Neural Processing Letters (2024) 56:47
https://doi.org/10.1007/s11063-024-11513-3

A Sparse Self-Attention Enhanced Model for Aspect-Level Sentiment Classification

P. R. Joe Dhanith[1] · B. Surendiran[2] · G. Rohith[3] · Sujithra R. Kanmani[1] · K. Valli Devi[1]

Accepted: 24 November 2023 / Published online: 16 February 2024

Abstract

Aspect based sentiment analysis (ABSA) manifests the well refined work as Aspect-level sentiment classification (ASC) due to recent high attention and profound outcomes. This paper reveals the exorbitant relation of concerned aspect in determining the sentiment polarity of a sentence in addition to their content. Considering an instance, "The food is tasty but restaurant is untidy", aspect food reveals the polarity as positive, while in case of restaurant, the polarity seems negative. Hence to scout relation between an *aspect* and *context* words present in sentence is much vital. In spite of the exceptional progress, the ASC experiences certain pitfalls as (1) The current attention based methods makes the given aspect to falsely relate syntactically unrelated words as related ones. (2) Single context-independent representation is only achieved by traditional Word2Vec or GloVe based embedding vectors. (3) Sentiments for multiple words that are inconsecutive remain insufficient for CNN based models. A Sparse-Self-Attention-based Gated Recurrent Unit with Aspect Embedding (SSA-GRU-AE) implementing BERT for ASC is proposed to solve these issues. The proposed SSA-GRU-AE mechanism is centralized on various portions of sentence as multiple aspects are taken as input. The experimental analysis on ASC datasets has proved that proposed model enhanced performance on ASC.

✉ P. R. Joe Dhanith
joe.dhanith@gmail.com

B. Surendiran
surendiran@gmail.com

G. Rohith
rohithg1106@gmail.com

Sujithra R. Kanmani
sujithrakanmani@gmail.com

K. Valli Devi
vallidevi.k@vit.ac.in

[1] School of Computer Science and Engineering, Vellore Institute of Technology Chennai Campus, Chennai 600127, India

[2] Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal 609609, India

[3] School of Electronics and Communication Engineering, Vellore Institute of Technology Chennai Campus, Chennai 600127, India

Keywords ASC · ABSA · Self-attention · Sparse mechanism · BERT

1 Introduction

Sentiment Analysis (SA) [1] which is a sub-branch of Natural Language Processing (NLP) expresses emotion or view on the given text or speech or any other communication. ABSA [2] extends to be well refined in nature to sentiment analysis providing crucial information for NLP tasks. Aspect-category (ACSA) and Aspect-term (ATSA) forms two sub-tasks of ABSA [3, 4]. ACSA predicts the $sentimentpolarity$ for a given $aspect$ in predefined sets hidden in sentence.“The food is tasty but restaurant is untidy”, the ACSA could predict the $sentimentpolarity$ for aspect "eatable" yet not available in sentence. On the other hand, ATSA predicts the $sentimentpolarity$ of aspect term as a sequel to sentence, where the “The food” form the example sequel to the above sentence [5].

Sentiment polarity differs for various independent aspects. For an ABSA task, the pivotal point is the given aspect term. Furthermore, many words in the sentence are futile on sentiment prediction for a given term. Taking into consideration the given aspect as "restaurant" for weakly associative words, certain words like "food" and "tasty" is impertinent for the sentiment prediction and produces inconsistent results.

Albeit of the excellent performance of Neural Networks (NN) [6–8] in various research domains such as $languagetranslation$, $paraphraserecognition$, $questionanswering$ and $textsummarization$, they are still in nonage with ASC. Certain works in target dependent classification gets benefitted with only target information but not aspect information which forms crucial part to evaluate the ASC.

The original ASC task proposed with pre-trained word embedding and task-centric neural framework faces a bottleneck even with improved accuracy or F1-score. This bottleneck is caused due to task-agnostic embedding layer initialized with Word2Vec [9–11] and GloVe [12] which is insufficient to capture intricate semantic relevance in the sentence, providing only context- independent word-level features. The limitation in the dataset size to train task-centric frameworks is solved by a deep LSTM [13, 14] with pre-trained word embedding layer.

Overriding the promising nature of attention based models [15–17] the insufficiency to seize dependencies between $context$ and $aspect$ present in a sequence occurs, which then leads to the given aspect which mistakenly attend syntactically irrelevant context words as $descriptors$. The example, “Its model is ideal and function is excellent." explains that excellent is mistakenly taken as descriptor of aspect model. Certain models does not fully exploit syntactical structure but rather imposed syntactical constraints on attention-weights.

Issues of attention-based mechanisms is solved by Convolution Neural Networks (CNN) [18–20] by finding features as continuous words and using convolution operations over word sequences to predict sentiment of an aspect but failed to capture sentiment polarity for multiple words that are inconsecutive to one another. The sentence, "The workers should bit more work sincerely", makes incorrect prediction for CNN based model by considering "workers" as the aspect and "more work sincerely" as descriptive phrase which reverses the effect on the sentiment.

$BERT$ [21–23] overcomes the bottlenecks with the information from entire sentence as input to calculate token level representation while $Word2Vec$ or $GloVe$ based embedding provides single $context-independent$ representation alone. The modeling power of $BERT$ is investigated on ASC in this paper and is considered as well-known pre-trained model equipped with Transformer. The SSA-GRU-AE is proposed to impose the model by attending an important part of sentence with respect to a particular aspect.

Principal benevolent features are:

(1) The SSA-GRU-AE algorithm is proposed to achieve ASC by examining various portions of the sentence when several aspects are considered.
(2) The input and aspect embedding are generated by the BERT pre-trained model.
(3) Aspect which plays a major role in this work is employed in two ways, where in the first part, $aspectembedding$ vector is concatenated with the input embedding vector and the second part with the hidden vector of the sentence is generated by GRU to compute attention weights which efficiently extracts contextual semantic relationship between aspect and sentence.
(4) Sparse self-attention mechanism, proposed in our method can effectively filter out the unimportant words in sequence that does not contribute to sentiment analysis and also learns sentiment aware word embedding by applying weights on word embedding of input and aspect terms.
(5) Implementation of L_1-regularize on the attentions makes sure that minimal words influence semantic and sentiment of sequence.
(6) Results on ASC datasets reveals that proposed model outperformed the $state-of-the-art$ models discussed in the literature.

The remainder of this article is structured as, Sect. 2 discusses the related work, Sect. 3 discloses the proposed model, Sect. 4 analyses the result and finally concluded in Sect. 5 (Table 1).

2 Related Work

The various models that are put forth in the literature survey is studied for ASC task. ASC is a classification technique which is a well refined task in ABSA. Several of the modern approaches could identify polarity of entire sequence even in the unavailability of aspects. Conventional approaches are utilized to design a $bag-of-words$ and $sentimentlexicons$ as features which trains SVM classifier [24] to perform ABSA. In spite of intensive labor in feature engineering, the results are greatly dependable on the caliber of features.

