Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6658520, 6 pages
https://doi.org/10.1155/2021/6658520

Research Article

Multi-Self-Attention for Aspect Category Detection and Biomedical Multilabel Text Classification with BERT

Xuelei Zhang, Xinyu Song, Ao Feng, and Zhengjie Gao

Chengdu University of Information Technology, Chengdu, China

Correspondence should be addressed to Ao Feng; fengao@cuit.edu.cn

Received 13 October 2020; Accepted 20 November 2021; Published 30 November 2021

Academic Editor: Serdar Ulubeyli

Multilabel classification is one of the most challenging tasks in natural language processing, posing greater technical difficulties than single-label classification. At the same time, multilabel classification has more natural applications. For individual labels, the whole piece of text has different focuses or component distributions, which require full use of local information of the sentence. As a widely adopted mechanism in natural language processing, attention becomes a natural choice for the issue. This paper proposes a multilayer self-attention model to deal with aspect category and word attention at different granularities. Combined with the BERT pretraining model, it achieves competitive performance in aspect category detection and electronic medical records' classification.

1. Introduction

The multilabel classification (MLC) problem is designed to determine whether an input instance belongs to one or more predefined categories and is applied to various scenarios [1], such as electronic medical records' classification (EMRC) in the medical domain and aspect category detection (ACD) in aspect-level sentiment analysis.

Sentiment analysis is the heart of many business and social applications [2, 3]. ACD is a subtask of aspect category sentiment analysis [4] and can be treated as an MLC task. Its main purpose is to locate aspect information in comments. Comment "*While it was large and a bit noisy, the drinks were fantastic, and the food was superb*" evaluates the restaurant's "environment" and "food," which belong to the predefined categories in the dataset. Mining such information from comments is of great help to improve user experience and product quality. Therefore, ACD is a very meaningful task in comment analysis.

With the continuous development of computers, modern biomedical research often faces the problem of underutilized text data, such as electronic medical records. Electronic medical records are common text data in clinical medicine, which have great value when correctly utilized. Automatic classification of electronic medical records not only improves the work efficiency of medical workers but also has a great effect on the research of various diseases. An electronic medical record may include diagnosis, treatment, surgery, and many other aspects, such as "*Before admission, the patient was admitted to our hospital due to increased bowel movements. Colonoscopy showed that rectal polyps and recommended rectal polypectomy. The patient did not undergo surgery... Today, the patient was admitted to our hospital for rectal polypectomy and was admitted our hospital for treatment. Since the onset of the disease, the patient has normal appetite, conscious mind, good spirits, good sleep, normal bowel movements, normal urination, and no significant changes in weight.*" This record includes "diagnosis" and "surgical treatment" of the patient, two of the six predefined categories in the dataset. Classifying electronic medical records is also a multilabel text classification task.

In previous work, the earliest approach to solve MLC is to convert it into multiple single-label binary classification problems [5], but it ignores the correlation between tags. To retain correlation information, the classifier chain [6] is applied to the MLC problem. When the data volume is large, the calculation cost of the classifier chain will also be quite high. Besides, some machine learning algorithms are revised

to adapt to MLC, such as multilabel K-nearest neighbor (ML-KNN) [7] and RANK support vector (RANK SVM) [8]. With the development of deep neural networks, some representative deep learning models are also applied in MLC, especially after the introduction of attention mechanism [9]. Its excellent feature extraction capabilities are widely used in various fields of natural language processing. Most recently, pretrained language models including ELMo [10], OpenAI-GPT [11], and BERT [12] have shown their effectiveness to simplify the effort of feature engineering. However, direct use of the pretrained BERT model in the MLC task does not show significant improvement. We believe that the vanilla BERT model is unable to capture key information in each category, especially when the correlation between each label is strong.

In this paper, we propose a BERT-based multi-self-attention model (BERT-MSA) for MLC. The self-attention mechanism is used to capture the information of each category. Although a single attention head can obtain part of the important information in the text, a sentence often belongs to multiple categories in MLC. So, it is necessary to use multiple attentions to obtain relevant information of multiple categories, and neural network is an efficient tuning framework.

