IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 2, April, 2021, pp. 167-176
Journal homepage: http://iieta.org/journals/ria

A Multilabel Classifier for Text Classification and Enhanced BERT System

Bhavana R. Bhamare[1,2*], Jeyanthi Prabhu[3]

[1] Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, India
[2] Department of Information Technology, International Institute of Information Technology (I[2]IT), Pune 411057, India
[3] Department of Information Technology, Sathyabama Institute of Science and Technology, Chennai 600119, India

Corresponding Author Email: bhavanak@isquareit.edu.in

https://doi.org/10.18280/ria.350209

Received: 30 November 2020
Accepted: 25 March 2021

Keywords:
multilabel classifier, Bidirectional Encode Representation from Transformers (BERT), Binary Relevance (BR), Classifier Chains (CC), Label Powerset (LP), Aspect Based Sentiment Analysis (ABSA)

ABSTRACT

Now-a-day, a vast variety of reviews are published on the web. As a result, an automated system to analyze and extract knowledge from such textual data is needed. Sentiment analysis is a well-known sub-area in Natural Language Processing (NLP). In earlier research, sentiments were determined without considering the aspects specified in a review instance. Aspect-based sentiment analysis (ABSA) has caught the attention of researchers. Many existing systems consider ABSA as a single label classification problem. This drawback is handled in this study by proposing three approaches that use multilabel classifiers for classification. In the first approach, the performance of a model with hybrid features is analyzed using the multilabel classifier. The hybrid feature set includes word dependency rule-based features and unigram features selected using the proposed two-phase weighted correlation feature selection (WCFS) approach. In the second and third approaches Bidirectional Encoder Representation from Transformers (BERT) language model is used. In the second approach, a BERT system is enhanced by applying max pooling on target terms which specify an aspect of a review instance and a multibit label is given as input to the BERT system. In the third approach, the basic BERT system is used for word embedding only and classification is done using multilabel classifiers. In all approaches, the label used for all training instances specifies aspects with its sentiments. The experimentation shows that the results gained using the system proposed in the first approach are comparable to the results gained using the BERT system. The experimental results depict that the Enhanced BERT system gives better results compared to the existing systems.

1. INTRODUCTION

A substantial amount of digital data is available on the web in the form of text, audio, and video. Due to the availability of such a large amount of multimedia data, machine learning based text analysis has caught the attention of researchers. There are various applications of text analysis like text classification, review analysis, automated question-answering, product recommendation, etc. Sentiment analysis is one of the important tasks in text analysis. Sentiment analysis can be applied for the analysis of news, social media contents, or reviews of products, movies, etc. The generic opinion gives incomplete insights about the product reviews. Thus, the research focus is on ABSA. In ABSA, the tasks included are aspect extraction, sentiment prediction, and sentiment prediction for the extracted aspects. In this paper, the focus is on aspect-based sentiment prediction. Many existing works consider ABSA as an individual task like aspect extraction only or sentiment prediction only. This study aims to consider it as an end-to-end problem, i.e., determine sentiment for the aspects specified in the review. Many existing strategies solve the sentiment analysis problems without considering the context and use single-label classifiers like binary or multiclass classifiers for classification. A review instance may contain multiple aspects and opinions about those aspects, so it is a multiclass classification problem. Below is a sample review instance for mobile phones: "the camera quality is awesome but unsatisfied with audio quality".

This sentence has two aspects, camera quality and audio quality. In it, the opinion for camera quality is positive while it is negative for audio quality. Single label classifiers fail to make correct predictions for such instances.

Along with classifiers, the feature set used for the class prediction task determines the accuracy of classification. NLP plays a vital role in the ABSA classification problem. In a machine learning based solution for ABSA, the feature extraction and selection algorithms decide the performance of classification. A relevant feature set leads to improve the accuracy of classification. Recently, the state-of-the-art pre-trained NLP models are used like Generative Pre-trained Transformer (OpenAI GPT), ELMo, and BERT. OpenAI GPT is a left-to-right model, ELMo is the concatenation of left to right and right to left model, and BERT considers the context of both left and right side tokens. As these models are pre-trained, they are least dependent on the labeled data. The language models like BERT perform word embedding as well as classification, so it doesn't require an extra feature selection phase. It considers the relations among the tokens in a sentence and computes the word embedding that is further used for classification. The following section briefs about the multilabel classification and BERT system.

1.1 Multi-label classification

The classification problem can be considered as a single label or multilabel classification problem. The single label classifier can be a binary classifier or a multiclass classifier. As the textual data may contain more than one class label, in such cases, single-label classifiers do not give complete classification results. Example- A mobile review like "Picture quality is good but the sound is not clear". This sentence contains two aspects, picture quality and audio, i.e., it has two class labels. To classify textual data with multiple labels, the multilabel classifier can be used as it predicts multiple labels in a given review instance. The multilabel classification has attracted attention recently. The multilabel classification techniques can be categorized as problem transformation strategy, algorithm adaptation, and ensemble strategies.

Problem transformation method

Under problem transformation, the possible strategies are BR, CC, and LP.

Binary Relevance (BR)

It is an ensemble of binary classifiers. If a dataset has l labels, then the BR approach divides the dataset into l subsets and one binary classifier works for one subset. The union of all classifiers output is declared as a class label. This approach doesn't consider class label dependencies.

Classifier Chains (CC)

In this method, the number of classifiers required is equal to the number of labels in a dataset. All classifiers work in a sequence where each classifier considers the predictions of its previous classifier too. It considers the label dependencies unlike, the BR approach.

Label Powerset (LP)

In a dataset, if there are l labels, then the label powerset approach requires 2^l classifiers. It considers all possible combinations of labels. The label powerset for a dataset with 4 labels is ({0000}, {0001}, {0010}, {0011}, {0100}, {0101}, {0110}, {0111}, {1000}, {1001}, {1010}, {1011}, {1100}, {1101}, {1110}, {1111}).

Adapted algorithm

As the name indicates, it doesn't transform the problem into subsets, instead the algorithm is adapted to work on the dataset for multilabel classification. MLkNN is a multilabel kNN (k-nearest neighbors) adapted algorithm.

Ensemble approach

A multilabel ensemble classifier is created by a set of multiclass classifiers or a set of multilabel classifiers. A voting method is used to select the labels in it. Random k label set (RAKEL) is an example of a multilabel ensemble classifier. Multiple LP classifiers are used in it. Each classifier is trained on a subset of actual dataset labels. The class label is decided by a voting method.

