Bidirectional Transformer-based Text Classification

[1]Rekha B, MCA student, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India.

[2]Dr. Raghavendra S P, Assistant Professor, MCA, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India

Abstract

Text categorization served as a significant issue in natural language processing, which is arguably as a method with broad variety of information retrieval application sentiment analysis, news classification, etc. Conventional techniques for machine learning that method include significant amounts of manual engineering of features cannot easily capture deep context meaning. Therefore, this study considers the idea of applying Transformer-Based Bidirectional Encoder Representation (BERT) text classification for accurate and efficient text classification to get past these restrict ions. An obvious strength of BERT's pre-trained language representations is its capability to understand semantic links through context from the combined left and right sentence context, improving on previous word embedding approaches for text classification. This study will fine-tune BERT using a multi-class dataset covering business, sports, technology, entertainment, politics, and international news. This study demonstrates that BERT serves as an effective framework for scalable and domain-independent text categorization purposes, and illustrates how transformer-based models cope with linguistic complexity. The evaluation results revealed that BERT achieves competitive classification accuracy and robustness throughout the significant number of categories, and improved the regularity of misclassifications when compared with tabular or rule based approaches.

Keywords: *Text categorization, Deep learning , Machine learning, and Bert*

1. Introduction

The growth of digital content in social media, news sites, and online platforms has led to a increasing demands for automated methods to organize and understand textual form of data. Numerous application are created using classifying textual inputs which can be labelled by pre-existing labels (text classification task), including recommendation systems, sentiment analysis, spam detection, and information retrieval tasks in a specific domain. While they provide a baseline binding classification, traditional methods using statistical models or a bag of words does not quite utilize or represent the language and the meaning being communicated. Though there have been some advancements on the aspect of apply deep learning methods , particularly convolutional neural networks(CNNs), for feature enhancement and recurrent neural networks (RNNs) there are difficulties in modelling polysemy and long-distance dependencies.

Advances in transformer architectures, such as Bidirectional Encoder Representation from Transformer (BERT), have a led to revolutionary advancements in natural language processing in recent years. BERT enables a better understanding of syntax and semantics and modeling complex relationships found in text through self-attention mechanism and bidirectional contextual learning. BERT provides contextual representations of words which allows for adapting to the surrounding context which improves performance on classification tasks compared to traditional embeddings which generate static word vectors. This study investigates the usefulness of optimizing BERT for multi-class text

classification in field such as business, sports, technology, entertainment, politics and world news. By taking advantage of BERT's contextual capability, This study will clarify how BERT is able to outperform other methods of text classification and demonstrate its potential as a scalable, domain-independent framework that could classify other sources of textual data.

2. Literature Survey

BERT-Based Aspect-Based Sentiment Analysis (Wang et al., 2020)[1] Problem: Sentiment analysis frequently has to go beyond general sentiment to pinpoint sentiment directed at particular features of a good, service, or encounter. The suggested method BERT is used in this work to analyze sentiment based on aspects. A statement and a list of pre-established parameters—such as "battery life" for a phone review—are fed into the model. In order to determine not only the sentiment communicated but also the precise features the sentiment is directed towards, BERT next examines the connections between the sentence's words and the previously indicated elements. Benefits: By identifying sentiment toward particular characteristics, this method provides a more sophisticated picture of user opinions. Businesses who wish to comprehend customers will find this knowledge useful.

Alagha (2022)[2], explored BERT utilizing multilayer attention mechanisms to categorize short Arabic texts and showed BERT could be better in terms of classification performance for morphologically rich languages, when domain-adapted, leveraging transformer representations and keeping in mind linguistic features.

Hamzaoui, et al. (2023)[3],proposed a hybrid BERT-BiLSTM architecture for hierarchical text classification. Their combination of BiLSTM (which could take into account sequential modeling) with BERT (which could use context aware embeddings) demonstrated that hybrid deep learning architecture can increase the accuracy of multi-level classification tasks.

The implementation of distributed word embeddings (Mikolov et al. 2013; Pennington et al. 2014)[4] provided low-dimensional, dense vector representations of words, which facilitated the classification process. Then, after that, text classification researchers implemented several deep learning structure derived from, for instance , the Convolutional Neural Networks (CNNs) (Kim, 2014) in addition to Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) to text classification since they showed a much greater ability to address word order and syntactic patterns. However, these models of deep learning struggled with long-range dependencies.

