNN model extracts phrase level features with the help of the convolution layer and the maxpooling layer at the top. The extracted feature vector is fed into the self-attention network in the middle. The attention network assigns high weights to the relevant terms in the sentence. The BGRU at the bottom captures the long-term dependency within the sentence.

Figure 1. Preprocessing steps applied to a movie review.



Given Review

A positively thrilling combination of ethnography and all the intrigue , betrayal , deceit and murder of a Shakespearean tragedy or a juicy soap opera .

Punctuation Removed

A positively thrilling combination of ethnography and all the intrigue betrayal deceit and murder of a Shakespearean tragedy or a juicy soap opera

Case Conversion

a positively thrilling combination of ethnography and all the intrigue betrayal deceit and murder of a shakespearean tragedy or a juicy soap opera

Stop Word Removal

positively thrilling combination ethnography intrigue betrayal deceit murder shakespearean tragedy juicy soap opera

3.2.1 Word Embeddings

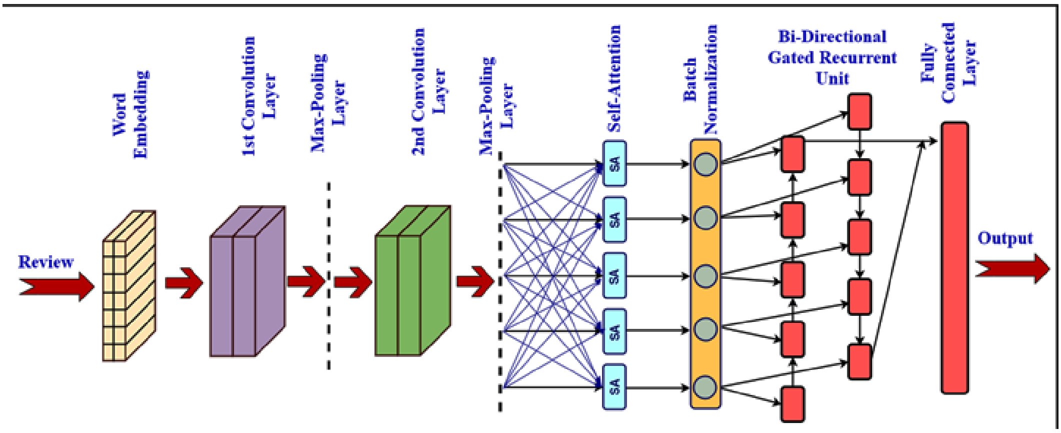
All machine learning and deep learning algorithms (A. M. Senthil Kumar et al., 2020) are operating on continuous value instead of plain text. To convert the plain text into vectors, each algorithm includes a word embedding method. One hot encoding converts the input into vectors whose size is equal to the whole vocabulary. The dimensionality space of the vector is significantly reduced when created by word embedding (S. Sivakumar et al., 2020). This forms the first layer in the proposed architecture.

Consider there are ‘N’ words in a sentence ‘S’. i.e., $S = \{W_1, W_2, W_3 \ldots W_N\}$. Let ‘d’ be the dimension of the word embedding. A single word W_t from the sentence ‘s’ is multiplied with embedding matrix W_e. The $W_e \varepsilon \Re^{dx|l|}$ represents word embedding matrix. Then, one hot encoding representation is obtained by the equation (1) as

$$e_t = W_e.W_t \tag{1}$$

A sentence ‘S’ is converted into a one-hot vector embedding as $S_e = \{e_1, e_2, e_3, \ldots, e_N\}$. In the sentence S_e, e_i represents one hot vector of word W_i.

Figure 2. Architecture diagram of proposed Attention based Convolution Bidirectional Recurrent Neural Network. (ACBRNN)



3.2.2 Convolution Layers

Let $C \ \varepsilon \ R^{d*l}$ be the output of the embedded matrix is given to the convolution layer as input. A new feature C_i is produced from a window $H \ \varepsilon \ \Re^{dxw}$ of 'W' words with the convolutional kernel. From a window of words $C\left[*, i : i + w\right]$ a new feature C_i is generated by the equation (2)

$$C_i = \sigma\left(\sum\left(C\left[*, i : i + w\right] oH\right) + b\right) \tag{2}$$

Here, o is a Hadamard product used in the matrices, σ is a ReLu non linear function, and $b \ \varepsilon \ R$ is a bias. For each possible word in a sentence, the kernel is applied with a given window to produce a new feature $C = \ \left[C_1, C_2, \cdots, C_{l-w+1}\right]$ with \$$C \ \varepsilon \ R^{l-w+1}$.

