Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Potential cyberbullying detection in social media platforms based on a multi-task learning framework

Guo Xingyi[a] and Hamedi Mohd Adnan[a*]

[a] *Department of Media and Communication Studies，Faculty of Arts and Social Sciences，Universiti Malaya, Malaysia*

C H R O N I C L E

Article history:
Received: July 10, 2023
Received in revised format: September 20, 2023
Accepted: October 20, 2023
Available online: October 20, 2023

Keywords:
Cyberbullying comment detection
Sentiment analysis
BERT
Multi-task learning
Attentional mechanism

A B S T R A C T

The proliferation of online violence has given rise to a spate of malignant incidents, necessitating a renewed focus on the identification of cyberbullying comments. Text classification lies at the heart of efforts to tackle this pernicious problem. The identification of cyberbullying comments presents unique challenges that call for innovative solutions. In contrast to traditional text classification tasks, cyberbullying comments are often accompanied by subtle and arbitrary expressions that can confound even the most sophisticated classification networks, resulting in low recognition accuracy and effectiveness. To address this challenge, a novel approach is proposed that leverages the BERT pre-training model for word embedding to retain the hidden semantic information in the text. Building on this foundation, the BiSRU++ model which combines attentional mechanisms is used to further extract contextual features of comments. A multi-task learning framework is employed for joint training of sentiment analysis and cyberbullying detection to improve the model's classification accuracy and generalization ability. The proposed model is no longer entirely reliant on a sensitive word dictionary, and experimental results demonstrate its ability to better understand semantic information compared to traditional models, facilitating the identification of potential online cyberbullying comments.

1. Introduction

The advent of the Internet and smartphones have ushered in a new era of social engagement, enabling people to connect with one another anytime and anywhere (Enke & Borchers, 2021). The convenience and anonymity of online communication have made it easier for social media users to express their emotions and, unfortunately, to attack others on the web. In the context of news that has a significant impact on public opinion, some individuals may follow the trend and post irrational feedback, or even launch a collective attack on a specific user, resulting in online violence (Castaño-Pulgarín et al., 2021). Compared to the social pressure exerted by traditional media in the past, cyberbullying has reached unprecedented levels, spilling over from virtual space into the real world and having a pernicious impact on social life. According to statistics from the Wikimedia Foundation in 2017, 54 percent of netizens have experienced online harassment firsthand (Wulczyn et al., 2017). A study by McAfee found 87 percent of teenagers have encountered information about online violence while browsing the internet (Mohammad, 2018). Women are particularly vulnerable to the negative impact of social media, as highlighted in (Chadha et al., 2020). Women aged 18 to 29 are more than twice as likely as men of the same age to experience sexual harassment online and are also more likely to receive physical insults and persistent harassing messages online (Lindsay et al., 2016). In the realm of digital interactions, young women in their late teens and twenties are often confronted with elevated frequencies of

* Corresponding author.
E-mail address: hamedi@um.edu.my (H. M. Adnan)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.10.021

online harassment across dating platforms and social media sites. Such forms of harassment encompass unsolicited explicit messages, derogatory comments, persistent pursuit, as well as focused and malicious targeting based on one's physical appearance or personal attributes (Valenzuela-Garcí a et al., 2023). These instances of online mistreatment not only impinge upon the well-being and safety of these individuals, but also underscore the pressing need for robust measures and safeguards to mitigate such pervasive challenges. The impact of online harassment on women is a critical matter that demands greater attention from policymakers and the wider public alike.

Online violence is a growing scourge that can manifest in numerous forms, from direct verbal attacks to reputation damage against the parties involved (Cobbe, 2020). Despite the often false and defamatory nature of such accusations, the effects can be irreparable, seriously damaging the livelihoods of victims. Moreover, in the era of digital lives, personal information is easily disclosed online, and the lack of protection of individuals' right to privacy can exacerbate the harm caused by cyberbullying. Even in cases where online verbal harassment or personal information disclosure has not yet infringed on the real lives of the parties involved, many may opt to block the internet and retreat to the safety of their physical lives. However, online violence can escalate and spill over from the virtual to the physical world, with the aid of behind-the-scenes pushers who orchestrate the harassment of not only the victims but also their relatives and friends (Brighi et al., 2019). This can have a pernicious effect on their daily lives, causing untold harm and distress.

As social networks have become a ubiquitous medium for online communication, the sheer volume of commenting information has rendered manual screening of malicious comments inadequate to foster a healthy online ecosystem. Advanced natural language processing (NLP) models are required to achieve automatic detection of such comments (Jahan & Oussalah, 2023). Traditional lexicon-based methods for identifying inappropriate comments have relied heavily on sensitive word dictionaries. However, if these dictionaries cannot be updated in a timely manner, the recall rate for identifying inappropriate feedback will continue to decrease. Additionally, with the rise of anime culture, various words are given new meanings, and some common words can become a means of attacking or insulting others in specific contexts (Ahn & Yoon, 2020). Traditional methods are insufficient to effectively identify such new inappropriate words in context. For example, seemingly harmless comments such as "Nobody cares about what you have to say, you're irrelevant" and "You're a waste of space, nobody would care if you were gone" do not contain explicit attack words, but can nevertheless be used to hurt or intimidate others.