1.2 Bidirectional Encoder Representations from Transformers (BERT)

It is an NLP model and is available in versions like $BERT_{BASE}$ and $BERT_{LARGE}$.

$BERT_{BASE}$: 12 transformer blocks, 110 million parameters, and 12 attention heads.

$BERT_{LARGE}$: 24 transformer blocks, 340 million parameters, and 16 attention heads.

BERT is an encoder stack of transformer architecture. Pre-trained BERT models can be used for NLP tasks. As these are pre-trained models, it doesn't require a large amount of labeled data to understand the context and can be used for many text classification tasks without any changes in the architecture. It considers the context of both left as well as right side tokens. BERT models can be used for various applications like sentiment analysis, next sentence prediction. The model considers [CLS] as a first input followed by the tokens in a sentence and [SEP] as a separator in sentences. The input to the BERT is a summation of token embedding, segment embedding, and position embedding. Position embedding captures the information related to the position of a token in a sentence, segment embedding is useful for question answering applications, and token embedding uses WordPiece vocabulary. This summative form of embedding contains useful information. This embedding is the input to the fully connected network which further does the classification. Figure 1 shows the word embedding layer in a BERT system.

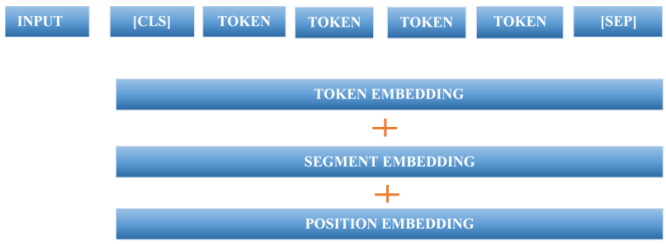

Figure 1. Word embedding in a BERT system

This study uses two different methodologies to accomplish the ABSA task. In the first methodology, a feature set dependent algorithm is used and in the second, the BERT system which is independent of the feature selection technique is used. The ABSA system proposed here is end-to-end ABSA, which determines sentiment for the extracted aspects.

The objectives of this study are:

1. To perform end-to-end ABSA using the hybrid feature set that combines word dependency relation based features and unigram features selected using the proposed WCFS algorithm. Furthermore, to test the performance of this system on the multilabel classifier.
2. To perform end-to-end ABSA using an enhanced BERT system in which BERT is used for embedding and classification.
3. To analyze the performance of an end-to-end ABSA model using the basic BERT system (for embedding) and multilabel classifiers.

This paper is organized like: Section 1 is an introduction, Section 2 is related work, Section 3 describes the proposed system, results and discussion are presented in Section 4, and concluding remarks are given in Section 5.

2. RELATED WORK

Related work on ABSA is discussed in this section. This paper focuses on an end-to-end ABSA, so to evaluate the research in this area is the aim of this section. This section concentrates on systems that use the machine learning method, CNN and BERT to solve the ABSA problem.

Deng et al. [1] suggested some methods for overcoming challenges related to domain heterogeneity by considering content domain and language domain. The method of lexicon expansion was used to improve the classification of sentiment by analyzing domain data. With it, as seed and baseline, two large unannotated developing corpora's and five existing sentiment lexicons were used. The results show that the expanded lexicon greatly improved the performance of the sentiment classification compared with the seed lexicon.

Deep neural network [2-5] is used. Specifically, Lee et al., and Tao et al. [2, 3] used supervised learning with deep learning to identify sentiments, aspects, and keywords. Lee et al. [2] suggested a method to classify keywords that distinguish between positive and negative phrases using a weakly supervised method of learning based on a convolutional neural network (CNN). Each word is represented as a continuous value vector in the model and each phrase is represented as a matrix whose rows correspond to the word vector used in the phrase. Using these sentence matrices as inputs and the sentiment labels as the output, the CNN model is trained. This proposed classification and localization model based on the class activation map (CAM^2) uses zero paddings compared to the previous CNN-based text classification model to help CNN recognize every word equally regardless of its place in the sentence. Tao et al. [3] presented an aspect-based sentiment dynamic of online reviews by proposing a semi supervised, deep learning facilitated analytical pipeline. This method examines deep learning techniques for text representation and classification. Additionally, building on previous studies that address aspect extraction and sentiment identification in isolation, they address both aspect and sentiment analysis simultaneously. Rida-E-Fatima et al. [4] proposed a cascaded feature selection system and classifier ensemble using particle swarm optimization (PSO) for ABSA. They used features that are described on the basis of characteristics of various classifiers and domains. Three classifiers, namely Maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM) were used. Yu et al. [5] presented a study that adopts and refines the existing context-based word embedding which results in words with similar vector representation with a much improved refinement model. The idea of this model is to improve each word vector such that it can be closer to both semantically and sentimentally similar words in the lexicon. In this model, the words which are similar are kept as neighbors with higher rank and words which are dissimilar are given lower ranks. Benefit of this system is that it is applicable to any pre-trained word embedding.

CNN is used for ABSA [6-9]. Specifically, the researches [6-8] used the long short-term memory (LSTM) technique for ABSA. Tang et al. [6] showed a comparative study for sentimental sentence classification. The comparisons were made with LSTM, Target Dependent LSTM, and Target Connection LSTM. All models were trained in a supervised learning framework. The Target connection LSTM has proven to be more effective than several other methods. Meng et al. [7] presented ABSA comprising of 2 subtasks: description of aspect identification and sentiment prediction. Therefore, to learn the relationship between aspect and sentiment, a new model, Feature Enhanced Attention CNN-BiLSTM (FEA-NN) is used. The technique involves word embedding, called Improved Word Vector (IWV). Liu and Guo [8] highlighted problems such as high dimensionality and sparsity of text data in text classification. In this strategy, BiLSTM is utilized to get the previous and succession context representations. Ishaq et al. [9] has presented an effective method to analyze the sentiments. It operates by combining three distinct operations like semantic mining, features transformation of extracted corpus using word2vec, and CNN implementation for mining opinions. To extract opinions, CNN was utilized. The CNN parameters were tuned using a genetic algorithm.

For ABSA, Akhtar et al. [10] presents a cascaded architecture of feature selection and classifier ensemble using PSO. Authors designed a PSO-based ensemble and cascaded it after the feature selection module. Pham and Le [11] proposed a multilayer architecture for the representation of customer reviews. The representation learning techniques including word embedding and compositional vector were used. These representations are further integrated into a neural network and a backpropagation algorithm was used for training a model for aspect rating prediction as well as generating aspect weights. Experimental results have shown that the proposed model outperforms the other popular methods.