The invention of learning distributed representations has made the Neural Network (NN) approaches become popular for ASC. Classical models such as RecNN [25, 26], CNN [27, 28], RNN, LSTM [29–31], GRU [32], GCN [33] and Tree-LSTMs [34] were applied for ASC. Syntax structures of sequences used by Tree-based LSTMs though proved as an effective approach to solve ASC problems also has failed due to syntax parsing errors commonly found in resource lacking languages. LSTMs and GRUs have achieved great success in ASC. Liang et al. introduced a $deeptransition$ model called as AGDT [35] using GRU encoder to effectively utilize the aspect from the scratch to improve the feature selection and extraction. The utilization of target information by TD-LSTM and TC-LSTM [36] has made remarkable achievement in ASC task where the target vector attained by TC-LSTM is acquired, by computing the average of word vectors related to aspect, however is insufficient to capture its semantics, resulting in lousy performance.

Attention or gating mechanism is implemented on several existing models that could capture context related features. Ma et al. [37] elucidated a hierarchical $attention$ model which primarily visits aspect words then goes through the whole sentence and later integrates this information with external practical knowledge using sentic-LSTM which resolves the word

Table 1 List of abbreviations

Abbreviation	Name
RecNN	$RecursiveNeuralNetwork$
RNN	$RecurretNeuralNetwork$
LSTM	$LongShortTermMemory$
GRU	$GatedRecurrentUnit$
GCN	Graph Convolution network
SVM	Support vector machine
AGDT	Aspect guided deep transition
HAPN	Hierarchical attention based position aware network
DSMN	Deep selective memory network
CNN	Convolution neural networks
ASGCN	Aspect specific GCN
ABSA	Aspect based sentiment analysis
BERT	Bidirectional encoder representations from transformers
R-GAT	Relational gated attention
IAN	Interactive attention network
MemNet	Memory network
AOA	Attention over attention
TD-LSTM	$TargetDependentLSTM$
TC-LSTM	$TargetConnectionLSTM$
SAGCN	Selective attention graph convolution network
KGCapsAN	Knowledge guided capsule attention network
IGCN	Interactive gated convolution network
CMA-MemNet	Convolution multi-hop attention memory network

conflict problems caused by the basic LSTM models. Position information of aspect plays a key role for ASC task which is realized by [38] by proposing a HAPN capable of learning the position information of the aspect followed by fusing the aspects and contexts to bring about the final sentence representation. Zheng et al. [39] proposed a rotatory $attention$ mechanism based neural network which associates the aspect and the left/right contexts implementing three LSTMs one for the $leftcontext$, second for the $targetphrase$ and third for the $rightcontext$. Laddha et al. [40] introduced an $attention$ based $Bi-LSTM$ model that productively seizes relation between the multiple aspects and the context words by ignoring the effect of one $aspect$ on another. Finally a CRF is used to model dependencies among output labels. Lin et al. proposed a DSMN [41] to guide multi-hop attention mechanism by computing distance between an aspect and its context to capture aspect aware context information.

Numerous existing models combines CNN with LSTM to seize context related word-level information. Liu et al. [19] proposed a model by combining both regional CNN and $Bi-LSTM$ which obtains the context information as well as the relationship between $aspect$ and $context$ and also gate mechanism which improves word vector representation to make model language independent. Zhang et al. [27] proposed a CMA-MemNet to obtain semantic information from aspects and sentences. Convolution proposed in this model captures context related information and multi-head self-attention obtains semantic information. CNN based models inadequately determines sentiments for multiple words that are not consecutive.

Graph Convolution Network (GCN) models are developed to effectively capture dependency relation between $aspect$ and $context$ due to lack in performance of existing attention based models. Zheng et al. [42] proposed ASGCN which has an LSTM and a multi-layer GCN, to fully leverage the syntactical dependency structure within a sentence. The LSTM generates contextual word embedding for word orders and the multi-layer GCN filters out unimportant words leaving back the important information which is then fed into the attention based LSTM to generate aspect based features to predict sentiments. Hou et al. [43] proposed SAGCN with self-attention, effectively enables the interaction between $aspect$ and its $opinionwords$ even if aspect term is far away from it, which later considers the connection between target and its syntactic neighbors. Sun et al. [33] proposed a $convolutionoverdependencytree$ which utilizes the Bi-LSTM and captures the important features from the sentence that could be fed as input to GCN which then transfers information from the opinion words to aspect words. Liang et al. [44] proposed an interactive $multitask$ learning model by incorporating a new message passing mechanism which utilizes dependency relation embedded GCN to completely exploit the syntactic knowledge for end-to-end ABSA. Wu et al. [45] introduced a GCN with $attention$ utilizing BERT to capture the relation between $aspect$ and its $context$, where attention controls information flow in the GCN.

The influence of the overall contextual score is worsened due to failure of dependency tree based models to capture hidden vector representation based on aspect. Veyseh et al. [46] elucidated a $graph-basedmodel$ with gate vectors that could customize hidden vectors towards aspect terms and also a $dependencytree$ based mechanism to acquire importance scores for every word in sequence.

In spite of existing GCN models captures the entire tree which makes it complex during optimization, it is required that only a small part of the $dependencytree$ is necessary in ASC task. Wang et al. [47] reshaped and pruned only the important part of the dependency tree to specifically focus on target aspects. Then this pruned tree is fed into the $R-GAT$ to encode the $dependencyrelations$ and also establishes connections between the $aspects$ and $contexts$.

There are few issues observed in the existing models,

- Overriding the promising nature of attention based models the insufficiency to capture dependencies between $context$ and $aspect$ present in a sequence occurs, which then leads to the given aspect which mistakenly attend syntactically irrelevant context words as descriptors.
- Certain models does not fully exploit syntactical structure but rather imposed syntactical constraints on attention-weights.
- Large-scale corpus training improves neural network models. Manually labeling aspect targets to generate aspect-level training data is difficult.
- As comments and other corpora with $document-level$ sentiment labels are hard to obtain, gathering users' preferences about multiple aspect categories becomes infeasible.

To differentiate various sentiment polarities at a well refined aspect level, is highly demanding in spite of the efficiency in all the methods. Hence a design of a dominant neural network which could completely engage aspect information for ASC is vital. To address these issues this work proposes a novel SSA-GRU-AE to effectively classify the sentiments with respect to the aspects. Especially the Sparse self-attention mechanism introduced in our work devalues the unimportant words and computes the important words related to the aspect words and also helps to outperform the existing models available in the literature.

3 Proposed Work

This paper focuses on ASC for the given input sequence which employs BERT pre-trained model to compute contextualized word embedding vectors for sentence as well as the aspect terms which then forms the input to the SSA-GRU-AE to classify the aspect-level sentiments.

