**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

# Bidirectional Encoder Representations from Transformers (Bert) And Serialized Multi-Layer Multi-Head Attention Feature Location Model Foraspect-Level Sentiment Analysis

**I. Anette Regina**
Research Scholar, Department of Computer Science, Periyar University, Salem
anette1967cs@gmail.com

**Dr. P. Sengottuvelan,**
Associate Professor, Department of Computer Science, P.G Extension Centre,
Periyar University, Dharmapuri.
sengottuvelan@gmail.com

**ABSTRACT**

With the popularity of internet social platforms, sentiment analysis has become one of the hottest topics in Natural Language Processing (NLP). It has seen a lot of attention in the last few years. The purpose of Aspect-level Sentiment Classification (ASC) is to expose the sentiment polarity of users' opinions on a specific aspect in the text. ASC has two distinct parts such as Aspect Extraction (AE) and labeling the Aspects with Sentiment Polarity (ALSA). However, the existing ALSA methods mainly focus on attention mechanisms and recurrent neural networks. They lack emotional sensitivity to the position of aspect words and tend to ignore long-term dependencies. In this paper, argue that the prediction of aspect-level sentiment polarity depends on both context and target. Bidirectional Encoder Representations from Transformers (BERT) and Serialized Multi-layer Multi-Head Attention (SMMHA) based feature location method is proposed to solve the problem of ALSA. Specifically, a pretrained BERT model is proposed to mine more aspect-level auxiliary information from the comment context. For the sake of learning the expression features of aspect words and the interactive information of aspect words' context, SMMHA feature extraction method is introduced for ASC. Amazon customer review data is proposed which focuses on finding aspect terms from each review, applying classification algorithms to find the score of each review. In this method, SMMHA is introduced to better capture the sentiment features in short texts. "Amazon Customer Review Dataset" which is collected from Amazon. The study shows that assigning higher results in accuracy, precision, recall, and F1-score of the sentiment prediction when compared to existing methods.

**INDEX TERMS:** Sentiment Analysis, feature extraction, Aspect-level Sentiment Classification (ASC), Aspects with Sentiment Polarity (ALSA), Bidirectional Encoder Representations from Transformers (BERT), Serialized Multi-layer Multi-Head Attention (SMMHA), and Amazon dataset.

## 1. INTRODUCTION

E-commerce is a thriving industry with increasing importance to the global economy. Particularly with the rapid development of social media, more and more users begin to express their sentiments on various online platforms. These comments reflect the sentiments of users and consumers and provide sellers and governments with a lot of valuable feedback on the quality of goods or services [1–3]. Governments and companies can collect a large number of public comments directly from the Internet and analyze users' opinions and satisfaction from them, so as to meet their needs. Therefore, as a basic and key work of Natural Language Processing (NLP), sentiment analysis has attracted widespread attention from the theoretical and practical circles [4].

SA is the process of manipulating textualmedia and extracting the subjective value from the text. It determines the reviewauthor's attitude towards a text: whether it is positive, negative, or indifferent. SA iscurrently being used all over the internet for various purposes such as political profiling,recommendation engines, fact checking, spam filtering, etc. It

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

has rapidly generated a lotof attention among researchers working with machine learning methods. However, the classic SA task can only determine the users' sentiment polarities (e.g., positive, negative, and neutral) of the product or event from the entire sentences and cannot determine the sentiment polarity of a particular aspect of the sentence, let alone identify the multiple sentiments existing in a single sentence. In contrast, aspect-based sentiment analysis is a more fine-grained classification task, which can identify the sentiment polarities of multiple aspects in a sentence. Aspect-Based Sentiment Analysis (ABSA) aims to determine sentiment polarity with respect to a specified aspect term in a piece of text [5,6].

In the ABSA process, the concerned target on which the sentiment is expressed shifts from an entire sentence or document to an entity or a certain aspect of an entity. ABSA is thus the process of building a comprehensive opinion summary at the aspect level, which provides useful fine-grained sentiment information for downstream applications. In ABSA process, three processing steps can be distinguished when performing aspect-level sentiment analysis: identification, classification, and aggregation [7]. While in practice, not every method implements all three steps or in this exact order, they represent major issues for aspect-level sentiment analysis. The first step is concerned with the identification of sentiment-target pairs in the text. The next step is the classification of the sentiment-target pairs. The expressed sentiment is classified according to a predefined set of sentiment values, for instance positive and negative. Sometimes the target is classified according to a predefined set of aspects as well.

Aspect-level Sentiment Classification (ASC) is an interesting and challenging research task to identify the sentiment polarities of aspect words in sentences [8].ASC, machine learning models a series of features, e.g., a set of words and sentiment dictionaries, were set up to train classifiers [9]. Their classification effect heavily depended on the features' quality.However, those methods rely on carefully designed manual features on large-scale datasets, resulting in a lot of waste of manpower and time [10, 11]. The neural network model can automatically learn the low dimensional representation of reviews without relying on artificial feature engineering. Recently, neural network methods have dominated the study of ABSA since these methods can be trained end-to-end and automatically learn important features [12].

In recent, the recurrent neural network (RNN) and its variant models have been widely used in ASC tasks [13]. Lai et al [14] used a two-way loop structure to obtain text information. Compared with traditional window-based neural networks, their method reduced more noise. For targeted sentiment classification, Gan et al [15] put forward a sparse attention mechanism based on a separable dilated convolution network. Their method is superior to the existing methods. Tang et al [16] proposed a Target-Dependent Long-term Short-Term Memory network (TD-LSTM). This network is modeled by the contexts before and after the target word. By combining the information of the two LSTM hidden layer states, they further achieved the ASC tasks. Compared with the RNN model, the performances of these RNN variant models have small improvements on the ASC task. Besides, the neural network is difficult to capture long-term dependencies between aspect words and context, which causes a loss of valuable information. Even if the attention mechanism [17] can be positioned in the right context to alleviate this problem, but the problem still remains and limits their performance.

For the sake of solving the aforementioned problems, on the basis of Serialized Multi-layer Multi-Head Attention with Bidirectional Encoder Representations from Transformers (SMMHA-BERT) method is introduced for feature extraction. The core idea of the SMMHA-BERT approach is to recognize the emotion of different aspect words in the text, consider the contextual interaction information of aspect words, and reduce the interference of irrelevant words, thus forming an effective aspect-based sentiment analysis framework. Extensive experiments on Amazon Customer Review Dataset are conducted, and their model is evaluated by using the Precision, Recall, F1-Score, and Accuracy.

## 2. LITERATURE REVIEW

Li et al [19] exploited a new direction named coarse-to-fine task transfer, which aims to leverage knowledge learned from a rich-resource source domain of the coarse-grained AC task, which is more easily accessible, to improve the learning in a low-resource target domain of the fine-grained AT task. Multi-Granularity Alignment Network (MGAN) proposed both the aspect granularity inconsistency and feature mismatch between domains. In MGAN, a novel Coarse2Fine attention guided by an auxiliary task can help the AC task modeling at the same finegrained level with the AT task. To alleviate the feature false alignment, a contrastive feature alignment method is adopted to align aspect-

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

specific feature representations semantically. In addition, a large-scale multi-domain dataset for the AC task is provided. Empirically, extensive experiments demonstrate the effectiveness of the MGAN.

Zeng et al [20] proposed new attentive Long Short-Term Memory (LSTM) model, dubbed Position ATTention (PosATT-LSTM), which not only takes into account the importance of each context word but also incorporates the position-aware vectors, which represents the explicit position context between the aspect and its context words.The relationships are applied to the calculations of the attention weights. Position-aware influence vectors are appended into the hidden representations of the context words on top of LSTM layer. At last, the ultimate aspect-specific attentive representations are obtained via computing attention weights between the aspect embedding and the concatenated representations. Conduct substantial experiments on the SemEval 2014 datasets, and the encouraging results indicate the efficiency of proposed approach.