Cyberbullying can take many forms, and any behavior intended to harm or intimidate another person online is unacceptable. In a broader context, online violence refers to acts of violence carried out by internet users, manifesting social violence within the realm of the internet. In a narrower sense, online violence represents acts of soft violence directed towards individuals, empowered using online media, inflicting profound emotional distress upon the victims (Kiritchenko et al., 2021). The specific manifestations of online violence encompass the dissemination of rumors, engaging in “doxing” to violate personal rights, and the practice of "media trials" based on moral judgments as the highest standard (Eckert & Metzger-Riftkin, 2020). The channels of online violence can be further categorized into various aspects, such as mobile communication, video platforms, and email (Jones, 2013). Overall, previous studies examining the manifestations of online violence have identified common terms such as threats, harassment, and rumors as frequently encountered phenomena.

To address the difficulties, the BERT pre-training model is leveraged for word embedding to preserve the hidden semantic information in text. The convolutional layer is utilized to capture multi-scale local semantic features of the text, while the bidirectional simple recurrent unit (Bi-SRU++) with built-in attentional mechanism is used for context semantic modeling at different levels. The output features obtained from different scales are concatenated, and the attentional mechanism assigns higher weights to key features that contribute more to the classification results. Moreover, a multi-task learning (MTL) framework is introduced for joint training of sentiment analysis and cyberbullying detection, to enhance the model's classification accuracy and generalization ability. This reduces the model's dependence on sensitive words in detecting online violence comments. By adopting this approach, the model can effectively capture the nuances and implicit meanings of comments, enabling it to identify cyberbullying in its various forms.

2. Related Work

Malicious comment identification is a crucial task in text classification, a cornerstone of NLP that automatically categorizes texts into predefined groups. Text classification has found widespread use in recommendation systems, spam email filtering, and other domains. Traditional methods of text classification have largely revolved around rule-based and statistical-based models, such as sentiment-based classification or Naive Bayes classifier (Hartmann et al., 2019). Support vector machines (SVMs) trained on local, emotional, and contextual features of comments were employed to enhance the detection of online harassment (Yin et al., 2019). SVMs based on the features of comment content, malicious language, and user features were constructed to determine the malicious intent of a comment (Dadvar et al., 2013). In another study, various Machine Learning (ML) methods were used to identify potential malicious language in online comments (Muneer & Fati, 2020). In addition, syntactic features of vocabulary were extracted to identify potentially malicious words in online comments (Chen et al., 2012).

While rule-based and statistical-based methods have been widely used in text classification, they are limited in their ability to account for the complex connections between words and sentences. Traditional sentiment-based classification methods, for example, often rely on calculating the sentiment polarity of emotional words that appear in the text to determine whether it qualifies as online violence. However, the text data of today is no longer confined to obvious sensitive words, and the meaning of a sentence is often obscured in deeper expressions (Ma et al., 2020). As such, it is crucial to consider the context semantics of text when classifying it, rather than merely looking at individual words in isolation. Despite their success in various fields, traditional classification methods are ill-suited to the demands of modern text data.

To enhance model stability under complex scenes, researchers have increasingly turned to deep learning (DL) models. For instance, an ensemble of LSTM-based classifiers was employed, with each classifier receiving vectorized tweets and behavioral features as input (Pitsilis et al., 2018). These inputs encompassed user-associated features, such as their inclination towards discriminatory or biased behavior. To model the input tweets, they used word-based frequency vectorization, which indexes the occurrence frequency of words or phrases in a specific corpus and uses the index value of each word or phrase in the tweet as one of the vector elements describing the tweet. This model was evaluated on a labeled public dataset and demonstrated superior classification quality. Similarly, statistical topic modeling techniques were leveraged to preprocess the original tweet corpus using lexical collocation of offensive language. They then utilized a reliable dictionary in a semi-supervised ML framework to detect potentially offensive tweets (Xiang et al., 2012). On the other hand, a Convolutional Neural Networks (CNN) model was designed to automatically identify and classify malicious online comments into six overlapping subcategories. By leveraging the power of DL models, significant strides have been made in the detection of online violence and cyberbullying.

In the early days of NLP studies, one-hot encoding was a common technique used for text content encoding. However, it fails to account for word order and contextual continuity relationships. Furthermore, when the dictionary is too large, high dimensionality can be a problem. To address these shortcomings, the Word2Vec model which considers contextual word order relationships was proposed, supporting the calculation of word vector similarities after training, thus taking into account contextual relationships in the form of word order (Mikolov, 2013). However, it still struggles to associate the semantic relationship between words. FastText, on the other hand, uses all words and their corresponding N-gram features as input to learn a vector representation that can represent the sentence and then predicts the sentence category. Its advantage is that it does not require pre-trained word vectors and considers word order information through N-gram features (Young & Rusli, 2019). More recently, The BERT model was proposed which pre-trains on a large scale of unmarked text based on Transformer models (Devlin et al., 2019). Its word embedding contains more text information, enabling BERT to make word vectors that are semantically close or related to each other closer in space, while word vectors that are semantically opposite have a symmetrical distribution. This approach unlocks a wealth of text information, making it possible to boost the efficiency of NLP models. More specifically, BERT was utilized to detect hate speech on Twitter (Zhu et al., 2019). They used the original BERT to split sentences into word units and then performed word embedding operations in order.

