Cyberbullying Detection With Side Information: A Real-World Application of COVID-19 News Comment in Chinese Language [⋆]

Jian Xing[1,2,3(✉)], Xiaoyu Zhang[1,2], Lin Chen[1,2], Yu Ding[1,2(✉)] Yaru Zhang[1,2], Wei Hu[3], Zhicheng Jin[3], Jingya Wang[3], Yaowei Chen[3], and Yi Hong[3]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, university of Chinese Academy of sciences, Beijing, China
[3] National Computer Network Emergency Response Technical Team/Coordination Center of China Xinjiang Branch, Urumqi, China
{xingjian,zhangxiaoyu,chenlin,dingyu,zhangyaru}@iie.ac.cn

Abstract. Cyberbullying is an aggressive and intentional behavior committed by groups or individuals, and its main manifestation is to make offensive or hurtful comments on social media. The existing researches on cyberbullying detection underuse natural language processing technology, and is only limited to extracting the features of comment content. Meanwhile, the existing datasets for cyberbullying detection are non-standard, unbalanced, and the data content of datasets is relatively outdated. In this paper, we propose a novel Hybrid deep Model based on Multi-feature Fusion (HMMF), which can model the content of news comments and the side information related to net users and comments simultaneously, to improve the performance of cyberbullying detection. In addition, we present the JRTT: a new, publicly available benchmark dataset for cyberbullying detection. All the data are collected from social media platforms which contains Chinese comments on COVID-19 news. To evaluate the effectiveness of HMMF, we conduct extensive experiments on JRTT dataset with five existing pre-trained language models. Experimental results and analyses show that HMMF achieves state-of-the-art performances on cyberbullying detection. To facilitate research in this direction, we release the dataset and the project code at https://github.com/xingjian215/HMMF.

Keywords: Cyberbullying detection · Side Information · New Benchmark Dataset · COVID-19 · Nature Language Processing · Chinese Language Processing.

1 Introduction

Cyberbullying, which means to bully or harass others by online comments on social media, has become a widely discussed problem in recent years [1, 2] . The

[⋆] Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant U2003111, 61871378 and Grant U1803263).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

anonymity and concealment of the mobile Internet have accelerated cyberbullying into a widespread social phenomenon. Meanwhile, the global pandemic of COVID-19 in recent years has exacerbated the anomie of such online comments [3]. The anomie phenomenon of online comments is mainly manifested in the use of insulting and discriminatory language, such as abuse, slander, contempt and ridicule without the constraints of moral norms and laws. It makes others suffer mental and psychological violations and damage by language violence. In order to create a harmonious network atmosphere and purify the language environment for comments, it is necessary to effectively detect and analyze cyberbullying on social media.

As is known to all, detecting cyberbullying on social media is a difficult and challenging problem that needs much efforts to be devoted [4]. The reasons are two-fold.

Firstly, the task of extracting and identifying such language is generally attributed to the field of natural language processing(NLP) [5, 6]. However, with the flexibility and irregularity of language comments, it is difficult to directly find and deal with cyberbullying in time. For instance, comments that do not directly contain malicious words, sarcastically asked comments, and comments that quote questionable statements. Furthermore, most existing detection methods of cyberbullying mainly focus on the modeling of comment content, while ignoring the rich side information in social comments [4]. Significantly, the popularity of social media enables us to collect relevant side information from the perspective of net users, which helps us capture rich information except the content of comments. Another is that, with the explosive growth of Internet information in modern society, manually checking users' comments by the administrators of social media platforms is completely inadequate for cyberbullying detection. Therefore, the application of machine learning technology has become a practically feasible approach for automated cyberbullying detection[7–9].

Secondly, most existing researches mainly focus on English social media platforms, such as twitter and Instagram [10–13]. But at the same time, various Chinese pre-trained models provide a basis for us to conduct in-depth research on automated detection technology of cyberbullying based on Chinese [14–18], and creating a baseline dataset is the priority of this research [1].

