They provided methods for using BERT for classification tasks

A study of Chinese Text Classification based on a new type of BERT pre-training

Youyao Liu
Xi'an University of Posts and Telecommunication
School of Electronic Engineering
Xi'an, China
lyyao2002@xupt.edu.cn

Haimei Huang
Xi'an University of Posts and Telecommunication
School of Electronic Engineering
Xi'an, China
3395177679@qq.com

Jialei Gao
Xi'an University of Posts and Telecommunication
School of Electronic Engineering
Xi'an, China
gjl_fkl@163.com

Shihao Gai
Xi'an University of Posts and Telecommunication
School of Electronic Engineering
Xi'an, China
13201805502@163.com

2023 5th International Conference on Natural Language Processing (ICNLP) | 979-8-3503-0221-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICNLP58431.2023.00062

***Abstract*—Chinese Text Classification (TC) is the process of mapping text to a pre-given topics category. In recent years, it has been found that TC is mainly based on RNN and BERT, so the development of different novel pre-training applied to Chinese TC is described as based on BERT pre-training. BERT combined with convolutional neural network is proposed to extend the BERT-CNN model for the problem of lack of semantic knowledge of BERT to derive a good classification effect. The second RoBERTa model performs feature extraction and fusion to obtain the feature word vector as the text output vector, which can solve the problem of insufficient BERT extracted features. The BERT-BiGRU model, on the other hand, mainly avoids the increase in the number of texts leading to long training time and overfitting, and uses a simpler GRU bi-word network as the main network to fully extract the contextual information of Chinese texts and finally complete the classification of Chinese texts; at the same time, it makes an outlook and conclusion on the new pre-training model for Chinese TC.**

Keywords*—*Chinese TC, BERT model, RoBERTa, BERT-BiGRU

I. Introduction

The Internet provides a platform for providing information, but the vast amount of online textual information needs to be regulated, and proper Text Classification (TC) not only facilitates management, but also selects reading content based on the user's personal interests. Compared to English TC [1], Chinese text processing is relatively complex because there is no spacing between words in Chinese, and individual Chinese characters have weaker meanings than phrases. At the same time, the sheer volume and semantic complexity of textual information on the Internet makes the study of TC even more difficult. Therefore, the study of Chinese TC is a hot topic in current research.

According to current research, there are two main types of TC, namely applying traditional machine learning methods and building deep neural network model methods. Traditional machine learning methods [2] will extract the term frequency-inverse document frequency (TF-IDF) or bag-of-word structure for training when classifying, such as Support Vector Machines [3], Logistic Regression and XGBoost [4]. The basic processes are data acquisition, data preprocessing, feature extraction and model training and prediction. However, both TF-IDF and bag-of-word structures require manual lexicon construction, vocabulary counting and relevant order calculation, so both methods suffer from high complexity, poor interpretability and unclear semantics. With the use of deep learning in architectures such as convolutional neural networks, recurrent neural networks, graph neural networks, solutions to the TC represention problem are more efficient. One of the current models for solving the semantic complexity of text is the pre-trained method represented by BERT [5], which reduces dependence on large datasets and obtains the most comprehensive global and local feature sequences available, taking into account contextual information, to effectively solve the problem of multiple meanings of a word.

The core idea of the pre-training approach is to abandon random re-training of the model with a large corpus to obtain a generic language representation containing contextual information, and then fine-tune the corresponding downstream tasks. Pre-training techniques can obtain more generic information, and these linguistic representations of generic information restart downstream tasks, not only to obtain better performance but also to accelerate training downstream tasks. The emergence of a series of large corpus-based pre-training techniques such as GPT, BERT and XLNet has established the mainstream status of pre-training and fine-tuning models, which has led to a leap forward in pre-training techniques.