The use of attentional mechanisms has emerged as an effective way for selecting important information and achieving better results. By combining attentional mechanisms with neural networks, researchers are unlocking new levels of performance in NLP. For instance, a dual attention mechanism was designed, and the STCKA model, a short text classification model with knowledge-driven attention was proposed (Chen et al., 2019). It enables the model to identify and focus on salient information, leading to superior classification accuracy. Similarly, by combining attention mechanisms, CNN, and BiLSTM, the AC-BiLSTM text classification model was proposed (Liu & Guo, 2019). The model has been shown to effectively improve text classification performance.

3. BERT-based Cyberbullying Detection Framework

In this paper, a cyberbullying speech recognition framework based on BERT and MTL is proposed, as shown in Fig. 1. The framework process mainly involves word vector transformation, deep text feature extraction and classification. The word vector transformation module is a pre-trained model based on BERT, which converts text into word vectors and performs preliminary feature processing. The feature extraction module is responsible for extracting contextual information from different levels of text and obtaining the contribution degree of each feature to the classification results through an attention mechanism, assigning higher weights to key features. At the same time, an MTL framework is designed, introducing a text sentiment classification task as an auxiliary source, and jointly training the cyberbullying speech classification model to optimize detection accuracy and generalization ability. To identify potential cyberbullying comments that do not contain obvious vulgar words, features are extracted based on semantic associations, and the semantic similarity is determined by the relative positions of various types of word vectors in the space of the pre-trained model. Ambiguous vocabulary is trained in different contexts to enable the model to judge whether it is offensive based on the context.

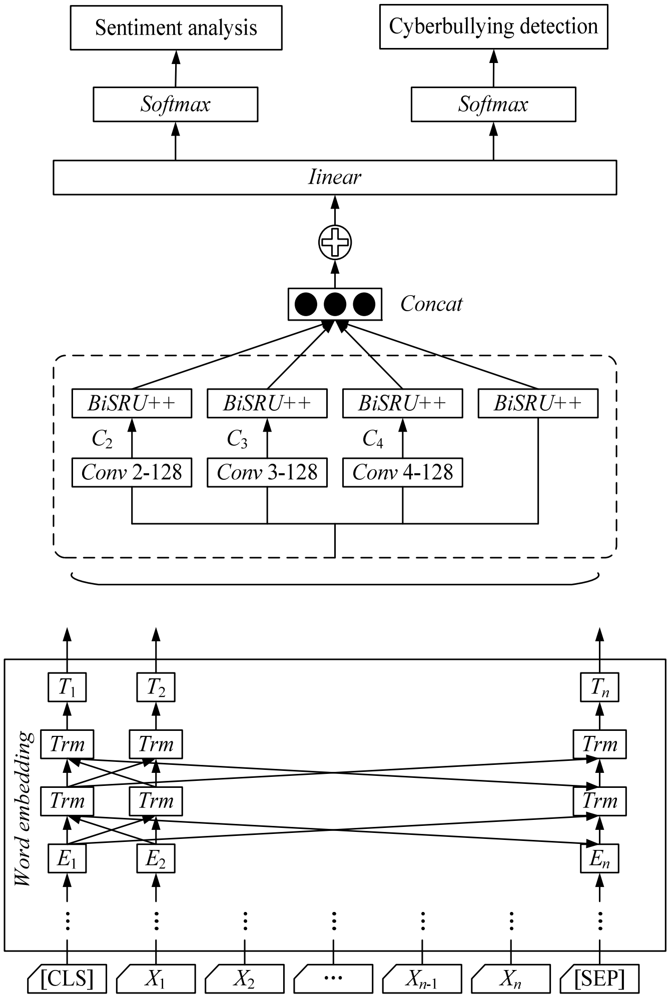


Fig. 1. BERT-based cyberbullying detection framework

3.1. Preliminary feature extraction

The preliminary feature extraction process involves transforming the text into word vectors composed of Token Embeddings, Segment Embeddings, and Position Embeddings.