3.1 Input Layer

The given input sentence $S = \{s_1, s_2, s_3, \ldots, s_a, s_{a+1}, s_{a+2}, \ldots s_{a+(m-1)}, \ldots, s_N\}$ of length N with m aspect words $\{s_a, s_{a+1}, s_{a+2}, \ldots s_{a+(m-1)}\}$ has to be recast into contextualized word embedding vectors by using $pre - trained$ BERT model. The input format of the given sentence S to the BERT is given by the split up of "[CLS] + sentence + [SEP] + aspectterm + [SEP]". Thus this input format extracts the overt interactions between sentence and aspect term. The embedding information of sub-words generated by BERT is subjected to average pooling which produces ultimate embedding vector $X \in R^{N \times d}$, where d is dimension of BERT output. The vector X is then represented as vector sequence, $\{w_1, w_2, \ldots w_n\}$ for sentence and $\{v_a, v_{a+1}, v_{a+2}, \ldots v_{a+(m-1)}\}$ for aspect terms respectively.

3.2 Gated Recurrent Unit

The $vanillaRNN$ suffers from vanishing or exploding $gradient$ problem for long sequences due to replacement of entire sequence in hidden state during every time step t. The Gated Recurrent Unit (GRU) solves this issue by the inclusion of two additive gates in their architecture. The GRU contains two gates where one is $updategate$ and other is $resetgate$. The gates present in GRU removes the unimportant and retains only the important information. The reset gate combines current input with the important part of the previous $hiddenstate$ to produce a new $hiddenstate$. The update gate determines the amount of information from current hidden state to be included with final hidden state. This allows network to retain long-term dependencies. The workflow diagram of GRU has been illustrated in Fig. 1.Calculation process of GRU is given from Eq. (1) to (5):

$$g_r = \sigma(W_{ir} \cdot x_t + W_{hr} \cdot h_{t-1}) \tag{1}$$

$$r = \tanh\left(g_r \odot (W_h \cdot h_{t-1}) + W_x . x_t\right) \tag{2}$$

$$g_u = \sigma(W_{iu} \cdot x_t + W_{hu} . h_{t-1}) \tag{3}$$

$$u = g_u \odot h_{t-1} \tag{4}$$

$$h_t = r \odot \left(1 - g_u\right) + u \tag{5}$$

The class labels represented in the proposed work {positive, negative, neutral}.

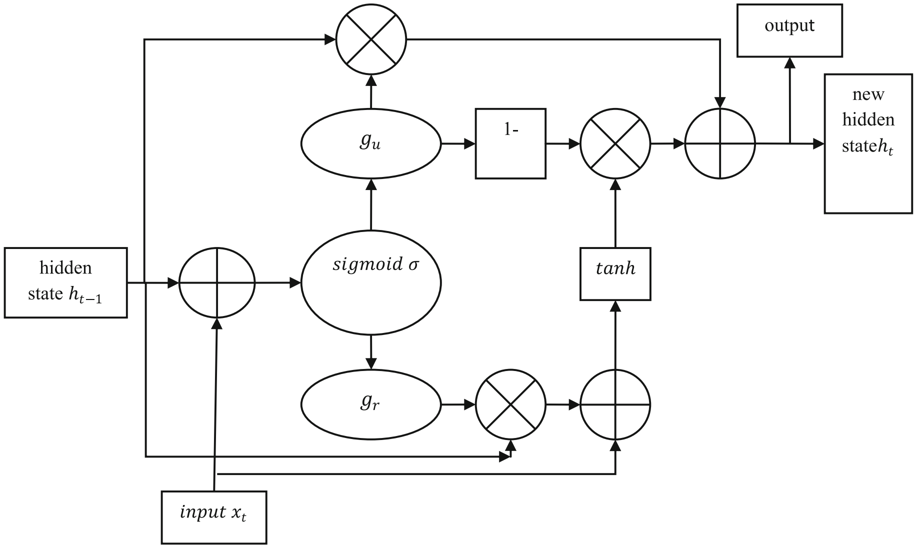

Fig.1 Gated recurrent unit

3.3 Self-Attention Based GRU with Aspect Embedding

The general GRU struggles to identify the key part of a sentence for the ASC task. To overcome this problem, this work proposes a novel SSA-GRU-AE. This work integrates sparse self-attention mechanism with GRU to capture a relevant part of sentence with respect to a given aspect. Aspect information play key role to classify polarity of given sentence. To utilize aspect information effectively this work proposes to generate embedding vector using BERT for each aspect. Here $V_{a_i} \in \mathbb{R}^{d_a}$ is embedding of aspect i, where d_a is dimension of aspect embedding. $H \in \mathbb{R}^{d \times N}$ is matrix formed by hidden vectors $[h_1, h_2, h_3, \ldots, h_N]$ that GRU generated. Here d is size of the hidden layers and N is the length of given sentence. $e_N \in \mathbb{R}^N$ is vector of ones. An attention weight vector α and a hidden weight vector r is produced by self-attention mechanism, as specified in Eqs. (6) to (8).

$$M = \tanh\left(\begin{bmatrix} w_h H \\ W_v V_a \otimes e_N \end{bmatrix}\right) \tag{6}$$

$$\alpha = \mathrm{softmax}\left(W^T M\right) \tag{7}$$

$$r = H\alpha^T \tag{8}$$

where $M \in \mathbb{R}^{(d+d_a)\times N}$, $\alpha \in \mathbb{R}^N$, $r \in \mathbb{R}^d$. $W_h \in \mathbb{R}^{d\times d}$, $W_V \in \mathbb{R}^{d_a \times d_a}$ and $W \in \mathbb{R}^{d+d_a}$ are weight parameters. ⊗ is a concatenation operator, which repeatedly concatenates V for N times. $W_V V_a \otimes e_N$ repeats the linearly tranformed V_a as many times till the last word in the given sentence and eventually represented in the Eq. (9),

$$h^* = \tanh\left(W_p r + W_x h_N\right) \tag{9}$$

where $h^* \in \mathbb{R}^d$, W_p and W_x are weight parameters.

The self-attention mechanism captures significant part of sentence with respect to an aspect. Here h* represents feature of a particular sentence with respect to an aspect. Then a linear layer is added to transform sentence vector e, which is a vector of length which equals class number |C|. Finally a softmax layer is applied to transform e to conditional probability distribution as specified in Eq. (10).

$$y = \mathrm{softmax}\left(W_s h^* + b_s\right) \tag{10}$$

where W_s and b_s are the weight and bias parameter of softmax layer. Further L_1 regularize is applied to make sure only minimal words of the sentence contributes to the semantic and sentiments of the sentence. The sparse self-attention mechanism is depicted in Eq. (11).

$$|y|_{L_1} = \left|\mathrm{softmax}\left(W_s h^* + b_s\right)\right| \tag{11}$$

The proposed sparse self-attention mechanism effectively removes the words which are unimportant to predict the sentiment and calculates the key part of the sentence. After computation of the important part of the sentence, a weighted summation is performed to predict the sentiment polarity which is specified in Eq. (12).

$$d^k = y_1 x_1 + y_2 x_2 + \ldots + y_n x_n \tag{12}$$

where x_i denotes embedding of i^{th} word in sentence, n is length of sentence. Finally output of sparse self-attention layer is obtained using Eq. (13).

$$\hat{y} = \mathrm{softmax}\left(W d^k + b\right) \tag{13}$$

where $\hat{y}$ is a predicted sentiment polarity which is a $2 - D$ vector where (1, 0) and (0, 1) are *positive* and *negative* labels respectively.