In this paper, we study the problem of cyberbullying detection with side information. Particularly, the goal is to focus on how to effectively use the rich side information generated by social media for performance improvement of cyberbullying detection. To cope with this, we propose a novel Hybrid deep Model based on Multi-feature Fusion framework, called HMMF, for the problem of cyberbullying detection. The main contributions of our work are summarized as follows:

(1) Model-oriented: We propose a novel framework namely HMMF, which is the first to integrate diverse Chinese pre-trained models with Transformer for automated cyberbullying detection. The HMMF can perform the mutual fusion of multiple features and improve the performance of cyberbullying detection.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

(2) Feature-oriented: We mine more effective features from the side information related to net users and comments. The HMMF can effectively extract sentence embeddings of the comment content through the comprehensive application of existing Chinese pre-training models. Meanwhile, it can learn more useful social features, attribute features and interaction features from side information through Transformer model.

(3) Data-oriented: We publish a new benchmark dataset based on Chinese social media for cyberbullying detection, called JRTT. It is the first publicly available dataset based on Toutiao with side information. The data content of JRTT set is closely related to COVID-19 and can better reflects the characteristics of the current era of cyberbullying.

(4) We conduct in-depth research in real-world applications of COVID-19 news comments. Experimental results on real-world dataset show that the HMMF achieves better performance than previous methods in cyberbullying detection.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 gives a formal definition of cyberbullying detection. Section 4 describes the proposed model in detail. Section 5 introduces the dataset and the experimental settings, present the contrast models and shows the experimental results. Section 6 concludes the paper with discussion.

2 Related Work

Cyberbullying detection on social media is a new research field. The existing researches on cyberbullying detection still have some limitations.

In Model-oriented, the existing researches mainly utilized traditional machine learning methods and deep learning algorithms to detect cyberbullying. Algorithms such as Support Vector Machine [19, 20], Naive Bayes [21–23], Random Forest [20, 24], Decision Tree [25], and Logistic Regression [19, 21] were used to construct cyberbullying detection models. Models such as CNN [26], BiLSTM [27], and C-LSTM [28] were used to detect cyberbullying. However, few researches used knowledge-based advanced NLP models to perform automated cyberbullying detection[29]. In recent years, a large number of researches[14–18] had shown that the pre-trained models based on large corpus can learn the general language representation, which is conducive to the downstream NLP tasks. It can also avoid training the model from scratch.

In Feature-oriented, the existing researches mainly focused on identifying aggressive language through text analysis. They used feature representation methods, such as sentiment analysis[30], TF-IDF[10, 13, 31], and word embedding vector[2]. The existing researches invested a lot of energy to model text content generated by net users[32]. Both binary classification[12] and fine-grained classification[33, 34] methods mainly focused on the features based on text content, while ignoring the side information. Therefore, it is a new research direction to improve the performance of cyberbullying detection through side information[35–37].

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

In Data-oriented, at present, there is no standard dataset for cyberbullying detection. Most studies[28, 38, 39, 11, 40] independently created datasets from the social media platforms (such as twitter and Youtube) by using public APIs . However, these datasets could not be compared with each other and were quite unbalanced. Less than 20% of the available samples were classified as cyberbullying. Most existing datasets are in English and a few of them are in other languages. For example, Dutch [20], Arabic [41] and Bangla [42]. Due to some problems in Chinese processing, such as polysemy and word vector pre-training, there are relatively few researches using Chinese datasets[43]. The existing publicly available datasets mainly include: Ask.fm[28], Formspring[38], Myspace[39], Twitter[11], and Toxicity[40]. The above datasets were produced before 2019, and they could not reflect the situation and characteristics of cyberbullying incidents in the current network society.

The above analysis shows that automated cyberbullying detection on social media is still an extremely challenging task. Aiming at the limitations of existing researches, we first construct a manually labeled Chinese cyberbullying detection dataset, and then propose a hybrid deep model with side information based on multi-feature fusion for cyberbullying detection.