The Transformer was introduced by Vaswani et al. (2017) and subsequently BERT (Devlin et al., 2019)[5], there was a paradigm shift. BERT was extremely powerful because of its next sentence prediction (NSP) and masked language modelling (MLM) goals that allow pretraining from large corpura and then fine-tuning up to a particular assignment for example, it has been demonstrated in the literature, that BERT demonstrates state-of-the-art performance on GLUE benchmark tasks, including text classification and other downstream tasks. Research has shown BERT was modified in a number of papers for several domain-specific tasks, for example, sentiment classification, news categorization, and biomedical text mining (BioBERT). Given these developments, this project seeks to maximize performance in general text categorization by using optimized BERT models.

3. Proposed methodology

The proposed methodology consists of a tuned BERT model and Gradio for text categorization. A brief explanation of the recommended steps is provided below:

1. Data Preprocessing
2. Model Fine-Tuning
3. Classification Head
4. Interface Integration
5. Performance Evaluation

3.1 Data Preprocessing: We will use BERT's WordPiece tokenizer to tokenize the incoming text and convert it to input embeddings.

3.2 Model Fine-Tuning: We will fine-tune a pre-fine-tuned BERT model to domain-specific class-level training data (Sports, politics, Sci/Tech, Entertainment, Lifestyle, Business and World).

3.3 Classification Head: We will concatenate the BERT [CLS] representation with dense layer that is totally connected with softmax activation to predict probability of the classes.

3.4 Interface Integration: Using the Gradio package we will create an interactive interface where users will be able to see real-time performance metrics and put text as input to get predicted categories,.

3.5 Performance Evaluation: Performance of the prediction system will be evaluated on accuracy, precision, recall, F1 score, and visual accuracy evaluation by confusion matrix.

3.6 Block Diagram

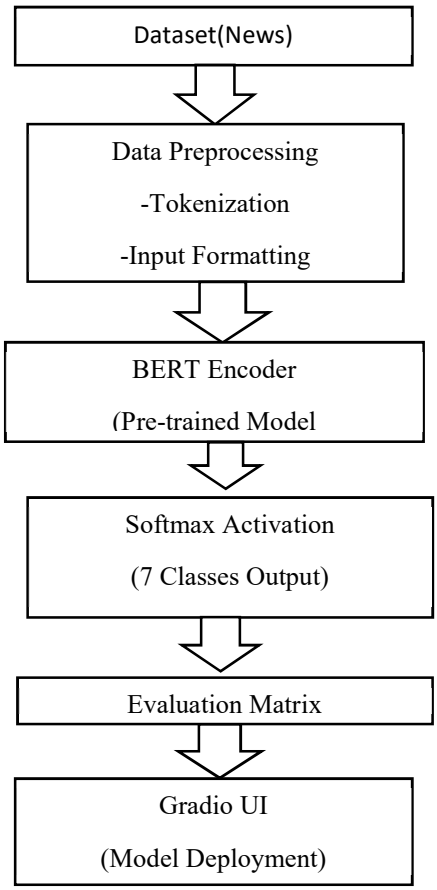

Figure 3.6.1 Block diagram of Text Classification Using BERT

The Figure 3.6.1 the project starts with a raw text dataset from a news source. Before putting the text into a BERT encoder, which a pre-trained model that'll generate contextual word embeddings, the text needs to be converted to a machine-readable format using tokenization and input formatting. Next the embeddings will flow through a Softmax activation layer which enables the generation of a probability distribution across the seven output classes such as a sports, business, Politics, World, Lifestyle, Entertainment, Sci/Tech ; an evaluation matrix used to indicate model performance, indicating performance levels of accuracy and reliability; finally, the trained model will be placed into a Gradio user interface (GUI) to enable accessibility and interactive output monitoring, taking all of the hassle of installation and coding away, allowing the user to enter new text and automatically see the classification prediction.