The model is trained using back propagation where *crossentropy* loss function is used to minimize error between y and $\hat{y}$ for all sentences which is specified in Eq. (14).

$$\mathrm{loss} = -\sum_i \sum_j y_i^j \log \widehat{y}_i^j \tag{14}$$

where i and j is index of sentence and class respectively.

Then an Adagrad optimization [48] is adopted to train the model over mini batches. This optimizer improves the strength of SGD on learning process and also performs increase in updates to learning rate for unusual parameters and decrease in updates to learning rate for usual parameters. The workflow diagram of proposed SSA-GRU-AE is illustrated in Fig. 2.

4 Experiments

The proposed SSA-GRU-AE model performs the ASC task in which both the *word* and *aspect* embedding vectors are formed by pre-trained BERT and length of attention weights equals length of input sentence. The hidden size (dim_h) of BERT is 768 and transformer layers (L) are 12. The pre-trained BERT model initializes word and aspect embedding vectors while initialization of all other weight parameters are done by random sampling with normal distribution $\mathcal{N}(0; 0 : 0.2)$. The attention weights are the same length as the sentences. To implement SSA-GRU-AE, there are a total of 8 attention heads and 12 training layers are utilized. Implementation of proposed and all baseline models are performed by Tensorflow and hyper parameters applied are specified in Table 2.

Averaging of 3 runs applying random initialization considering evaluation metrics as *Accuracy* as well as $F1-score$ gives the required results is shown in Table 3. A Friedman − *test* was done on both the *Accuracy* and $F1 - score$ to check the competency of our model with baseline models as shown in Table 10.

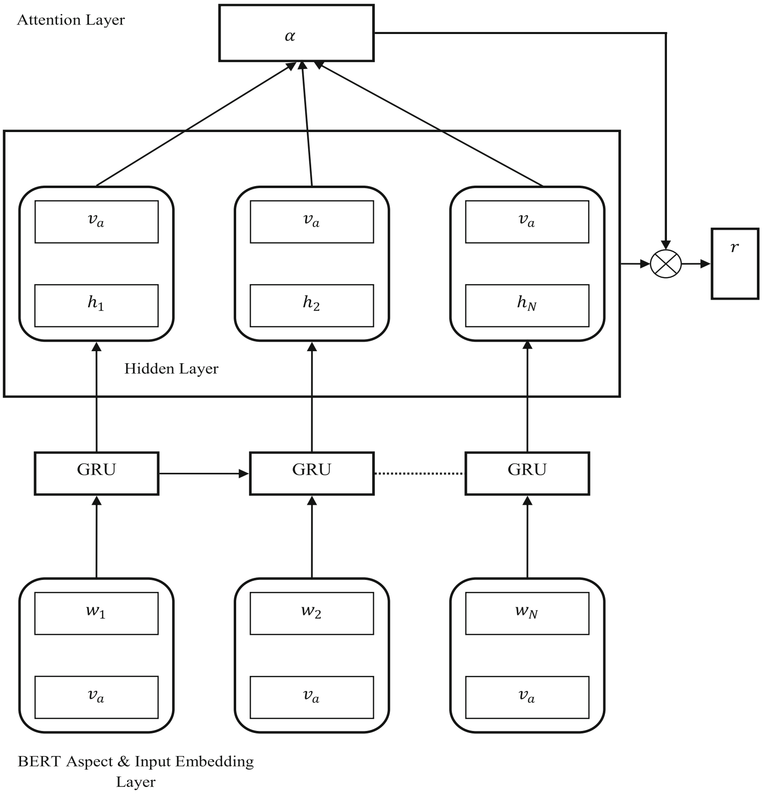

Fig.2 Proposed SSA-GRU-AE model

Table 2 Hyper parameters for baseline models

Batch size	25
Momentum	0.9
L2-regularization weight	0.001
$LearningRate$	0.01
$DropoutRate$	0.5

Table 3 Evaluation metrics

Metrics	Formula
Accuracy (a)	$a = \frac{T_P+T_N}{T_P+T_N+F_P+F_N}$
Precision (p)	$p = \frac{T_P}{T_P+F_P}$
Recall (r)	$r = \frac{T_P}{T_P+F_N}$
F1-Score (f)	$f = \frac{2\times p\times r}{p+r}$

Table 4 Statistical representation of dataset

Name	Train/Test	Positive	Neutral	Negative
Twitter [49]	Train	1559	3119	1559
	Test	169	339	169
LAP14 [50]	Train	989	459	869
	Test	339	163	129
REST14 [50]	Train	2159	629	811
	Test	729	197	197
REST15 [51]	Train	911	37	257
	Test	327	37	179
REST16 [52]	Train	1239	67	438
	Test	468	29	116

4.1 Datasets

5 datasets (twitter, LAP14, REST14, REST15 and REST16) are employed to prove that the proposed method mastered the existing baseline models.The 5 datasets contains reviews of users with a set of aspects with its polarities from which the sentences with contradictory polarities or inexplicit aspects are removed. The purpose of the proposed work is to ascertain aspect polarity of a sentence with respect to aspect. The dataset details are exhibited in Table 4.

4.2 Baselines and Experimental Setting

Ten sentiment classification models are implemented as baseline models for comparison with proposed model is studied as:

1. **SVM**: SVM [24] identifies many qualities of aspects are pointed out, and a quantitative look at each part is effective.
2. **LSTM**: LSTM [36] models the sentiment representation and prediction.
3. **Interactive attention network** (**IAN**): IAN model [53] utilizes two LSTMs one for the target and the other for context that are used to extract meaningful information independently with an interactive attention mechanism which later concatenates them to classify the sentiments.
4. .**Memory Network** (**MemNet**): MemNet [54] is a combination of attention mechanism and explicit memory. The multi-hop attention mechanism proposed in this model helps to improve sentiment classification.
5. **AOA** [55]: This model learns the aspects and the sentence representation jointly and also explicitly captures the interaction between them.
6. **ASGCN** [42]: This model utilizes LSTM and generates contextual information followed by a GCN to obtain aspect specific features which are then fed to a masking mechanism to remove non-aspect words and later fed back to another LSTM to predict the sentiment.
7. **SAGCN** [43]**:** SAGCN utilizes GCN to find correlation between aspect and sentence using dependency tree.

8. **IGCN** [28]**:** A bidirectional gating proposed in IGCN computes the relation between the *aspect* and its *context*.
9. **DSMN** [41]**:** The multi-hop attention is guided by dynamically selected context memory which then integrates the *aspect* information with the memory networks.
10. **CMA-MemNet** [27]: The rich semantic information between the *aspect* and the *sentence* is extracted by this memory networks.