3 Problem Definition

A Definition of Cyberbullying: In the past few years, many social and psychological researches have attempted to give an accurate definition of cyberbullying. However, there is no consensus on the definition of cyberbullying. Based on various concepts proposed in existing literature [36, 41], we define cyberbullying as aggressive and intentional behavior committed by groups or individuals. The typical feature is to make offensive or hurtful remarks on social media, which are specifically manifested in abuse, slander, threat and ridicule. In this paper, we focus on cyberbullying, i.e. language violence.

Cyberbullying Detection: Generally, cyberbullying detection can be defined as automated detecting cyberbullying through deep learning technology using text features or other higher-order information features in social media data [36, 37]. In this paper, cyberbullying detection on social media is defined as a binary classification problem, because our work focuses on providing a general detection benchmark rather than a multi-class classification.

In this section, we detail the mathematical definition of cyberbullying detection on social media. First, we introduce the mathematical definitions of the main components of comment and side information. Second, we give a formal definition of cyberbullying detection by referring to the mathematical definitions given by existing researches. We define the basic notations as follows.

- Let w denotes the *comment content*. It only consists of a major component: user's comments. In general, it refers to Chinese short news comments on social media.
- Let s denotes the *side information*. It contains two main components: side information of comments and side information of users. Side information

of comment s_c consists of a series of properties that contain the comment interaction, such as the number of people who agree with comments and the degree of comments interaction. Side information of user s_u contains a list of key characteristics that describe net users, such as user name, user personal description, users' social history behavior statistics, and user interaction data.

- Let $E = \{e_1, e_2, ..., e_T\}$ denotes a corpus of T social media sessions. Each social media session contains one comment $C = \{w_1, w_2, ..., w_m\}$ and the corresponding side information $S = \{s_1, s_2, ..., s_n\}$, where m and n represent the number of words in the comment and the number of types of side information respectively. We treat the cyberbullying detection as the binary classification problem. Each social media session e_i is associated with a binary label $y \in \{0, 1\}$, where 0 represents non-bullying session, and 1 represents bullying session.

Definition. *Given the side information of session s and corresponding comment w, the target of cyberbullying detection is to automatically predict y for unlabeled comment w, i.e., $F : s\&w \rightarrow y$ such that,*

$$F(s, w) = y \tag{1}$$

where F is the target model we want to obtain.

4 The Proposed Model

In this section, we fully illustrate the proposed **H**ybrid deep **M**odel based on **M**ulti-feature **F**usion (HMMF) for cyberbullying detection, which is composed of three modules: (1) a *semantic context encoding* module that encodes text content of comments and text category side information of session, (2) a *digital data encoding* module for encoding digital side information of session, (3) a *session prediction* module that integrates semantic context information features with side information features into the final session representation, and uses it to predict whether the session is bullying or non-bullying. The overall framework of HMMF is shown in Figure 1. Specifically, HMMF consists of the following three parts:

(1) Semantic Context Encoding: Pre-trained model of BERT or its variants are used to extract sentence embeddings of comment content and word embeddings of text category side information. The above sentence embeddings and word embeddings are concatenated as the representation vector t.

(2) Digital Data Encoding: A fully connected layer followed by a Transformer model are applied to encode digital side information. Finally, Max-Pooling operation is applied to output a high-level representation vector d of digital data.

(3) Session Prediction: We directly connect the output vectors of above two modules to form the final feature vector v, and make the final prediction through a fully connected layer.

Next, we introduce the detailed construction procedure of each module in HMMF, and then present the training process for the whole framework.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