A bias neuron in each layer of the CNN is set to one and it is connected to next layer neuron. The output layer of the convolution network doesn't have a bias neuron. To calculate the output width and height of the convolution layer, the width 'W', height 'H', filter width 'F_w' and filter height 'F_h' of the input is used. It is shown in the equation (3) and (4).

$$\text{Output width} = \frac{W - F_w + 2P}{S_w + 1} \tag{3}$$

$$\text{Output height} = \frac{H - F_h + 2P}{S_h + 1} \tag{4}$$

where 'p' refers to the zero padding. In convolution, the horizontal and vertical strides are represented as S_w and S_h respectively. The padding scheme helps in determining the output size of the layer. The *'SAME'* and *'VALID'* are the standard padding schemes available for convolution operation. This architecture uses *'VALID'* padding scheme. In equation (5) and (6), the output width and height of the architecture are given with zero padding.

$$\text{Output width} = \left(\frac{W - F_w + 1}{S_w}\right) \tag{5}$$

$$\text{Output height} = \left(\frac{H - F_h + 1}{S_h}\right) \tag{6}$$

The ReLu function used in this architecture is shown below. The derivative of this function is also monotonic. This function is used in most of the Deep Learning Models (T. R. Kumar et al., 2020a). The output of the function given in equation (7) will be zero for negative input. Otherwise, the output of the function will be same as input.

$$f\left(y\right) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \tag{7}$$

3.2.3 Attention Layer

The attention layer creates a dependence between distinct parts of a sentence to create a relation at different times. This layer chooses the pertinent parts of the sentence to improve the prediction process. A sentence from a movie review may contain Parts Of Speech (POS) like noun, adjective, verb, preposition, adverb, determiner, etc. It assigns high weight to the adjective and verb, while it imparts a lower weight to irrelevant terms like determiner and prepositions. Dzmitry Bahdanau et al., 2014 has designed an attention model that calculates a vector c_t based on the weighted mean of the sequence 'h' and it is given in equation (8)

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{8}$$

The weighted mean is produced by a hidden concealed state h_t. Here, T is the time step needed in preparing the input sequence and α_{ij} is a weight measured at time 't' on input for a state h_j. The new state 's' measures the context vector, where s_t is confide on 's_{t-1}', 'c_t' and the output at time 't-1'. The weight αe_{tj} are calculated by the equation (9)

$$e_{tj} = a\left(s_{t-1}, h_j\right), \alpha e_{tj} = \frac{exp\left(e_{tj}\right)}{\sum_{k=1}^{T} exp\left(e_{tk}\right)} \tag{9}$$

The preceding state s_{t-1} and h_j are used in the calculation of 'a' learning function. This function is helpful in the calculation of h_j. It produces a fixed length 'c' vector in the equation (10) with the help of weighted mean of 'h'.

$$e_t = a\left(h_t\right), \alpha_t = \frac{exp\left(e_t\right)}{\sum_{k=1}^{T} exp\left(e_k\right)}, c_t = \sum_{t=1}^{T} \alpha_t h_t \tag{10}$$

When the input is not changing with time 'T', the entire network will function in the absence of self attention. The entire self attention functionality will be used, when there is a varying sequence in the input. The un-weighted average of h_t is helpful in calculating 'c' is given in equation (11)

$$c_t = \frac{1}{T} \sum_{t=1}^{T} h_t \tag{11}$$

3.2.4 BGRU

GRU has few gating units than LSTM. It has only two gates viz., update and reset gate. This gate helps to overcome vanishing gradient issue. The reset gate decides a method to combine the new input with memory, while update gate determines the amount of previous memory to remember for future. A BGRU is composed of twso GRU connected in reverse direction. The gates are designed in BGRU to store information in both directions for longer time. It provides better performance than convolution network. In a sequence of input, this neural network can use both future and past context.

BGRU is expressed in equation (12) with forward block $\vec{h}_t$ and with backward block $\overleftarrow{h}_t$ as

$$h_t = \left[\vec{h}_t, \overleftarrow{h}_t\right] \tag{12}$$

The final output y_t of the BGRU in equation (13) at time 't' can be written as

$$y_t = \sigma_y\left(w_y h_t + b_y\right) \tag{13}$$

where σ is the activation function, w_y is the weight matrix and b_y is the bias. The update gate z_t is given in the equation (14) as:

$$z_t = \sigma_y\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{14}$$

where w_x , u_z and b_z are parameter matrix and bias. The reset gate r_t is written in equation (15) as

$$r_t = \sigma_y\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{15}$$

Where w_r , u_r and b_r are parameter matrix and bias. At time t, the output vector h_t is given by the equation (16) as

$$h_t = z_t h_{t-1} + \left(1 - z_t\right) \odot_h \left(W_h x_t + U_h\left(r_t \otimes h_{t-1}\right) + b_h\right) \tag{16}$$