Text Convolutional Neural Network (TextCNN) was first proposed by Kim et al. in literature [6], which combines natural language processing with convolutional neural networks in TC models. TextCNN is very capable of acquiring features on text through one-dimensional convolution. TextCNN is poor in information extraction and insensitive to discourse order, so it is not suitable for use in long TCs; literature [7] uses multiple filters to construct a multi-channel TextCNN network structure to extract features from data in many ways, capturing more hidden text information; literature [8] uses graph convolutional neural networks for text content; literature [9] uses graph convolutional neural networks for multiple filters to construct a multi-channel TextCNN network structure, which extracts features from data in many aspects and captures more hidden text information; literature [10] proposes a TC model of recurrent neural network, but the RNN structure is a serial structure, and if the distance between two words is too long it will lead to the problem of gradient disappearance and gradient

979-8-3503-0221-9/23/$31.00 ©2023 IEEE
DOI 10.1109/ICNLP58431.2023.00062

explosion due to the lack of semantic learning ability; Liu et al [11] proposed a hierarchical model structure that can improve the quality of Chinese TC. The model uses both Long Short Term Memory (LSTM) and temporal convolutional networks (TCN) to extract contextual information in Chinese text in a sequential manner; Shao et al [12] proposed a new method for large-scale multi-label TC with a classification step in which the text content is first changed into a graphical structure, then features are extracted, using convolutional neural network, and finally the attention mechanism is used to derive the full functionality of the text. In subsequent studies, LSTM and Gate Recurrent Unit (GRU) models have been applied to natural language processing tasks to effectively alleviate gradient disappearance and burst problems; in literature [13], the Long Short Term Ordered Neurons Memory (ONLSTM) structure has been proposed, and the improvement over LSTM is that the additional layer structure can extract hierarchical information from the text.

Based on the above research, this paper outlines different Chinese TC approaches based on the BERT model, analyses the advantages and disadvantages of each model, and proposes the BERT-CNN model, which is obtained by combining and extending BERT and convolutional neural network, and can solve the problem of lack of semantic knowledge in text classification. To solve the problem of insufficient feature extraction in the BERT model, the RoBERTa model is used for feature extraction and fusion to obtain the feature word vector as the text output vector, which can be better generalized to downstream tasks. The BERT-BiGRU model, on the other hand, uses a simpler two-word GRU network as the main network to fully extract the contextual information of Chinese texts and finally complete the classification of Chinese texts.

II. Deep learning methods

A. BERT Structure

Among TC's deep learning methods, feedforward neural networks and recurrent neural networks can improve performance compared to traditional machine learning methods. The emergence of BERT, which can generate up and down word vectors, is an important turning point in the development of TC and other NLP technologies. The emergence of BERT, which generates contextual word vectors, was an important turning point in the development of TC and other NLP techniques. BERT acts as a pre-trained language model by continuously adjusting the model parameters so that the semantic features output by the model describe the essence of the language as closely as possible. The BERT model is trained on a large unlabeled corpus to obtain features that contain semantic information about the text. The initialization methods provided in the pre-training process not only help to improve the generalization ability of the model, but also accelerate the convergence of the model.

The BERT model differs from other word vector models in that its input is the sum of three vectors: a word vector, a sentence vector, and a position vector; the BERT model uses Word Embeddings to encode the word vector, Segment Embeddings to encode the sentence vector, and Position Embeddings to encode the position vector after the text is obtained. The input to the BERT encoder is a set of three vectors that incorporating semantic information from the full text.

BERT continues to use the Transformer architecture's encoder module, which provides a multi-layer bi-directional encoding capability based on the Transformer's encoder module, using the bi-directional Transformer encoding layer for text to feature extract information from text. The Self-Attention Mechanism and feedforward neural network form the core of the Encoder, which obtains a bi-directional representation of words by collating the relationships between words and words in the same sentence, with full knowledge of the contextual information of words. The structure of the BERT model is shown in Figure 1. Where $E_1, \cdots, E_n$ represents the input vector of the model, with a multilayer bidirectional Transformer feature extractor in the middle.$T_1, \cdots, T_n$ denotes the output vector of the model, which is used to obtain word vectors that can be applied in subsequent TCs.