5 Experimental Results

In the experiments, *accuracy* and $F1 - score$ are the metrics which evaluates performance of proposed method. To evaluate stability of model, the method is run thrice and the *meanaccuracy* and *standarddeviation* are reported in Table 5, 6, 7, 8, 9. The Friedman test verifies the magnitude in differences between the proposed and the other approaches with $p - value$ of 0.05.

Table 5 *Accuracy* and $F1 - score$ on Twitter dataset

Model	Accuracy	F1-score
SVM	63.40	63.30
LSTM	69.56	67.70
IAN	71.48	69.90
MemNet	72.30	70.20
AOA	72.50	70.81
ASGCN	72.15	70.40
SAGCN	72.3	70.2
IGCN	73.1	71.1
DSMN	73.6	71.4
CMA-MemNet	73.9	71.9
Proposed (A-GRU-AE)	74.9	73.0

Table 6 *Accuracy* and $F1 - score$ on Lap14 dataset.

Models	Accuracy	F1-score
SVM	70.49	67.32
LSTM	69.28	63.09
IAN	72.05	67.38
MemNet	70.64	65.17
AOA	72.62	67.52
ASGCN	75.55	71.05
SAGCN	74.9	67.3
IGCN	76.8	68.9
DSMN	77.1	70.1
CMA-MemNet	77.6	71.3
Proposed (A-GRU-AE)	78.3	73.4

Table 7 *Accuracy* and $F1 - score$ on Rest14 dataset.

Models	Accuracy	F1-Score
SVM	80.16	73.4
LSTM	78.13	67.47
IAN	79.26	70.09
MemNet	79.61	69.64
AOA	79.97	70.42
ASGCN	80.86	72.19
SAGCN	77.14	67.16
IGCN	74.13	65.18
DSMN	78.17	70.13
CMA-MemNet	79.14	71.23
Proposed (A-GRU-AE)	80.90	73.6

Table 8 *Accuracy* and $F1 - score$ on Rest15 dataset.

Models	Accuracy	F1-Score
SVM	76.3	56.17
LSTM	77.37	55.17
IAN	78.54	52.17
MemNet	77.31	58.28
AOA	78.17	57.02
ASGCN	79.89	61.89
SAGCN	76.19	59.73
IGCN	80.3	62.3
DSMN	79.16	61.7
CMA-MemNet	80.9	62.9
Proposed (A-GRU-AE)	81.3	63.1

Table 9 *Accuracy* and $F1 - score$ on Rest16 dataset.

Models	Accuracy	F1-Score
SVM	83.17	61.39
LSTM	86.80	63.88
IAN	84.74	55.21
MemNet	85.44	65.99
AOA	87.50	66.21
ASGCN	88.99	67.48
SAGCN	88.3	67.4
IGCN	87.42	68.76
DSMN	86.7	67.4
CMA-MemNet	87.13	68.76
Proposed (A-GRU-AE)	89.76	71.43

5.1 Discussion

The result analysis of proposed work with eighteen baseline models clearly proves that proposed work outperformed the existing models in terms of $accuracy$ and $F1 - score$. The SSA-GRU-AE model consistently performed better than all the baseline models in all the five datasets ($twitter$, $LAP14$, $REST14$, $REST15andREST16$) whereas SVM and LSTM performed poorly in all five datasets consistently due to the manual feature engineering of SVM and the lack of aspect information in LSTM for ASC.

The IAN and AOA model is implemented using attention mechanism to solve ASC task by attending all the aspect and context words that helps to outperform both the SVM and baseline LSTM models. The experimental results on all the five datasets proved that both the IAN and AOA consistently performed well. The issue with the attention mechanism is that if the dataset is noisy or it contains multiple aspects this mechanism falsely assigns high scores to irrelevant words and also these attention based models attend all the aspect words with different weights which incorrectly guides $aspectterm$ to focus on syntactically $unrelatedwords$. These issues causes AOA and IAN models performs moderately on all five datasets, in terms of $accuracy$ and $F1 - score$.

GCN based models effectively captures both syntactic word dependencies and long range word relations. These models worked well with datasets which are rich in syntax information with good grammatical structure. Especially all the GCN based models effectively worked on $LAP14$, $REST15andREST16$ datasets and also improved both the $accuracy$ and $F1 - score$. These models failed to work well with the datasets which has less grammatical information and less sensitive syntax information. That is on both the Twitter and REST14 datasets these GCN models produced less $accuracy$ and $F1 - score$. The results shown in Table 5, 6, 7, 8, 9 and 10 clearly proved that proposed work outperformed existing GCN based models in terms of $accuracy$ and $F1 - score$.

Memory network based models worked well on all five datasets continuously in terms of $accuracy$ and $F1 - score$. The base memory network model effectively captured aspect-sequence modeling well, but it failed to capture context and sequence information and hence decreased performance of the MemNet model in all the five datasets. But CMA-MemNet and DSMN captured both the information effectively, which helps them to perform better in all five datasets. The results on all five datasets showed that CMA-MemNet and DSMN performed well on all five datasets which outperformed all basic, $attention$ and GCN models. Although these models performed well on all datasets, still it failed to recognize all aspects correctly, which led to performance loss.

The proposed novel sparse self-attention GRU with aspect embedding implementing BERT outperformed all the baseline models on all five datasets in terms of accuracy and

Table 10 Friedman's test

Source	SS	df	MS	Chi-square	Prob. > Chi-Square
Columns	117.9	5	18.79	24.63	0.00031
Error	20.3	23	0.8341		
Total	137	32			
P-value	$2.731e^{-04}$				

F1 − score. The BERT model used in our work effectively captured contextual information which helped to capture semantic information. Also sparse self-attention mechanism introduced in our work effectively removed unimportant words and captured only important words related to sentence. Also L1-regularize applied on attentions helped to ensure that only a few words contributed to sentiment of sentences. This helped proposed model to perform better on datasets which are grammatically poor and noisy (Twitter and REST14). These advantages helped our model to outperform the existing baseline models on all five data sets where especially our model consistently performed well on Twitter and REST14 datasets too. Figures 3, 4, 5, 6, 7 shows the comparative results of all the baseline models with the proposed model on five different datasets.