4. Mathematical Model

Mathematical Model and performance metrics of text classification Using BERT

1. **Input Representation:**
   Word Piece tokenization is used to tokenize each input text sequence $S = \{\omega_1, \omega_2 \ldots\ldots, \omega_n\}$ into tokens $T = \{t_1, t_{2,\ldots\ldots,} t_m\}$ where $m \geq n$.
   An embedding vector is assigned to each token:
   $$E_i = E_{token}(t_i) + E_{position(i)+}E_{\text{segment(i)}}$$
   where:
   - Token embedding = E token
   - Position= E position is equivalent to positional encoding.
   - Segment embedding = E segment
2. **BERT Encoding(Transformer Layer):**
   BERT uses multi-head self-attention to record context in both directions:
   $$\text{Attention } (Q, K, V) = \ softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right)v$$
   where:
   - Q, K, and V are query, key, and value matrix derived from E_i
   - d k = the dimensionality of the keys

   Attention to several heads:
   $$MHA(H) = concat(head1,,,,head\ h)w^0$$
   Every head:
   $$head_j = Attention(HW_j^Q, HW_j^k, HW_j^{v:})$$
3. **Sequence classification Head:**
   The aggregate sequence representation is the unique classification token $[CLS]$
   $$h_{CLS} = BERT(S)_{CLS}$$
   This vector is mapped to class logits by a fully connected (dense) layer:
   $$z = W_c h_{CLS} + b_c$$
4. **Output Layer(Softmax) :**
   For C categories, class probabilities are calculated as follows:

$$P(y = c \ \backslash S) = \frac{exp(Zc)}{\Sigma_{j=1}^{c} exp(Z_j)}$$

5. Loss Function:

Cross-entropy loss is utilized to train the model:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{c} y_{i,c}\, log\, P(y = c|S_i)$$

5. Graphs

5.1 Training and Validation Accuracy

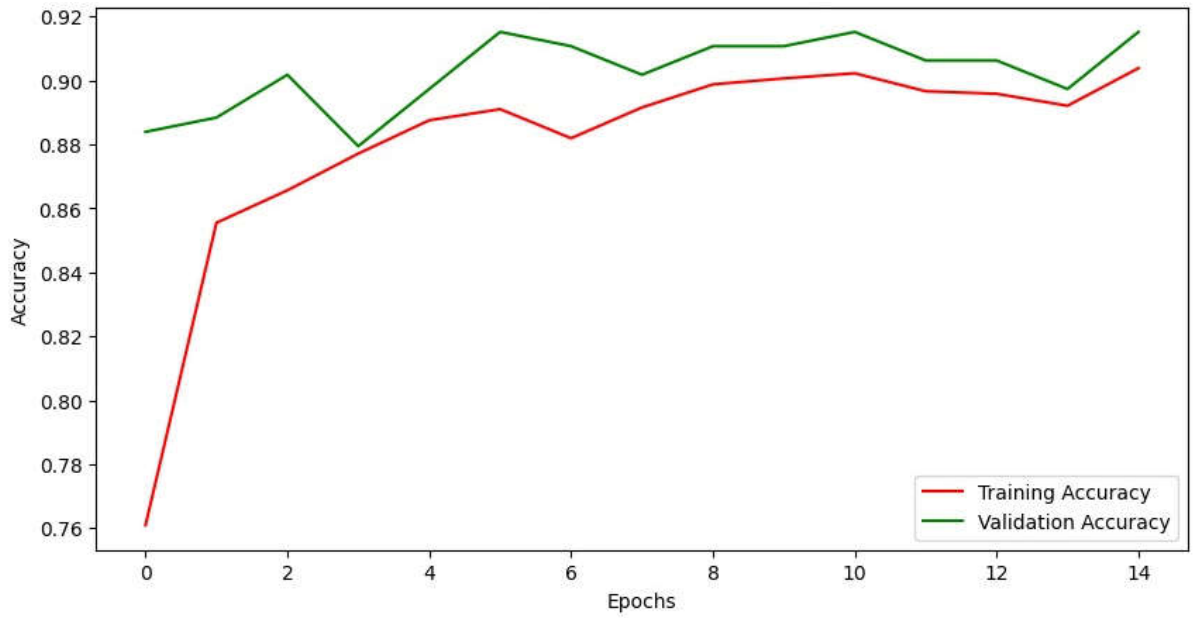

Fig 5.1.1: Training And Validation Accuracy

The Figure 5.1.1 shows the changes in the training accuracy and validation accuracy of a BERT-based text categorization model across 15 epochs. The green line is the validation accuracy and the red line is the training accuracy. When the model first begins to learn from the dataset, the training accuracy is quite low (~76%). After the initial epochs, the training accuracy increases significantly and eventually, the training accuracy continues to improve until after epoch, it is above 90%.