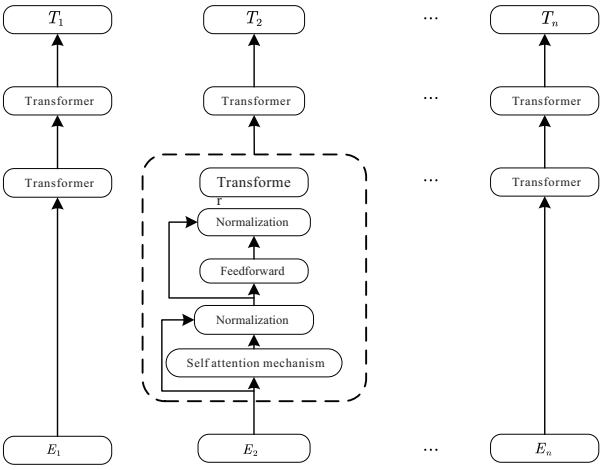

Fig. 1. BERT model structure diagram

B. LSTM Networks

Early pre-trained language models [14] aimed at training word embeddings, and while word vectors can capture the semantics of words, they are not perceptive of the corresponding contexts. As a result, the model could not fully grasp the concept of a text if the same text in different contexts expressed different syntax and semantics. With the development of computational power, scholars Melamud, Dagan and Goldberger [15] from BarIlan University have successfully demonstrated that on the basis of a large amount of unlabeled text data, it is possible to train to obtain context-informed word embeddings by incorporating contextual semantic information into word embeddings through a bidirectional LSTM approach for interpreting the relationship between word embeddings and language models. The structure of the LSTM network is shown in Figure 2.

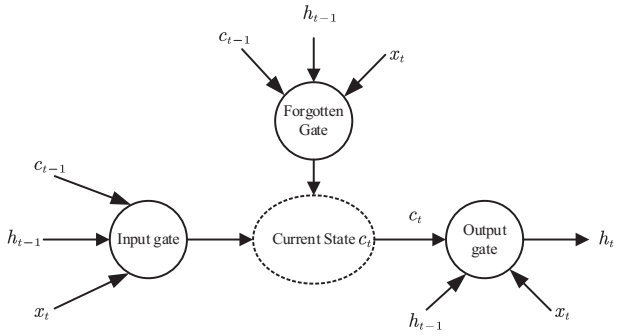

Fig. 2. LSTM Unit Structure Diagram

Researchers have attempted to use RNN-based methods for processing textual data precisely because text is processed into a longer sequence by taking full account of the connections between textual contexts. The LSTM[16] was proposed to address the shortcomings of RNN methods that exhibit gradients that tend to disperse. The LSTM consists of four components: an input gate, a cell unit, a forgetting gate and an output gate. The input vector to the information is processed, retaining important features in the text and removing irrelevant content. A basic LSTM model is constructed based on word vectors and long-short-term memory neural networks, where syntactic and semantic features of the user's problem are represented and then pre-processed for operation, and then transformed into vectors by Word2Vec as input to the LSTM model.

III. CHINESE TC BASED ON NEW-STYLE PRE-TRAINING

A. BERT-CNN Based Chinese TC

The CLS tags in the BERT output can achieve good classification results, but the rich semantic knowledge contained in BERT is not fully utilized, so a fusion CNN is proposed to extend BERT. After the BERT model receives the processed text, the text information passes through a two-layer Transformer mechanism [17] to obtain vector representation. The vectorized representation of the integrated semantic information at the output of the model is composed of individual word vectors, sentence vectors, and position vectors in the text, and then the output of the model is fed into a convolutional neural network that uses three different convolutional kernels to capture different feature information. The purpose of convolution kernels is to obtain the same feature map and different feature scales through feature alignment. After the convolution operation is completed, the feature matrices of the three channels are stitched together to obtain the dimensional vectors. The convolutional neural network model is then connected by a fully connected layer of word vector mapping to obtain further semantic information about the text. The structure of the convolutional neural network is shown in Figure 3.