Liu and Chen [12] presented a multi-label classification-based approach for sentiment analysis. The proposed prototype has three main components: text segmentation, feature extraction, and multi-label classification. The features used in this paper included raw segmented words, sentiment features based on three different sentiment dictionaries HowNet Dictionary (HD), National Taiwan University Sentiment Dictionary (NTUSD), Dalian University of Technology Sentiment Dictionary (DUTSD), and the bag of words. DUTSD has the best performance among the three separate dictionaries of sentiment. BERT representation technique is used for ABSA [13, 14]. Many previous methodologies treated labels as symbols without semantics and ignored the relation among labels, which caused information loss. This problem is handled by Cai et al. [13]. In this approach, the hybrid BERT model incorporates label semantics via ajustive attention, which searches and identifies semantic dependencies of label space and text space simultaneously. BERT is used in ABSA and it needs input in a word sequence form which does not provide extra context information. Li et al. [14] suggested a GBCN procedure that uses the gating components with context-aware aspect embedding to control and upgrade BERT presentation for ABSA.

Kang and Zhou [15] proposed unsupervised rule-based techniques (RubE), which extracted objective and subjective characteristics from reviews. In this, the authors detected objective features by integrating review-specific patterns and relations. Further they extracted the subjective features by ranging double propagation with indirect dependency and comparative construction. Findings indicate that RubE is much more advanced in the extraction technologies of product characteristics. Liu et al. [16] proposed a methodology AS-Reasoner to mitigate issues related to precise sentiment expression. AS-Reasoner appoints significance degrees to various words in a sentence to catch important sentiment expressions of a particular aspect. Jia et al. [17] proposed position-aware hierarchical gated deep memory network. This system embeds position information as a feature in the sentence representation. Afzaal et al. [18] suggested a multilabel classifier for tourist reviews. In this work, coreferential aspects are identified using co-occurrence information of aspects and sentiments. In addition to coreferential aspects implicit aspects are extracted and this

system is tested using multilabel classifier. Many authors used BERT system for text clasiification [19].

Sun et al. [20] proposed the construction of an auxiliary sentence to transform ABSA to a sentence-pair classification task. The BERT model pre-trained on the sentence pair classification task is fine-tuned and new state-of-the-art results are obtained.

For sentiment classification from text, Sultana and Islam [21] used seven separate and popular supervised machine learning algorithms. On them, an ensemble technique (boosting, stacking, bagging) was used. Finally, the proposed ensemble solution was compared to the individual prediction accuracy of these classifiers. The final sentiment class estimation was determined using the Majority voting (stacking) process, which takes into account the classification outcomes of these classifiers. In order to increase the performance of these learners, bagging and boosting methods were added to SVM and Stochastic Gradient Descent (SGD) classifiers. Mohammadi & Shaverizade [22] proposed aspect-based sentiment analysis using deep ensemble learning. Four deep learning models, namely LSTM, CNN, BiLSTM, and Gated Recurrent Unit (GRU) were used. The outputs of these models were then collectively used in a stacking ensemble technique, with logistic regression serving as the meta-learner. When compared to basic deep learning approaches, the effects of applying the proposed approach to actual datasets show that the method improved the precision of aspect-based prediction by 5% to 20%. Current aspect based approaches are not able to adapt to the general lexicons and hence yield poor results. Mowlaei et al. [23] focused on the development of specialized methods for generation of dynamic lexicons. These generated lexicons are then fused with commonly used static lexicons (Bing Liu's Opinion lexicon, MPQA Subjectivity lexicon, and SentiWordNet) to compensate for the weaknesses of each type of lexicon with the other and achieve the best performance. A Multi Attention Network (MAN) approach for ABSA is proposed by Xu et al. [24]. This model uses an inter-level and intra-level attention mechanism. In the inter-level attention mechanism, a transformer encoder is employed which encodes the input sentence in parallel, using CNN, hence covering the drawback of the sequence model and reducing training time. It also preserves long-distance sentiment relations. In the intra-level attention mechanism, the global and local attention module is used to capture differently embedded information between aspect and context. Global attention captures coarse-grained i.e. the whole interaction, whereas local attention captures fine-grained i.e. the word-level interaction between aspect and context words.

Many machine learning based methodologies discussed in this section used a single label classifier for classification. Moreover, it shows that multi-label classifier performance can be boosted using the relevant feature set. As a result, the proposed system fixes these drawbacks by conducting ABSA with a hybrid model that uses a feature selection approach and a multilabel classifier for classification. Additionally, this study proposes an enhanced BERT model for ABSA. For all experiments conducted here, a multi-bit label is used during training which details about aspects and its corresponding sentiments.

3. PROPOSED SYSTEM

This study proposes three methodologies for an end-to-end ABSA task. The objectives of this study are: 1. (Approach 1)- To propose a hybrid model WCFS for ABSA using multilabel classifiers. 2. (Approach 2)- To analyze the performance of ABSA with the proposed enhanced BERT system. 3. (Approach 3)- To analyze ABSA performance using basic BERT (for word embedding) and multilabel classifiers (for classification). Figure 2 shows the proposed system architecture for approach 1.

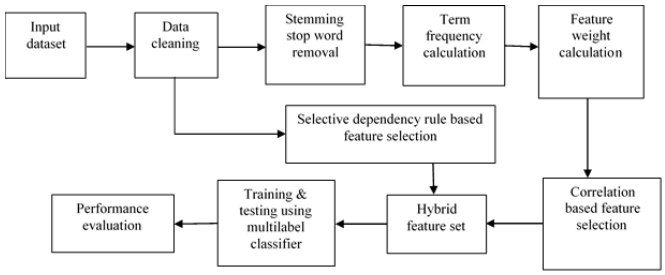

Figure 2. Architecture of the proposed hybrid model with WCFS approach for ABSA

The datasets used for this work are SemEval 2014 restaurant review dataset [25] with 3044 review instances and SemEval 2015 laptop review dataset [26] with 1399 review instances. The main focus of this study is to perform sentiment prediction concerning the aspect specified in the review instance. In the restaurant review dataset, there are five aspect categories for which the sentiments are specified. The aspect categories in the restaurant review dataset are food, price, service, ambience, and miscellaneous. In this dataset, 11% of review instances have 2 aspect categories, 2% instances have more than 2 aspect categories, and 87% instances have 1 aspect category. Figure 3 shows the snippet of the review instance from the restaurant review dataset. The sentiments specified in this dataset are positive, negative, conflict, and neutral. The sentiments specified in the laptop review dataset are positive, negative, and neutral. For the laptop review dataset, 9 aspect categories are considered like general, operation_performance, design_features, usability, portability, price, quality, miscellaneous, and connectivity.

```
<sentence id="425">
    <text>The price is reasonable although the service is poor.</text>
  - <aspectTerms>
        <aspectTerm to="9" from="4" polarity="positive" term="price"/>
        <aspectTerm to="44" from="37" polarity="negative" term="service"/>
    </aspectTerms>
  - <aspectCategories>
        <aspectCategory polarity="negative" category="service"/>
        <aspectCategory polarity="positive" category="price"/>
    </aspectCategories>
</sentence>
```