Two tasks, ACD and EMRC, are applied to verify the effectiveness of our model. For the ACD task, with extensive experiments on subtask 3 of SemEval-2014 task 4 (http://alt.qcri.org/semeval2014/task4/) [4], the results indicate that our BERT-MSA model is superior to other baseline methods in aspect category sentiment analysis. For the EMRC task, we use subtask 1 of CCKS-2019 task 1 (http://www.ccks2019.cn/?page_id=62) [13]. The results show that our model can still obtain results that exceed the benchmark scores on the medical datasets. Good performance in two completely different fields proves the generalization ability of our model.

2. Related Work

Early methods of MLC are mostly based on traditional machine learning [7, 8]. Recently, some neural network models have also been applied in the MLC task and have made important progress. For example, Zhi-Hua Zhou and Zhou [14] apply a fully connected neural network with a paired ranking loss function, Kurata et al. [15] recommend convolutional neural networks (CNN) for classification, Kurata et al. and Chen et al. [15, 16] use both CNN and long short-term memory networks (LSTM) [17] to capture the semantic information of text.

Different models are used to improve the quality of text feature extraction. With the emergence of pretrained language models and attention mechanism [9], they are soon adopted in MLC due to their excellent representation and feature extraction capabilities. These general MLC models have also been applied to specific fields, such as medical record processing and aspect-level sentiment analysis.

In aspect-level sentiment analysis, ACD aims at identifying aspects about which users express their sentiments. A popular aspect detection method is based on the single noun and compound noun frequency method [18]. In addition to focusing on frequency, syntax-based methods are also used to detect aspects through syntactic relations [19, 20]. In general, this kind of model operates with an unsupervised learning manner.

To improve the performance of aspect detection, some deep learning methods based on word embedding [21] are applied to aspect detection, using the grammatical and semantic information embedded in the distributed representation [22]. CNN is also used in aspect detection due to its excellent feature extraction capabilities [23, 24]. LSTM with attention (LSTM-Attention) [25] applies the attention mechanism in sequential text input for accurate aspect detection. Ensemble CNN-RNN networks are applied to process MLC task [16]; the networks capture both the global and the local textual semantics and model high-order label correlations. Recently, CNN-stacked bidirectional LSTM networks with a multiplicative attention mechanism are proposed to process MLC task [26].

More recently, a hybrid Siamese-convolutional neural network [27] with additional technical attributes is applied to the MLC task. It is based on single and Siamese multitask architecture networks and calculates the category-specific similarity in the Siamese structure. Besides, it is based on RNN and a tree structure [28] to represent the relationships among labels, consequently developing an efficient max-product algorithm for exact inference of label prediction for the MLC task.

3. Model

3.1. Input and Embedding Layers. Both ACD and EMRC can be formulated as a multilabel text classification problem. A sentence usually consists of a series of words: $x = (w_1, w_2, \ldots, w_n)$. In ACD, we need to predict the sentence category $Y = \{\text{food, price, service, ambience, anecdote}\}$. In EMRC, the sentence category includes $Y = \{\text{diagnosis and disease, image inspection, laboratory inspection, surgery, medical treatment, anatomy}\}$.

BERT is a new language representation model, which uses bidirectional transformers [9] to pretrain a large corpus, and fine tunes the pretrained model on other tasks. To obtain a fixed-dimensional pooled representation of the input sequence, we use the BERT fine tuning final hidden state as the input. The vector is denoted as $S = (e_1, e_2, \ldots, e_n)$.

3.2. Model Description. As shown in Figure 1, our model contains BERT-MSA as the key component. First, we use the hidden layer vector output by BERT to capture the important information about each tag in the sentence and then reuse the attention score and attention output to obtain the most important information in the sentence.