The validation accuracy starts off higher than expected (~88%) as a result of BERT's powerful pretrained language representations that already include significant contextual information. There are very slight changes between 88% and 92%, so the validation accuracy stays high throughout training, often staying higher than the training accuracy. This indicates that the model isn't only learning well but is also generalizing to fresh information using with little, if any, overfitting.

5.2 Training and Validation Loss

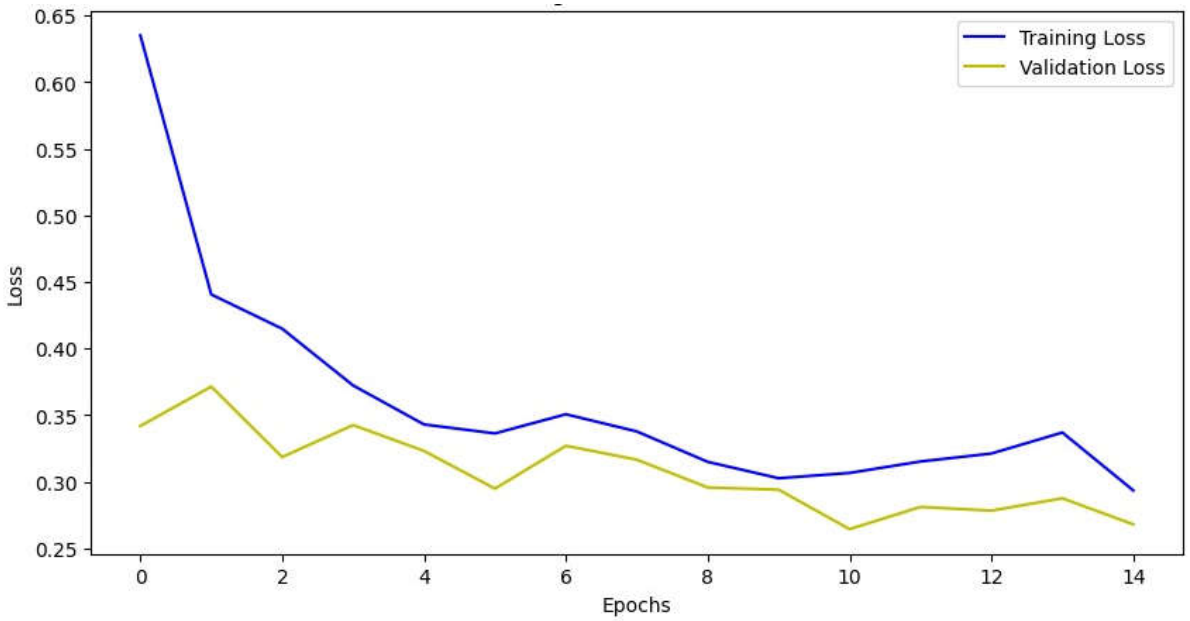

Figure 5.2.1 Training And Validation loss

The Figure 5.2.1 below displays the training and validation loss trajectories for the model's BERT-based text classification architecture over 15 epochs. The blue curve indicates the training loss, and the yellow curve indicates the validation loss. There is a jump in the model's training loss near the start of the training, as represented in the figure, of about 0.63. This early difficulty to optimize network parameters shows some initial size in the model's overall training loss (but the training loss gets bad in training). The training loss went down through training.

There is a relatively steep decline in training loss during the early epochs of training, after which it starts to level off beginning around epoch 8 and generally reaches a training loss of around 0.29.

The validation loss starts from a lower position of roughly 0.34 but continues a similar downward path due to the strong pretrained representations that BERT provides. During training the validation loss fluctuates quite a bit, but ends up, at the end of the last epoch at around 0.27 and stays below the training loss. This demonstrates that while the model clearly learns well from the training set, it is also keeping some kind of strong ability to broaden to the unknown data.

Because the training and validation losses are so near to one another, the model is well-regularized and does not overfit.The consistent decrease in both loss curves indicates that the optimization approach has adapted BERT for the classification task and this has been shown to converge. Coupled with the accuracy graph, this loss curve illustrates how reliable and successful BERT is at tackling text classification tasks. It makes sure that the model maintains a high degree of predictive performance based on contextual dependencies.

6. Experimental results and Discussion:

The process of using the BERT model is shown on the interface as a junction of research and applied use. Many of the BERT model features are abstracted on a GUI interface so that users can just engage in the task of categorizing text even if the back-end deep learning processes are more complex. The

goal is to offer the more advanced uses of BERT and artificial intelligence techniques in a way that technical users do not need deep machine learning knowledge to utilize it, e.g. journalists, and content moderators, and business analysts.

