2023 International Conference on Intelligent Management and Software Engineering (IMSE) | 979-8-3503-9414-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IMSE61332.2023.00042

Cyberbullying Detection Based on GECC-BiGRU

Xianbin Li[1,a]
[1] School of Information Engineering, China Jiliang University, Hangzhou 310018, China
[a]P21030854028@cjlu.edu.cn

Hangxia Zhou[2,b]
[2]School of Information Engineering, China Jiliang University, Hangzhou 310018, China
[b] zhx@cjlu.edu.cn

Chen Cui[3*]
[3]Big Data and Network Security Research Institute, Zhejiang Police College, Hangzhou 310053, China
*doublecwork@163.com

Abstract: **Aiming at the noise problem of cyberbullying text content and the limitations of traditional CNN models, we propose a cyberbullying detection model based on GECC-BiGRU. Among them, Gated Expanded Causal Convolution (GECC) not only captures the multi-granular semantic features of the text in a multi-scale perceptual field of view but also ensures that the convolution process of the model is carried out by the word order, which solves the problem of the traditional CNN model that has a small perceptual field of view and ignores the word order. Meanwhile, the bidirectional gated recursive unit (BiGRU) can be used to extract long-term dependencies in the text and adaptively focus on key features in the text using the attention mechanism. Experiments show that our method can effectively detect cyberbullying texts. In addition, the proposed method achieves the best performance on both Chinese and English datasets compared to other cyberbullying detection models.**

Keywords: Cyberbullying detection, gated expansion causal convolution, bidirectional gated recurrent units

I. INTRODUCTION

Cyberbullying[1] involves individuals or groups using social networks to send malicious behavior that attacks, harasses, threatens, or humiliates others. Adolescents who suffer from cyberbullying often develop serious psychological and behavioral problems, including depression, anxiety, fear, difficulty concentrating on their studies, and even suicidal tendencies[2]. Therefore, accurate detection of cyberbullying in social network comments has become an urgent problem.

In recent years, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been widely used for cyberbullying text detection. Kargutkar[3]found that CNNs could extract local text features. Sun[4] proposed that incorporating a gating mechanism in CNNs can improve effective information delivery. Dholvan[5]proposed that the CNN model ignores the temporal order of the text and therefore uses a temporal convolutional network (TCN) model. Although the CNN and its variants are effective in extracting local context sequence features, they are deficient in capturing global context sequence features. On the other hand, RNN show a strong ability to extract global context sequence features, and the effectiveness of LSTM[6] and GRU[7] in cyberbullying text detection is demonstrated. However, it is difficult for RNN and its variants to capture subtle semantic features. Therefore, Badri[8] proposed the CNN-BiGRU model, which utilizes CNN to extract semantic features from text and then captures sequential features via BiGRU. Some researchers have found that adding an attention mechanism can enhance the feature extraction ability of the model. NERGIZ[9] proposed the GCA model, which utilizes an attention mechanism to identify key information to improve the detection ability of the model. Kumar[10] proposed the Bi-GAC model, which replaces the CNN in GCA and utilizes the dynamic routing of Capsnet[11] to extract semantic features of the text. However, these models' structure has some limitations:(1) CNN doesn't strictly follow the word order in the convolution process and cannot capture the semantic relationship between words. (2) When extracting text features, CNN is limited by the size and manner of the convolution kernel, resulting in the loss of some edge information within the limited sensory field.

To address the above problems, we create a Chinese Weibo dataset specialized for cyberbullying text detection and propose a detection method based on GECC-BiGRU. The main contributions of this study are as follows:

- The GECC model is proposed, which realizes the extraction of local features under different sensory horizons at multiple scales while ensuring that the model strictly follows the word order for feature extraction.
- Using gating mechanisms, input data can be selectively filtered or ignored, thus improving the model's ability to convey relevant and valid information.
- Constructed a new Chinese dataset specialized for cyberbullying detection.