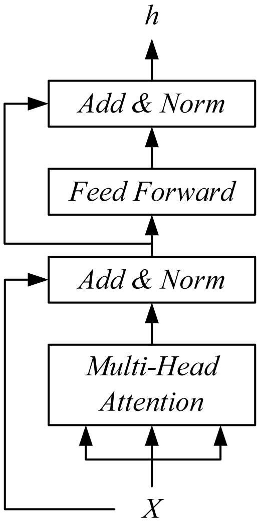


Fig. 2. Transformer structure

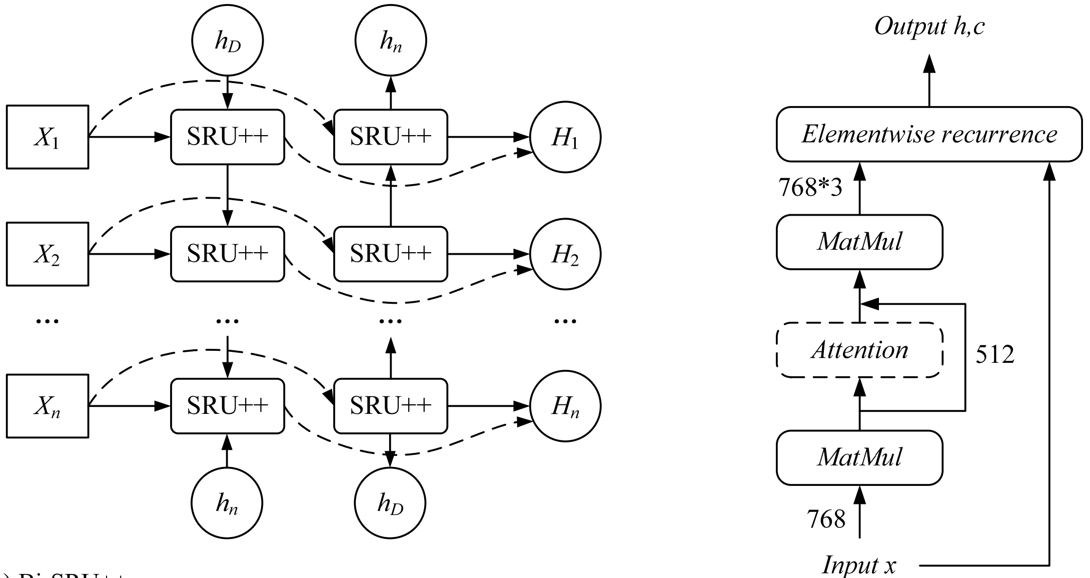
These vectors are then used as input into a Transformer structure consisting of self-attention mechanisms and feedforward neural networks (FFNN) for further operations. The pre-training of BERT on a large corpus allows it to convert semantically similar words or characters into one-dimensional word vectors that are close in distance in the feature vector space. This retains more useful information and emotional tendencies, making it easier to differentiate cyberbullying speech in subsequent tasks. Typically, the degree of association between negative words and personal pronouns in a sentence largely determines the level of malicious intent. For example, in the sentence "You're so boring, no wonder nobody wants to hang out with you", there are no obvious vulgar words, but the word "boring" has a strong negative emotion and is closely related to the personal pronoun "you", achieving the effect of malicious insult. Therefore, entering the transformed word vectors into the Transformer structure and calculating the degree of closeness between the two can help determine if the sentence is offensive. The obtained word vectors are then input into a multi-layered Transformer network, which includes self-attention mechanisms, normalization, and FFNN, as shown in Fig. 2.

3.2. Deep feature extraction

To enhance the extraction of different levels of features in short comments, multiple CNNs are used to capture local semantic features of words and phrases. The text features are convolved using convolutional layers:

$$\boldsymbol{c}_i = f(\boldsymbol{w} \otimes \boldsymbol{T}_{i;i+m-1} + \boldsymbol{b}) \tag{1}$$

where the convolution operation is denoted by $\otimes$, with a a sliding window of size m. $\boldsymbol{w}$ and $\boldsymbol{b}$ denote convolution kernel and bias, respectively. $\boldsymbol{T}_{i;i+m-1}$ represents the text vector from the i-th to the $(i+m-1)$-th row in $\boldsymbol{T}$. Non-linear function f is employed to transform negative values to 0 while maintaining positive values, with a single-sided suppression effect. By sliding window convolutions, a new feature vector $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_{n-m+1})$ can be obtained. Convolution kernels of sizes (2, 3, 4) are used to obtain feature vectors at different scales, represented by $\boldsymbol{C}_2$ 、$\boldsymbol{C}_3$ and $\boldsymbol{C}_4$, respectively. To maintain the efficient modeling ability of LSTM while avoiding the dependency upon the output state of previous time step, SRU++ is used (Pan et al., 2022). The bidirectional SRU++ (Bi-SRU++) model is built to reduce semantic loss, by stacking forward and backward SRU++ layers to extract the meaning of words in the context. The structures of Bi-SRU++ and SRU++ are given in Fig. 3a and Fig. 3b, respectively.