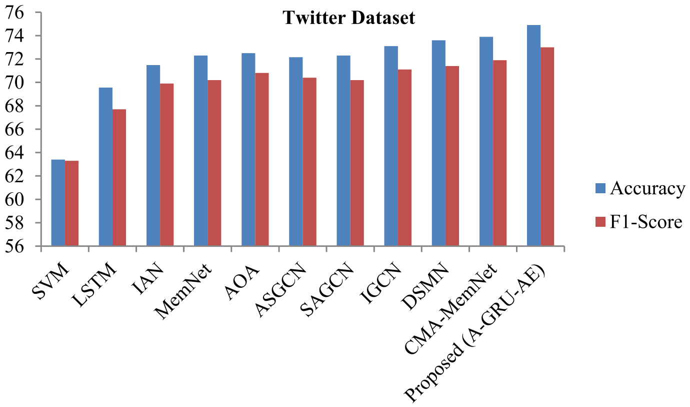

Fig. 3 Result analysis on Twitter dataset

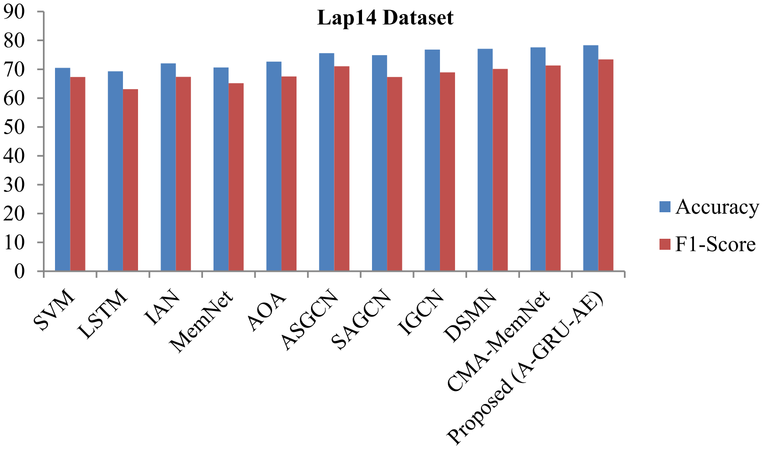

Fig. 4 Result analysis on Lap14 dataset

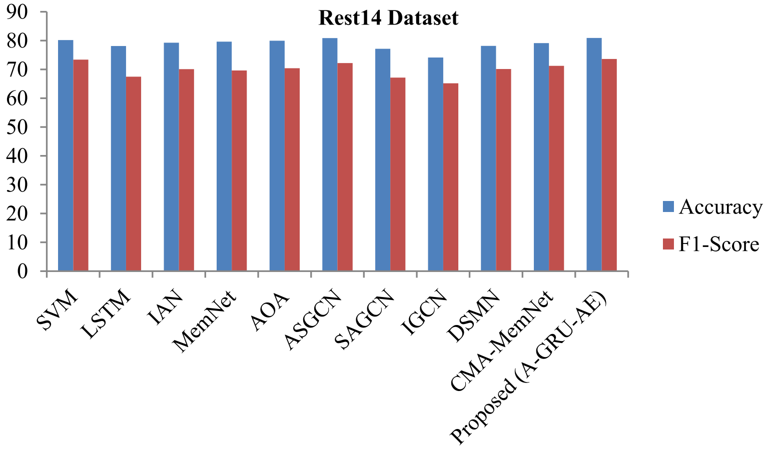

Fig. 5 Result analysis on Rest14 dataset

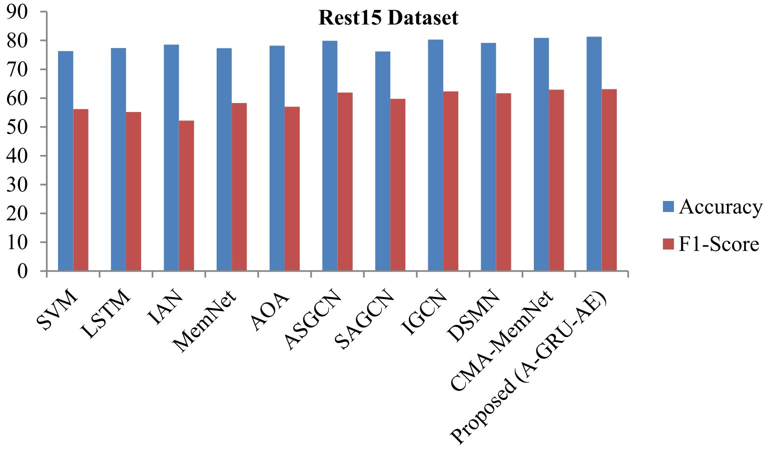

Fig. 6 Result analysis on Rest15 dataset

5.2 Ablation Study

An ablation research is carried out to examine the significance of each component in the proposed model. First pre-trained BERT model is replaced with GloVe model and then executed on five datasets. The proposed model without BERT performed poorly in terms $accuracy$ and $F1 - score$ compared to model with BERT. The model without BERT produced 2.9%, 2.7%, 1.6%, 1.4%, and 1.3% less $accuracy$ and 2.6%, 2.4%, 1.3%, 1.1% and 1% less $F1-score$ than model with BERT on $Twitter$, $REST14$, $LAP14$, $REST15andREST16$ datasets respectively. The proposed model without BERT performed poor when compared to some of GCN and memory network models too. The BERT model effectively captured semantic relation between aspect and context. This helped to improve performance of proposed model.

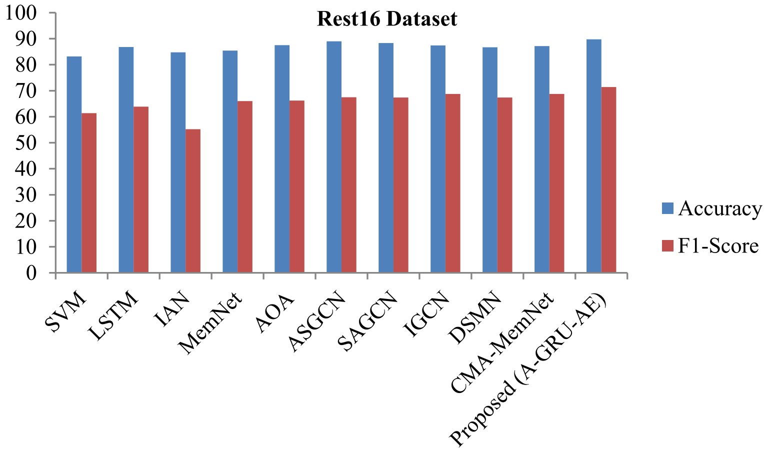

Fig. 7 Result analysis on Rest16 dataset

To identify importance of $sparseself - attentionmechanism$, a model without $sparseself - attentionmechanism$ is designed. Then model without sparse self-attention mechanism was executed on five datasets. The proposed model without sparse self-attention mechanism performed poorly in terms accuracy and F1 − score compared to model with model with sparse self-attention mechanism. The model without sparse self-attention mechanism produced 2.6%, 2.3%, 1.4%, 1.2%, and 1.1% less accuracy and 2.3%, 2.1%, 1.1%, 0.9% and 0.7% less F1 − score than model with sparse self-attention mechanism on Twitter, REST14, LAP14, REST15 and REST16 datasets respectively.The experimental results on five datasets proved that proposed sparse self-attention mechanism improved performance of proposed model. The sparse self-attention mechanism effectively captured the importance of each word in the context with respect to $aspect$. This helped to remove unimportant words from sentence and keep only important part of sentence. This helped to improve performance of proposed model over model without sparse self-attention mechanism.