II. METHODOLOGY

Extended causal convolution as a variant of CNN effectively combines dilation convolution and causal convolution. Expansion convolution expands the convolution field by different expansion rates; causal convolution ensures that the model strictly follows the word order, namely the semantic information of the text, during the convolution process. Meanwhile, there is some useless information in the text, and the addition of the gating mechanism can improve the delivery rate of effective information. Therefore, we adopt the GECC structure as a local feature extractor. GRU can effectively solve the long-term dependency problem of RNN by learning the long-term and short-term dependencies of local features. However, both the convolution process and the recurrent neural network transmission process are unidirectional, ignoring the effect of reverse text information on the detection results. Therefore, we use BiGRU to capture

979-8-3503-9414-6/23/$31.00 ©2023 IEEE
DOI 10.1109/IMSE61332.2023.00042

the long-term dependencies in the text. The cyberbullying detection model based on GECC-BiGRU includes the Input layer, GECC layer, BiGRU layer, Attention layer, and Output layer. The detection model is shown in **Fig. 1**.

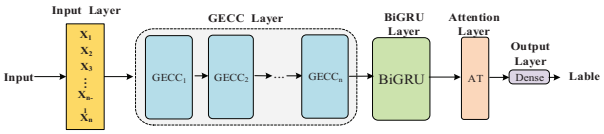

Fig. 1 Structure diagram of the GECC-BiGRU model

A. Input layer

The sentence sequence is passed through a pre-trained word vector model to generate the corresponding word vector matrix $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...\boldsymbol{x}_m]$. In this matrix, m represents the number of words in the sequence.

B. GECC layer

Expanded causal convolution is a variant of CNN that effectively combines causal and expanded convolution. Causal convolution strictly controls the order of the convolution process and can ensure that the convolution of the model depends on the output of the previous moment. Extended convolution increases the perceptual field of view of the model by varying the dilation rate. This improvement improves the feature extraction capability of the model and reduces the loss of edge information.

To cope with the problem of noisy and progressively ambiguous content in cyberbullying texts, we propose to utilize two layers of gated extended causal convolutions with different magnitudes of expansion rates (see **Fig. 2**). The first gated extended causal convolution extracts text features that may contain cyberbullying, with an expansion rate of 2^i; the second gated extended causal convolution performs deep feature extraction, using the information from the first layer to determine which features contain cyberbullying content, with an expansion rate of 2^{i+1}. This allows the model to capture the text features of cyberbullying texts more effectively.

To enhance the robustness and performance of the gated dilation causal convolution model, we use three identical gated dilation causal convolutions in each layer and employ an average pooling strategy to improve the performance of the model. In addition, we address the problem of information loss during model overlay by introducing residual connections.

$$\boldsymbol{O}_n^{\mathrm{j}}=(\boldsymbol{X}^{n-1}*_{\boldsymbol{d}}\boldsymbol{k})(\boldsymbol{s})*\sigma(\boldsymbol{X}^{n-1}*_{\boldsymbol{d}}\boldsymbol{k})(\boldsymbol{s}),1\le \mathrm{j}\le 3 \quad (1)$$

$$\boldsymbol{P}_n=\frac{1}{3}(\boldsymbol{O}_n^1+\boldsymbol{O}_n^2+\boldsymbol{O}_n^3) \quad (2)$$

$$\boldsymbol{T}_n^{\mathrm{j}}=(\boldsymbol{P}_n*_{\boldsymbol{d}}\boldsymbol{k})(\boldsymbol{s})*\sigma(\boldsymbol{P}_n*_{\boldsymbol{d}}\boldsymbol{k})(\boldsymbol{s}),1\le \mathrm{j}\le 3 \quad (3)$$

$$\boldsymbol{S}_n=\frac{1}{3}(\boldsymbol{T}_n^1+\boldsymbol{T}_n^2+\boldsymbol{T}_n^3) \quad (4)$$

$$\boldsymbol{X}^{\boldsymbol{n}}=\boldsymbol{S}_n+\boldsymbol{X}^{\boldsymbol{n}-1} \quad (5)$$

where n is the nth GECC unit, $\boldsymbol{X}^{n-1}$ is the input to the model, k is the convolutional kernel size, d is the expansion rate, $*$ is the convolution operation, $\sigma(\bullet)$ is the sigmoid function operation, $\boldsymbol{O}_n^j$ and $\boldsymbol{T}_n^j$ are the outputs of each gated convolutional neural network in the first and second layers, respectively. $\boldsymbol{P}_n$ and $\boldsymbol{S}_n$ are the average pooled outputs and the total output of the GECC unit.