(a) Bi-SRU++

(b)SRU++

Fig. 3. Structures of Bi-SRU++

The Bi-SRU++ model performs multi-scale contextual modeling of local features to extract contextual information at different levels:

$$\boldsymbol{H}_t = [\vec{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t] = \text{BiSRU} + +(\boldsymbol{X}_t) \tag{2}$$

where $\vec{\boldsymbol{h}}$ and $\overleftarrow{\boldsymbol{h}}_t$ denote the outputs of the forward and backward SRU++ at time step t, respectively. The output of BiSRU++

at time step t is denoted as $\boldsymbol{H}_t$. The local feature vectors obtained from the convolutional layer, represented by $\boldsymbol{C}_2$、$\boldsymbol{C}_3$ and $\boldsymbol{C}_4$, are input into the BiSRU++. The last hidden states $\vec{\boldsymbol{h}}_L^i$、$\overleftarrow{\boldsymbol{h}}_L^i$ from the forward and backward SRU++, respectively, are concatenated to obtain the feature representation of each scale. Meanwhile, the original context feature extraction channel obtains the last hidden state output of the BiSRU++. The above features are then concatenated to obtain the multi-scale feature representation:

$$\boldsymbol{H}_l = Concat(\boldsymbol{H}_L^1, \boldsymbol{H}_L^2, \boldsymbol{H}_L^3, \boldsymbol{H}_L^4) \tag{3}$$

3.3. Attention mechanism

Multi-scale feature representation $\boldsymbol{H}_l$ is fed into the attentional layer to assess significance of high-dimensional features at different scales for the classification result, assigning higher weights to key features. The calculation can be expressed as:

$$\boldsymbol{M} = \tanh(\boldsymbol{W}\boldsymbol{H}_L + \boldsymbol{b}) \tag{4}$$

$$\boldsymbol{V} = \boldsymbol{H}_L(\text{softmax}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}))^{\mathrm{T}} \tag{5}$$

where $\boldsymbol{W}$ represents a learnable parameter matrix. $\boldsymbol{b}$ is the bias. $\boldsymbol{V}$ represents the attentional output after weight allocation. Afterwards, the attentional output $\boldsymbol{V}$ is passed through a fully connected layer to obtain $\boldsymbol{Y}$. After mapping the feature vectors to the instance classification space, Softmax is applied to normalize the output, obtaining the probability distribution $\boldsymbol{P}$ for the text classification:

$$\boldsymbol{Y} = \tanh(\boldsymbol{W}_L\boldsymbol{V} + \boldsymbol{b}_L) \tag{6}$$

$$\boldsymbol{P}_s = \text{softmax（}\boldsymbol{W}_s\boldsymbol{Y} + \boldsymbol{b}_s) \tag{7}$$

3.4. Multi-task learning

MTL is a powerful technique whereby multiple tasks are jointly trained to improve the performance and generalization ability of individual tasks through information sharing among them. This approach has proven particularly useful in the realm of cyberbullying comment classification, where abusive comments often exhibit a certain emotional tendency (Venkit & Wilson, 2021). By introducing an MTL framework to train a text sentiment classification model, the classification accuracy and generalization ability of the online cyberbullying comment classification model can be improved through the sharing of network parameters. There are typically three methods for parameter sharing in MTL: hard sharing, soft sharing, and hierarchical sharing (Sun et al., 2020). We adopt the most effective and intuitive way, hard parameter sharing, as shown in Fig. 4.

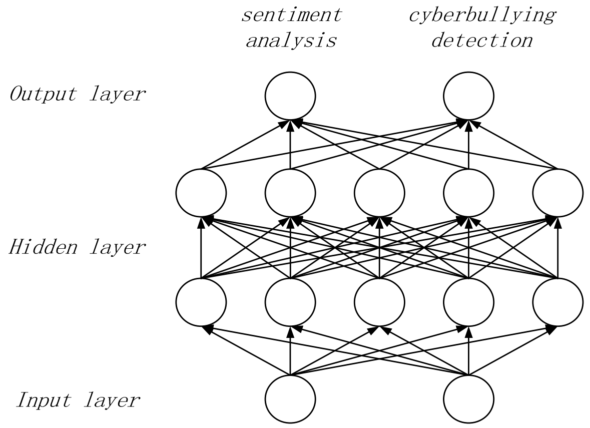


Fig. 4. Hard parameter sharing-based MTL mechanism

In the hard parameter sharing mode, the basic network structures of the two tasks are the same. The input layer and hidden layer are stored, and data of each task at the output layer is fitted. After calculating the loss, the network parameters are jointly adjusted. Introducing the text sentiment classification task can effectively mine shared features in similar tasks. Online cyberbullying comments may exhibit the following situations: 1) malicious comments with positive sentiment; 2) malicious comments with negative sentiment. Theoretically, most of the cyberbullying comments tend to have negative sentiment. Therefore, introducing the text sentiment classification task to train the model can enable it to judge the sentiment tendency of text and enhance the recognition ability of malicious comments with negative sentiment. In the training dataset of cyberbullying detection, the label of malicious comments is set to 1. And in the training dataset of sentiment analysis, the label of text with negative sentiment is also set to 1. In this way, the model is more likely to classify text with shared features as 1 during the learning process. However, the text sentiment classification model is only an auxiliary model. If the losses are directly added, it is not conducive to improving the performance of cyberbullying detection (such as malicious comments with positive sentiment or normal comments with negative sentiment). Dynamic task priority mechanism is designed to weight the calculated loss values, to better improve the over performance for cyberbullying detection:

$$w_i(t) = -(1-k_i(t))^{\gamma_i} \log k_i(t) \tag{8}$$

where $w_i(t)$ represents the weight of different tasks. $k_i(t)$ represents the performance metric. In this paper, the precision value of each epoch on validation set was used as the evaluation metric. γ_i is responsible for adjusting the weight for specific tasks, so as to give more weight to cyberbullying detection. If the precision of the next epoch is lower than that of the current epoch, the weight of sentiment classification for modifying the shared network parameters will be moderately reduced.