The most important aspect of Human-Computer Interaction (HCI), in regard to machine learning systems, is the deployment. If a model cannot be accessed and used by the intended end users, it is almost completely useless; and the above examples are a great demonstration of how NLP research can be common research documents that are transformed into interactive analysis programs to facilitate many different real-world implications.

- Dataset: 7,600 validation samples, and 120,000 training samples.
- Training: Loss steadily decreased, and validation loss followed suit.
- Accuracy: High, balanced performance by class.
- Classes: World, Sports, Business, Entertainment, Lifestyle, Sci/Tech, Politics

Fig 6.1 Home Page

The displayed Home page represents the front-end interface of a BERT-based text classifier, designed to provide a seamless way for users to interact with the model. It consists of a simple layout with a text box for entering input, an output box to display the predicted category, and control buttons such as Clear, Submit, and Flag to manage interactions. The clean and minimal design reflects a user-centered approach, making the deployment of advanced deep learning models both practical and accessible for real-time text classification.

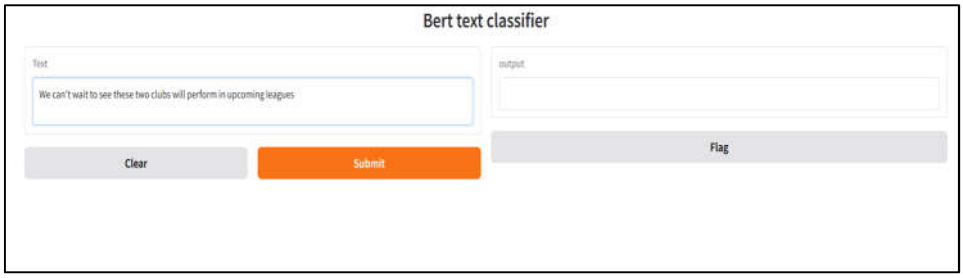

Figure 6.2 Input Page

The figure 3 shows what a BERT text classifier interface looks like, as well as the user input sentence that will be classified. When the user inputs "We can't wait to see which of these two clubs will perform the best in upcoming leagues" into the input box, it takes input from the user and utilizes the improved BERT model to analyze the input data and put forth a prediction for which class label will appear in the output box. The design showcases how the users may engage with the model, in near real-time, converting unstructured text into meaningful predictions, through a simple and intuitive interface. And this demonstrates the practical usefulness of the system and shows how intricate deep learning models.

we have described can be leveraged within real-world instances remote from the academic environment such as on news and sports data classification.

Figure 6.3 Output Page

The Figure 4 displays the functional output of the BERT text classifier interface whereby the system performed classification on an input sentence provided by the user. The user typed in the text of "We can't wait to see these two clubs will perform in upcoming leagues," thereby establishing the contextual meaning of the sentence that would ultimately be classified by the BERT text classifier model. Once the user submitted both the text's the user submits variables, by pressing the Submit button. The classifier can identify the the domain of the text it supplied it, and then ultimately the one sentence is classified under the domain of "Sports." This also serves as an example of how well the improved BERT model is able to identify semantic context and classify the user inputs into satisfactory and meaningful classes. The interface example demonstrates to the end user how machine learning can offer convoluted results in an easily understood example, by demonstrating how advanced natural language processing (NLP) models can be deployed as real-time applications.

7. Conclusion:

The BERT-based classifier's accuracy and overall performance measured strongly on the full range of metrics allowed for a robust classification performance for the news articles overall. Gradio deployment is a helpful method for model deployment, as the Gradio-based deployment of this research fostering the ease of model use in instantaneous was a definite advantage of the deployment option. Overall, the deployment, model enhancement, related evaluation, dataset acquisition and preparation were done for the BERT-based news article classifier. The research's paper future plans will include fidelity, improved interpretability and larger datasets.

8.Future enhancement:

Even if BERT has shown to be a strong text categorization architecture, there are several ways it will be advantageous to develop more flexible, effective, and useful systems in the future. One possibility is for the model to be resource-efficient and lighter as models like DistilBERT or TinyBERT seek to maintain BERTs context "awareness" while lowering computational costs. This will allow deployment on mobile and edge devices with limited hardware resource.