Figure 3. Restaurant review instance snippet

The methodologies attaining objectives are explained below:

3.1 A hybrid model with the proposed WCFS approach for ABSA using multilabel classifier

This is a machine learning based approach for ABSA. The accuracy of classifiers depends on the feature set used. In this approach, a feature selection strategy is proposed whose performance is analyzed using multilabel classifiers. This methodology gives a hybrid feature set that combines grammatical rule-based features and unigrams. The steps of this approach are explained below:

Preprocessing

Data cleaning: It includes removing punctuation and replacing abbreviations like don't with do not, can't with cannot, etc.

Dependency rule-based features

After data cleaning, a Stanford dependency parser is applied to extract word dependency rule-based features [27] from each sentence. The output of the parser is two word features that are contextually related and the name of the relationship between them. Below is an example of the output of the dependency parser for a laptop review instance.

"Picture quality is good but processing is slow".

The possible relationships that exist between two words in the given sentence are listed below.

compound(quality-2, picture-1)
nsubj(good-4, quality-2)
cop(good-4, is-3)
root(ROOT-0, good-4)
cc(slow-8, but-5)
nsubj(slow-8, processing-6)
cop(slow-8, is-7)
conj(good-4, slow-8)

Here, instead of all relationships, only selective relationships that cover nouns, adjectives, and adverbs are considered to extract features. The relationships considered in this study are: adjectival complement (acomp), adverbial clause modifier (advcl), adverb modifier (advmod), agent, adjectival modifier (amod), conjunct (conj), copula (cop), direct object (dobj), negation modifier, noun compound modifier (nn), nominal subject (nsubj), passive nominal subject (nsubjpass), relative clause modifier (rcmod), open clausal complement (xcomp), and nominal modifier (nmod). To select rule-based features, the frequency count is not considered. All features extracted after applying selective grammatical rules are considered for further processing.

WCFS approach to select unigrams

Algorithm 1 is used for unigram feature selection in each aspect category. It is a two-phase process of feature selection.

Phase 1:

1. Stemming is applied for all review instances and stop words are removed.
2. The frequency count of each unigram is calculated. It is calculated across the dataset and in the corresponding aspect category. Unigrams with a frequency count greater than 3 (across the dataset) are selected for further processing. A vector of unique unigrams is created for each aspect category with its frequency count across the dataset and in the corresponding aspect category.
3. In each aspect category, weight W_{fk} for each unigram is calculated using Eq. (1)

$$W_{fk} = \frac{Frequency\ count\ of\ unigram\ f\ in\ aspect\ category\ k}{Frequency\ count\ of\ unigram\ f\ in\ dataset} \quad (1)$$

4. In each aspect category, the unigram features are arranged in decreasing order of their weight.

Phase 2:

In this phase, unigrams are selected by applying correlation. In this approach, frequency count is used for feature extraction, while feature weight and correlation are used for feature selection. Feature weight helps to select relevant unigram features and correlation avoids redundancy among them. The following steps are applied for each aspect category to select features.

5. The feature having the maximum weight is added to the list of selected features.
6. For each unselected feature, its correlation is computed with the selected features. The computed correlation of each unselected feature is subtracted from its weight W_{fk} in the corresponding aspect category and W_{fk}_new calculated.
7. The features are arranged in decreasing order of W_{fk}_new.
8. The feature having the maximum value for W_{fk}_new is added in the list of selected features as its weight is more, which shows high relevancy and less correlation that helps to avoid redundancy.
9. Steps 6 to 8 are repeated until the required number of features are selected.

The proportion of features to select from each category is calculated using Eqns. (2), (3), and (4). This proportion is decided according to the number of instances of each category available in the dataset. Phase 2 is applied for each aspect category to select features.

$$total\_no\_of\_features = \sum_{i=j}^{k} total\ number\ of\ features\ in\ aspect\ category\ j \quad (2)$$

$$get\_Percent = \frac{total\_no\_of\_features\_to\_select}{total\_no\_of\_features} \quad (3)$$

$$select\_number\_of\_features_k = total\_no\_of\_features_{\ k} \times get\_Percent \quad (4)$$

Eq. (2) indicates the number of features extracted, which is the sum of features in each aspect category.

Eq. (3) depicts the fraction of features to select from the extracted features.

Eq. (4) shows the fraction of features to select from the aspect category k.

As the datasets used in this study do not have even distribution across all classes, so features are selected from each aspect category rather than select randomly from the dataset.

A hybrid feature set is generated which contains unigram features obtained after step 9 and dependency rule-based features. Hybrid features are used to train a multilabel classifier. The machine learning classifiers used for multilabel classification are BR, CC, and LP along with Naïve Bayes (NB), Support Vector Machine (SVM) baseline classifier. A cross-validation method is used during the training and testing phase. Each training review instance is labeled with a multibit label. The format of the multibit label is shown in Figure 4. Each bit in this label represents sentiment concerning the aspect category. The first four bits of the label show positive, negative, neutral, and conflict sentiments for food aspect category. The restaurant review dataset has 5 aspect categories and 4 sentiments (4-way), so the number of bits in the label are 20. If review instances with 3 sentiment classes (3-way) are considered for training and testing, then the number of bits in the label will be 15 and for binary sentiments it will be 10 bits.

Figure 4 shows the multibit label for the example review instance "The food is tasty but the surrounding is not pleasant". In this review, the sentiment for the food aspect category is positive and for the ambience it is negative. Therefore, the corresponding sentiment representing bits are set to 1 and the other bits are 0.