4. Experiment

4.1. Dataset description

The dataset used in the experiment was obtained from Kaggle toxic comment challenge (https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data). The training set comprises 159,571 comments, classified individually into six distinct categories. Upon analyzing the dataset, it can be found that each comment has a value of 0 or 1 for each category label, and some comments have a value of 1 for multiple category labels. Only 16,225 comments in the dataset have category labels, while the remaining 143,346 comments (about 90% of the original dataset) do not have category labels. These comments can be deemed harmless. The statistics of the experiment dataset are given in Fig. 5.

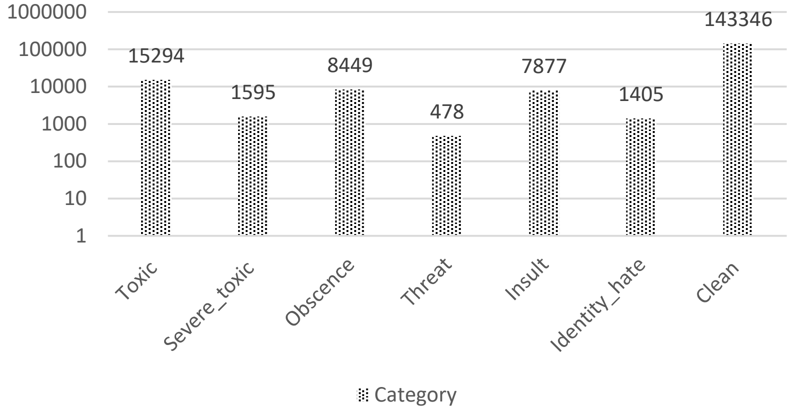


Fig. 5. Category distribution of the experiment dataset

4.2. Parameter setting

The experiment was programmed in Python 3.7 and based on the Pytorch 1.7.2 framework. During training, the NVIDIA GeForce RTX 2060 GPU was used. The dimension of word embeddings extracted from the BERT pre-training model was 768. The batch size was 64, and the pad size for text processing with BERT was set to 32. The convolutional kernel sizes were chosen as (2, 3, 4), and each convolutional kernel had 128 filters. The number of hidden layers in the BiSRU++ was 256, and the attention dimension was 512. To prevent overfitting, the dropout rate was set to 0.3 during training. The initial learning rate of the framework was 1e-5, and cross-entropy loss function was adopted. The RAdam optimizer Liu et al. (2019) was selected, which can adaptively adjust the learning rate.

4.3. Evaluation metrics

To compare performance of cyberbullying detection with different models, precision, recall, F1 score are used as primary metrics:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$F_1 = 2\times\frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

where *TP* and *TN* denote counts of comments properly categorized as malicious and non-malicious, respectively. *FP* and *FN* denote the count of comments mistakenly categorized as malicious and non-malicious, respectively. The F_1 score quantifies the balance between precision and recall in an evaluation, providing a single value that represents the overall effectiveness of a classification model. In addition, AUC (Area Under the Curve) is an indicator commonly employed in comment detection to assess performance. It is determined by calculating the area under the ROC (Receiver Operating Characteristic) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The AUROC (Area Under the Receiver Operating Characteristic curve) specifically refers to the numerical value representing the extent of the area under the ROC curve. It serves as a measure of the model's ability to distinguish between positive and negative comments, where a higher AUROC signifies improved discriminatory capability.

5. Results and Discussions

The paper analyzed the classification performance of five different single classifiers on cyberbullying detection tasks, with SVM as the baseline reference representing traditional methods. The results are shown in Table 1, and all DL models used BERT word embeddings. The results reveal several key observations. Firstly, all DL methods outperformed SVM, which can be attributed to the fact that SVM fails to capture the underlying correlations between words. Secondly, the use of attention mechanisms was found to effectively improve model performance. By leveraging the power of attention mechanisms, DL models can better determine the importance of words based on their context, leading to superior performance. Lastly, the model combining MTL mechanism achieved the best performance, demonstrating that the sentiment classification task can be used to assist in improving the performance of cyberbullying detection task.