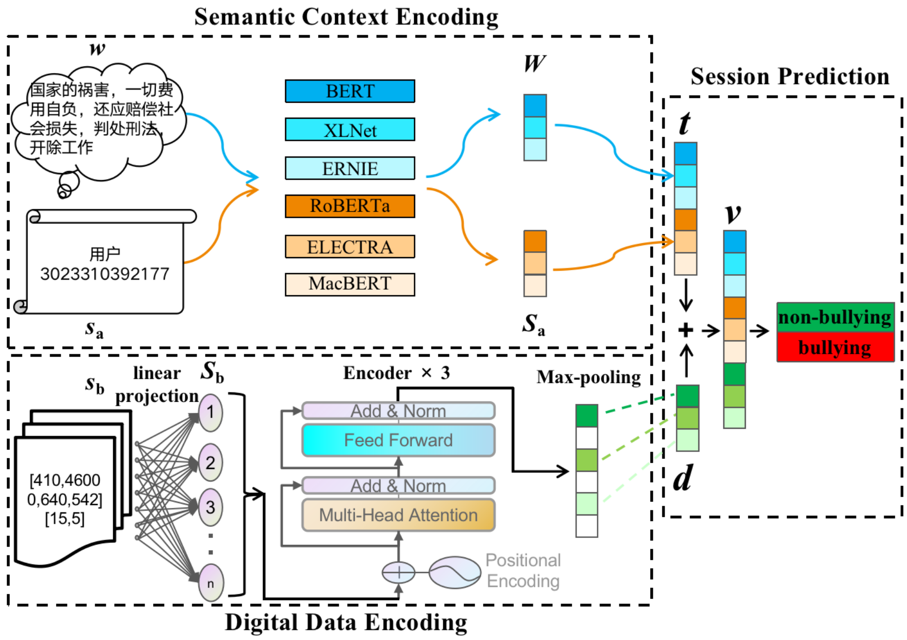

Fig. 1. The framework of HMMF.

4.1 Semantic Context Encoding

In the first module, the comment content w and text category side information s_a are used as input respectively, and then their pre-computed feature vectors are encoded by BERT or its variants. We get the feature vectors of dimension 768. The detailed process is:

$$t = Concat(M(w) \rightarrow W, M(s_a) \rightarrow S_a) \tag{2}$$

where W is a matrix of sentence embeddings representing comment content and S_a is a matrix of word embeddings representing text category side information. M is the pre-trained model.

4.2 Digital Data Encoding

In the second module, linear transformation is used to map digital side information s_b to higher dimensions. Then, we encode digital side information using Transformer, which uses a stack encoder consisting of two identical layers. Each layer includes a multi-head self-attention mechanism followed by a position-wise fully connected feedforward network. The formulation of multi-head self-attention mechanism with s_b can be defined as[44]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

$$MultiHead(S_b) = Concat(Attention_1, ..., Attention_h) \tag{4}$$

where S_b, represents the matrix vector of digital side information, which is obtained by linear transformation. Q, K, V are its different subspace matrices respectively. d_k represents the dimension of vector Q and h represents the number of heads in multi-head self-attention mechanism. Next, we perform a max-pooling operation on the output vector of the encoder. Finally, we get the final representation of digital side information d by a max-pooling operation:

$$d = Maxpooling(r_1, r_2, ..., r_j) \tag{5}$$

where r_i is the i-th column of output matrix of the encoder.

4.3 Session Prediction

Finally, we design a session prediction module that concatenates t and d to form the final session representation v for the final classification task. Then, v is sent to a fully connected layer to divide social media sessions into bullying sessions and non-bullying sessions.

$$Y = softmax(W^T v + b) \tag{6}$$

where, $Y \in \mathbb{R}^n$, n is the number of classes equals 2, respectively represent bullying and non-bullying.

4.4 Training process

As shown in Algorithm 1, we describe the HMMF training process. In each iteration of the algorithm, w and s_a are concatenated, and then the sentence embeddings W and word embeddings S_a are extracted by the pre-trained model of BERT or its variants (line 2 and 3, Eq. 2). After that, S_b is given by linear transformation (line 5). Representation vector d is computed through encoders (Eq. 3, Eq. 4, and Eq. 5). Representation vector t and d are concatenated to build the final session representation v (line 12). The final prediction Y is computed through fully connected layer (Eq. 6). Finally, once the training converges, the target model F is returned, which can be used for session prediction (line 16).

5 Experiments

In this section, we evaluate HMMF on a new publicly available benchmark dataset. We compare HMMF with a set of representative models. First, we introduce the real-world dataset JRTT, which is first published by us. Second, we describe the experimental settings including the contrast models and evaluation metrics. Finally, we present the experimental results and analyze them in detail from both macroscopic and microscopic perspectives.