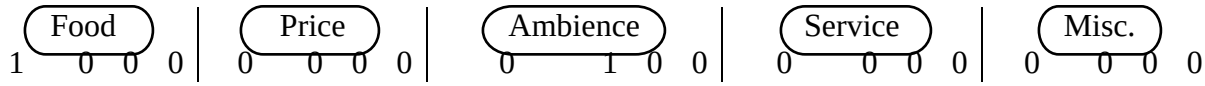

Figure 4. Multi bit label for multilabel classification

Table 1. Example of a multi bit label for review instances

Review	Label
the food is tasty but the surrounding is not pleasant	1000 0000 0100 0000 0000
awesome service with tasty food	1000 0000 0000 1000 0000
cheap restaurant with a fresh environment	0000 1000 1000 0000 0000

Table 1 shows the example review instances and corresponding multi-bit labels. This multibit label is used in all approaches proposed in this study.

Algorithm 1: WCFS approach for unigram feature selection

Input:

$F \leftarrow$ $\{number\ of\ features\ extracted\ from\ category\ k\}$

$count_k$ are the number of features to select from an aspect category k

Output:

$S \leftarrow \{selected\ features\ in\ an\ aspect\ category\ k\}$

Note: This algorithm is applied for each category k to select features

1. Begin Algorithm

$S \leftarrow \{\emptyset\}$ initially, no feature is selected from an aspect category.

$F = \{f1..n\}$ features extracted from aspect category k.

2. **Phase I**

Calculate weight wf_{jk} for each feature in F using Equation no. (1).

3. Sort features in decreasing order of weight.

4. $S \leftarrow S \cup f_j$

$where\ f_j\ is\ the\ feature\ with\ maximum\ wf_{jk}$

Remove f_j from F

5. **Phase II**

while $S.size() \neq count_k$

for each f_j in F

$Total\_Corr_j = 0$

for each f_i in S

calculate the correlation between features in S and F

Calculate $Correlation(f_j, f_i)$

$Total\_Corr_j = Total\_Corr_j + Correlation(f_j, f_i)$

end for

end for

for each f_j in F

$wf_{j\_new} = wf_{jk} - Total\_Corr_j$

$S \leftarrow S \cup f_j$

$where\ f_j\ is\ the\ feature\ with\ maximum\ wf_{j\_new}$

Remove f_j from F

end while

6. Return S, which contains selected features in aspect category k.

7. End algorithm

3.2 Extended BERT (for word embedding and classification) for ABSA

Figure 5 represents the proposed extended BERT system. Recently, BERT is used in many NLP applications. In approach 1, the feature selection strategy with multilabel classifiers is proposed. BERT model doesn't require feature selection. These are pre-trained models and can be used for any NLP application without modifications. In this work, an extended BERT system is proposed.

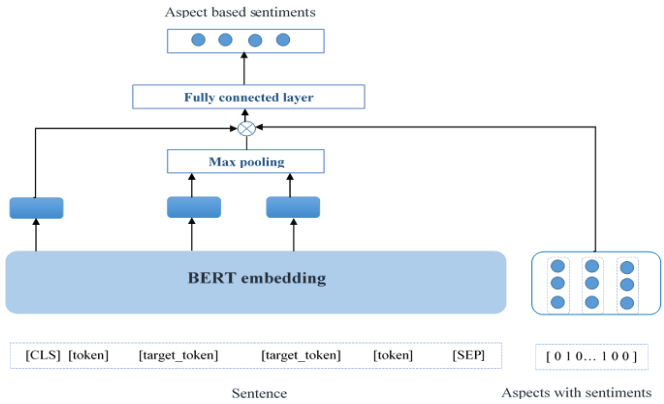

Figure 5. Architecture of the proposed enhanced BERT system

The BERT model is extended in two ways. Generally, the result generated at the CLS tag is given as input to the fully connected layer for ABSA. However, in this work, the max-pooling of target terms is taken and is concatenated with the result at CLS. This concatenated result is given as input to the fully connected layer. Furthermore, the BERT model is extended by giving enhanced multibit label input which depicts aspects and its sentiments. The multibit label format is depicted in Figure 4. This study also proposes approach 3 in which the basic BERT system is used for word embedding only and classification is done using the multilabel classifier. The results of all these approaches are presented and compared in the next section.

4. RESULTS AND DISCUSSION

4.1 Evaluation metrics for multilabel classification

In this experimentation, multilabel classifier is used for

classification which predicts multiple labels if present in an instance. Multiclass classifiers predict only one label for an instance.

The evaluation metrics used to measure the performance of multilabel classifiers are explained below:

- *Accuracy (per label):* The exact match between an actual set of labels and a predicted set of labels is determined by accuracy. Firstly, the accuracy per label is calculated for the whole dataset. Afterward, the average of per-label accuracy is calculated to obtain the final value.
- *Hamming loss:* Hamming loss shows how the relevance of an instance to a class label is wrongly estimated several times on average. In Eq. (5), L indicates the number of labels, p_i and a_i indicates predicted and actual labels.

The accuracy for binary classifiers is defined in Eq. (6). In Eq. (6), TP is true positive, TN is true negative, FP is false positive and FN is false negative. For multiclass classification, it is the average of accuracy of all classes. In multiclass classification, the most common approach used is one-vs-rest. In this approach, one classifier is trained for one class. At the time of testing, the class label with highest prediction is assigned as a label for a review instance.

$$Hamming\_Loss = \frac{1}{|X|}\sum_{i=1}^{X}\frac{XOR(p_{i,}\, a_i)}{|L|} \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

4.2 Results and discussion

This section gives a comparative analysis of the proposed model with other existing methods. This work proposed three approaches for ABSA. The first approach (Hybrid model with WCFS) suggests a feature selection methodology to select unigram features. These unigram features are combined with dependency rule-based features to generate the hybrid feature set. The multilabel classifier is trained on the hybrid feature set. The dependency rule-based features are selected using selective dependency rules. In dependency rule-based features, the extracted feature, i.e., a word pair, is related to each other with some meaningful grammatical relations. Such word pairs generally depict aspect category and sentiment. Therefore, such features are helpful to determine sentiment for aspects. In this approach, the unigram feature selection strategy is two-phase. The first phase selects the relevant features. In the second phase, features are selected based on correlation. The features having more weight value in the corresponding aspect category and less correlation are selected. Less correlation helps to avoid redundancy. A high correlation value means the features are redundant. In this approach, the multilabel classifier is used for classification. Table 2 shows the results obtained using this approach.