Tan et al [21] presented a learning method which trains aspect embeddings according to the relation between aspect-categories and aspect-terms. Cosine measure metric is successfully alleviated in the aspect embeddings which are trained by method. The trained aspect embeddings can be used as initialization in existing models to solve ACSA task. Experiments on SemEval datasets are used for ACSA task, and the results indicate that pre-trained aspect embeddings are capable of improving the performance of sentiment analysis.Xu et al [22] proposed a Multi-Attention Network (MAN) makes uses intra- and inter-level attention mechanisms. In the former, the MAN is employed a transformer encoder instead of a sequence model to reduce training time. The transformer encoder encodes the input sentence in parallel and preserves long-distance sentiment relations. In the latter, the MAN uses a global and a local attention module to capture differently grained interactive information between aspect and context. The global attention module focuses on the entire relation, whereas the local attention module considers interactions at word level; this was often neglected in previous studies. Experiments demonstrate that the proposed model achieves superior results when compared to the baseline models.

Zhang et al [23] proposed a Multi-head Attention (MHA) networks. First, the word embedding and aspect term embedding are pre-trained by Bidirectional Encoder Representations from Transformers (BERT). Second, make full use of MHA and convolutional operation to obtain hidden states, which is superior to traditional neural networks. Then, the interaction between context and aspect term is further implemented through averaging pooling and MHA. Extensive experiments are conducted on three benchmark datasets and the final results show that the Interactive Multi-head Attention Networks (IMAN) model consistently outperforms the state-of-the-art methods on ASC task.

Zhou and Wang [24] proposed a position and self-attention mechanism R-Transformer network (PSRTN) model. Firstly, obtaining the position-aware influence propagates between words and aspects by Gaussian kernel and generating the influence vector for each context word. Secondly, capturing global and local information of the context by the R-Transformer, and using the self-attention mechanism to obtain the keywords in the aspect. Finally, context representation of a particular aspect is generated for classification. In order to evaluate the validity of the model, conduct experiments on SemEval2014 and Twitter. Thus, it is clear that the position information needs to consider in the context attention calculation.

Zhou et al [25] proposed a Filter Gate Network based on Multi-head attention (FGNMH). First, train the context in a domain-specific corpus and integrate the part-of-speech features of the context to enrich the representation of the context. Second, use multi-head attention mechanism to model contextual semantic information. Finally, a filter layer is designed to remove context words that are irrelevant to current aspect. To verify the effectiveness of FGNMH, conduct a large number of experiments on SemEval2014, Restaurant15, Restaurant16 and Twitter.

Leng et al [26] proposed a new model combines a bidirectional Long Short-Term Memory network (BiLSTM) or a bidirectional Gated Recurrent Unit (biGRU) and an Enhanced Multi-Head Self-Attention mechanism. The Enhanced Multi-Head Self-Attention is a two-layer modified Transformer encoder. Through this attention, the inter-sentence information can be encoded. BiLSTM is better than biGRU by comparing the effect of them in the model. Finally, Bidirectional Encoder Representations from Transformers (BERT) is used in method instead of word2vec as a pre-

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

training structure. The movie review datasets (Internet Movie Database (IMDB) movie comment dataset and Stanford Sentiment Treebank v2 (SST-2) sentiment dataset) are used in experiments. Experiment results show that the proposed model performs better in terms of accuracy, precision, recall rate, and F1-scores comparing with the baseline models.Therefore, the attention mechanism is becoming more and more important in the ASC task.

Tang et al [27] proposed a deep Memory Network (MemNet) for aspect level sentiment classification. A sequential neural model by LSTM, this approach explicitly captures the importance of each context word when inferring the sentiment polarity of an aspect. Such importance degree and text representation are calculated with multiple computational layers, each of which is a neural attention model over an external memory. Experiments on laptop and restaurant datasets demonstrate that proposed approach performs comparable to state-of-art feature based SVM system, and substantially better than LSTM and attention-based LSTM architectures. On both datasets are shows that multiple computational layers could improve the performance. The deep memory network with 9 layers is 15 times faster than LSTM with a Central Processing Unit (CPU) implementation.

Song et al [28] proposed an Attentional Encoder Network (AEN) which eschews recurrence and employs attention based encoders for the modeling between context and target. Raise the label unreliability issue and introduce label smoothing regularization. Also apply pre-trained BERT to this task and obtain new state-of-the-art results. Experiments and analysis demonstrate the effectiveness and lightweight of model.Pang et al [29] proposed an effective aspect-level sentiment analysis- Bidirectional Encoder Representations from Transformers (ALM-BERT) by constructing an aspect feature location model. Pretrained BERT model is introduced to firstly to mine more aspect-level auxiliary information from the comment context. Secondly, for the sake of learning the expression features of aspect words and the interactive information of aspect words' context, construct an aspect-based sentiment feature extraction method.

## 3. PROBLEM FORMULATION

Aspect-based sentiment analysis refers to the process of outputting the sentiment polarity of each aspect word in a sentence with a sentence and some predefined aspect words as input data. The aspect-based sentiment analysis can be defined as follows,

Formally, give a comment sentence $S = \{w_1, w_2, \cdots, w_n\}$, where n is the total number of words in S. $A = \{a_1, \cdots, a_i, \cdots, a_m\}$ with length m represents an aspect vocabulary of length m, where $a_i$ denotes the $i^{th}$ aspect word in aspect vocabulary A, and A is a subsequence of sentence S. $P = \{p_1, \cdots, p_j, \cdots, p_C\}$ denotes the candidate sentiment polarities, where C denotes the number of categories of sentiment polarity and the $p_j$is the $j^{th}$ sentiment polarity. The goal of the aspect-based sentiment analysis model is to predict the most likely sentiment polarity ofspecific aspect word in a sentence, which can be formulated as follows,

Input: $\begin{cases} S = \{w_1, \ldots w_n\}, \\ A = \{a_1, \ldots . a_m\}, \end{cases}$

Output: $p_k = \phi_{max}(a_i, p_j|S)$,

Constraints: $A \in S, m \in [1, N]$

where $\phi$represents a function that quantifies the degree of matching between the aspect word $a_i$, A denotes the aspect vocabulary, and the sentiment polarity $p_j$in the sentence S. Finally, the model outputs the sentiment polarity with the highest matching degree to be the classification result.

## 4. PROPOSED METHODOLOGY

In this work, prediction of aspect-level sentiment polarity depends on both context and target. Bidirectional Encoder Representations from Transformers (BERT) and Serialized Multi-Layer Multi-Head Attention (SMMHA) method is proposed to solve the problem of ALSA. Aspect-location model is proposed based on SMMHA& BERT methods, which can mine different aspects of sentiment in Amazon review dataset. Amazon customer review dataset is used which focuses on finding aspect terms from each review, applying classification algorithms to find the score of each review. The overall framework of the proposed approach is shown in Figure 1, which mainly includes four parts: multiangle text vectorization mechanism, importantfeature extraction model, fusion layer, and sentimentpredictor.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