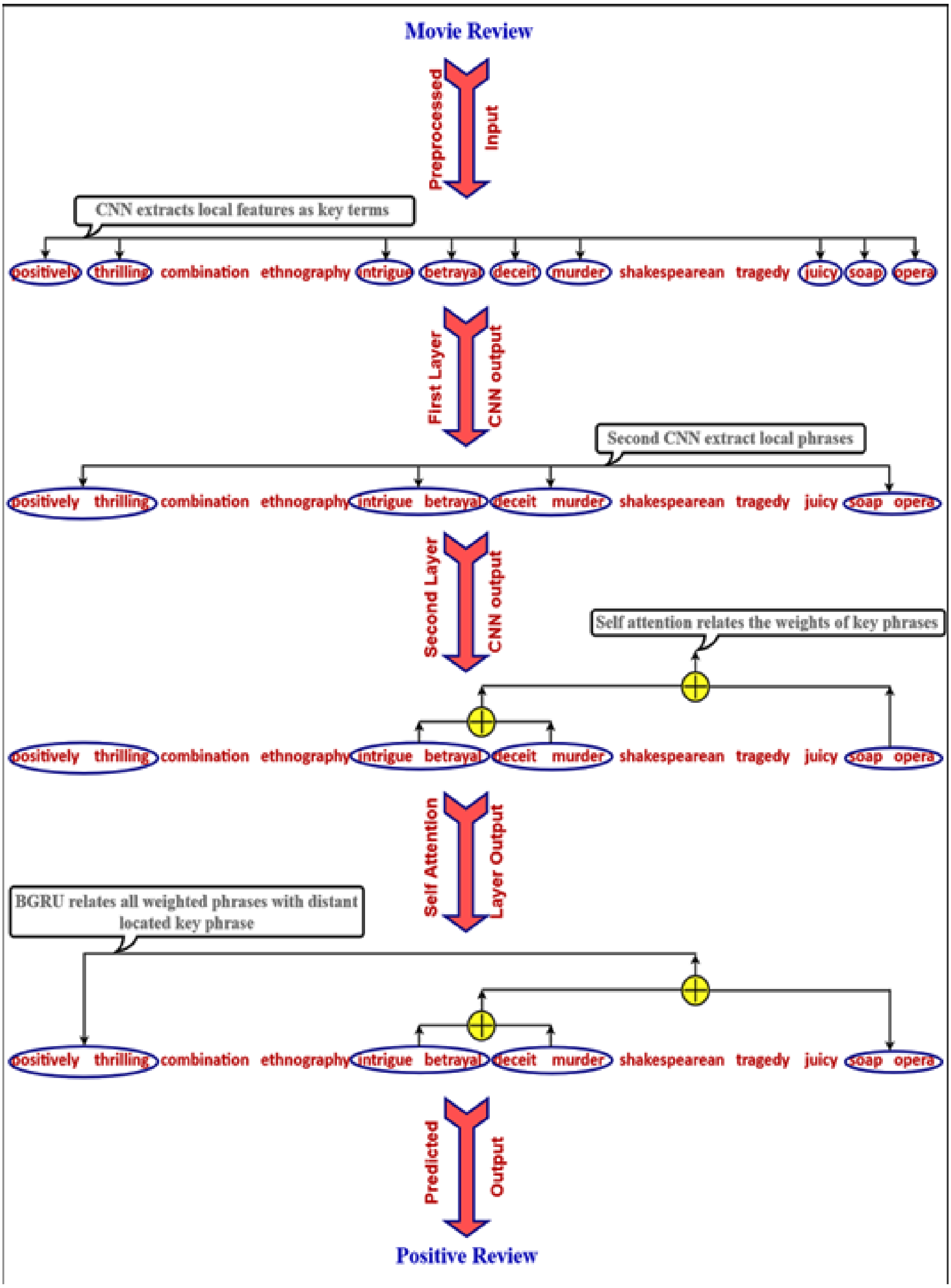
where $\otimes$ is an element wise multiplication and $\odot_h$ is hyperbolic tangent function.

3.3 Discussion

A sequence of steps involved in processing a movie review with the proposed architecture is shown in Figure 3. Preprocessed review is given as input to the first layer of CNN. A convolution layer primarily focuses on the local feature despite global feature. The first convolutional layer extracts the important key terms such as *"positive", "thrilling", "intrigue", "betrayal"* etc. from the movie review. Extracted key terms are passed through the second convolutional layer. Now in this layer the key terms are grouped into key phrases as a feature through the local spatial correlation. The key phrases formed in this layer are *"positive thrilling", "intrigue betrayal"* etc.

Self-Attention layer focuses on the pertinent part of the movie review like noun, verb, adverb despite irrelevant terms like preposition, determiner etc. The key phrases like *"intrigue betrayal", "deceit murder"* and *"soap opera"* are given high weights in the movie review by the Self-Attention

Figure 3. Sequence of steps involved in processing a movie review with the proposed architecture


Movie Review
Preprocessed
Input
CNN extracts local features as key terms
positively thrilling combination ethnography intrigue betrayal deceit murder shakespearean tragedy juicy soap opera
First Layer
CNN output
Second CNN extract local phrases
positively thrilling combination ethnography intrigue betrayal deceit murder shakespearean tragedy juicy soap opera
Second Layer
CNN output
Self attention relates the weights of key phrases
positively thrilling combination ethnography intrigue betrayal deceit murder shakespearean tragedy juicy soap opera
Self Attention
Layer Output
BGRU relates all weighted phrases with distant located key phrase
positively thrilling combination ethnography intrigue betrayal deceit murder shakespearean tragedy juicy soap opera
Predicted
Output
Positive Review

layer. BGRU as a BRNN has a special property of combining the distance related phrases in a sentence. It scans through the review in forward and reverse direction, that enables to understand the contextual information in a better way. It relates four key phrases together to predict (S. Nimmagadda et al., 2020) the whole sentence as a positive even though this feature are located far apart from one another.