In this first approach, the ML classifiers used are BR, CC, and LP with baseline classifiers NB and SVM. In 2-way consideration, the highest per-label accuracy obtained is 0.8926 using BR-SVM, in 3-way it is 0.9296 using BR-SVM, and in 4-way consideration it is 0.9341 using BR-SVM. The results obtained using 4–way sentiment consideration are better than 2–way and 3-way. As shown in Table 2, the results obtained using the proposed system (3-way) are comparable and improved compared to the system described by Afzaal et al. [18] for 3-way. Figure 6 shows the accuracy (per label) gained by the hybrid model with the WCFS method using a multilabel classifier. It shows that the results obtained using BR, CC classifiers are better than LP classifier for this approach. In Figure 6, (2) represents only 2 sentiment classes are considered along with aspect category labels. Therefore, the number of bits in a label of a training instance is 10 representing positive and negative sentiment for each of the five aspect categories. Similar to binary, (3) and (4) represent three sentiment classes and four sentiment classes are considered along with aspect category labels.

Table 2. Accuracy (per label) and Hamming loss obtained using different methods

Approach	ML Classifier	Baseline Classifier	Accuracy (per label)	Hamming Loss
Afzaal et al. [18] (3-way)	CC	SVM	0.92	0.08
Hybrid model with WCFS (2-way*)	LP	NB	0.8237	0.176
	CC	NB	0.8641	0.136
	LP	SVM	0.867	0.133
	BR	NB	0.8699	0.13
	CC	SVM	0.8794	0.121
	BR	SVM	0.8926	0.107
Hybrid model with WCFS (3-way**)	LP	NB	0.8845	0.115
	CC	NB	0.8885	0.111
	BR	NB	0.8927	0.107
	LP	SVM	0.8978	0.102
	CC	SVM	0.9145	0.086
	BR	SVM	0.9296	0.07
Hybrid model with WCFS (4-way***)	LP	NB	0.8934	0.107
	CC	SVM	0.90345	0.097
	CC	NB	0.90465	0.095
	LP	SVM	0.9089	0.091
	BR	NB	0.9093	0.091
	BR	SVM	0.9341	0.066

*2-way means only two sentiments, positive and negative, are considered.
**3-way means positive, negative, and neutral sentiments are considered.
***4-way means positive, negative, neutral, and conflict sentiments are considered.

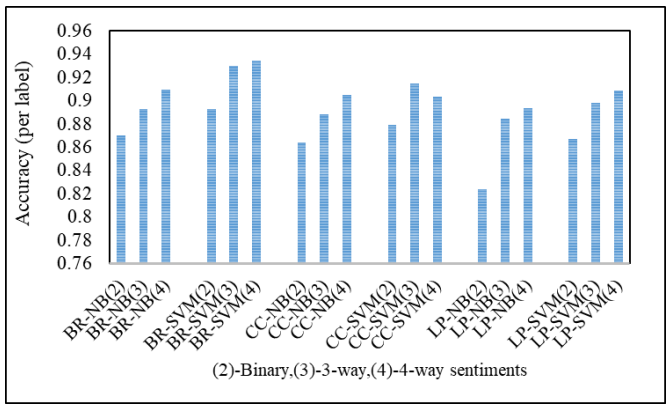

Figure 6. Accuracy (per label) gained by the proposed hybrid model with WCFS approach using different multilabel classifiers

In the second approach, a BERT language model-based solution is proposed for the ABSA task. Recently, for many NLP based problems BERT is used. As these are pre-trained models, it doesn't require large labeled data for training. These pre-trained language models can be directly applied for NLP based applications. BERT can be considered as a dynamic

approach of problem-solving for NLP problems. The BERT system is extended by applying max-pooling on target tokens and multibit class labels are given as input. The max-pooling supports to improve class prediction accuracy as it is applied to target token embedding which represents aspects many times. The output vector of max-pooling is an additional input for the fully connected layer. In this approach, extended BERT is used for both word embedding and classification. In the output layer, the SoftMax activation function is used. In this methodology, the hyper-parameters used are shown in Table 3.

Table 4 demonstrates the results obtained using this approach. These results are for epoch 6. It has gained 97.62% accuracy for binary sentiment along with aspect labels, which is better than the accuracy attained in the studies [14, 20] for the same dataset.

Table 4 depicts that the results gained using the enhanced BERT approach are better than the systems in the studies [7, 14, 20]. This methodology is also tested for the laptop dataset with a binary and 3-way sentiment. It has attained 96.03% accuracy for binary sentiment consideration. The accuracy obtained using this approach for 3-way is 95.77% which is better compared to the accuracy gained by Meng et al. [7] for the laptop dataset.

Figure 7 depicts that the accuracy gained using enhanced BERT approach is comparable and better than the other approaches mentioned in Table 4.

Table 3. Hyper parameter set for BERT system

Parameter	Value
$BERT_{BASE}$	
Dropout Rate	0.1
Batch Size	8
Learning Rate	2e-5
Max Epoch	6
Max Sequence Length	256
Optimizer	Adam

Table 4. Accuracy (%) gained using different approaches

Methodology	% Accuracy		
Restaurant dataset	**4-way**	**3-way**	**2-way**
Li et al.(2020) [14]	86.4	90.8	96.5
Meng et al.(2019) [7]	-	83.21	-
Sun et al. (2019) [20]	85.9	89.9	95.6
BERT (Epoch 6)	95.00	93.33	97.62
Laptop dataset			
Meng et al.(2019) [7]	-	78.55	-
BERT (Epoch 6)	-	95.77	96.03

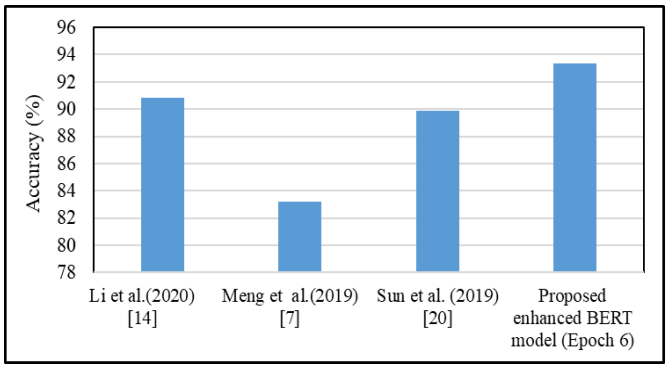

Figure 7. Comparison of % accuracy gained using different approaches

In the third approach, the BERT language model is used for word embedding and the multilabel classifier is used for classification. The BERT model used in this approach is a pre-trained model with no extension of max pooling on target terms. The classifiers used for this approach are BR, CC, LP with baseline classifiers NB and SVM. It is tested on both datasets. For the restaurant review dataset, the system is evaluated for 2-way, 3-way, and 4-way sentiment considerations. For the laptop review dataset, it is evaluated for 2-way and 3-way sentiment consideration. Table 5 shows the results obtained for this approach using the restaurant review dataset. For the restaurant dataset, the maximum per-label accuracy gained is 0.9459 for 4-way using BR-SVM classifier, and using 3-way it is 0.9331 which is more than the accuracy attained by Afzaal et al. [18]. This model has gained better results for the restaurant dataset for 4-way sentiment consideration for all classifiers. Table 6 represents the results gained using this approach for the laptop review dataset. For the laptop dataset, the maximum per-label accuracy attained is 0.9462 for 3-way using BR-SVM classifier.