The first step in calculating self-attention [29] is to create 3 vectors from the input of BERT fine tuning. For each word, we create a query vector, a key vector, and a value vector. These vectors are generated by multiplying the word embeddings by the three matrices created during our training process. $W^Q, W^K, W^V \in R^{d_x \times d_z}$ are our predefined model parameter matrices, randomly initialized before training,

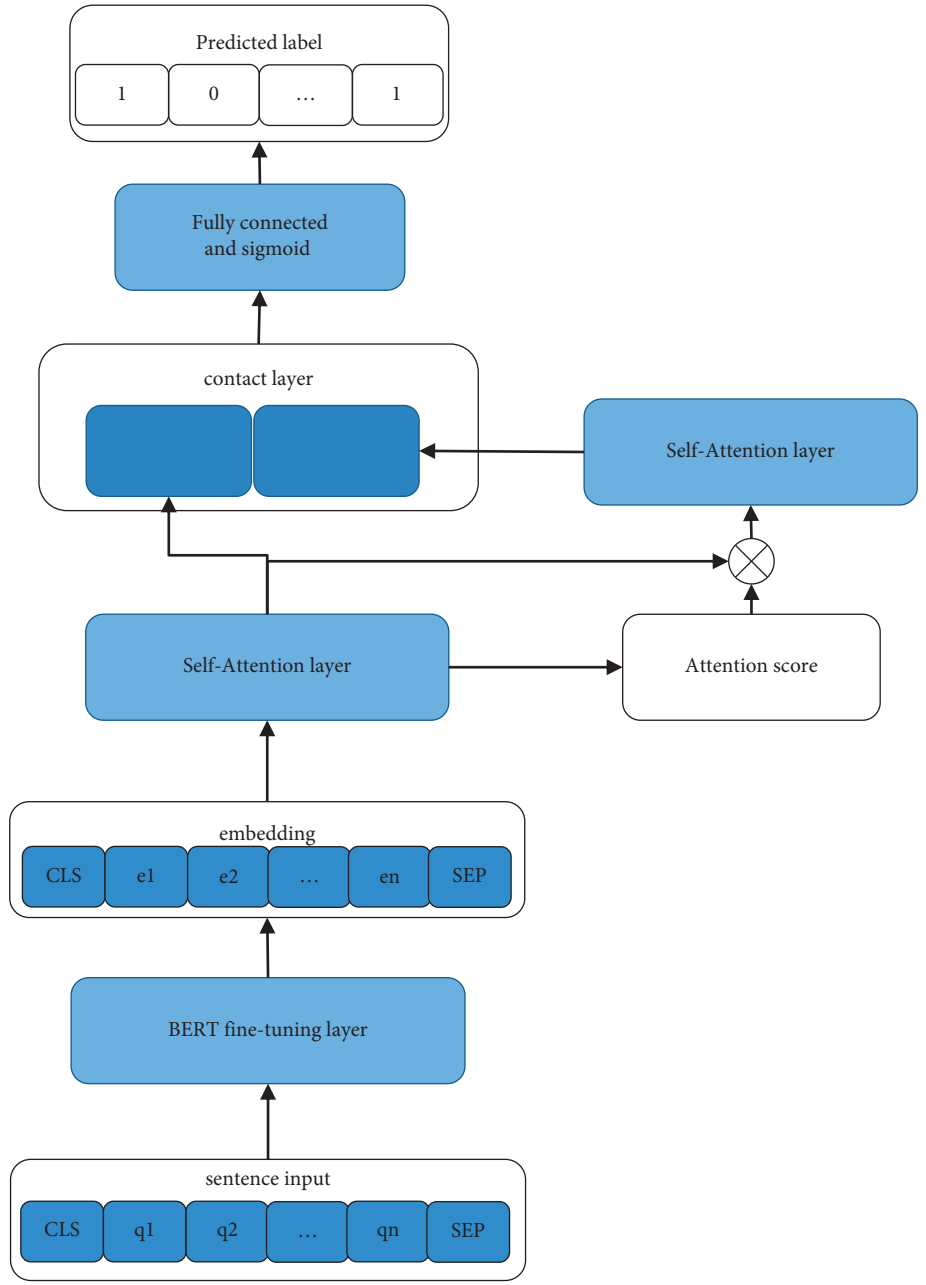

Figure 1: The overall architecture of BERT-MSA.

and continuously updated through gradient descent in the training process.

The input sentence $x = (x_1, x_2, \ldots, x_n)$ contains n elements, where $x_i \in R^{d_x}$, and a new sequence $z = (z_1, z_2, \ldots, z_n)$ is calculated with the same length, where $z_i \in R^{d_z}$.