5.3 Case Study

For instance, consider the following phrase: " $\boldsymbol{Evenifit'sagoodday, Idon'tfeelit. I'mreallymiserable.}$ " The terms "$miserable$," "$feel$," and "$don't$" are extremely significant in predicting the sentiment polarity of this sentence than the words "$good$" and "day". Many words, including "the," "in," "it," and "I'm," are also unimportant. Therefore, it's crucial to create a model that can accurately depict the significance of each word in the documents. Additionally, it must remain sparse enough that only a $fewwords$ can accurately categorize the sentiment labels of the sentence. The self-attention layer of the proposed SSA-GRU-AE was used to determine the significance of each word in sentence. In order to make sure that only a handful of words are needed to identify the sentiment polarities of sentences, an L1 regularization is then used for these weights.

Consider the following sentence," **The meal is tasty, but the restaurant is untidy.** ", the proposed model predicts $sentimentpolarity$ for the aspect "meal" as $positive$ and "restaurant" as $negative$. The self-attention layer in the proposed model accurately identifies the $aspect$ term "meal" and its $context$ as "tasty" and also the $aspect$ term "restaurant" and its $context$ as "untidy". The sparse nature of the $self - attention$ mechanism of the proposed model helps to retain only important words from the sentence such as "meal", "tasty", "restaurant" and "untidy" and also helps to avoid unimportant words such as "The", "is" and "but". The ability of sparsification in retaining only important words from the sentence helps to achieve ASC tasks efficiently. The self-attention mechanism serves as a sparsification mechanism in the proposed model, where it is regarded as a type of regularization that potentially enhances the quality of the model by efficiently diminishing noise within it. Due to the sparse nature of the proposed SSA-GRU-AE, neurons combine the output activations which are very similar, change their biases, and then rewire the network to reflect these changes. This sparsification enhances the efficiency of models by allowing them to function effectively in feature spaces with high dimensions. This also reduces the complexity of representation, wherein only a subset of dimensions is utilized at any given moment and decrease complexity by nullifying specific subsets of the model parameters. This resulted in reducing many useless words present in the sentence for $sentimentpolarity$ prediction with respect to an $aspect$. Because of these advantages, the proposed model identifies the $aspect$ phrases and its corresponding $context$ words as a sequel to sentences, such as "meal" and "tasty" in the above statement.

Another example is, " **The laptop's model is ok, and its performance is great.** ". The proposed SSA-GRU-AE model identifies the aspect term "model" and its context as "ok" correctly and predicts its sentiment polarity as neutral. But for the aspect term "performance" the context is identified as "great" and predicts its sentiment polarity as positive. The sparse-self-attention mechanism assigns high weights to the terms "laptop", "model", "ok", " performance", and "great" and low weights to the terms "The", "is", "its" and "and". This example shows the importance of sparsification in self-attention mechanism for ASC tasks. The L1-regularize applied on attentions helped to ensure that only a few words contributed to the $sentiment$ of sentences. This is due to the fact that when entire neurons or filters are eliminated, the principles of associativity and distributivity can be employed to convert a sparsified structure into a more compact, dense structure. Nevertheless, in the event that eliminating arbitrary components of a weight matrix, it becomes necessary to retain the indices corresponding to the non-zero items that remain. The process of model sparsification alters the model's characteristics, although it does not modify the sparsity pattern observed after successive inferences or forward passes. This helped proposed model to perform better on datasets which are grammatically poor and noisy.

Consider the following example, " **The workers should do more work truly.** " by considering "workers" as the aspect and "more work truly" as its context phrase to identify its sentiment polarity. The existing DNN models identify the aspect term as "workers" and its context term as "work truly" and predicts the sentiment polarity as positive. But actually the weight of the term "more" is important in this context. The sparse-self-attention in the proposed model helps to identify the context term "more work truly" to predicts its actual sentiment polarity "negative". This elucidates the importance of the sparse nature in self-attention mechanism. This also shows that the proposed model can able to capture the context words with implicit meaning (Fig. 8).

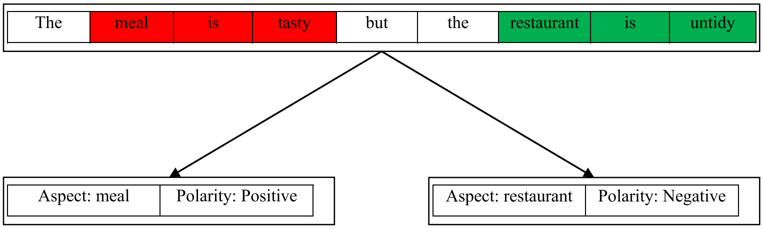

Fig.8 Case study example

6 Conclusion

SSA-GRU-AE is proposed to perform ASC task which contains three parts. The first part is BERT embedding layer, second part is GRU layer and third part is sparse self-attention layer. The BERT pre-trained model used in this work effectively captures contextual word embedding of both the sentence and aspect and also captures relation between them. The sparse self-attention mechanism proposed in our work effectively captures the important part of sentences with respect to aspect. The experimental results on 5 datasets (twitter, LAP14, REST14, REST15 and REST16) proved that proposed method outperformed the existing baseline models in terms of accuracy and F1 – score. Addition of ablation study and discussion has further demonstrated the proficiency of proposed model.

Acknowledgements Not applicable.

Authors contributions All authors equally contributed to this work. PRJD developed the idea and written the manuscript, BS supervised and reviewed this work, GR designed the experimental setup & analyzed the results and SRK & KVD prepared the figures and tables.

Funding Not applicable.

Data Availability Supporting data will be provided based on request.

Declarations

Competing interests The authors declare no competing interests.

Consent for Publication Not applicable.

References

1. Ambartsoumian A, Popowich F (2019) Self-attention: a better building block for sentiment analysis neural network classifiers. 130–139