Table 1
Cyberbullying detection performance under different models

Models	Precision	Recall	F1	AUROC
SVM	0.695	0.817	0.751	0.937
LSTM	0.728	0.833	0.777	0.979
Bi-LSTM	0.737	0.845	0.787	0.982
Bi-SRU++	0.746	0.869	0.803	0.986
Bi-SRU++&Attiention	0.829	0.878	0.853	0.988
Bi-SRU++&Attiention&MTL	0.841	0.881	0.861	0.989

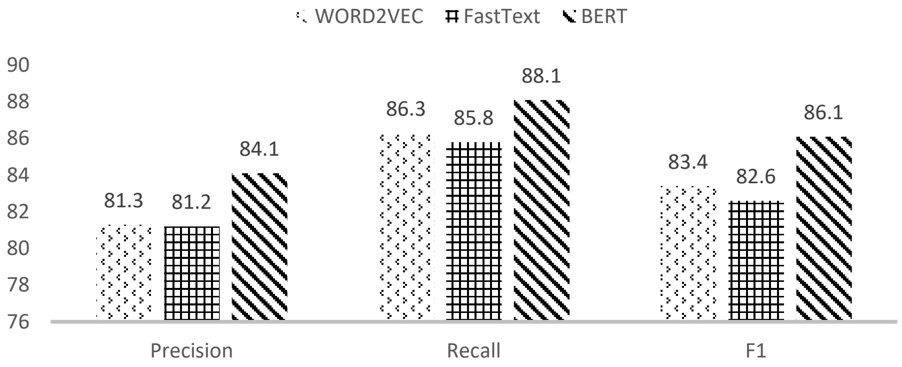


Fig. 6. Cyberbullying detection performance with different word embeddings

Fig. 6 displays the test results of the proposed framework using different word embedding models, revealing that BERT achieved the best performance. This pre-trained language model is uniquely capable of learning more complex semantic relationships, offering superior interpretability and the ability to be trained on large-scale unlabeled corpus. These advantages make BERT particularly well-suited for cyberbullying detection tasks. By leveraging the power of BERT, the proposed framework can learn the semantic relationships of words in their context, capturing more complex semantic information. In contrast, traditional word embedding models like Word2Vec and FastText are based on context-insensitive word representations and cannot capture the different meanings of words in different contexts. For potential cyberbullying comments, BERT can more accurately capture the potential maliciousness in the text, thereby improving the detection performance. The impact of different numbers and sizes of convolutional kernels on cyberbullying detection performance was thoroughly validated. Considering that too few convolutional kernels can lead to insufficient capture of local semantic features, ultimately reducing the model's detection ability. Conversely, using too many convolutional kernels will increase the time cost of model training and inference, without necessarily improving classification performance. Therefore, we set the number of kernels to 3, and tested the impact of different kernel sizes on the cyberbullying detection performance. The results are shown in Fig. 7. As the data reveals, the model's classification performance is best when using convolutional kernel sizes of (2,3,4). In a final analysis, the proposed method was compared with other DL models, revealing that the proposed framework outperformed all comparison methods across all evaluation metrics. Table 2 presents the findings. By leveraging BERT pre-trained model as the word embedding layer, the proposed method learned word representations that conformed to the context, effectively solving the problem of static word vectors' inability to represent polysemous words.

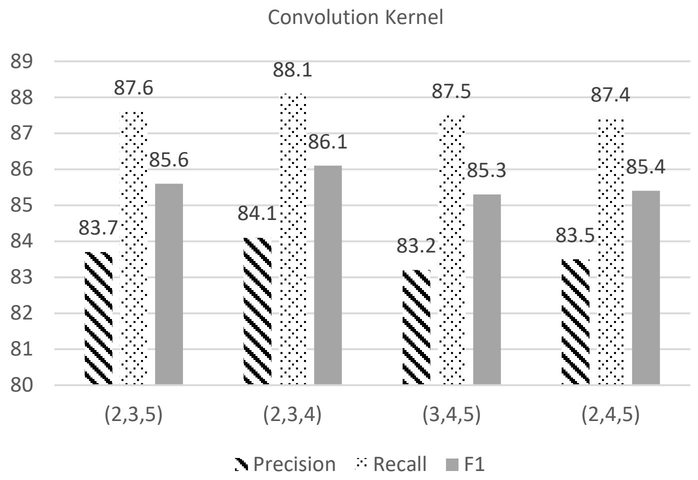


Fig. 7. Performance validation with different kernel sizes

Additionally, by considering position information of words in the sentence, the representation ability of words was enhanced. Furthermore, CNN-BiSRU++ was used as a secondary semantic extractor, deriving multiple local semantics through convolutional kernels of different sizes. Local semantic context modeling was performed through a simple recurrent unit with built-in bidirectional attention, resulting in more comprehensive text features. The attention module was able to focus on important features, thereby improving the classification performance of potentially cyberbullying comments.

Table 2
Comparison of different DL models

Models	Precision	Recall	F1	AUROC
G.K Pitsilis	0.701	0.829	0.760	0.945
G Xiang	0.746	0.835	0.788	0.982
Georgakopoulos	0.802	0.864	0.832	0.981
Zhu	0.837	0.873	0.855	0.987
Proposed	0.841	0.881	0.861	0.989

6. Conclusion

We present a novel framework to cyberbullying detection on social platforms by combining the BERT pre-trained model with MTL mechanism. The proposed algorithm outperforms traditional models by improving the detection performance of identifying potentially malicious comments. Thanks to the training based on millions of pre-training data, the word embedding retains more semantic information and avoids the dependence on sensitive word dictionaries. Furthermore, BERT also includes emotional tendencies in word vector distribution, enabling better detection of synonyms and improving the overall performance. By leveraging Bi-SRU++ for different levels of secondary context semantic modeling, the proposed approach fully explores multi-level features within texts. The attention mechanism is then used to allocate feature weights, further improving the detection accuracy of the overall model. Looking ahead, we plan to expand the experiment dataset, introduce constantly emerging new network words and phrases for repeated training of the model, and further improve the effectiveness of identifying cyberbullying comments.