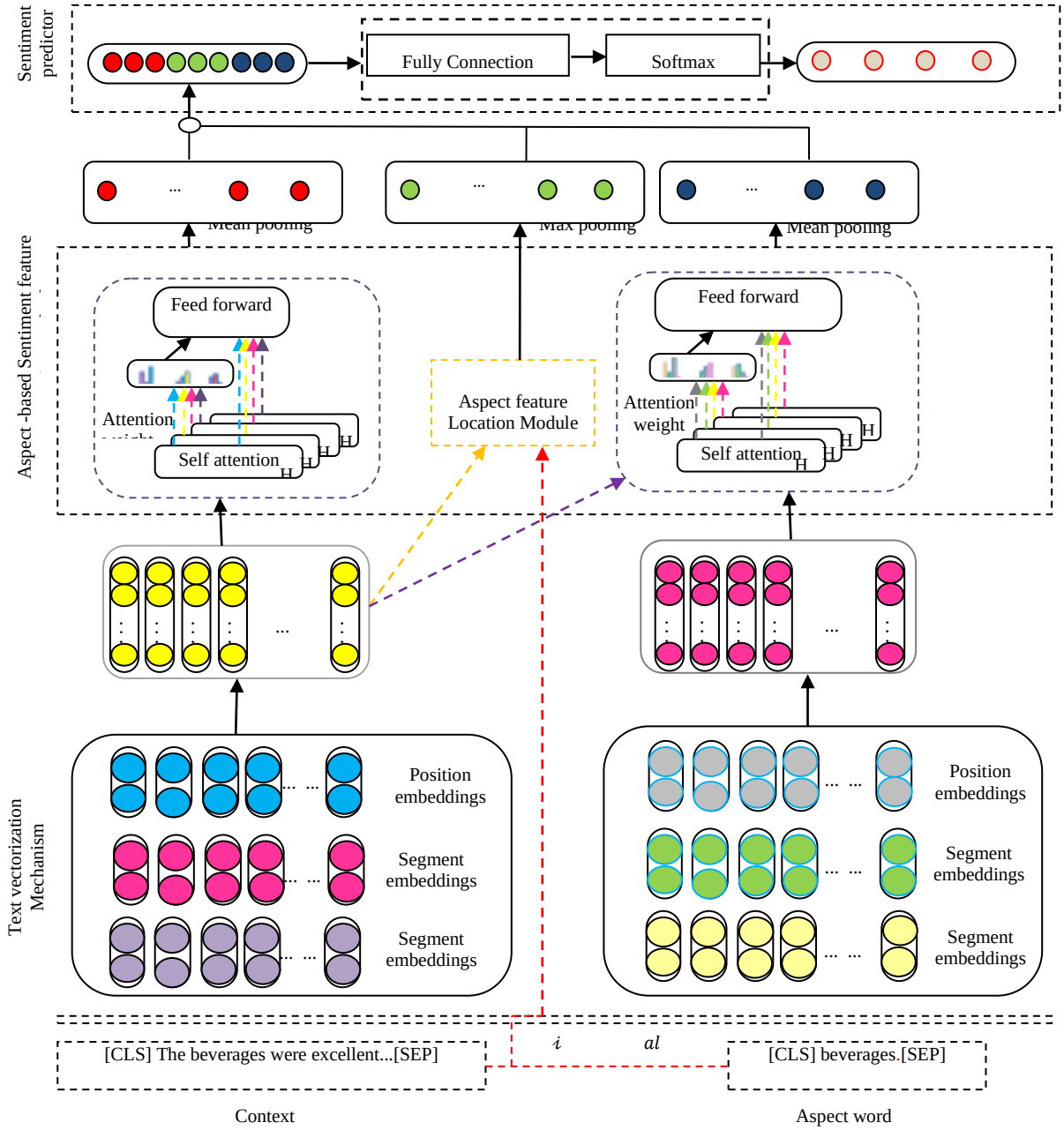



**FIGURE 1. OVERALL FRAMEWORK OF SMMHA-BERT**

Firstly, the pretrained model BERT is introduced to generate a high-quality word vector of sequence, which provides effective support for subsequent steps. Then, a new feature extractor of Serialized Multi-Layer Multi-Head Attention (SMMHA) mechanism and position feedforward network is introduced to extract important context and target information and build an aspect feature location model, which can select information related to aspect words from context feature representation.

### 4.1. Multiangle Text Vectorization Mechanism

The word embedding maps each word to a high-dimensional vector space, which mainly assists machines in understanding natural language. Its mainstream methods include Word2vec and Glove. Both of these methods belong to context-based word embedding models and have achieved good performance in aspect-level sentiment analysis tasks.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

However, previous research has already demonstrated that these two word embedding models cannot capture the enough information in the text [30], which leads to poor classification accuracy and reduces the performance of the aspect-based sentiment analysis model. Therefore, a high-quality word embedding model has an important influence on improving the accuracy of classification results [31]. The key of aspect-level sentiment analysis is to understand NLP effectively. BERT is a language pretraining model that can effectively use unlabeled text. The model utilizes a method of randomly covering some words, utilizes a multilayer two-way converter encoder to extract a general NLP recognition model from a large amount of unlabeled text, and further uses a small amount of labeled data for fine-tuning to generate high-quality text feature vectors. Inspired by this idea, the proposed approach adds special word breaker tags [CLS] and [SEP] at the beginning and end of a given word sequence, respectively, and finally divides a given sequence into different segments. That is, the word embedding vector input in this way includes generating vectors such as token embeddings, segmentation embedding, and position embedding for different segments. In the proposed approach, convert the comment text and aspect word into the form of "[CLS] + comment text + [SEP]" and "[CLS] + target + [SEP]", correspondingly. Finally, obtain the context representation $E_c$and aspect representation $E_a$,

$$E_c = \{we_{|CLS|}, we_1, we_2, \ldots., we_{|SEP|}\} \quad (1)$$

$$E_a = \{ae_{|CLS|}, ae_1, ae_2, \ldots., ae_{|SEP|}\} \quad (2)$$

where $we_{|CLS|}, ae_{|CLS|}$denotes the vector of classification mark [CLS], and the $we_{|SEP|}$,and $ae_{|SEP|}$expressions the vector of separator [SEP].

### 4.2. Aspect-Based Sentiment Feature Extraction Method

In order to extract the implicit features of the aspect words and their context and to consider the auxiliary information contained in the aspect words, design an aspect-based sentiment feature extraction method inspired by a transformer encoder [32]. The basic idea of this method is to integrate the information of aspect words and context and to model the interaction between context and target words. Furthermore, we hold the opinion that the accuracy of sentiment classification can be improved by capturing the feature information of aspect words in context.Among them, the different aspects include query sequence (Q), key-value pairs (K and V).

$$f_s(Q, K, V) = \sigma\big(f_e(Q, K)\big)V \quad (3)$$

where $\sigma(.)$ stands for the normalized exponential function, and $f_e(.)$ is the energy function to learn the correlation features between K and Q, which can be calculated by using the following equation (4),

$$f_e(Q, K) = \frac{QK^T}{\sqrt{d_k}} \quad (4)$$

where $\sqrt{d_k}$denotes the scale factor, and the $d_k$is the dimension of the query and key vectors.The attention score of Serialized Multi-Layer Multi-Head Attention (SMMHA) $f_{mmah}(.)$ is obtained by concatenating attention score of self attention mechanism,

$$f_{mmah}(Q, K, V) = \left[a^1; a^2; a^3; \ldots.; a^i; \ldots.; a^{n-head}\right]W_d \quad (5)$$

$$a^i = f_s^i(Q, K, V) \quad (6)$$

where $a_i$represents the $i^{th}$ attention score, [; ] denotes concatenates of the vector, and $W_d$is the weight matrix.

#### 4.2.1. Statistics pooling

Let $h_{tc}$be the latent context vector at the output of the attention network with time. By statistics pooling, compute the mean and standard deviation of $h_{tc}$ along the context aware information with time, t=1,….,T. In particular, the first and second-order statistics are computed as follows,

$$\mu = \frac{1}{T}\sum_{t=1}^{T} h_{tc} \quad (7)$$

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

$$\sigma = \frac{1}{T}\sqrt{\sum_{t=1}^{T} h_{tc} \odot h_{tc} - \mu \odot \mu} \tag{8}$$

where equal weights of $\alpha_t = \frac{1}{T}$ are assigned to all features with query sequence (Q). The operator $\odot$ represents element-wise multiplication. The mean and standard deviation are concatenated as a fixed-dimensional representation and mapped to the feature embedding vector, typically, implemented with a Fully Connected (FC) layer.