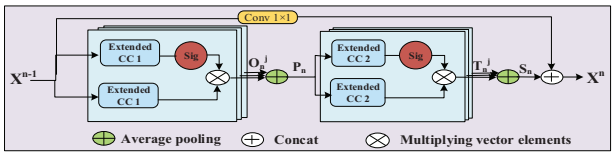

Fig. 2 GECC unit structure diagram

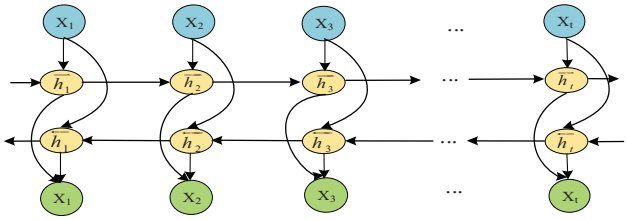

Fig. 3 BiGRU structure diagram

C. BiGRU Layer

The BiGRU model (see **Fig. 3**) consists of two different directional GRUs, each of which processes the input sequence independently, and finally combines the outputs of the two directions to obtain a state that contains information about the previous time step and the future time step. The output of the GECC model is fed into the BiGRU to obtain the global context feature $\boldsymbol{H}$.

$$\vec{\boldsymbol{h}}_t=\overrightarrow{\boldsymbol{GRU}}(\boldsymbol{X}_t^n,\vec{\boldsymbol{h}}_{t-1}) \quad (6)$$

$$\overleftarrow{\boldsymbol{h}}_t=\overleftarrow{\boldsymbol{GRU}}(\boldsymbol{X}_t^n,\overleftarrow{\boldsymbol{h}}_{t-1}) \quad (7)$$

$$\boldsymbol{h}=[\vec{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t] \quad (8)$$

$$\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,...,\boldsymbol{h}_t\} \quad (9)$$

D. Attention layer

Among the features extracted by the BiGRU model, not all features contribute equally to the detection process. Therefore, we introduce an attention mechanism that dynamically assigns weights based on the relevance of the extracted features to the detection results. The output characteristics of the BiGRU model $\boldsymbol{H}$ are obtained through the attention mechanism by using the $\boldsymbol{V}$.

$$\boldsymbol{Score}_i=\frac{\exp(\boldsymbol{e}_i^{\,T}\boldsymbol{u}_s)}{\sum_i\exp(\boldsymbol{e}_i^{\,T}\boldsymbol{u}_s)} \quad (10)$$

$$\boldsymbol{e}_i=\tanh(\boldsymbol{W}_s\boldsymbol{h}_i+\boldsymbol{b}) \quad (11)$$

$$\boldsymbol{V}=\sum_i\boldsymbol{Score}_i\boldsymbol{h}_i \quad (12)$$

Where $\boldsymbol{W}_s$ and $\boldsymbol{u}_s$ are the weight matrices, $\boldsymbol{e}_i$ is the correlation, $\boldsymbol{h}_i$ is the output feature of BiGRU, $\boldsymbol{Score}_i$ is the attention weight and b is the bias term.

E. Output layer

The Softmax function takes the output V from the attention mechanism to generate the detection results.

III. EXPERIMENTS AND RESULTS

A. Dataset

The datasets used in the experiments include a public Twitter dataset obtained from GitHub and a newly constructed Chinese Weibo dataset specialized in detecting cyberbullying texts. The Twitter dataset has 4,817 data on cyberbullying and 6,037 data on non-cyberbullying; the Weibo dataset has 3,000 data on cyberbullying and 3,000 data on non-cyberbullying. Sample datasets are shown in **Table 1**.