References

Ahn, J., & Yoon, E. (2020). Between love and hate: The new Korean wave, Japanese female fans, and anti-Korean sentiment in Japan. *Journal of Contemporary Eastern Asia*, *19*(2), 179–196. https://doi.org/10.17477/jcea.2020.19.2.179

Brighi, A., Menin, D., Skrzypiec, G., & Guarini, A. (2019). Young, bullying, and connected: Common pathways to cyberbullying and problematic Internet use in adolescence. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01467

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Chadha, K., Steiner, L., Vitak, J., & Ashktorab, Z. (2020). Women's responses to online harassment. *International Journal of Communication*, *14*(1), 239-257.

Chen, J., Hu, Y., Liu, J., Xiao, Y., & Jiang, H. (2019). Deep short text classification with knowledge powered attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6252–6259. https://doi.org/10.1609/aaai.v33i01.33016252

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety*. https://doi.org/10.1109/socialcom-passat.2012.55

Cobbe, J. (2020). Algorithmic censorship by social platforms: power and resistance. *Philosophy & Technology*, *34*(4), 739–766. https://doi.org/10.1007/s13347-020-00429-0

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. In *Lecture Notes in Computer Science* (pp. 693–696). https://doi.org/10.1007/978-3-642-36973-5_62

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.18653/v1/n19-1423

Eckert, S., & Metzger-Riftkin, J. (2020). Doxxing, privacy and gendered harassment. *The shock and normalization of veillance cultures. M&K Medien & Kommunikationswissenschaft*, *68*(3), 273-287.

Enke, N., & Borchers, N. S. (2021). Social nedia influencers in strategic communication: A conceptual framework for strategic social media influencer communication. In *Routledge eBooks* (pp. 7–23). https://doi.org/10.4324/9781003181286-2

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Jahan, S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Jones, L. M., Mitchell, K. J., & Finkelhor, D. (2013). Online harassment in context: Trends from three youth internet safety surveys (2000, 2005, 2010). *Psychology of violence*, *3*(1), 53.

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, *71*, 431-478.

Lindsay, M., Booth, J. M., Messing, J. T., & Thaller, J. (2016). Experiences of online harassment among emerging adults: Emotional reactions and the mediating role of fear. *Journal of interpersonal violence*, *31*(19), 3174-3195.

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

Liu, L., Jiang, H., & He, P. (2019). *On the variance of the adaptive learning rate and beyond.* arXiv preprint arXiv:1908.03265

Ma, D., Liu, H., & Song, D. (2020). Word Graph Network: Understanding obscure sentences on social media for violation comment detection. In *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-030-60450-9_58

Mikolov, T., Chen, K., & Corrado, G. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781

Mohammad, F. (2018). *Is preprocessing of text really worth your time for online comment classification?* arXiv.org. https://arxiv.org/abs/1806.02908

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, *12*(11), 187.

Pan, J., Lei, T., Kim, K., Han, K. J., & Watanabe, S. (2022). SRU++: Pioneering fast recurrence with attention for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp43922.2022.9746187

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). *Detecting offensive language in tweets using deep learning*. arXiv preprint arXiv:1801.04433

Sun, T., Shao, Y., Li, X., Liu, P., Yan, H., Qiu, X., & Huang, X. (2020). Learning sparse sharing architectures for multiple tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8936–8943. https://doi.org/10.1609/aaai.v34i05.6424

Valenzuela-García, N., Maldonado-Guzmán, D. J., García-Pérez, A., & Del-Real, C. (2023). Too Lucky to Be a Victim? An Exploratory Study of Online Harassment and Hate Messages Faced by Social Media Influencers. *European Journal on Criminal Policy and Research*, 1-25.

Venkit, P. N., & Wilson, S. (2021). *Identification of bias against people with disabilities in sentiment analysis and toxicity detection models.* arXiv preprint arXiv:2111.13259, 2021.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina. *Proceedings of the 26th International Conference on World Wide Web*. https://doi.org/10.1145/3038912.3052591

Xiang, G., Fan, B., Wang, L., Hong, J. I., & Rosé, C. P. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus*. https://doi.org/10.1145/2396761.2398556

Yin, D., Xue, Z., Hong, L., Davison, B.D., & Edwards, L. (2009). *Detection of harassment on Web 2.0. Proceedings of the Content Analysis in the WEB, 2*(0), 1-7.

Young, J. C., & Rusli, A. (2019). *Review and visualization of Facebook's FastText pretrained Word Vector model.* https://doi.org/10.1109/icesi.2019.8863015

Zhu, J., Tian, Z., & Kübler, S. (2019). *UM-IU@LING at SEMEval-2019 Task 6: Identifying offensive tweets using BERT and SVMs*. https://doi.org/10.18653/v1/s19-2138