Algorithm 1: Training procedure of HMMF.

```
Input: comment content $w$ and side information $s$, and labels $y = \{0, 1\}$
Output: target model $F$
for number of epoch do
    $w \rightarrow W$ through Pre-trained model of BERT or its variants;
    $s_a \rightarrow S_a$ through Pre-trained model of BERT or its variants;
    compute $t$ according to Eq. 2: $t = Concat(M(w) \rightarrow W, M(s_a) \rightarrow S_a)$;
    $s_b \rightarrow S_b$ linear transformation;
    foreach encoder do
        compute $Q, K, V$ according to Eq. 3 and Eq. 4:
        $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$ ;
        $MultiHead(S_b) = Concat(Attention_1, ..., Attention_h)$;
    end
    compute $d$ according to Eq. 5: $d = Maxpooling(r_1, r_2, ..., r_j)$;
    integrate $t$ and $d$ into $v$: $v = t + d$;
    compute $Y$ according to Eq. 6: $Y = softmax(W^T v + b)$;
end
if the training converges then
    $Calculate - Centers(s\&w : F)$;
end
return $F$;
```

5.1 JRTT: a New Benchmark Dataset

We collect, process and publish a new benchmark dataset, called JRTT. It is the first publicly available Chinese dataset based on Toutiao, which is the largest information platform in China. It ranks second in news Apps and has the most downloads. Therefore, it can be well used for cyberbullying detection research. The JRTT dataset includes 4016 manually annotated news comments from Toutiao COVID-19 special column. These comments come from 3875 different net users. There are two types of comments labels: bullying and non-bullying. The distribution of labels in the JRTT dataset is relatively well-balanced: 1833 bullying sessions and 2183 non-bullying sessions. The comments dates are from January 2021 to December 2021. Following the standard setting, we divide the dataset into three subsets, i.e., training set, validation set, and test set. They respectively account for 80%, 10%, and 10% of the entire dataset. The data of JRTT are described in detail as follows.

- **Comment content.** It includes only one data field: comment. The comments are mainly short sentences from net users, and they only involve one topic: COVID-19. We model the comments using the architecture described in Section 4.1.
- **Side information of comment.** It consists of two data fields: the number of likes on comments and the number of replies on comments. They represent whether other users agree with the comments and the level of interactions on

comment. It is digital side information about the comment. We combine and model them using the architecture described in Section 4.2.

- **Side information of user.** It consists of five data fields: username, number of Toutiao posted by user, the total number of likes, Number of fans and Number of concerns. They represent user’ s characteristics and profile information. It is primarily the digital side information about user. We model username using the architecture described in Section 4.1. We combine other four data fields and model them using the architecture described in Section 4.2.

To make a fair comparison, we run a series of comparative experiments on the real-world dataset JRTT, which is a new publicly available benchmark dataset for cyberbullying detection.

5.2 Experimental settings

We use several Chinese pre-trained models, including BERT[14], XLNet[15], ERINE[16], RoBERTa[17], and MacBERT[18], to initialize sentence embeddings and word embeddings. The padding size is 256 and batch size is 10. The dropout rate is 30% and the learning rate is 0.00005. The optimizer selects Adam optimizer. The epoch parameter is set to 30. In Digital Data Encoding, the number of layers and the number of heads are set to 3.

Experimentally we compared HMMF with five representative models:

BERT. A new language representation model. The full name is Bidirectional Encoder Representations from Transformers. BERT is a language transformation model introduced by Google. It is a general language model trained on a very large corpus, and then used for NLP tasks. Therefore, using BERT requires two steps: pre-training and fine tuning. The pre-trained BERT model can create the most advanced model for eleven tasks. It uses the same architecture in different tasks.

XLNet. A generalized autoregressive pre-trained method for NLP that significantly. It improves upon BERT on 20 tasks and achieves the current state-of-the-art on 18 tasks. XLNet is a variant of BERT released by CMU and Google brain team in 2019. It learns the bidirectional contexts over all permutations of the factorization order and integrates the idea of the current optimal autoregressive model Transformer-XL.