Table 5. Accuracy (per label) and hamming loss gained using BERT + Multilabel classifier for restaurant review dataset

	ML Classifier	Baseline Classifier	Accuracy (per label)	Hamming Loss
BERT + Multilabel classifier				
Restaurant dataset (2-way)	BR	NB	0.8707	0.129
	LP	NB	0.8991	0.101
	CC	NB	0.9003	0.1
	BR	SVM	0.9023	0.098
	LP	SVM	0.9065	0.093
	CC	SVM	0.9282	0.072
Restaurant dataset (3-way)	LP	NB	0.9148	0.085
	LP	SVM	0.9178	0.082
	CC	NB	0.9196	0.08
	BR	NB	0.9224	0.078
	CC	SVM	0.9317	0.068
	BR	SVM	0.9331	0.067
Restaurant dataset (4-way)	BR	NB	0.9184	0.082
	CC	NB	0.9190	0.081
	LP	SVM	0.9240	0.076
	LP	NB	0.9328	0.067
	CC	SVM	0.9457	0.054
	BR	SVM	0.9459	0.054

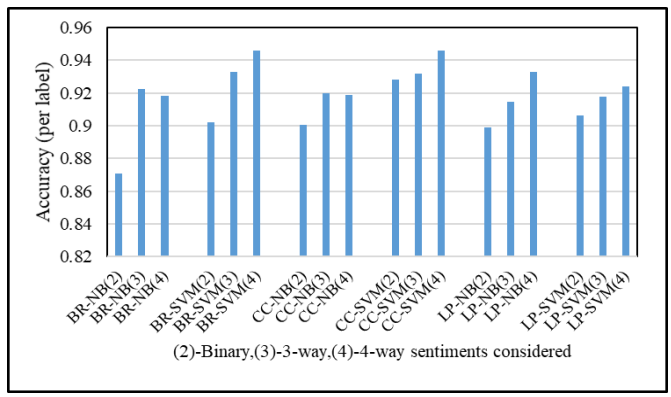

Figure 8. Accuracy (per label) gained by BERT+ multilabel classifier approach for restaurant review dataset

Table 6. Accuracy (per label) and hamming loss gained using BERT + Multilabel classifier for laptop review dataset

	ML Classifier	Baseline Classifier	Accuracy (per label)	Hamming Loss
		BERT + Multilabel classifier		
Laptop dataset (2-way)	LP	SVM	0.9089	0.091
	BR	NB	0.9119	0.088
	CC	NB	0.9135	0.086
	LP	NB	0.9135	0.086
	CC	SVM	0.9255	0.074
	BR	SVM	0.9276	0.072
Laptop dataset (3-way)	LP	SVM	0.9318	0.068
	LP	NB	0.9352	0.065
	BR	NB	0.9355	0.064
	CC	NB	0.9369	0.063
	CC	SVM	0.9447	0.055
	BR	SVM	0.9462	0.054

The proposed first approach, i.e., the hybrid model with the WCFS method, has gained 0.9341 per label accuracy using the BR-SVM classifier (4-way), which is very close to the accuracy obtained using approach 3. Figure 10 shows the accuracy obtained using the system specified in the study [18], the hybrid model with WCFS, and BERT + multilabel classifier. Accuracy (per label) gained using the hybrid model with WCFS and BERT + multilabel classifier is better than the accuracy attained in the study [18]. This experiment demonstrates that the results gained using the first approach are comparable to the results obtained using the third approach, i.e., BERT + multilabel classifier. The system proposed in approach 1 is highly dependent on the training dataset and the feature set. The system proposed in approach 2, i.e., the extended BERT model, gives significant and better results compared to existing methodologies for both datasets. Approach 1 system can further be extended by adding other grammatical rule-based features and testing them on other multilabel classifiers for different datasets. Figures 6, 8, and 9 depict that there is an increase in per-label accuracy from 2-way to 4-way sentiment class consideration. The 4-way sentiment class consideration add more details about the sentiments specified in a review instance, so it increases the classification accuracy.

5. CONCLUSIONS

This work proposed an end-to-end system for ABSA where the sentiments are determined for the predicted aspects instead of treating sentiment classification and aspect category prediction as two separate tasks. Many earlier systems used single label classifiers for ABSA. This work handled this drawback by using multilabel classifiers. Here, three approaches are proposed for ABSA. In the first approach, a hybrid feature set is used which contains grammatical rule-based features and unigram features selected using the proposed WCFS algorithm. In another approach, the BERT system is used for word embedding and the multilabel classifier is used for classification. The per-label accuracy gained using the hybrid approach is comparable to BERT + Multilabel classifier approach. In the third experimentation, the enhanced BERT system is used for word embedding as well as classification. This system has achieved better results compared to existing systems. This experimentation shows that the BERT system achieves significant results as it considers the bidirectional context of tokens in a sentence. The per-label accuracy (3-way) gained using the hybrid model + WCFS approach is 0.9296, using BERT + multilabel classifier (3-way) it is 0.9331 and the % accuracy gained using the enhanced BERT system (3-way) is 93.33. These results are for the restaurant review dataset. The datasets used in this experimentation are unbalanced. The results of the hybrid model can be improved for balanced datasets. The hybrid model + WCFS approach achieved comparable results as it includes features containing word pairs that are related by some meaningful grammatical relations. This experimentation proves that the machine learning solutions to the text classification problems need to consider features that are related by some meaningful grammatical relations.