Table 1. Example sample of the datasets

Twitter dataset		Weibo dataset	
Text	Label	Text	Label
I didn't think there were any of those people left...	0	用蛆来形容你真的一点都不过分	1
My job gives me a constant stream of endorphins.	0	很高兴以这种方式认识你	0
These girls are fucktards!!	0	长的跟猪一样	1
Stick to your day jobs girls	1	这是哪里我在干什么⊙?⊙?	0

The Weibo dataset collects comments by crawling data from Chinese Weibo. The data collection focused on comments related to events involving 15 celebrities with damaged reputations and trending topics. The collected data were screened and the bit of cyberbullying text was judged based on the following criteria: (1) Containing sexism, racial or geographical discrimination. (2) Using swear words or humiliating someone. (3) Expressing violent tendencies or curses towards others. (4) Attacks on another person's appearance, body, or family members.

B. Assessment indicators

Accuracy and F1-score were used as evaluation indicators, and the formula was calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$F1 - Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{14}$$

C. Comparative experiments and analysis

To verify the model classification effect, it is compared with CNN[6], MGCNN[4], TCN[5], CNN-BiGRU[8], GCA[9], and Bi-GAC[10], and the experimental results are shown in **Table 2**.

Table 2. Comparative experimental results of models with Chinese and English datasets

Model	Weibo dataset		Twitter dataset	
	Accuracy(%)	F1-Score (%)	Accuracy(%)	F1-Score (%)
CNN	74.67	74.61	82.40	82.03
MGCNN	78.08	78.11	82.73	82.42
TCN	78.58	78.58	83.51	82.91
CNN_BiGRU	80.00	80.00	83.97	83.70
TCN-BiGRU	81.42	81.39	84.39	83.92
GCA	80.44	80.44	84.39	83.92
Bi-GAC	81.40	81.40	84.02	83.44
Proposed	**83.33**	**83.31**	**85.67**	**85.34**

From **Table 2**, it can be seen that the accuracy and F1 score of our method is optimal on both datasets. Compared with CNN, MGCNN, and TCN which have the highest accuracy in these two datasets, the accuracy is improved by 4.75% and 2.11%, and the F1 is improved by 4.75% and 2.15%, respectively. This indicates that the combined model based on CNN and RNN can more fully mine the information features of the text and make the detection results more accurate. Compared with the CNN-BiGRU and TCN-BiGRU models, the attentional mechanism plays a significant role and the accuracy is improved by 2.41% and 1.28%, respectively. Meanwhile, the accuracy is improved by 1.93% and 1.28% compared to the models with the highest accuracy in the two datasets with GCA and Bi-GAC, which proves that GECC can extract local features under a larger sensory field of view when extracting local features. The accuracy improvement of the model is not significant due to the class imbalance problem in the Twitter dataset, but the method proposed in this paper still achieves the highest accuracy.

D. Comparative experiments and analysis

To verify the effect of each layer in the GECC-BiGRU model on the classification effect, ablation experiments were conducted, and the experimental results are shown in **Table 3**.

- NoGECC: Remove the GECC layer.
- NoGate: Remove the gating mechanism in the GECC layer.
- NoBiGRU: Remove the BiGRU layer. Use GECC for feature extraction.
- NoAtt: Optimizing the output without Attention.

- OneLayer: Use only one layer of gated dilation causal convolution.
- ThreeLayer: uses three layers of gated expanded causal convolution.