ERNIE. Enhanced Representation through Knowledge Integration. ERNIE is the model first released by Baidu in 2019. It surpasses BERT and XLNet in 16 Chinese and English tasks and achieves state-of-the-art effect, especially in Chinese NLP tasks. ERNIE’s advantage is that it learns the semantic representation of complete concepts in the real world through the learning of entity concept knowledge, and it extends training corpus, especially the introduction of forum dialogue corpus, enhances the semantic representation ability of the model.

RoBERTa. A Robustly Optimized BERT Pretraining Approach. Roberta has mainly improved BERT in three aspects: one is to improve the optimization function; the other is to use the dynamic mask to train the model, which proves the shortcomings of NSP (next sense prediction) training strategy and adopts a

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

larger batch size; the third is to use a larger dataset for training and BPE (byte pair encoding) to process text data.

MacBERT. A new pre-trained language model, which replaces the original MLM task into MLM as correction (Mac) task and mitigates the discrepancy of the pretraining and fine-tuning stage. MacBERT is the model released by Harbin Institute of Technology SCIR Laboratory. It achieves state-of-the-art performances on many NLP tasks.

Performance metrics: To evaluate the model, we utilize the standard metrics in classification, i.e., accuracy, precision, recall and F1-score. Existing researches[35, 37, 41] mainly use F1-score as an evaluation metrics to evaluate cyberbullying detection models. For binary classification problems, such metrics are easy to obtain. In order to compare the models with each other, we also use AUC score (Area Under Curve) to evaluate the models. Most researches use AUC scores[28], which is the criteria specified in many classification challenges.

5.3 Experimental results and analysis

In this section, we empirically evaluate the proposed model HMMF on JRTT, a new publicly available benchmark dataset. Through a series of experiments, we demonstrate the effectiveness of the model based on multi-feature fusion. Particularly, we answer the following research questions.

- *Q1*. How does the proposed model perform on cyberbullying detection?
- *Q2*. Can the various features generated by side information improve the detection performance?

To answer questions *Q1* and *Q2*, we compare HMMF with above five representative models. The comparison results on JRTT are shown in Table 1 to Table 5. Among them, data sources D1 is comment content, D2 is user name, and D3 includes all other side information. For a fair experimental comparision, HMMF and the corresponding comparison model use the same pre-trained model.

For question *Q1*, compared with five representative models, HMMF achieves the best detection performance in accuracy, precision, recall, F1-score and AUC. Among them, HMMF based on MacBERT and ERNIE pre-trained model perform best. For instance„ the F1-score of HMMF increased by an average of 2.64%, among which the HMMF based on RoBERTa pre-trained model increased the most, reaching 6.50%.

For question *Q2*, the experimental results fully demonstrate that the fusion of comment content features and side information features can improve the performance of cyberbullying detection. Specifically, when we fuse D2 and D3 based on D1, the detection performance of all models becomes better. The reasons for the better performance of HMMF are as follows: Firstly, HMMF can extract sentence embeddings of comment content and word embeddings of text category side information. Secondly, using Transformer model, we can make better use of digital side information. Thirdly, besides news comment, the side information of session contains effective features, which can be used to improve the performance of cyberbullying detection. Next, we study a specific case to illustrate the importance of side information features.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_42

A case study: RoBERTa mispredicts some comments as non-bullying, which are:” 为人师表, 只是口头上的, 实际做法大相径庭!” , ” 公布信息吧, 让大家看看这两位老师是怎么为人师表的” , and ” 现在的老师有几个有师德的” . However, HMMF correctly predicts them as bullying through the features generated by side information. The above comments do not contain malicious words, but are comments in the form of irony or rhetorical questions. Therefore, it is difficult to predict cyberbullying only by comment content. At this time, side information is needed to aid detection. Moreover, through this case, we can find that these bullying comments do not concern COVID-19, but express dissatisfaction with specific identity groups.

Table 1. Performance comparison based on BERT pre-trained model.