#### 4.2.2. Attentive statistics pooling

Attentive statistics pooling method aims to capture the context information focusing on the importance of features [33]. An attention model works in conjunction with the original embedding neural network and calculates a scalar score $e_t$ for each feature with aspect word, as follows,

$$e_t = V^T f(Wh_{tc} + b) + k \tag{9}$$

where $f(.)$ is a non-linear activation function, such as tanh or ReLU. The scores are normalized over all features with a softmax function as follows,

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau}^{T} \exp(e_\tau)} \tag{10}$$

The normalized scores are then used as the weights in the pooling layer to calculate a weighted mean $\tilde{\mu}$ and a weighted standard deviation $\tilde{\sigma}$,

$$\tilde{\mu} = \frac{\exp(e_t)}{\sum_{\tau}^{T} \exp(e_\tau)} \tag{11}$$

$$\tilde{\sigma} = \sqrt{\sum_{t=1}^{T} \alpha_t h_t \odot h_t - \tilde{\mu} \odot \tilde{\mu}} \tag{12}$$

Let $(v_{tc}, k_{tc}, q)$ be the {value, key, query} tuple. Here, $v_{tc}$ is the value vector with $d_v$ dimensions, q is a time-invariant query with $d_q$ dimensions, and $k_{tc}$ is the key vector with $d_k$ dimensions. The $(v_{tc}, k_{tc})$ is derived from different layers of the feature processor network, while q is a trainable parameter. The query vector maps the key vector sequence $[k_1, k_2, \ldots, k_{Tc}]$ to the weights $[k_1, k_2, \ldots, k_{Tc}]$ via scaled dot-production attention and softmax function,

$$\alpha_t = \text{softmax}\left(\frac{q.k_{tc}}{\sqrt{d_k}}\right) \tag{13}$$

Note that the softmax is performed along the time. Finally, the weighted mean $\tilde{\mu}$ and weighted standard deviation $\tilde{\sigma}$ are computed in the same way as equation (11) and equation (12) with the weights $\alpha_{tc}$ applied on the value vectors $v_t$.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

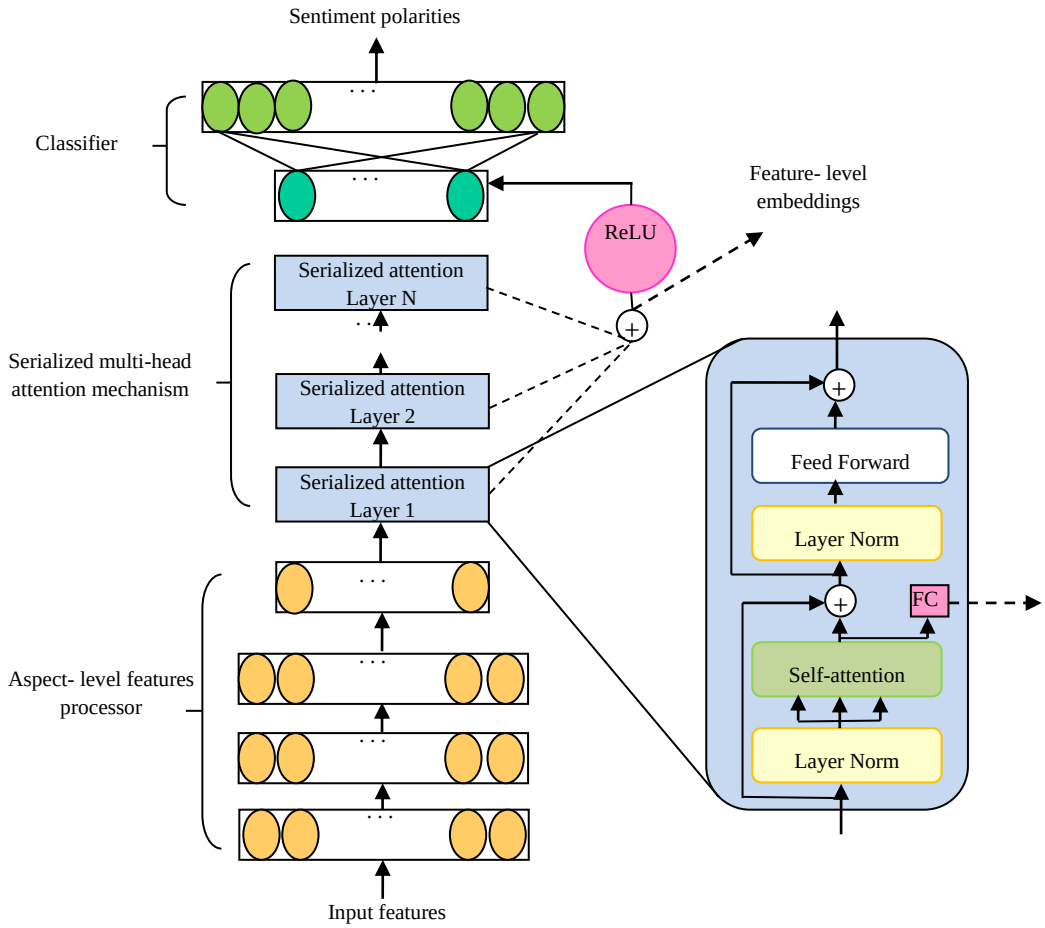



**FIGURE 2. SERIALIZED MULTI-LAYER MULTI-HEAD ATTENTION (SMMHA) BASED FEATURE LOCATION MECHANISM**

Serialized multilayer multi-head attention mechanism is proposed in this work. As depicted in Figure 2, the embedding neural network consists of three main stages, namely, an aspect-level feature processor, a serialized attention mechanism, and an aspect classifier. The aspect-level feature processor is the same as that in the x-vector [29]. The middle part of Figure 2, a serialized attention mechanism is used to aggregate the variable-length feature sequence into a fixed-dimensional representation. The top part of Figure 2 is feed-forward classification layers. Similar to the x-vector, the entire network is trained to classify input sentence into aspect classes.

### 4.2.3. Serialized attention

The serialized attention mechanism consists of a stack of N identical layers, and each layer is composed of two modules stacked together, i.e, a self-attention module and a feed forward module. A residual connection is employed around each of these modules. As in [34], layer normalization is applied on the input before the self-attention module and feed-forward module, separately. That is, the output of each sub-layer is x + Layer Norm (Sublayer(x)). Instead of having multi-head attention is proposed in parallel to aggregate and propagate the information from one layer to the next in a serialized manner with stacked self-attentionmodules. In the original multi-head attention, the input sequence is split into several homogeneous sub-vectors called heads. However, a deeper architecture of the aggregation network will increase the representational capacity; with more discriminative features can be learned and aggregated at different levels. In the proposed serialized attention mechanism, self attention module is performed in a serialized manner, allowing the model to aggregate information with temporal context from deeper layers. Specifically, from the nth self attention module ($n^{th}$ $[1, \ldots, N]$), the weighted mean $\tilde{\mu}$ and weighted standard deviation$\tilde{\sigma}$are obtained. After transformed by an affine transformation, it is converted to a feature-level vector, also seen as a serialized head from layer n. The final aspect-level embedding is then obtained with the summation of the feature-level vectors from all heads. After passing through a ReLU activation and Batch Normalization, it is then fed into classifier layers.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

#### 4.2.4. Input-aware self-attention

The attention function is mapping a query and a set of key-valuepairs to an output [32]. Instead of using a fixed query for all features, employ an input-aware unique queryfor each feature. Considering that mean and standard deviationare capable of capturing the overall information, statistics pooling is used togenerate the query.