Multilingual and cross-lingual classification is another improved area as BERT could be extended for dialects and languages besides English. A single model could manage classification for texts of various

linguistic contexts via multilingual embeddings, making it useful for wide-reaching applications such as social media monitoring or multilingual news classification.

In addition, using explainable AI (XAI) techniques with BERT classification will mitigate the "black-box" risk that comes from using transformers. This enables users and domain experts to have confidence in the model, since they can gain an understanding and close the interpretability gap through attention visualization or feature importance mapping for predicted classifications.

Also, incorporating continual and active learning methods is another beneficial improvement. With continual learning, a model can learn from new, incoming data without being retrained from scratch and is valuable when it comes to evolving domains such as 24-hours news. Alternatively, with active learning, the learning model actively seeks feedback from the user to learn and adapt to the new data. This is an excellent tool when considering a medical or real-time social media social trend model.

Finally, modalities need to be combined with BERT - multimodal - when looking at sentences containing formats other than text such as images and audio. This will provide a deeper learning model that can improve classification accuracy (e.g., news articles with accompanying images; video with captions; etc.).

In summary, the future of BERT text classification improvements includes modelling that is faster, multilingual, interpretable, adaptive, and multimodal; and still ensures we and society stay true to delivering relevant academic research, and ultimately turns research into ongoing and real-world applications.

References:

[1] A. McCallum and K. Nigam, “A Comparison of Event Models for Naïve Bayes Text Classification,” in Proc. AAAI Workshop, 1998.

[2] A. Pilicita and E. Barra, “Using Models for Transformer in Text classification in Mobile Educational Applications,” IEEE Latin America Transactions, vol. 21, no. 6, pp. 730–736, Jun. 2023.

[3] A. Rashid, S. Ahmed, and N. Khan, “Email Spam Detection using Fine-tuned BERT Models,” IEEE Access, vol. 10, pp. 9511–9520, 2022.

[4] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, “Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification,” arXiv, Aug. 2019.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need,” in Proc. NeurIPS, pp. 5998–6008, 2017.

[6] B. Hamzaoui, D. Bouchiha, A. Bouziane, and N. Doumi, “BERT-BiLSTM Model for Hierarchical Text Classification,” in Proc. ICCS, 2023.

[7] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?” in Proc. CCL, 2019.

[8] I. Alagha, “Leveraging Knowledge-Based Features with Multilevel Attention Mechanisms for Short Arabic Text Classification,” IEEE Access, vol. 10, pp. 51908–51921, 2022

[9] J. Bai and X. Li, “Chinese Multilabel Short Text Classification Method Based on GAN and Pinyin Embedding,” IEEE Access, vol. 12, pp. 83323–83329, 2024.

[10] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” in Proc. EMNLP, Doha, Qatar, 2014.

[11] M. Muņikar, S. Shakya, and A. Shrestha, “Fine-Grained Sentiment Classification using BERT,” EasyChair Preprint 1762, 2019.

[12] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] T. Joachims, “Text Categorization with Support Vector Machines,” in Proc. ECML, pp. 137–142, 1998.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv:1301.3781, 2013;

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT: A Distilled Version of BERT,” arXiv preprint arXiv:1910.01108, 2019.

[16] V. Singh, V. Tibrewal, C. Verma, Y. R. Singh, T. Sinha, and V. K. Shrivastava, “A BERT Model-Based Sentiment Analysis on COVID-19 Tweets,” in Soft Computing: Theories and Applications, LNNS, vol. 425, Springer, Singapore, 2022, pp. 641–652.

[17] W. Wang, Y. Lu, and C. Zhai, “BERT-Based Aspect-Based Sentiment Analysis,” in Proc. COLING, Barcelona, Spain, 2020.

[18] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang, “Domain Adaptation with BERT-based Domain Classification and Data Selection,” in Proc. DeepLo Workshop, Deep Learning Approaches for Low-Resource NLP, Hong Kong, China, Nov. 2019.

[19] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained Models for Natural Language Processing: A Survey,” Science China Technological Sciences, vol. 63, no. 10, pp. 1872–1897, 2020.

[20] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in Proc. EMNLP, Doha, Qatar, 2014.

[21] Y. Yang, Z. Qi, and H. Zhang, “TabBERT: Pretrained Contextual Embedding for Structured Text Classification,” in Proc. ACL, 2020.