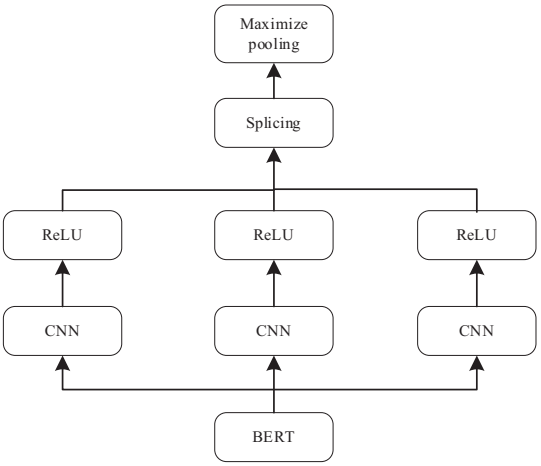

Fig. 3. BERT-CNN model structure

In the Chinese classification process of the BERT-CNN model, the Chinese text is first pre-processed with data, and then the dataset is input to the BERT model. The word vectors in the extracted dataset are used to extract the deep semantic information of the corresponding text using the extended neural network model to improve the generalization ability of the model, and finally the linear classifier is used to classify the obtained deep semantic information.

For the output of the modelh_i , use a linear classifier

$$classifier = f(h_i, w, b) = wb_i + b \tag{1}$$

$classifier$ The output is represented as a vector, with each column in the vector indicating the likelihood that each classification label is correctly classified, with the largest column indicating the classification result. For example, a vector$A = \{x_1, x_2, x_3, x_4\}$, assumes that$\max\{x_1, x_2, x_3, x_4\} = x_i$, the predicted result is therefore labeled i.

B. Chinese TC based on RoBERTa

The RoBERTa model is proposed to avoid the problems of not using full word coverage and insufficient feature extraction capability in the pre-training phase of BERT. Compared with BERT, RoBERTa uses a larger dataset, not only inherits BERT's bidirectional encoder, the input sentence representation is the sum of word vector, sentence vector and position vector, and the vectorized representation of the text is obtained through a multilayer bidirectional Transformer encoder. The RoBERTa model pre-trains the input Chinese text, and obtains the sentence vector and word vector containing contextual semantics. The word vectors are input into the DPCNN [18] feature extraction layer and the improved gated neural network respectively, and then the two feature word vectors are fused using the Attention Mechanism to obtain the feature word vectors containing deep structure and local information; the sentence vectors are fused with the word vectors to obtain the final semantic vector representation, and the final result is output after the softmax activation layer. The structure of the RoBERTa model is shown in Figure 4.

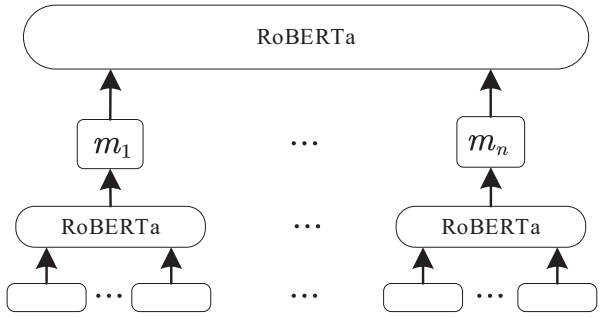

Fig. 4. Structure of the RoBERTa model

Suppose the input sentence is X, $X = \{x_1, x_2, \cdots, x_n\}$, and n represents the number of words in the sample sentence, the words are represented by codes with dimension k, and $Y = \{y_1, y_2 \cdots, y_n\}$ represents the word embedding matrix corresponding to X, and x_i. The corresponding vector is denoted as y_i. Q(Query), K(Key), V(Value) matrices can be obtained by training the model.d_k denotes the dimension size of the column vector in K, and thus the attention value is calculated as

$$A_{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

The projections of Q, K and V are made by h different linear transformations, and finally the combination of the different$A_{attention}$ results to obtain the multi-headed attention value of

$$M_{multihead}(Q, K, V) = C_{concat}(h_1, \cdots, h_n)W^0 \quad (3)$$

$$h_i = A_{attention}\left(QW_i^Q, KW_i^K, VK_i^V\right) \quad (4)$$

Where W_i^Q ,W_i^K ,K_i^V denote the Q, K, V weight matrices for the h, respectively.W^0 denotes the matrix of additional weights. $C_{concat}(\cdot)$ denotes the connection function.