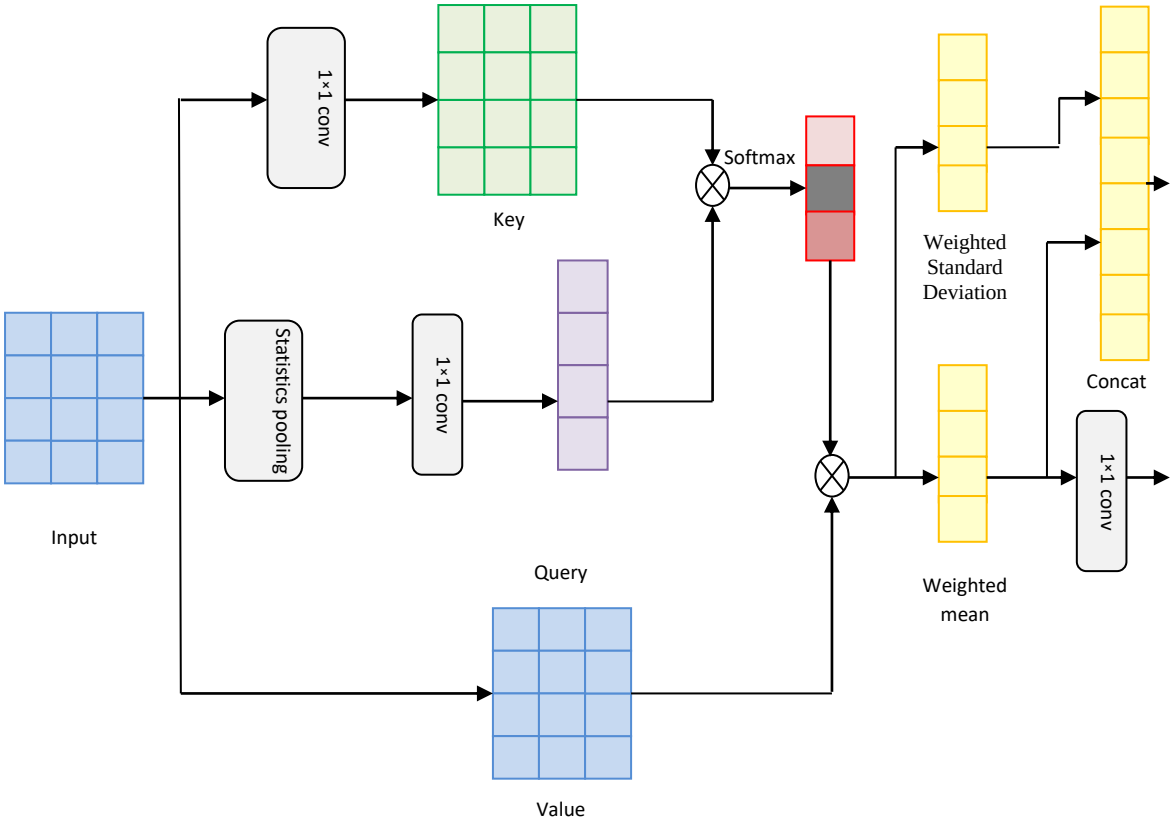



**FIGURE 3. SELF-ATTENTION MECHANISM WITH INPUT-AWARE QUERY**

As shown in Figure 3, consider an input sequence$[h_1,\ldots, h_{Tc}]$ with $h_{tc} \in \mathbb{R}^d$, where T is the length ofthe input sequence. The model transforms the input sequenceinto the query q as follows,

$$q = W_q g(h_{tc}) \tag{14}$$

where $g(.)$ is statistics pooling is applied to calculate $[\mu, \sigma]$ with equation (7) and equation (8), and $W_q \in \mathbb{R}^{d_k \times 2d}$ is a trainable parameter.As for key-value pairs, in order to reduce the number ofmodel parameters, the input sequence $[h_1,\ldots, h_{Tc}]$ is directlyassigned to the value sequence $[v_1,v_2,\ldots,v_{Tc}]$ of d dimensionswithout any extra computation. The key vector $k_{tc}$is obtained by a linear projection with a trainable parameter$W_k \in \mathbb{R}^{d_k \times d}$.

$$k_{tc} = W_k h_{tc} \tag{15}$$

With $(v_{tc}, k_{tc}, q)$ as $\{value, key, query\}$ tuple, the weights arecomputed via scaled dot-product attention as in equation (13). The firstand second-order statistics are calculated the same as in equation (11) andequation (12). The weighted mean vector $\tilde{\mu}$ is then added to all featuresafter an affine transformation in the residual connection.

#### 4.2.5. Serialized multi-head embedding

After the self-attention module, the output from each selfattention layer is fed into a feed-forward module, which is to process the output to better fit the input for the next selfattention layer [32]. It consists of two linear transformations with ReLU activation in between.

$$FFW(h) = W_2 f(W_1 h + b_1) + b_2 \tag{16}$$

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

where h is the input, f(.) is a ReLU function, and the linear transformations are different from layer to layer. $W_1 \in \mathbb{R}^{d_{ff} \times d}, W_2 \in \mathbb{R}^{d \times d_{ff}}$with inner dimension $d_{ff}$, which can also be described as two convolutions with kernel size 1. The feature-level embedding of serialized attention mechanism is fed into one fully-connected layer and a standard softmax layer. The softmax layer with each of the nodes corresponds to the class labels in the training set.

### 4.3. Aspect Feature Location Model

Feature extraction model captures the long-term dependence of the context and also generates the interactive semantic information between the aspect word and the context. On this basis, in order to further highlight the importance of different aspect words, we build an aspect feature positioning model based on the maximum pooling function (which is shown in Algorithm 1). This model divides the extracted aspect words and their context hiding features into multiple regions (i.e., line 3) and selects the maximum value in each region to represent the region (i.e., lines 4-5). In this way, the model can also locate core features and reduce the influence of noise words that are not related to aspect words, thereby improving the integrity of aspect word information. In other words, capturing aspect features and the different importance of aspect features can further improve the accuracy of aspect-level emotion classification. Specifically, combining the characteristics of the position and length of the aspect word, the feature location algorithm extracts the most important relevant information of the aspect word af from the context representation $e_c$. Moreover, max-pooling is applied to af to get the most important features AF.

$$\mathrm{AF} = \mathrm{Maxpooling}(\mathrm{af}, \mathrm{dim} = 0) \tag{17}$$

Afterwards, perform a dropout operation on AF and obtain the important features $h_{af}$of the aspect word in the context representation.

**ALGORITHM 1: ASPECT FEATURE LOCATION ALGORITHM**

**REQUIRE:** the context representation $e_c$, the position $i$ of aspect words in a sentence, the length $al$ of aspect words; the batch size $bs$.