4. EXPERIMENT AND RESULT

4.1 Experimental Setup

We performed a comparative analysis of various Deep Learning Methods along with the proposed method ACBRNN. For experimental setup, we used the system Intel Core i3 2.40 GHz, 6GB RAM with 70GB hard disk in Ubuntu 16.04 environment. We used the IMDB dataset from Stanford University. All the experiments were performed using Python 3.5 version, *Tensorflow* 1.3 version and *Keras* 2.0.8 version. *Keras* is a deep learning library that contains all Deep Learning Methods. Pickle 1.3 library is used to save and load the preprocessed object. The *sklearn* library is used for retrieving the confusion matrix. The embedding layer parameters for top word, max length and embedding dimension chosen for the IMDB dataset (E. Sreedevi et al., 2021) are 110000, 5000 and 30 respectively for all the models. The parameter used in the convolutional layer of all model filters, kernel size and activation function is 250, 3 and ReLu respectively. Adam and Sigmoid are the optimizer and activation functions used in all the models. A dropout of 0.1 is used in the output of the embedding layer and BGRU. BGRU uses 128 memory cells.

4.2 IMDB Dataset

It is a binary class dataset (A. L. Maas et al., 2011). The dataset contains 25,000 (12500 positive and 12500 negative) instances in training and 25,000 (12500 positive and 12500 negative) instances in the testing set. This dataset also contains 50,000 unlabeled movie reviews. Each movie review has several sentences, which is considered a key aspect of this dataset. It has a maximum collection of 30 reviews from a single movie. The average length of the documents is 255 words and the maximum length of the document is 300 words for the IMDB dataset. Total there are 239232 words in the vocabulary of IMDB. The train- validation-test split are used to divide the dataset as 25,000, 5000 and 20000 in the ratio of 50%, 10% and 40% for train, validation and test split respectively.

4.3 Metrics

The goal of the metrics is to measure and understand the effectiveness of the proposed architecture. The performance of the models is evaluated using Accuracy, Precision, Recall and F_1 Score. The four components of the confusion matrix that are used in the above performance metrics are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP refers to the positive movie reviews that are detected as positive. TN means the negative movie reviews that are classified as negative. FP means the negative movie reviews that are detected as positive. FN refers to the positive review that is wrongly classified as a negative movie review.

4.3.1 Accuracy

It is a proposition of perfectly categorizing the movie review to the total samples of movie review. It is interpreted in the below equation (17).

$$\text{Accuracy}=\frac{\left(TP+\text{TN}\right)}{\left(TP+\text{FP}+\text{TN}+\text{FN}\right)} \tag{17}$$

4.3.2 Precision

It conveys the percentage of positive movie review that are predicted as positive. This metric helps to determine the effectiveness of the model and it is given in equation (18).

$$\text{Precision} = \frac{TP}{\left(TP + \text{FP}\right)} \tag{18}$$

4.3.3 Recall

It relates the percentage of real positive movie review correctly predicted as positive by the system. Recall metric is defined in the equation below (19).

$$\text{Recall} = \frac{TP}{\left(TP + \text{FN}\right)} \tag{19}$$

4.3.4 F_1 Score

It is a harmonic mean of Precision and Recall. This metric is used to measure the performance of two classifiers and it is interpreted in equation (20).

$$F_1 = \frac{2 * \text{TP}}{\left(2 * TP + \text{FP} + \text{FN}\right)} \tag{20}$$

4.4 Result and Discussion

Various experiments are conducted to highlight the importance of the proposed architecture in choosing the significance terms viz., unidirectional (CNN, LSTM and GRU), bidirectional (BLSTM and BGRU) and hybrid models (CNN+LSTM and CNN+BGRU) including the proposed architecture (ACBRNN). The metrics that are used in comparing the performance of the model are Mean Squared Error, Accuracy, Precision, Recall and F_1 score. The experimental results of the baseline models and proposed model are discussed in the following section.

4.4.1 Unidirectional Neural Network

Consider a negative review from IMDB dataset:

Review 4: *"The completeness of a scene and sequence of transition from one scene to another is poorly pictured in the movie."*

Review four represents a negative polarity from the IMDB dataset. The phrase *"poorly pictured"* represents the nature of shooting the scene by a director. The CNN is looking for similar and local patterns from the input. This network is more suitable for image and video processing. It obtains an F_1 score of 82.53% and it is shown in Table 2. RNN is a special type of neural network suitable for sequence processing like text and speech. It has a memory cell to remember the previous sequence in time. Gated Recurrent Unit achieved an F_1 score of 85.66%. Long Short Term Memory has an additional gate named forget gate. This forget gate enables the network to remember the important parts in the sequence.

The CNN network captures the local features like *"completeness", "sequence", "poorly",* etc. There are two issues in the review that enables it to be negative. The first issue is *"completeness of the scene",* while the next issue is *"transition from a scene"*. GRU network can address only the

Table 2. Comparison of various metrics for unidirectional neural network on IMDB dataset.