Compatibility function e_{ij} is calculated with the scaled dot product, which compares the relationship or similarity of two input elements. Through linear transformation of the input, more expressive power is added to the input:

$$e_{ij} = \frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T}{\sqrt{d_z}}. \tag{1}$$

We get the attention score a_{ij} of each word in the sentence; with using the soft-max function,

$$a_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{n} \exp\left(e_{ik}\right)}. \tag{2}$$

Each output element z_i is calculated by linearly transforming the weighted sum of the attention score calculated by soft-max and the sentence vector output by BERT fine tuning:

$$z_i = \sum_{j=1}^{n} a_{ij}\left(x_j W^V\right). \tag{3}$$

After calculating the attention output vector of the BERT's output and the attention scores, we calculate the product of the first-layer attention output vector and attention score to make up for the lack of the first attention to capture text features. The result is fed into the second self-attention layer, and the corresponding attention output vector is calculated again. The results of the two attention's output vectors are concatenated after that. The two attention layers are calculated in the same way, using different parameters.

After that, the spliced vector is converted into the same dimension of the label size with a fully connected layer. Finally, the activation goes through a sigmoid function [20, 30] to generate the probability if a sample belongs to the corresponding class. The instance is assigned to the category if the probability is over threshold 0.5:

$$Y(x) = \frac{1}{1 + e^{-x}}. \tag{4}$$

Loss function is calculated with binary cross entropy between the predicted probability and the true label:

$$\text{loss} = -\sum_{i=1}^{n} \widehat{y}_i \log(y_i) + (1 - \widehat{y}_i)\log(1 - \widehat{y}_i). \tag{5}$$

4. Experiments

4.1. Datasets. In the ACD experiment, we use the SemEval-2014 ABSA challenge dataset [4] for performance comparison. Table 1 shows the details of the dataset, including the number of samples in each category. The dataset is from subtask 3 of SemEval-2014 Task 4.

We use the Chinese Electronic Medical Record for the medical text classification experiment. This dataset is from subtask 1 of CCKS2019 Task 1 [13]. Table 2 shows the details of the dataset, with statistics separated by categories.

4.2. Baseline Methods. Using the datasets described above, BERT-MSA is compared to some baseline systems using the standard evaluation metrics from SemEval-2014 task4 and CCKS2019 Task 1.

RANK SVM [8] uses TF-IDF to extract text features. It is a basic machine learning algorithm based on pointwise sorting.

ML-KNN [7] is also based on TF-IDF features. It finds K-nearest neighbor samples and uses Bayesian conditional probability formula to calculate probability of the current label.

TEXTCNN [23] uses Glove vector [31] for text representation. It applies multiple convolution kernels to extract text features and then input them to the linear transformation layer. The sigmoid function is used to output probability distribution on the label space.

Bi-LSTM [11, 12] represents the basic bidirectional LSTM model.

Bi-LSTM-Attention is based on the basic Bi-LSTM. The hidden layer output of LSTM is fed to self-attention for classification.

Table 1: Details of SemEval-2014 task 4 Subtask 3 dataset: ACD TASK.

Category	Training	Testing	Total
Food	1166	402	1568
Price	304	80	384
Service	562	167	729
Ambience	384	105	489
Anecdote	1102	219	1321
Total	3518	973	4491

Table 2: Details of CCKS2019 Task 1 dataset: Chinese EMRC task.

Category	Training	Testing	Total
Diagnosis and disease	981	359	1340
Image inspection	452	199	651
Laboratory inspection	224	157	381
Surgery	792	115	907
Medical treatment	548	187	735
Anatomy	993	375	1368
Total	3990	1392	5382

XRCE [32] achieved the highest score in the SemEval-2014 competition.

Attention-XML [33] uses the label tree-based deep learning model for multilabel text classification.

BERT-base feeds the hidden layer of BERT's pretraining output to a fully connected layer.

BERT-Attention applies a one-layer attention network [9] over BERT to obtain information in each text category.