Table 3. Results of ablation experiments

Model	Weibo dataset		Twitter dataset	
	Accuracy(%)	F1-Score(%)	Accuracy(%)	F1-Score(%)
NoGECC	80.58	80.58	84.02	83.50
NoGate	81.92	81.91	84.29	83.63
NoBiGRU	82.17	82.12	84.39	84.05
NoAtt	81.75	81.75	84.34	83.92
OneLayer	82.25	82.25	85.08	84.73
ThreeLayer	82.00	82.00	84.66	84.44
Proposed	**83.33**	**83.31**	**85.67**	**85.34**

From **Table 3**, the following conclusions can be drawn: ① The method in this paper improves the accuracy by 1.16% and 1.38% compared to the two models with the highest accuracy in the two datasets, NoGECC, NoGate, NoBiGRU, and NoAtt, respectively. It shows that all modules in our method play a role, with the GECC having the greatest impact on the model. ② Compared with the OneLayer and ThreeLayer methods that have the highest accuracy in the two datasets, the accuracy is improved by 1.08% and 0.59%, respectively, indicating that building a two-layer GECC improves the performance of the model and extracts the most effective information.

IV. CONCLUSION

To address the problems of noisy text content of cyberbullying and the limitations of traditional CNN models, this paper proposes a GECC-BiGRU-based cyberbullying detection method. The method captures the semantic features of text at multiple granularities under multi-scale perceptual view by the GECC model, then extracts the long-range dependencies in the text by BiGRU, and finally assigns different weights by attention, to extract key features for final classification. The experiments show that the method in this paper outperforms the benchmark model, but the accuracy of the model is not high due to the small size of the dataset and the valid source of the data, so in the future, we will enhance and add several different domains for the data to improve the generalization ability of the model.

ACKNOWLEDGMENTS

This work was mainly supported by the Open Project of the Key Laboratory of Informatization of the Ministry of Public Security Based on Big Data Architecture (2021DSJSYS004) and the Public Welfare Technology and Industry Project of Zhejiang Provincial Science Technology Department (LGF21F020006).

REFERENCES

[1] Macaulay P J R, Betts L R, Stiller J, et al. Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. Computers in Human Behavior,113, 107238(2022).

[2] Pabian, Sara. An investigation of the effectiveness and determinants of seeking support among adolescent victims of cyberbullying. The Social Science Journal 56(4), 480-491(2019).

[3] S. M. Kargutkar and V. Chitre. A Study of Cyberbullying Detection Using Machine Learning Technique. 2020 Fourth International Conference on Computing Methodologies and Communi-cation (ICCMC), pp. 734-739, IEEE, Erode, India (2020).

[4] Sun, J., Jin, R., Ma, X., Park, Jy., Sohn, Ka., Chung, Ts. Gated Convolutional Neural Networks for Text Classification. In: Park, J.J., Fong, S.J., Pan, Y., Sung, Y. (eds) Advances in Computer Science and Ubiquitous Computing. Lecture Notes in Electrical Engineering, vol 715, pp. 309-316. Springer, Singapore (2021).

[5] Dholvan M, Bhuvanagiri A K, Bathina S M, et al. Offensive text detection using temporal convolutional networks. Int. J. Adv. Sci. Technol 29(6), 5177-5185(2020).

[6] Paruchuri V L, Rajesh P. CyberNet: a hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks. Evolutionary Intelligence (2022), 1-15(2022).

[7] Iwendi, C., Srivastava, G., Khan, S. et al. Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems 29, 1839–1852 (2023).

[8] Badri, N., Kboubi, F., Habacha Chaibi, A. (2022). Towards Automatic Detection of Inappropriate Content in Multi-dialectic Arabic Text. In: Bădică, C., Treur, J., Benslimane, D., Hnatkowska, B., Krótkiewicz, M. (eds) Advances in Computational Collective Intelligence. ICCCI 2022. Communications in Computer and Information Science(CCIS), vol 1653, pp. 84-100. Springer (2022).

[9] NERGİZ G, AVAROĞLU E. Türkçe Sosyal Medya Yorumlarındaki Siber Zorbalığın Derin Öğrenme ile Tespiti. Avrupa Bilim ve Teknoloji Dergisi (31), 77-84(2021).

[10] Kumar A, Sachdeva N. A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web 25(4), 1537-1550(2022).

[11] Kumar, A., Sachdeva, N. Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimedia Systems 28, 2043–2052 (2022).