At the same time, RoBERTa eliminates the BERT's NSP (Next Sentence Prediction) task and uses a larger dataset for training, which better represents the semantic and syntactic information of words and has a better representation of text vectors. roBERTa modifies the key hyperparameters in BERT to use a larger batch approach and learning rate for training sequences. The results show that RoBERTa can be generalized better to downstream tasks than BERT.

C. Chinese TC Based On BERT-BiGRU

In order to solve the problems of long training time and overfitting due to the increasing number of texts, a simple structured GRU neural network model was proposed [19]. Compared with the LSTM structure, the GRU neural network combines three selective pass units into two selective pass units, which are update gate and reset gate, and the model has fewer parameters and lower training overhead, which can adaptively capture the information transfer relationship between sequence data. The basic unit of the GRU neural network is shown in Figure 5.

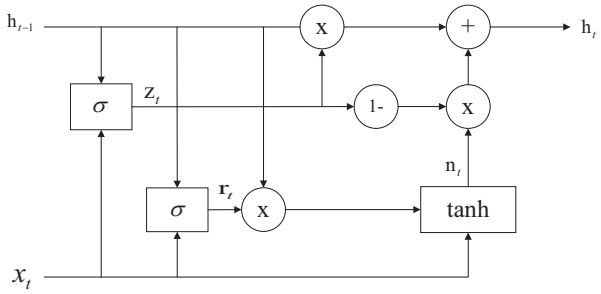

Fig. 5.GRU neural network unit diagram

The BERT model transforms the Chinese text into an augmented semantic vector after semantic characterization x_t by using the previous moment state information h_{t-1} and the current input x_t to obtain the updated gate state z_t and reset the gate state r_t The BERT model

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \quad (5)$$

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (6)$$

where: r_t is the reset gate output; and z_t is the update gate output. h_t is the candidate hidden state. W_{xr} and W_{xz} denote respectively the input x_t and the weight matrix between the reset gate and the update gate.W_{hr} and W_{hz} represents the weight matrices between h_{t-1} the weight matrices between the reset gate and the update gate. b_r and b_z are the corresponding biases. σ is the Sigmoid function that converts the calculation into a range of [0,1]. The closer the update gate is to 1, the more status information has been introduced from the previous moment. The closer the value of the reset gate is to 1 the less state information is written in the previous moment. The current state of information can be calculated h_t

$$n_t = \tanh\left(W_{xn}x_t + b_{xn} + r_t^{\circ}(W_{hn}h_{t-1} + b_{hn})\right) \quad (7)$$

$$h_t = (1 - z_t)^{\circ}n_t + z_t h_{t-1} \quad (8)$$

Where.°denotes the Hadamard product. W_{xn} , W_{hn} and b_{xn} are the weight matrix and bias respectively; the activation function tanh is used to scale the data to [-1,1]. GRU retains important information in the augmented semantic vector through a gating mechanism to ensure the effective transfer of contextual information. GRU is able to capture the overall features of the text excellently while reducing computational effort. In a single-layer GRU network the state is propagated unidirectionally, so a bidirectional GRU network is built to make the most of the textual contextual relationships. Bidirectional GRUs process the input text vectorized semantic representation features sequentially in the temporal dimension, in sequential and inverse order respectively, and stitch the GRU output of each word into the final output. The output node of each word thus contains the complete semantic information of the current word in both sequential above and inverse order

below, and the structure of the bidirectional GRU network model is shown in Figure 6.

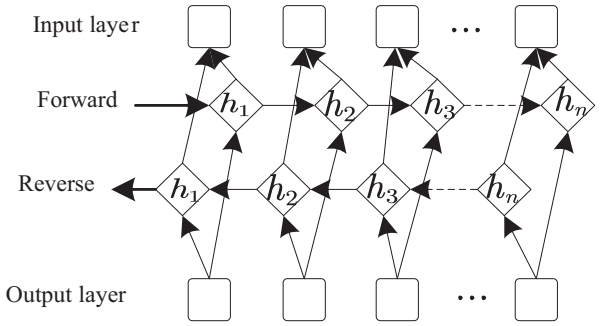

Fig. 6. Structure of the bidirectional GRU network model

BERT-BiGRU is further improved in recognition accuracy. The introduction of a two-way loop mechanism in BERT-BiGRU considers the connection between each word and its context in both sequential and inverse order, providing a better characterization of the logicality of Chinese text and a better description of the differences between different text categories.