4.3. Hyperparameters. In the ACD task, we use the English pretrained BERT-based (https://huggingface.co/BERT-base-uncased) model for fine tuning to solve ACD the task and use the Chinese pretrained BERT-based (https://github.com/ymcui/Chinese-BERT-wwm) model for fine tuning to solve the EMRC task. The number of the Transformer blocks is 12, the size of the hidden layer is 768, the number of self-attention heads is 12, and the total number of parameters of the pretraining model is about 110M. The learning rate is set to $5e-5$, and the batch size is 32, with maximum sentence length 80. Adam optimizer is used to tune the model. For the EMRC task, most parameters are the same, but the batch size is set to 16. Sentences in the CCKS2019 collection are relatively long, and BERT supports a maximum sentence length of 512, so we truncate the sentences to size 512.

4.4. Evaluation Metrics. We use precision, recall, and Micro-F1 [34] to evaluate the performance in both tasks.

4.5. Results and Analysis. Table 3 shows that BERT-MSA achieves clear improvements in the ACD dataset over BERT-base, BERT-Attention, and other based convolutional and LSTM networks. Introduction of self-attention improves performance, while multiple self-attention shows a much larger gain. We believe that the multiple self-attention mechanism makes up for the defects caused by the feature

Table 3: The test set results for SemEval-2014 task 4 Subtask 3: ACD TASK

Model	Precision	Recall	F1
ML-KNN	70.77	56.06	62.09
Rank SVM	86.57	70.48	77.41
TextCNN	85.27	78.77	79.44
BiLSTM-Attention	81.20	80.67	79.28
XRCE	83.23	81.37	82.29
Attention-XML	84.73	82.24	83.43
BERT base	85.20	90.43	85.81
BERT Att	85.82	90.46	86.29
BERT-MSA	**85.88**	**91.11**	**86.61**

Table 4: The test set results for the Chinese EMRC task.

Model	Precision	Recall	F1
ML-KNN	71.85	84.09	76.94
Rank SVM	81.62	84.83	82.68
TextCNN	95.20	92.60	93.00
BiLSTM-Attention	92.55	91.95	89.22
Attention-XML	92.64	92, 55	89.78
BERT base	95.55	93.33	93.41
BERT Att	94.97	**95.28**	93.48
BERT-MSA	**96.75**	94.57	**93.94**

extraction layer and helps the model to find words that best reflect the category.

At the same time, Table 4 shows that our multi-self-attention mechanism still obtains promising results on medical datasets. Its recall is slightly lower than BERT-Attention, but precision is much higher, resulting in a small increase in the F1 value.

Compared with the EMRC task, improvement in the ACD task is small, even when the baseline is much lower. We argue that the restaurant dataset has shorter sentences, leaving less context for the attention mechanism. Improvement in the feature extraction method is required to achieve better results in such small text pieces.

5. Conclusion

In order to provide a general solution of the MLC task in multiple applications, we propose the BERT-MSA model. It introduces a BERT-based multiple self-attention mechanism, which can obtain more comprehensive features of each word in a sentence. As self-attention has the ability to update representation of the current word from any context word in the sentence, it can potentially learn full semantic information in a sentence. For its applications, we select the ACD task in the online review domain and the EMRC task in the medical field, two areas with huge differences in text representation. Experiments on these diverse areas show that our model has achieve gratifying results. In some datasets, our model has achieved state of the art in comparison to baseline models. Moreover, our model has shown good results in MLC tasks in Chinese and English. This shows that our model is suitable for processing MLC tasks in a specific field.

As a recent concept, the prompt framework [35] for text classification is proposed, which converts the text classification task into a cloze task, making full use of the masked language model's representation power. It is an innovative paradigm in the pretrained language model, and we will try modeling the MLC task in that framework for richer in-depth semantic representation.

Data Availability

All data supporting this systematic review are from previously reported studies and datasets, which have been cited within the article. The processed data are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Research Innovation Team Fund (Award no. 18TD0026) from the Department of Education, Sichuan Province, China, and Sichuan Science and Technology Program (Project no. 2020YFG0168).

References

[1] G. Tsoumakas and I. Katakis, "Multi-label classification," *International Journal of Data Warehousing and Mining*, vol. 3, pp. 1–13, 2007.

[2] H. Peng, E. Cambria, and A. Hussain, "A review of sentiment analysis research in Chinese language," *Cognitive Computation*, vol. 9, no. 4, pp. 423–435, 2017.