	CNN	GRU	LSTM
MSE	0.1785	0.0503	0.1169
Val_Acc(%)	82.54	85.18	87.01
Test_Acc(%)	89.15	99.64	95.50
Precision	82.59	88.43	90.87
Recall	82.46	88.53	82.28
F_1 Score	82.53	85.66	86.36

second issue and it is very close to the sentiment *"poorly"*. LSTM has a forget gate that remembers the first issue and second issue of the review. The first issue is located far from the sentiment word *"poorly"*. This network achieves a better performance than the other two unidirectional networks. Due to this reason, LSTM has an increase in F_1 score of 0.70% then GRU.

4.4.2 Bidirectional Neural Network

The bidirectional architecture processes the inputs in two ways, from past to future and future to past. Thus, a bidirectional network has more learning capability than a unidirectional network. They preserve the information from the future during the backward run. A positive review is taken from the dataset and given below:

Review 5: *"It is a traditional romantic Indian movie, despite breath taking action stunts. Spanish girl has fallen in love with the Indian guy. This movie is interesting and affectionate to watch under any age from a family."*

Two bidirectional architectures are employed viz., BLSTM and BGRU. In review 6, a unidirectional network cannot understand the reason why the movie is liked by everyone. But, a bidirectional network can learn the contextual information more in-depth from the movie review due to forward and backward scanning of the network. The BLSTM and BGRU networks provide a better F_1 score of 86.52% and 87.09% respectively when compared to unidirectional architecture and it is displayed in Table 3.

4.4.3 Hybrid Neural Network

To attain the best performance in neural network and to better understand the complex inputs a hybrid approach is introduced. Two important hybrid models are used in the experiment viz., CNN

Table 3. Comparison of various metrics for BRNN on IMDB dataset.

	BLSTM	BGRU
MSE	0.1259	0.1393
Val_Acc(%)	86.50	86.71
Test_Acc(%)	94.00	93.89
Precision	86.39	84.62
Recall	86.66	89.72
F_1 Score	86.52	87.09

Table 4. Comparison of various metrics for hybrid neural network on IMDB dataset.

	CNN+LSTM	CNN+BGRU
MSE	**0.1091**	**0.1220**
Val_Acc(%)	85.11	87.33
Test_Acc(%)	88.51	94.89
Precision	77.75	85.59
Recall	98.48	87.63
F_1 Score	86.90	87.62

with LSTM and CNN with BGRU. In the first model (T. R. Kumar et al., 2020b), the convolution layer excerpt the local features at the top and LSTM connects the local features that are located in a long distance in the prediction. This model links the phrase *"romantic indian movie"* at the left side of review 7 with the phrase *"under any age"* at the right side. It obtained an F_1 score of 86.90% and it is showed in Table 4. The second model has two convolution layers at the top that extract the phrase level aspect and BGRU at the bottom perform a contextual based scanning. The phrase aspect extracted by CNN is located in the distinct parts of the review. These features are strongly related to each other through the contextual scanning of BGRU. It increases the prediction score of a review. Due to this integrated approach, it scored an F_1 value of 87.62%.

4.4.4 Attention based Neural Network (Proposed)

To find the relative terms and avoid less important terms in the prediction process attention layer is used. The attention layer finds the pertinent parts of the sentence and assigns a high score to that sentence. This proposed architecture consists of two convolution layers at the top, a Self-Attention layer in the middle and a Bidirectional GRU at the bottom. A positive review is taken from the IMDB dataset and it is given below:

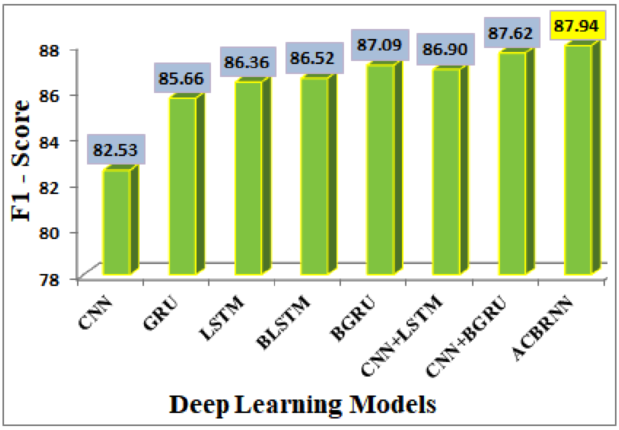
Review 6: *"In the drastic climate change of mount Everest, John Hawkes health becomes worst and struggles a lot for survival. Rob calls Helen to bring oxygen and water to the current location. He handles the unpredictable situation in a better way, which reflects the real act of friendship in the movie."*

The convolution layer finds the local feature in a given review. A combination of two convolution layers extracts phrase-level features. The two-layer CNN extracts phrases like *"worst struggles"*, *"drastic climate"*, *"oxygen water"* etc. from the review. Self-Attention relates the distinct parts of the

Table 4. Comparison of various metrics for proposed Attention based Hybrid Neural Network on IMDB dataset.

	ACBRNN
MSE	0.0942
Val_Acc(%)	87.45
Test_Acc(%)	94.91
Precision	88.27
Recall	87.61
F_1 Score	87.94

Figure 4Shows the comparison of the F1 score of proposed ACBRNN with other baseline models on the IMDB dataset.