IV. Development prospects

In TCs, methods incorporating deep learning in semantic comprehension literal meaning understanding far exceed human reading comprehension, and their ability to continuously improve accuracy metrics for most classification tasks, but cannot understand text at a semantic level as humans do. In response to the current high demand for individual models with large areas, long training cycle times and hardware configuration requirements, it is hoped that a language model that can be applied to each Chinese TC can be proposed. Secondly, the natural language still needs to adjust the parameters of the overall structure to improve the model effect, so that the model can better solve the tasks of Chinese TC in natural language processing.

V. Summary

This paper focuses on the comparative analysis of TC pre-training methods in the field of NLP. The pre-training model represented by BERT is constantly updated, and the analysis of Chinese TC based on three models, BERT-CNN, RoBERTa and BERT-BiGRU, solves the problems of long training time and overfitting phenomenon in the BERT base model, avoids the problem of insufficient feature extraction ability and enriches the knowledge of semantic information in the text. These three models use word embedding and multi-head attention mechanism to extract contextual information of words in Chinese web information text, and use various new types of pre-trained Chinese BERT models to complete the classification of Chinese web information text. However, in the face of more accurate extraction and classification Chinese text information, continuous optimization is still needed.

Reference

[1] Kowsari, Kamran, et al. "Text classification algorithms: a survey." information 10.4 (2019): 150.

[2] Kamath, Cannannore Nidhi, Syed Saqib Bukhari, and Andreas Dengel. "Comparative study between traditional machine learning and deep learning approaches for text classification." Proceedings of the ACM Symposium on Document Engineering 2018. 2018.

[3] Chen, Pai-Hsuen, Chih-Jen Lin, and Bernhard Schölkopf. "A tutorial on v-support vector machines." Applied Stochastic Models in Business and Industry 21.2 (2005): 111-136.

[4] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system. "Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[6] Kim, Taehoon, and Jihoon Yang. "Abstractive text classification using sequence-to-convolution neural networks." arXiv preprint arXiv:1805.07745 (2018).

[7] CHEN Ke, LIANG Bin, KE Wende, et al. "Chinese Weibo Sentiment Analysis Based on Multi-channel Convolutional Neural Network." Journal of Computer Research and Development, 2018, 55(5):945-957.

[8] Yao, Liang, Chengsheng Mao, and Yuan Luo. "Graph convolutional networks for text classification." Proceedings of the AAAI conference on artificial intelligence. vol. 33. no. 01. 2019.

[9] Linmei, Hu, et al. "Heterogeneous graph attention networks for semi-supervised short text classification." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

[10] Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang. "Recurrent neural network for text classification with multi-task learning." arXiv preprint arXiv:1605.05101 (2016).

[11] Liu, Jingang, et al. "Hierarchical comprehensive context modeling for Chinese text classification." IEEE Access 7 (2019): 154546-154559.

[12] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).

[13] Shen, Yikang, et al. "Ordered neurons: Integrating tree structures into recurrent neural networks." arXiv preprint arXiv:1810.09536 (2018).

[14] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063 (2019).

[15] Melamud, Oren, Jacob Goldberger, and Ido Dagan. "context2vec: learning generic context embedding with bidirectional lstm. "Proceedings of the 20th SIGNLL conference on computational natural language learning. 2016.

[16] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." neural computation 31.7 (2019): 1235-1270.

[17] Girdhar, Rohit, et al. "Video action transformer network." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[18] Li, Xuewei, and Hongyun Ning. "Deep Pyramid Convolutional Neural Network Integrated with Self-attention Mechanism and Highway Network for Text Classification." Journal of Physics: Conference Series. vol. 1642. no. 1. IOP Publishing, 2020.

[19] Dey, Rahul, and Fathi M. Salem. "Gate-variants of gated recurrent unit (GRU) neural networks." 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE, 201