[3] H. Peng, "Phonetic-enriched text representation for Chinese sentiment analysis with reinforcement learning," *Information Fusion*, vol. 70, pp. 88–99, 2021.

[4] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 27–35, Association for Computational Linguistics, Dublin, Ireland, August 2014.

[5] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[7] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[8] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," *Advances in Neural Information Processing Systems*, vol. 14, pp. 681–687, 2002.

[9] V. Ashish, S. Noam, and N. Parmar, "Attention is all you need," pp. 5998–6008, 2017, arXiv.

[10] M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," arXiv preprint arXiv: 1802.05365, 2018.

[11] R. Alec, N. Karthik, and S. Tim, "Improving language understanding by generative pre-training," arXiv, 2018.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-trainingof deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[13] L. Nan, L. Luo, and Z. Ding, *DUTIR at the CCKS-2019 Task1: Improving Chinese Clinical Named Entity Recognition Using Stroke ELMo and Transfer Learning*, CCKS, Nanjing, China, 2019.

[14] M.-L. Zhi-Hua Zhou and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[15] G. Kurata, B. Xiang, and B. Zhou, "Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 521–526, San Deigo, CA, USA, 2016, June.

[16] G. Chen, D. Ye, and Z. Xing, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2377–2383, IEEE, Anchorage, AK, USA, May 2017.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence, Conference on Innovative Applications of Artificial Intelligence*, pp. 755–760, San Jose, CF, USA, July 2004.

[19] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and ranking product features in opinion documents," in *Proceedings of the COLING, Posters*, pp. 1462–1470, Beijing, China, August 2010.

[20] Y. Zhao, B. Qin, and H. Shen, "Generalizing syntactic structures for product attribute candidate extraction," in *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 377–380, Los Angeles, CF, USA, June 2010.

[21] M. Tomas, S. Ilya, and K. Chen, "Distributed representations of words and phrases and their compositionality," pp. 3111–3119, 2013, arXiv.

[22] S. A. Razavi and M. Asadpour, "Word embedding-based approach to aspect detection for aspect-based summarization of Persian customer reviews," in *Proceedings of the IML*, Liverpool, UK, October 2017.

[23] K. Yoon, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[24] Z. Toh and J. Su, "Nlangp at semeval-2016 task 5: improving aspect based sentiment analysis using neural network features," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Deigo, CF, USA, June 2016.

[25] H.-j. Park, M. Song, and K.-S. Shin, "Deep learning models and datasets for aspect term sentiment classification: implementing holistic recurrent attention on target-dependent memories," *Knowledge-Based Systems*, vol. 187, Article ID 104825, 2020.

[26] T. T. Esther and C. Erik, "A convolutional stacked bidirectional LSTM with a multiplicative attention mechanism for aspect category and sentiment detection," *Cognitive Computation*, vol. 20, pp. 1–10, 2021.

[27] W. Yang, J. Li, and F. Fukumoto, "MSCNN: a monomeric-siamese convolutional neural network for extremely imbalanced multi-label text classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6716–6722, Online, November 2020.

[28] Z. Peng, A. Behnoush, and M. Xie, "Multi-label classification of short texts with label correlated recurrent neural networks," in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 119–122, Canada, July 2021.

[29] H. Zhang, I. Goodfellow, and A. Metaxas Dimitris, "Self-attention generative adversarial networks," in *Proceedings of the International Conference on Machine Learning. PMLR*, pp. 7354–7363, Long Beach, CF, USA, June 2019.

[30] X. Yin, G. Jan, and A. Lantinga Egbert, "A flexible sigmoid function of determinate growth," *Annals of Botany*, vol. 91, no. 3, pp. 361–371, 2003.

[31] J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.

[32] C. Brun, D. N. Popa, and C. Roux, "Xrce: hybrid classification for aspect-based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 838–842, Doha, Qatar, October 2014.

[33] R. You, Z. Zhang, and Z. Wang, "Attentionxml: label tree-based attention-aware deep model for high-performance extreme multi-label text classification," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5820–5830, 2019.

[34] C. S. Saranyamol and L. Sindhu, "A survey on automatic text summarization," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7889–7893, 2014.

[35] P. Liu, W. Yuan, and J. Fu, "Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing," arXiv preprint arXiv:2107.13586, 2021.