Model	BERT		HMMF	
Data source	D1	D1+D2	D1+D3	D1+D2+D3
Accuracy	91.78%	95.47%	94.62%	**95.75**%
Precision	91.79%	95.55%	94.62%	**95.77**%
Recall	91.78%	95.47%	94.62%	**95.75**%
F1-score	91.78%	95.46%	94.61%	**95.75**%
AUC	91.73%	95.36%	94.62%	**95.78**%

Table 2. Performance comparison based on XLNet pre-trained model.

Model	XLNet		HMMF	
Data source	D1	D1+D2	D1+D3	D1+D2+D3
Accuracy	92.35%	93.20%	92.63%	**93.77**%
Precision	92.38%	93.24%	92.66%	**93.77**%
Recall	92.35%	93.20%	92.63%	**93.77**%
F1-score	92.35%	93.19%	92.63%	**93.77**%
AUC	92.28%	93.12%	92.67%	**93.73**%

Table 3. Performance comparison based on ERNIE pre-trained model.

Model	ERNIE		HMMF	
Data source	D1	D1+D2	D1+D3	D1+D2+D3
Accuracy	96.32%	**97.17**%	96.60%	**97.17**%
Precision	96.35%	**97.19**%	96.63%	97.17%
Recall	96.32%	**97.17**%	96.60%	**97.17**%
F1-score	96.31%	**97.17**%	96.60%	**97.17**%
AUC	96.25%	**97.21**%	96.64%	97.16%

Table 4. Performance comparison based on RoBERTa pre-trained model.

Model	RoBERTa		HMMF	
Data source	D1	D1+D2	D1+D3	D1+D2+D3
Accuracy	89.80%	94.05%	95.75%	**96.32**%
Precision	89.95%	94.07%	95.88%	**96.35**%
Recall	89.80%	94.05%	95.75%	**96.32**%
F1-score	89.81%	94.05%	95.75%	**96.31**%
AUC	89.90%	94.08%	95.85%	**96.25**%

Table 5. Performance comparison based on MacBERT pre-trained model.

Model	MacBERT		HMMF	
Data source	D1	D1+D2	D1+D3	D1+D2+D3
Accuracy	95.47%	95.47%	97.17%	**97.45**%
Precision	95.50%	95.53%	97.17%	**97.49**%
Recall	95.47%	95.47%	97.17%	**97.45**%
F1-score	95.47%	95.47%	97.17%	**97.45**%
AUC	95.51%	95.53%	97.16%	**97.39**%

5.4 Repeatability

The experimental equipment is configured as 128GB memory and a GeForce RTX 2080 GPU. Source code is available at https://github.com/xingjian215/HMMF.

6 CONCLUSION

In this paper, we present a new hybrid deep model HMMF for cyberbullying detection, which can learn useful representations from both comment contents and side information to boost the detection performance. Inspired by the widely used pre-trained model in NLP technology, we utilize several Chinese pre-trained models for encoding the semantic context and utilize Transformer for encoding the digital data synthetically to mining more effective features contained in side information. We introduce JRTT, a new benchmark dataset for automated cyberbullying detection. JRTT’ s authentic, real-world comments on COVID-19 from diverse net users can fully reflect the characteristics of cyberbullying and promote the further development of research. Experimental results further demonstrate that HMMF achieves new SOTA performance in the real-world application of COVID-19 news comments. In the future, we can mine more side information features, such as net user profile, gender, historical comment information, etc.