1. Repeat
2. foreach $e_c \in bs$ do
3. Select lines ($i$ + 1 and $i$ + 1+ $al$) of $e_c$to obtain aspect feature $af$
4. Calculate the most important features $AF$ according to equation (17)
5. Apply the dropout operation to all the important features to get the $h_{af}$
6. end for
7. Untilmetrics to tend to be stable

### 4.4. Sentiment Predictor

One of the cores of SMMHA-BERTis to utilize multiple self-attention mechanisms to obtain multiangle text hidden expression features, and after processing byaspect feature positioning models, we have obtained a wealthof aspect-level auxiliary features and contextual interactionof aspect word information. In order to effectively utilizethese complete and rich features, this paper uses fully connectionlayer to fuse and preprocess the features in advanceand uses the softmax function to map the features to the[0,1] interval, so as to achieve effective mapping from featuresto sentiment classification. Specifically, concatenate the$h_{cm}$, $h_{am}$, and $h_{af}$ first to obtain the comprehensive representationr, which is shown as follows,

$$\mathrm{r} = [\mathrm{h_{cm}}; \mathrm{h_{am}}; \mathrm{h_{af}}] \tag{18}$$

Subsequently, a linear function is used to preprocess the data of r, as shown in the following,

$$\mathrm{x} = \mathrm{W_u r} + \mathrm{b_u} \tag{19}$$

where $W_u$represents the weight matrix, and $b_u$denotes the bias.At last, softmax function is used to compute the probability Pr that the sentiment polarity of the aspect word a in a sentence is p, as shown in the following,

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

$$\Pr(a = p) = \frac{\exp(x_p)}{\sum_{i=1}^{C} \exp(x_i)} \tag{20}$$

where C denotes the number of categories of sentiment polarity. On the whole, the SMMHA-BERT approach is an end-to-end computing process. Moreover, in order to optimize the parameters of the proposed approach, so as to minimize the loss between the predicted sentiment polarity y and the correct sentiment polarity $\hat{y}$, cross-entropy with $L_2$ regularization is used as the loss function to train proposed model, which is defined as follows,

$$\text{Loss} = -\sum_{j}\sum_{i} y_i^j \log \hat{y}_i^j + \lambda\|\theta\|^2 \tag{21}$$

where D means all training data, and j and i denote the index of a training data sample and a sentiment class, respectively. $\lambda$represents the factor for $L_2$ regularization, and $\theta$denotes the parameter set of the model.

## 5. EXPERIMENTAL EVALUATION

For the sake of evaluating the rationality and effectiveness of the SMMHA-BERT approach, this section describes the details of experiment settings and designs comparative experiments. Moreover, also analyze the experimental results.

### 5.1. Dataset

For experiments, conduct experiments on Amazon customer review dataset. This dataset consists of over 34,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database. The dataset includes basic product information, rating, review text, and more for each product. It has been widely used in aspect-based sentiment analysistasks.

### 5.2. Baselines and Evaluation Metrics

In order to verify the effectiveness of model, compare the proposed approach with many popular aspect-based sentiment analysis models, as listed in the following:

MemNet [27] is a data-driven model that utilizes multiple attention-based computational layers to capture the importance of each context word.

AEN-BERT [28] is a model based on attention mechanism and BERT and shows good performance in aspect-based sentiment analysis tasks.

ALM-BERT [29] is a model and it is performed based on the effective aspect-level sentiment analysis approach by constructing an aspect feature location model.

For the sake of measuring the performance of the model fairly, extend the MemNet, AEN, and ALM models by replacing the embedding layer of these models with BERT.

In addition, in order to objectively evaluate the performance of the proposed model, similar to existing aspect level sentiment analysis tasks, metrics like precision, recall, macro-F1 score (F1), and accuracy as evaluation indicators.

Macro-F1 is used to truly reflect the performance of the model, which is the weighted average of precision and recall. The macro-F1 is calculated according to the following equation (22),

$$\text{Pre}_{C_i} = \frac{T_{c_i}}{T_{c_i} + FP_{C_i}} \tag{22}$$

$$\text{Re}_{C_i} = \frac{T_{c_i}}{T_{c_i} + FN_{C_i}} \tag{23}$$

$$macro - F1 = \frac{1}{C}\left(\sum_{i=1}^{C}\left\{\frac{(2 * Pre_{C_i} * Re_{C_i})}{(Pre_{C_i} + Re_{C_i})}\right\}\right) \tag{24}$$

where T represents the number of samples correctly classified as sentiment polarity i, FP denotes the number of samples incorrectly classified as sentiment polarity i, FN represents the number of samples whose sentiment polarity i is misclassified as other sentiment polarities, C denotes the number of categories of sentiment polarity, $Pre_{C_i}$indicates the precision of sentiment polarity i, and $Re_{C_i}$denotes the recall of sentiment polarity i.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

Accuracy (Acc) is calculated according to the following equation (25),

$$Acc=SC/N \quad (25)$$

As shown in Table 1, 2, and 3, the results of sentiment classification methods. In the table 1 shows the performance comparison of classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), XGboost, and SMMHA. Table 2 and 3 easily observe from theexperimental results that the accuracy, macro-F1, precision, and recall of SMMHAare significantly higher than those of MemNet, AEN, and ALM based models.

**TABLE 1. RESULTS COMPARISON OF SENTIMENT ANALYSIS METHODS**

| Classifiers | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **LR** | 0.7780 | 0.7810 | 0.7795 | 0.795 |
| **SVM** | 0.8721 | 0.8678 | 0.8699 | 0.878 |
| **XGBoost** | 0.8872 | 0.8724 | 0.8798 | 0.885 |
| **SMMHA** | 0.9314 | 0.9242 | 0.9278 | 0.927 |

**TABLE 2. RESULTS COMPARISON OF ASPECT-BASED SENTIMENT ANALYSIS METHODS**

| Classifiers | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **MemNet** | 0.8527 | 0.8618 | 0.8572 | 0.8611 |
| **AEN** | 0.8912 | 0.8821 | 0.8866 | 0.8827 |
| **ALM** | 0.9124 | 0.9052 | 0.9088 | 0.9071 |
| **SMMHA** | 0.9314 | 0.9242 | 0.9278 | 0.9270 |

**TABLE 3. RESULTS COMPARISON OF ASPECT-BASED SENTIMENT ANALYSIS METHODS BY BERT**

| Classifiers | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| **MemNet-BERT** | 0.8712 | 0.8665 | 0.86885 | 0.8871 |
| **AEN-BERT** | 0.9051 | 0.9106 | 0.90785 | 0.9167 |
| **ALM-BERT** | 0.9271 | 0.9215 | 0.9243 | 0.9324 |
| **SMMHA-BERT** | 0.9416 | 0.9358 | 0.9387 | 0.9514 |