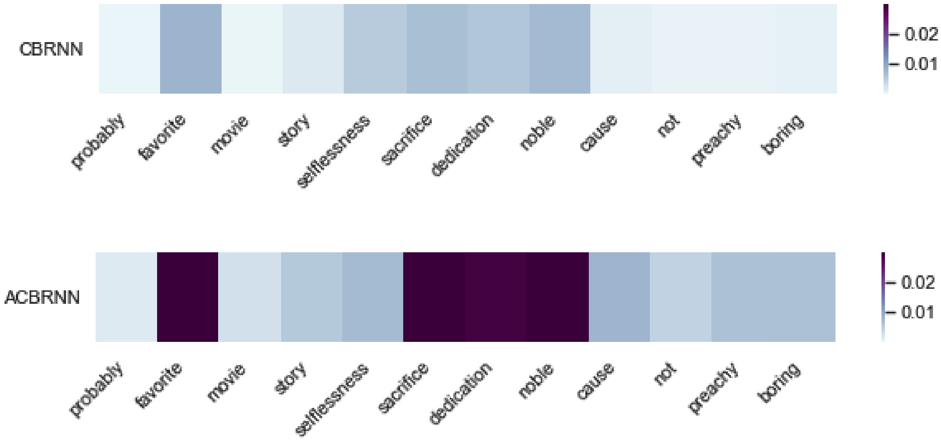


review to compute the strength of a term or a phrase in the review. To compute the weight of "survival", the Self-Attention considers the weight from various terms like *"drastic change"*, *"health"*, *"worst"* and *"struggles"*. The bidirectional network at the bottom scans the review in a forward and backward direction. It better understands the *"friendship"* between *"John Hawks"* and *"Rob"* by scanning the multiple sentences in the review and predicting the movie review as positive. Thus, the proposed network outperforms all conventional unidirectional, bidirectional and hybrid neural networks by obtaining an F_1 score of 87.94% and it is shown in Table 4. It has obtained an increase in F_1 score of 1.58%, 0.85% and 0.32% than LSTM (Unidirectional), BGRU (Bidirectional) and CNN+BGRU (Hybrid) neural networks respectively and it is represented in the Figure. 4.

4.4.5 Attention Visualization

A part of the positive movie review is taken from the IMDB dataset for Attention visualization. The Figure. 5. displays the attention visualization of CBRNN and the proposed CBRNN model. The influencing terms that impact the review as positive are *"favorite"*, *"sacrifice"*, *"dedication"* and *"noble"*. In the visualization of the CBRNN model the terms, *"favorite"*, *"sacrifice"* and *"noble"*

Figure 5Attention visualization of CBRNN and proposed CBRNN model.



are having light-dark colors. The terms *"favorite"*, *"sacrifice"*, *"dedication"* and *"noble"* are having an increase in weight due to the utilization of self-attention in CBRNN. This is reflected in the visualization of ACBRNN with an increase in the color of these influencing terms. There is also a slight increase in the weight of *"selflessness"*, *"cause"*, *"preachy"* and *"boring"*. It is also reflected in the visualization.

4.5 Comparative Study

We have compared the proposed method with the state-of-the-art method for IMDB and Polarity dataset based on accuracy.Table 5. shows the comparison of the proposed method for the IMDB dataset. The LSTM is a sequence model which gives an accuracy of 83.96%, which is less than the LSA feature extraction methods. The multiple branches Convolution LSTM surpass the shallow network, but the multi-task sharing bidirectional LSTM architecture provides better accuracy of 91.30%. The proposed architecture ACBRNN provides a good accuracy of 94.91% than cBLSTM LM because there is a keyword extraction CNN layer at the top, the Self-Attention layer assigns high weight to the significant terms in the middle and BGRU performs a contextual scanning of the review in the bottom.

4.6 Limitations of Research

The prediction of the proposed ACBRNN can be improved with the integration of external sources through a knowledge graph. Input sources for the external integration will be from different domains that help to improve the performance of the model. Further, we have planned to apply intra and inter attention layers on various deep learning architectures with different datasets. We also would like to explore a variety of hybrid deep learning methods along with attention layers on different datasets and domains.

5. CONCLUSION

In sentiment analysis, the significance terms that impact the polarity of the review have not been contemplated. To overcome these restrictions, an attention-based hybrid convolution bidirectional recurrent neural network is introduced. The proposed architecture is divided into three sections, the two convolution layers excerpts the phrase, while the self-attention in middle accredits high weights to the relevant term and BGRU at the bottom is responsible for inferring the context of the review. Several experiments are carried with the IMDB dataset to compare the proposed model with other baseline models. From the experiment results, we conclude that the proposed architecture ACBRNN has achieved an F_1 score of 87.94%. It has obtained an increase in F_1 score of 1.58%, 0.85% and 0.32% than LSTM (Unidirectional), BGRU (Bidirectional) and CNN+BGRU

Table 5. Comparison of Proposed method with state of art methods for IMDB dataset.