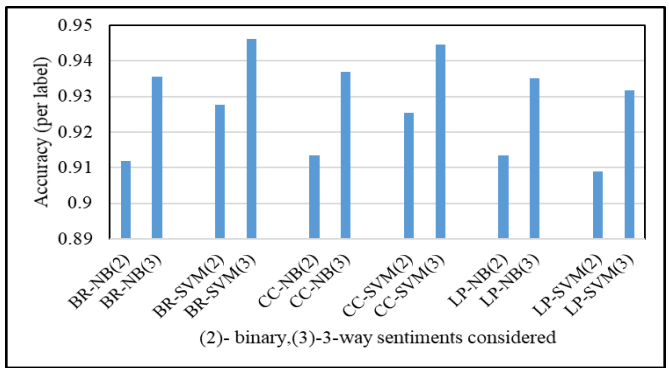

Figure 9. Accuracy (per label) gained by BERT+ multilabel classifier approach for laptop review dataset

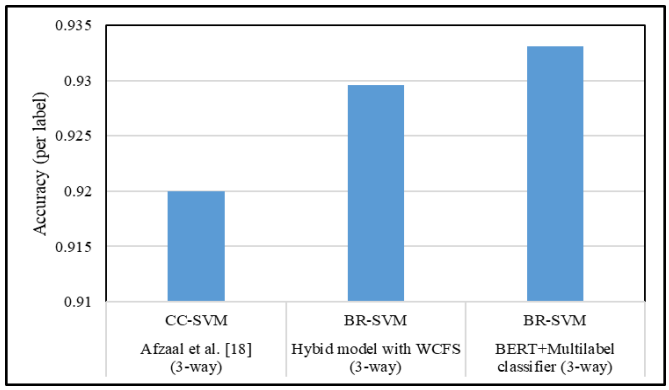

Figure 10. Comparison of Accuracy (per label) gained using different methodologies

REFERENCES

[1] Deng, S., Sinha, A.P., Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. Decision Support Systems, 94: 65-76. https://doi.org/10.1016/j.dss.2016.11.001

[2] Lee, G., Jeong, J., Seo, S., Kim, C.Y., Kang, P. (2018). Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowledge-Based Systems, 152: 70-82. https://doi.org/10.1016/j.knosys.2018.04.006

[3] Tao, J., Zhou, L., Feeney, C. (2019). I understand what you are saying: leveraging deep learning techniques for aspect based sentiment analysis. Proceedings of the 52nd Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2019.057

[4] Rida-E-Fatima, S., Javed, A., Banjar, A., Irtaza, A., Dawood, H., Dawood, H., Alamri, A. (2019). A multi-layer dual attention deep learning model with refined word embeddings for aspect-based sentiment analysis. IEEE Access, 7: 114795-114807. https://doi.org/10.1109/ACCESS.2019.2927281

[5] Yu, L.C., Wang, J., Lai, K.R., Zhang, X. (2017). Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(3): 671-681. https://doi.org/10.1109/TASLP.2017.2788182

[6] Tang, D., Qin, B., Feng, X., Liu, T. (2016). Effective LSTMs for target-dependent sentiment classification. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical, pp. 3298-3307.

[7] Meng, W., Wei, Y., Liu, P., Zhu, Z., Yin, H. (2019). Aspect based sentiment analysis with feature enhanced attention CNN-BiLSTM. IEEE Access, 7: 167240-167249. https://doi.org/10.1109/ACCESS.2019.2952888

[8] Liu, G., Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing, 337: 325-338. https://doi.org/10.1016/j.neucom.2019.01.078

[9] Ishaq, A., Asghar, S., Gillani, S.A. (2020). Aspect-based sentiment analysis using a hybridized approach based on CNN and GA. IEEE Access, 8: 135499-135512. https://doi.org/10.1109/ACCESS.2020.3011802

[10] Akhtar, M.S., Gupta, D., Ekbal, A., Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. Knowledge-Based Systems, 125: 116-135. https://doi.org/10.1016/j.knosys.2017.03.020

[11] Pham, D.H., Le, A.C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. Data & Knowledge Engineering, 114: 26-39. https://doi.org/10.1016/j.datak.2017.06.001

[12] Liu, S.M., Chen, J.H. (2015). A multi-label classification based approach for sentiment classification. Expert Systems with Applications, 42(3): 1083-1093. https://doi.org/10.1016/j.eswa.2014.08.036

[13] Cai, L., Song, Y., Liu, T., Zhang, K. (2020). A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. IEEE Access, 8: 152183-152192. https://doi.org/10.1109/ACCESS.2020.3017382

[14] Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., Liu, Q., Xiang, T. (2020). Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. IEEE Access, 8: 46868-46876. https://doi.org/10.1109/ACCESS.2020.2978511

[15] Kang, Y., Zhou, L. (2017). RubE: Rule-based methods for extracting product features from online consumer reviews. Information & Management, 54(2): 166-176. https://doi.org/10.1016/j.im.2016.05.007

[16] Liu, N., Shen, B., Zhang, Z., Zhang, Z., Mi, K. (2019). Attention-based sentiment reasoner for aspect-based sentiment analysis. Human-centric Computing and Information Sciences, 9(1): 1-17. https://doi.org/10.1186/s13673-019-0196-3

[17] Jia, Z., Bai, X., Pang, S. (2020). Hierarchical gated deep memory network with position-aware for aspect-based sentiment analysis. IEEE Access, 8: 136340-136347. https://doi.org/10.1109/ACCESS.2020.3011318

[18] Afzaal, M., Usman, M., Fong, A.C.M., Fong, S. (2019). Multiaspect-based opinion classification model for tourist reviews. Expert Systems, 36(2): e12371. https://doi.org/10.1111/exsy.12371

[19] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[20] Sun, C., Huang, L., Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588.

[21] Sultana, N., Islam, M.M. (2020). Meta classifier-based ensemble learning for sentiment classification. In Proceedings of International Joint Conference on Computational Intelligence, Springer, Singapore, pp. 73-84. https://doi.org/10.1007/978-981-13-7564-4_7

[22] Mohammadi, A., Shaverizade, A. (2021). Ensemble deep learning for aspect-based sentiment analysis. International Journal of Nonlinear Analysis and Applications, 12: 29-38. https://doi.org/10.22075/IJNAA.2021.4769

[23] Mowlaei, M.E., Abadeh, M.S., Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. Expert Systems with Applications, 148: 113234. https://doi.org/10.1016/j.eswa.2020.113234

[24] Xu, Q., Zhu, L., Dai, T., Yan, C. (2020). Aspect-based sentiment classification with multi-attention network. Neurocomputing, 388: 135-143. https://doi.org/10.1016/j.neucom.2020.01.024

[25] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27-35.

[26] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 486-495.

[27] De Marneffe, M.C., Manning, C.D. (2008). Stanford typed dependencies manual. Technical report, Stanford University.