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

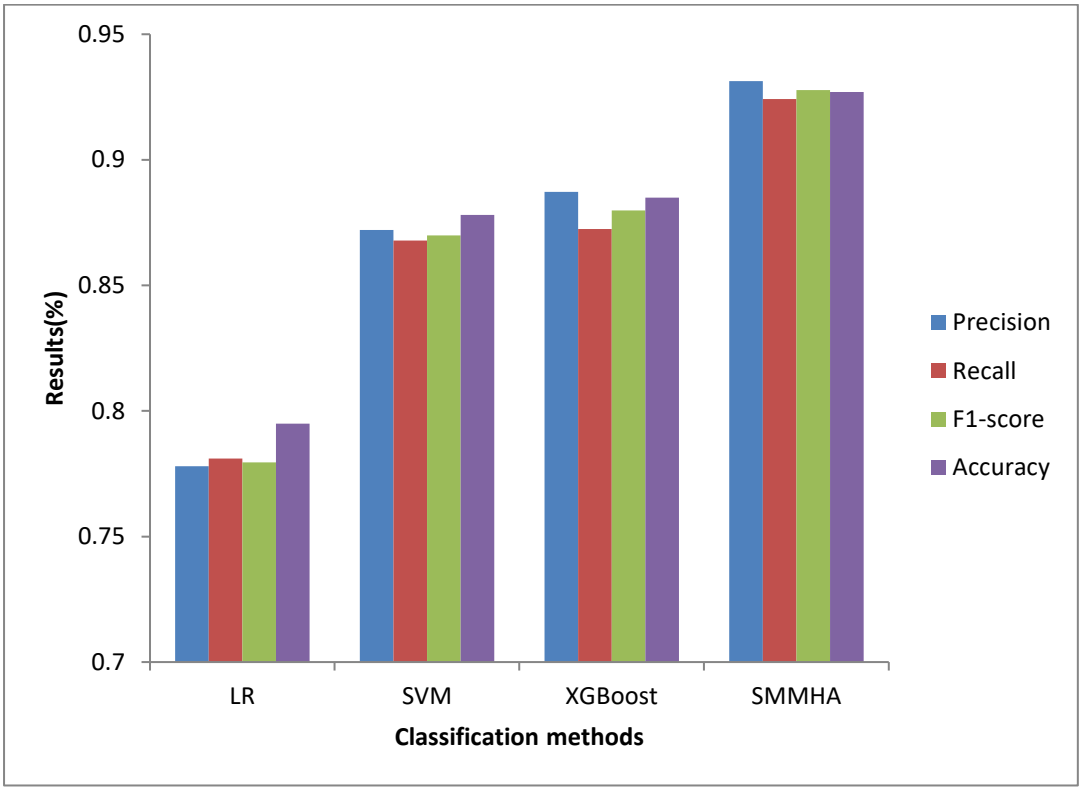
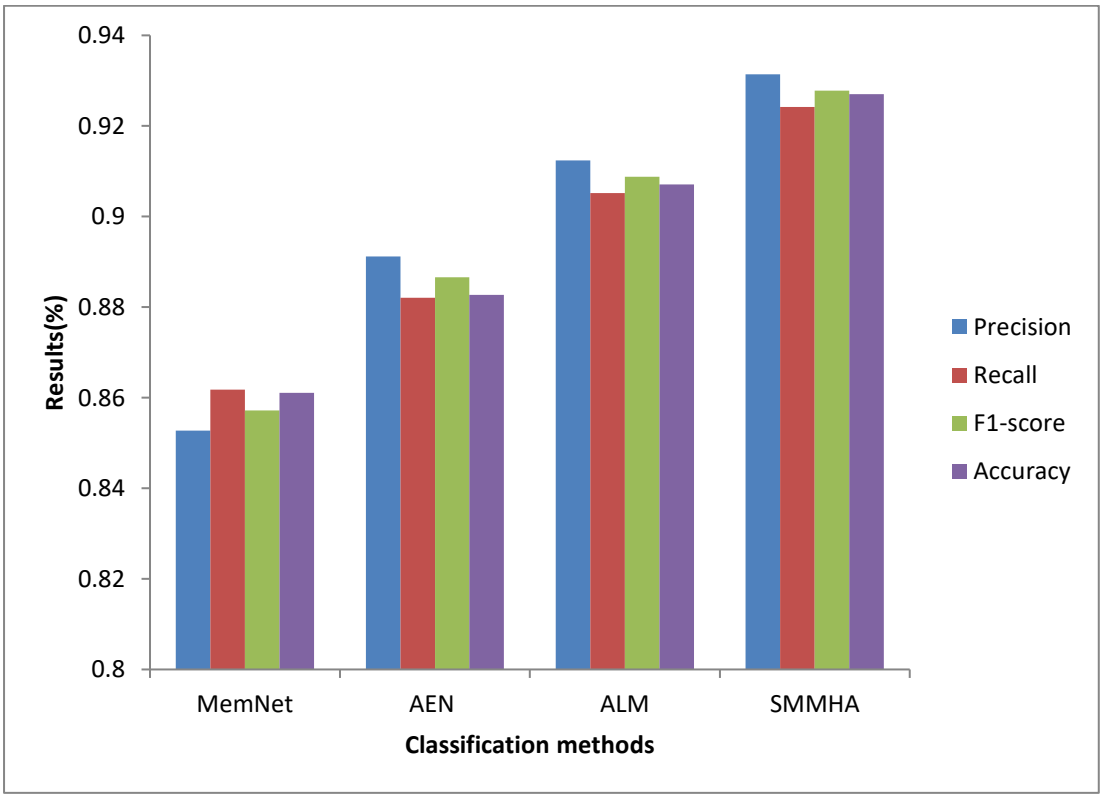



**FIGURE 4. EVALUATION METRICS FOR SENTIMENT ANALYSIS CLASSIFICATION METHODS**

As shown in Figures 4, the proposed classification approachobtains higher precision, recall, f1-score, and accuracy when compared to other classifiers onthe whole, which means that SMMHAcan simulatethe implicit relationship between contexts better thanexisting classifiers. In addition, compared with LR, asshown in Figure 4, the prediction results of SMMHAmodel will gives 16.46983%, 15.4944%, 15.9840%, and 14.23948% for precision, recall, F1-score, and accuracy, respectively.



**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

**FIGURE 5. EVALUATION METRICS FOR ASPECT-BASED SENTIMENT ANALYSIS METHODS**

Proposed aspect based sentiment analysis approachobtains higher precision, recall, f1-score, and accuracy when compared to other aspect based analysis methods. In addition, the proposed SMMHA model will giveaccuracy of 7.1089%, 4.77%, and 2.14671% for MemNet, AEN, and ALM respectively.

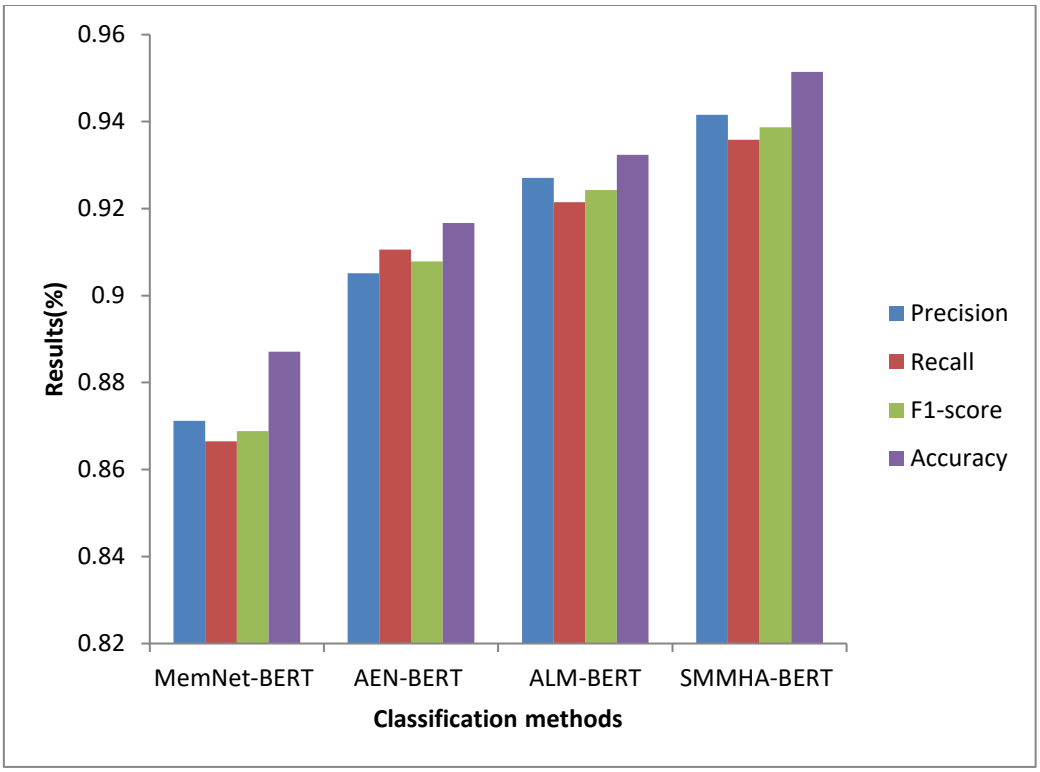



**FIGURE 6. EVALUATION METRICS FOR ASPECT-BASED SENTIMENT ANALYSIS METHODS BY BERT**

Figure 6 shows the performance comparison of aspect based sentiment analysis methods by BERT with respect to precision, recall, f1-score, and accuracy. Proposed aspect based sentiment analysis approach with BERT obtains higher precision, recall, f1-score, and accuracy when compared to other aspect based analysis methods. In addition, the proposed model will gives accuracy of 6.7584%, 3.647%, and 1.997% for MemNet, AEN, and ALM respectively.