Model	IMDB (Acc%)
LSA (A. L. Maas et al., 2011)	83.96
LSTM (Yuzhen Lu et al., 2017)	85.48
Multiple Branch CNN+LSTM (Alec Yenter et al., 2017)	89.50
RNN with Shared Layer Architecture (Pengfei Liu et al., 2016)	91.30
cBLSTM LM + BLSTM Binary Classifier (A. E. D. Mousa et al., 2017)	92.83
ACBRNN (Proposed method)	94.91

(Hybrid) neural networks respectively. In the future, multilingual and multi-domain reviews will be considered for sentiment analysis.

FUNDING AGENCY

This research received no specific grant from any funding body in the public, commercial, or not-for-profit sectors.

REFERENCES

Agarap, A. F. (2017). A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 26-30. doi:10.5281/zenodo.1045887

Asif, M., Zafar, R., & Zaib, S. (2020, October). False Fire Alarm Detection Using Data Mining Techniques. *International Journal of Decision Support System Technology*, *12*(4), 21–35. doi:10.4018/IJDSST.2020100102

Bahdanau, Cho, & Bengio. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate.* CoRR, abs/1409.0473.

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving Sentiment Analysis via Sentence Type Classification Using BiLSTM-CRF and CNN. *Expert Systems with Applications*, *72*, 221–230. doi:10.1016/j.eswa.2016.10.065

Dhrisya, K.J., Remya, G.R., & Mohan, A. (2020). Fine-grained entity type classification using GRU with self-attention. *International Journal of Information Technology*, 1-10.

Gokula Kaveeya, S., Gomathi, S., Kavipriya, K., Kalai Selvi, A., & Sivakumar, S. (2017). Automated unified system for LPG using load sensor. *2017 International Conference on Power and Embedded Drive Control (ICPEDC)*, 59-462. doi:10.1109/ICPEDC.2017.8081133

Goldberg, Y. (2015). A Primer on Neural Network Models for Natural Language Processing. *Clinical Orthopaedics and Related Research*, *1510*, 00726.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276

Hu, J., Jin, F., Zhang, G., Wang, J., & Yang, Y. (2017). A user profile modeling method based on word2vec. *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 410–414. doi:10.1109/QRS-C.2017.74

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Studies in Health Technology and Informatics*, *235*, 246–250. doi:10.3233/978-1-61499-753-5-246 PMID:28423791

Khamparia, A., & Pandey, B. (2020). A novel integrated principal component analysis and support vector machines-based diagnostic system for detection of chronic kidney disease. *Int. J. Data Analysis Techniques and Strategies*, *12*(2), 99–113. doi:10.1504/IJDATS.2020.106641

Kumar, Videla, SivaKumar, Gupta, & Haritha. (2020b). Murmured Speech Recognition Using Hidden Markov Model. *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, 1-5. doi:10.1109/ICSSS49621.2020.9202163

Le, T.-T.-H., Kim, J., & Kim, H. (2019). Classification performance using gated recurrent unit recurrent neural network on energy disaggregation. *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, 105–110. doi:10.1109/ICMLC.2016.7860885

Lee, J. Y., & Dernoncourt, F. (2016). Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 515–520. doi:10.18653/v1/N16-1062

Leng, X. L., Miao, X. A., & Liu, T. (2021). Using recurrent neural network structure with Enhanced Multi-Head Self-Attention for sentiment analysis. *Multimedia Tools and Applications*, *80*(8), 12581–12600. doi:10.1007/s11042-020-10336-3

Liang & Zhang. (2016). *AC-BLSTM: Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification.* ArXiv, abs/1611.01884.

Liu, P., Qiu, X., & Huang, X. 2016. Recurrent neural network for text classification with multi-task learning. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*, 2873–2879. doi:10.5555/3060832.3061023

Lu & Salem. (2017). *Simplified gating in long short-term memory (LSTM) recurrent neural networks.* CoRR, abs/1701.03441.

Luo, X., Zhou, W., Wang, W., Zhu, Y., & Deng, J. (2018). Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. *IEEE Access: Practical Innovations, Open Solutions*, *6*, 5705–5715. doi:10.1109/ACCESS.2017.2785229

Maas, Daly, Pham, Huang, Ng, & Potts. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 142–150. doi:10.5555/2002472.2002491

Meti, S. A., & Sangam, V. G. (2019). Enhanced auto associative neural network using feed forward neural network: An approach to improve performance of fault detection and analysis. *International Journal of Data Analysis Techniques and Strategies*, *11*(4), 291–309. doi:10.1504/IJDATS.2019.103754

Mousa, A. E.-D., & Schuller, B. W. (2017). Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. doi:10.18653/v1/E17-1096

Nimmagadda, S., Sivakumar, S., Kumar, N., & Haritha, D. (2020) Predicting Airline Crash due to Birds Strike Using Machine Learning. *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, 1-4. doi:10.1109/ICSSS49621.2020.9202137

Patel, N. V., & Chhinkaniwala, H. (2019). Investigating Machine Learning Techniques for User Sentiment Analysis. *International Journal of Decision Support System Technology*, *11*(3), 1–12. doi:10.4018/IJDSST.2019070101

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. *ICLR 18. Clinical Orthopaedics and Related Research*, *1705*, 04304.