2. Wang P et al (2023) A novel adaptive marker segmentation graph convolutional network for aspect-level sentiment analysis. Knowledge-Based Syst 270:110559
3. Nazir A, Rao Y, Wu L, Sun L (2020) Issues and challenges of aspect-based sentiment analysis: a comprehensive survey. IEEE Trans Affect Comput 3045:1–20
4. Dhanith PRJ, Prabha KSS (2023) A critical empirical evaluation of deep learning models for solving aspect based sentiment analysis. Artif Intell Rev
5. Wang Y, Huang M, Zhao L, Zhu X (2016) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
6. Yao X (1999) Evolving artificial neural networks. Proc IEEE 87(9):1423–1447
7. Zhou J, Huang JX, Chen Q, Hu QV, Wang T, He L (2019) Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. IEEE Access 7:78454–78483
8. Liu H, Chatterjee I, Zhou M, Lu XS, Abusorrah A (2020) Aspect-based sentiment analysis: A survey of deep learning methods. IEEE Trans Comput Soc Syst 7(6):1358–1375
9. Goldberg Y, Levy O (2014) word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722..
10. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26
11. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: 1st international conference on learning representation ICLR 2013 - Work. Track Proc., pp 1–12
12. Pennington J, Richard S (2017) GloVe: global vectors for word representation Jeffrey. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
13. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 4(January):3104–3112
14. Al-Smadi M, Talafha B, Al-Ayyoub M, Jararweh Y (2019) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybern 10(8):2163–2175
15. Yadav RK, Jiao L, Goodwin M, Granmo OC (2021) Positionless aspect based sentiment analysis using attention mechanism[Formula presented]. Knowledge-Based Syst 226:107136
16. Wu H, Zhang Z, Shi S, Wu Q, Song H (2022) Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis. Knowledge-Based Syst 236:107736
17. Su J et al (2021) Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning. Artif Intell 296:103477
18. Shu L, Xu H, Liu B (2019) Controlled CNN-based sequence labeling for aspect extraction.
19. Liu G, Huang X, Liu X, Yang A (2020) A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network. Comput J 63(3):410–424
20. Wang X, Li F, Zhang Z, Xu G, Zhang J, Sun X (2021) A unified position-aware convolutional neural network for aspect based sentiment analysis. Neurocomputing 450:91–103
21. Karimi A, Rossi L, Prati A (2020) Improving BERT performance for aspect-based sentiment analysis.
22. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics vol 1, pp 4171–4186
23. Liang Y, Meng F, Zhang J, Chen Y, Xu J, Zhou J (2021) A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. Neurocomputing 454:291–302
24. Kiritchenko S, Zhu X, Cherry C, Mohammad S (2015) NRC-Canada-2014: detecting aspects and sentiment in customer reviews. No. SemEval, pp 437–442
25. Nguyen TH, Shirai K (2015) PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis. In: Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process., no. September, pp. 2509–2514, 2015.
26. Luo H, Li T, Liu B, Wang B, Unger H (2019) Improving aspect term extraction with bidirectional dependency tree representation. IEEE/ACM Trans Audio Speech Lang Process 27(7):1201–1212
27. Zhang Y, Xu B, Zhao T (2020) Convolutional multi-head self-attention on memory for aspect sentiment classification. IEEE/CAA J Autom Sin 7(4):1038–1044
28. Kumar A, Teja Narapareddy V, Aditya Srikanth V, Bhanu Murthy Neti L, Malapati A (2020) Special section on advanced data mining methods for social computing aspect-based sentiment classification using interactive gated convolutional network. pp 22445–22453
29. Ma Y, Peng H, Cambria E (2018) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. 32nd AAAI Conf Artif Intell AAAI 2018:5876–5883
30. Giannakopoulos A, Musat C, Hossmann A, Baeriswyl M (2017) Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. arXiv preprint arXiv:1709.05094..

31. Zeng J, Ma X, Zhou K (2019) enhancing attention-based LSTM with position context for aspect-level sentiment classification. IEEE Access 7:20462–20471
32. Setiawan EI, Ferry F, Santoso J, Sumpeno S, Fujisawa K, Purnomo MH (2020) Bidirectional GRU for targeted aspect-based sentiment analysis based on character-enhanced token-embedding and multi-level attention. Int J Intell Eng Syst 13(5):392–407
33. Sun K, Zhang R, Mensah S, Mao Y, Liu X (2019) Aspect-level sentiment analysis via convolution over dependency tree. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing *(EMNLP-IJCNLP)*, pp 5679–5688
34. Tai KS, Socher R, Manning CD (2015) Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of 45th annual meeting of the association for computational linguistics (ACL), vol 1, 1556–1566
35. Liang Y, Meng F, Zhang J, Xu J, Chen Y, Zhou J (2020) A novel aspect-guided deep transition model for aspect based sentiment analysis. In: Proceedings of the 2020 conference on empirical methods in natural language processing, pp 5569–5580
36. Tang D, Qin B, Feng X, Liu T (2015) Effective LSTMs for target-dependent sentiment classification. In: COLING 2016 - 26th international conference on computing linguistics, pp 3298–3307
37. Ma Y, Peng H, Khan T, Cambria E, Hussain A (2018) Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cognit Comput 10(4):639–650
38. Li L, Liu Y, Zhou A (2018) Hierarchical attention based position-aware network for aspect-level sentiment analysis. In: Proceedings of the 22nd conference on computational natural language learning, pp 181–189
39. Zheng S, Xia R (2018) Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention
40. Laddha A, Mukherjee A (2019) Aspect specific opinion expression extraction using attention based LSTM-CRF network. International conference on computational linguistics and intelligent text processing 2018 Mar 18. Springer, Cham, pp 442–454
41. Lin P, Yang M, Lai J (2021) Deep selective memory network with selective attention and inter-aspect modeling for aspect level sentiment classification. IEEE/ACM Trans Audio Speech Lang Process 29:1093–1106
42. Zhang C, Li Q, Song D (2020) Aspect-based sentiment classification with aspect-specific graph convolutional networks. In: EMNLP-IJCNLP 2019–2019 conference on empirical methods in natural language processing, pp 4568–4578
43. Hou X, Huang J, Wang G, Qi P, He X, Zhou B (2021) Selective attention based graph convolutional networks for aspect-level sentiment classification. 83–93
44. Wu C et al (2021) Multiple-element joint detection for aspect-based sentiment analysis. Knowledge-Based Syst 223:107073
45. Ben Veyseh AP, Nouri N, Dernoncourt F, Tran QH, Dou D, Nguyen TH (2020) Improving aspect-based sentiment analysis with gated graph convolutional networks and syntax-based regulation. Find Assoc Comput Linguist EMNLP 2020:4543–4548
46. Wang K, Shen W, Yang Y, Quan X, Wang R (2020) Relational graph attention network for aspect-based sentiment analysis. 3229–3238
47. Duchi J, Hazan E, Singer Y (2010) Adaptive subgradient methods for online learning and stochastic optimization. In: COLT 2010 - 23rd conference on learning theory, pp 257–269
48. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: 52nd annual meeting association computing, linguistics ACL 2014, vol 2, pp 49–54
49. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2015) SemEval-2014 Task 4: aspect based sentiment analysis. no. SemEval, pp 27–35
50. Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I (2015) SemEval-2015 task 12: aspect based sentiment analysis, 486–495
51. Pontiki M et al (2016) SemEval-2016 task 5: aspect based sentiment analysis. In: ProWorkshop on semantic evaluation, pp 19–30
52. Ma D, Li S, Zhang X, Wang H (2017) Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the twenty-sixth international joint conference artificial intelligence, pp 4068–4074
53. Tang D, Qin B, Liu T (2016) Aspect level sentiment classification with deep memory network. In: EMNLP 2016 – conference on empirical methods natural language processing, 214–224
54. Huang B, Ou Y, Carley KM (2018) Aspect level sentiment classification with attention-over-attention neural networks. In: Social, cultural, and behavioral modeling: 11th international conference, SBP-BRiMS 2018, Washington, DC, USA, July 10–13, 2018, Proceedings 11