DOI: 10.1007/978-3-031-08751-6_42

References

1. Hee, C., et al. Automatic Detection of Cyberbullying in Social Media Text. PLoS ONE, 2018.
2. Rosa, H., et al. Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 2019.
3. Lin, Y., et al. Psychological intervention for three COVID patients who suffered online violence, Chin J Psychiatry, 2021.
4. Emmery, Chris., et al. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation , 2020.
5. Parma, N., et al. How Bullying is this Message: A Psychometric Thermometer for Bullying. COLING, 2016.
6. Walisa R., et al. Automated cyberbullying detection using clustering appearance patterns. IEEE International Conference on Knowledge and Smart Technology, 2017.
7. Gutiérrez-Esparza., et al. Classification of Cyber-Aggression Cases Applying Machine Learning. Applied Sciences, 2019.
8. Ducharme, and Daniel, N. Machine Learning for the Automated Identification of Cyberbullying and Cyberharassment. Open Access Dissertations, 2017.
9. Haidar, B., et al. A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. Advances in Science, Technology and Engineering Systems Journal, 2017.
10. Sugandhi, R., et al. Automatic monitoring and prevention of cyberbullying. International Journal of Computer Applications, 2016.
11. Zhao, R., et al. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Transactions on Affective Computing, 2016.
12. Zhao, R., et al. Automatic detection of cyberbullying on social networks based on bullying features. International conference on distributed computing and networking, 2016.
13. Hosseinmardi, H., et al. Prediction of cyberbullying incidents in a media-based social network. International Conference on Advances in Social Networks Analysis and Mining, 2016.
14. Devlin, J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, 2018.
15. Yang, Z., et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237, 2020.
16. Zhang, Z., et al. ERNIE: Enhanced Language Representation with Informative Entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
17. Liu, Y., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.
18. Cui, Y., et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. EMNLP, 2020.
19. Chavan, V.S., et al. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. ICACCI, 2015.
20. Van Hee, C., et al. Detection and fine-grained classification of cyberbullying events. RANLP, 2015.
21. Mangaonkar, A., et al. Collaborative detection of cyberbullying behavior in Twitter data. EIT, 2015.

22. Sanchez, H., et al. Twitter bullying detection. journal of the japanese association of periodontology, 2011.
23. Dinakar, K., et al. Modeling the detection of Textual Cyberbullying. ICWSM, 2011.
24. García-Recuero, Á. Discouraging Abusive Behavior in PrivacyPreserving Online Social Networking Applications. In Proceedings of the 25th International Conference Companion on World Wide Web. 2016.
25. Reynolds, K., et al. Using machine learning to detect cyberbullying. ICMLA, 2011.
26. Akhter, Muhammad Pervez., et al. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems, 2021.
27. AGRAWAL, S., et al. Deep learning for detecting cyberbullying across multiple social media platforms. In Advances in Information Retrieval, 2018.
28. ROSA, H., et al. A deeper look at detecting cyberbullying in social networks. IJCNN, 2018.
29. Hong, F.,et al. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. Electronics, 2021.
30. Walaa M., et al. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 2014.
31. Perera, A., et al. Accurate Cyberbullying Detection and Prevention on Social Media. Procedia Computer Science, 2021.
32. Dani, H., et al. Sentiment informed cyberbullying detection in social media. ECML-PKDD, 2017.
33. HEE, C., et al. Detection and Fine-Grained Classifification of Cyberbullying Events. In International Conference Recent Advances in Natural Language. RANLP, 2015.
34. HONNIBAL, M., et al. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
35. Maral D., et al. Improving Cyberbullying Detection with User Context. European Conference on Information Retrieval, 2013.
36. Ge, S., et al. Improving Cyberbully Detection with User Interaction. arXiv:2011.00449, 2020.
37. Chen, H. Y., et al. HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media. EMNLP, 2020.
38. Zhang, X., et al. Cyberbullying detection with a pronunciation based convolutional neural network. ICMLA, 2017.
39. Sugandhi, R., et al. Automatic monitoring and prevention of cyberbullying. International Journal of Computer Applications, 2016.
40. THAIN, N., et al. Wikipedia Talk Labels: Toxicity.
41. Mk, A , C., et al. Automatic Detection of Cyberbullying and Abusive Language in Arabic Content on Social Networks: A Survey - ScienceDirect. Procedia Computer Science, 2021.
42. Ahmed, M. F., et al. Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language, 2021.
43. Liu, Z., et al. Detection and Analysis of CyberneticsBullying Language on Common Chinese Social Network Platforms. Journal of Southwest China Normal University, 2021.
44. Ashish, V., et al. Attention Is All You Need. arXiv:1706.03762, 2017.