Prusa, J. D., & Khoshgoftaar, T. M. (2017). Improving deep neural network design with new text data representations. *Journal of Big Data*, *4*(1), 1. doi:10.1186/s40537-017-0065-8

Rajesh Kumar, T., Velu, C. M., Karthikeyan, C., Sivakumar, S., Nimmagadda, S., & Haritha, D. (2020a). Taylor Dirichlet Process Mixture for Speech PDF Estimation and Speech Recognition. *Advances in Mathematics: Scientific Journal*, *9*(10), 8675–8683. doi:10.1109/ICSSS49621.2020.9202163

Rose, S. L., Venkatesan, R., Pasupathy, G., & Swaradh, P. (2018). A lexicon-based term weighting scheme for emotion identification of tweets. *Int. J. Data Analysis Techniques and Strategies*, *10*(4), 369–380. doi:10.1504/IJDATS.2018.095216

Saranu, P. N., Abirami, G., Sivakumar, S., Ramesh, K. M., Arul, U., & Seetha, J. (2018). Theft Detection System using PIR Sensor. *2018 4th International Conference on Electrical Energy Systems (ICEES)*, 656-660. doi:10.1109/ICEES.2018.8443215

Senthil Kumar, A. M., Krishnamoorthy, P., Soubraylu, S., Venugopal, J. K., & Marimuthu, K. (2020). Efficient Task Scheduling Using GWO-PSO Algorithm in a Cloud Computing Environment. *International Conference on Intelligent Computing, Information and Control Systems*, *1272*, 8675–8683. doi:10.1007/978-981-15-8443-5_64

Si, N., Wang, H., & Shan, Y. (2018). Exploring global sentence representation for graph-based dependency parsing using BLSTM-SCNN. *Pattern Recognition Letters*, *105*, 96–104. doi:10.1016/j.patrec.2017.11.015

Sindoori, K. B. A., Karthikeyan, L., Sivakumar, S., Abirami, G., & Durai, R. B. (2017). Multiservice product comparison system with improved reliability in big data broadcasting. *Third International Conference on Science Technology Engineering Management (ICONSTEM)*, 48-53. doi:10.1109/ICONSTEM.2017.8261256

Sivakumar, S., Rajalakshmi, R., Prakash, K. B., Kanna, B. R., & Karthikeyan, C. (2020). Virtual Vision Architecture for VIP in Ubiquitous Computing. In Technological Trends in Improved Mobility of the Visually Impaired. Springer. doi:10.1007/978-3-030-16450-8_7

Sivakumar, S., Videla, L. S., Rajesh Kumar, T., Nagaraj, J., Itnal, S., & Haritha, D. (2020). Review on Word2Vec Word Embedding Neural Net. *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 282-290. doi:10.1109/ICOSEC49089.2020.9215319

Soniya, V., Sri, R. S., Titty, K. S., Ramakrishnan, R., & Sivakumar, S. (2017). Attendance automation using face recognition biometric authentication. *2017 International Conference on Power and Embedded Drive Control (ICPEDC)*, 122-127. doi:10.1109/ICPEDC.2017.8081072

Sreedevi, E., & Prasanth, Y. (2019). A Novel Ensemble Feature Selection and Software Defect Detection Model on Promise Defect Datasets. *International Journal of Recent Technology and Engineering*, *8*(1), 3131–3136.

Sreedevi, E., PremaLatha, V., Prasanth, Y., & Sivakumar, S. (2021). A Novel Ensemble Learning for Defect Detection Method With Uncertain Data. *Applications of Artificial Intelligence for Smart Technology*, 1-13. 10.4018/978-1-7998-3335-2.ch005

Sun, C., Jin, X., Lei, L., Zhao, Y., & Wang, X. (2015). Convolutional Neural Networks for Correcting English Article Errors. *Natural Language Processing and Chinese Computing*, *9362*, 102–110. doi:10.1007/978-3-319-25207-0_9

Tan, Z., Wang, M., Xie, J., Chen, Y., & Shi, X. (2017). Deep semantic role labeling with self-attention. *Clinical Orthopaedics and Related Research*, *1712*, 01586.

Videla & Ashok Kumar. (2020). Fatigue Monitoring for Drivers in Advanced Driver-Assistance System. In *Examining Fractal Image Processing and Analysis*. IGI Global. 10.4018/978-1-7998-0066-8.ch008

Xiao, L., Hu, X., & Chen, Y. (2020). Multi-head self-attention based gated graph convolutional networks for aspect-based sentiment classification. Multimed Tools Appl. doi:10.1007/s11042-020-10107-0 doi:10.1007/s11042-020-10107-0

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1480–1489. doi:10.18653/v1/N16-1174 doi:10.18653/v1/N16-1174

Yenter, A., & Verma, A. (2017). Deep CNN-LSTM with combined kernels from multiple branches for imdb review sentiment analysis. 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 540–546.

Zhang, L., Li, J., & Wang, C. (2017). Automatic synonym extraction using word2vec and spectral clustering. *2017 36th Chinese Control Conference (CCC)*, 5629–5632.

Zhao, Y., Yang, R., Chevalier, G., Shah, R., & Romijnders, R. (2017). Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction. *Clinical Orthopaedics and Related Research*, *1708*, 05824.