## 6. CONCLUSION AND FUTURE WORK

In this paper, propose a method based on deep learning to identify the sentiment polarity of opinion words expressed on a specific aspect of a sentence. Bidirectional Encoder Representations from Transformers (BERT) and Serialized Multi-layer Multi-Head Attention (SMMHA) is introduced to feature extraction for a sentence. Transformer encoder based onBERT is proposed to capture the long-term dependencies of the contextand generate the interactive semantic information betweenaspect words and context.BERT-SMMHAalgorithm, aims at sentiment analysis of entity, aspect combinations, making the well-studied ASC task a special case of it.SMMHA is proposed in parallel to aggregate and propagate the information from one layer to the next in a serialized manner with stacked self-attentionmodules. In the proposed serialized attention mechanism, self attention is performed in a serialized manner, allowing the model to aggregate information with temporal context from deeper layers. Proposed approach propagates both contextual and dependency information from opinion words to aspect words, offering discriminative properties for supervision. Experimental results rankproposed approach as the new state-of-the-art in aspect-based sentiment classification.The results achieved good results with respect to precision, recall, F1-Score, accuracy which shows promise for deployment in an integrated ASA system.Future work focuses on finding the embedding conclusions of the words with semantic relationships.

**JOURNAL OF ALGEBRAIC STATISTICS**
Volume 13, No. 2, 2022, p. 1391 - 1406
https://publishoa.com
ISSN: 1309-3452

**REFERENCES**

1. Mowlaei M. E., M. Saniee Abadeh, and H. Keshavarz, "Aspect based sentiment analysis using adaptive aspect-based lexicons," Expert Systems with Applications, vol. 148, no. 113234, pp.1-13, 2020.

2. Cai Z. and Z. He, "Trading private range counting over big IoT data," 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp. 144–153, Dallas, TX, USA, 2019.

3. Lin Y., X. Wang, F. Hao et al., "Dynamic control of fraud information spreading in mobile social networks," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 6, pp. 3725–3738, 2021.

4. Yanglan O., B. Huang, and K. M. Carley, "Aspect level sentiment classification with attention-over-attention neural networks," in Social, Cultural, and Behavioral Modeling. SBPBRiMS 2018, R. Thomson, C. Dancy, A. Hyder, and H. Bisgin, Eds., vol. 10899 of Lecture Notes in Computer Science, pp. 197–206, Springer, Cham, 2018.

5. Ganesh, D, T P Kumar, and M S Kumar. "Optimised Levenshtein centroid cross-layer defence for multi-hop cognitive radio networks." IET Communications 15, no. 2 (2021): 245-256.

6. Davanam, Ganesh, T. Pavan Kumar, and M. Sunil Kumar. "Novel Defense Framework for Cross-layer Attacks in Cognitive Radio Networks." In International Conference on Intelligent and Smart Computing in Data Analytics, pp. 23-33. Springer, Singapore, 2021.

7. Balaji, K., P. Sai Kiran, and M. Sunil Kumar. "Resource Aware Virtual Machine Placement in IaaS Cloud using Bio-Inspired Firefly Algorithm." Journal of Green Engineering 10 (2020): 9315-9327.

8. Xu, H., Liu, B., Shu, L. and Yu, P.S., 2019. A failure of aspect sentiment classifiers and an adaptive re-weighting solution. arXiv preprint arXiv:1911.01460, pp.1-12.

9. Sangamithra, B., P. Neelima, and M. S Kumar. "A memetic algorithm for multi objective vehicle routing problem with time windows." In 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE), pp. 1-8. IEEE, 2017.

10. Cai Z. and Z. Xu, "A private and efficient mechanism for data uploading in smart cyberphysical systems," IEEE Transactions on Network Science and Engineering, vol. 7, no. 2, pp. 766–775, 2018.

11. Mowlaei, M.E., Abadeh, M.S. and Keshavarz, H., 2020. Aspect-based sentiment analysis using adaptive aspect-based lexicons. Expert Systems with Applications, 148, pp.1-13.

12. Zhou, J., Huang, J.X., Chen, Q., Hu, Q.V., Wang, T. and He, L., 2019. Deep learning for aspect-level sentiment classification: survey, vision, and challenges. IEEE access, 7, pp.78454-78483.

13. Liu, B., and Lane, I. (2016). "Attention-based recurrent neural network models for joint intent detection and slot filling," in Proceedings of the 17th Conference on International Speech Communication Association (San Francisco, CA), pp.685–689.

14. Lai, S., Xu, L., Liu, K. and Zhao, J., 2015, Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence, pp.2267–2273.

15. Gan, C., Wang, L., Zhang, Z., and Wang, Z. (2020). Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis. Knowledge Based Syst. 188, pp.1–10.

16. Tang, D., Qin, B., Feng, X., and Liu, T. (2016a). "Effective LSTMs for targetdependent sentiment classification," in Proceedings of the 26th Conference on International Conference on Computational Linguistics (ICCL) (Osaka), 3298–3307.

17. Dong, D., Wu, H., He, W., Yu, D. and Wang, H., 2015, Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1723-1732.

18. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, pp.1-16.

19. Li, Z., Wei, Y., Zhang, Y., Zhang, X. and Li, X., 2019, Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 4253-4260).

20. Zeng, J., Ma, X., and Zhou, K. (2019). Enhancing attention-based LSTM with position context for aspect-level sentiment classification. IEEE Access 7, 20462–20471.

21. Tan, X., Cai, Y., Xu, J., Leung, H.F., Chen, W. and Li, Q., 2020. Improving aspect-based sentiment analysis via aligning aspect embedding. Neurocomputing, 383, pp.336-347.

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

23. Zhang, Q., Lu, R., Wang, Q., Zhu, Z. and Liu, P., 2019. Interactive multi-head attention networks for aspect-level sentiment classification. IEEE Access, 7, pp.160017-160028.

24. Zhou, Z. and Wang, Q., 2019. R-transformer network based on position and self-attention mechanism for aspect-level sentiment classification. IEEE Access, 7, pp.127754-127764.

25. Zhou, Z., 2021. Filter gate network based on multi-head attention for aspect-level sentiment classification. Neurocomputing, 441, pp.214-225.

26. Leng, X.-L., Miao, X.-A., and Liu, T. (2021). Using recurrent neural network structure with enhanced multi-head self-attention for sentiment analysis. Multimedia Tools Appl. 80, pp.12581–12600.

27. Tang, D., Qin, B. and Liu, T., 2016. Aspect level sentiment classification with deep memory network. arXiv preprint arXiv:1605.08900, pp.1-11.

28. Song, Y., Wang, J., Jiang, T., Liu, Z. and Rao, Y., 2019, Targeted sentiment classification with attentional encoder network. In International Conference on Artificial Neural Networks (pp. 93-103). Springer, Cham.

29. Pang, G., Lu, K., Zhu, X., He, J., Mo, Z., Peng, Z. and Pu, B., 2021. Aspect-Level Sentiment Analysis Approach via BERT and Aspect Feature Location Model. Wireless Communications and Mobile Computing, vol.2021, no. 5534615, pp.1-13.

30. Yu L.-C., J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 671–681, 2018.

31. Rida-E-Fatima S., A. Javed, A. Banjar et al., "A multi-layer dual attention deep learning model with refined word embeddings for aspect-based sentiment analysis," IEEE Access, vol. 7, pp. 114795–114807, 2019.

32. Vaswani A., N. Shazeer, N. Parmar et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 5998–6008, Long Beach, CA, USA, 2017.

33. Huang, T., Deng, Z.H., Shen, G. and Chen, X., 2020. A window-based self-attention approach for sentence encoding. Neurocomputing, 375, pp.25-31.

34. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L. and Liu, T., 2020, On layer normalization in the transformer architecture. In International Conference on Machine Learning ,pp. 10524-10533.