Multi-modal cyberbullying detection on social networks

Kaige Wang, Qingyu Xiong*, Chao Wu, Min Gao and Yang Yu
State Key Laboratory of Power Transmission Equipment and System Security and New Technology
School of Big Data and Software Engineering, Chongqing University
Chongqing, China, 401331
Email: kgwang@cqu.edu.cn, *cquxqy@163.com

***Abstract*—Because social networks have become a vital part of people's lives, cyberbullying becomes the most common risk encountered by young people on social networking platforms and raised serious concerns in society. Over the past few decades, most existing work on cyberbullying has focused on text analysis. Yet, the cyberbullying develops into multi-objective, multi-channel, and multi-form. Traditional text analysis methods cannot satisfy the diversity of bullying data in social networks. To deal with the new type of cyberbullying, we propose a multi-modal detection framework that takes into multi-modal information(e.g., image, video, comments, time) on social networks. Specifically, we not only extract textual features but also use the hierarchical attention networks to capture the session feature in social networks and encode several media information(e.g., video, image). Based on these features, we model the multi-modal cyberbullying detection framework to solve the new form of cyberbullying. Experimental analysis on two real-world datasets shows that our framework outperforms several existing state-of-the-art models.**

***Index Terms*—Cyberbullying, Multi-Modality, Social Media, Hierarchy Attention**

I. Introduction

Social networking is the main way for young people to socialize. However, the cyberbullying on the social network is happening all the time and has a serious impact on young people [22]. According to the statistics of the American Psychological Association and the White House [9], more than 40% of teenagers in the United States have been suffered cyberbullying in social media. Recent studies reported by the British indicated that the ratios bullied in the network are much larger than in the real-world, 12% of teenagers have been bullied. The social cyberbullying incidents occurred frequently and increased year by year. Bullying behavior gradually develops into multi-objective, multi-channel, and multi-form. The victims have suffered a severe negative impact on their physical and mental health [18], even made suicidal thoughts. It is not just a nightmare for the victims, but also a critical national health concern. Hence, it has stimulated research upsurge in the fields of psychology and computer science, aimed at understanding cyberbullying characteristics to identify bullying in social networks.

In the field of automatic cyberbullying detection, as malicious verbal attacks are a typical manifestation of cyberbullying, the existing efforts mainly focus on the analysis of text features. Many text classification methods have been introduced for cyberbullying detection. Cyberbullying can be defined as repeated sending of hostile or aggressive information by any individual or group through electronic devices or digital media to cause harm or discomfort to others. Text-only feature analysis faces several challenges. It isn't very easy to determine whether the content ·targets a specific person and/or group, without contextual information. Besides, the normal textual content with offensive visual information is still a potential danger on social networks. Hence, it is necessary to pay more attention to critical information included in the various social media, such as image, video, comments, and social relationships.

Existing efforts for multi-modal information pay attention to a single modality. The comments are considered a short conversation about the topic. The study [17] used contextual information to understand the entire context better and thus determine the behavior. Even though the study attempted to learn the relationship of comments, but ignored the effect between each comment. Soni [23] combined visual features to complement the lack of textual features. Although these methods have better performance than text analysis, they can't solve the limitations of single-mode information.

In addition, Cyberbullying has other essential characteristics of the persistence and repetition of aggressive acts over time [9]. A new challenge is how to stop the discussion of cyberbullying and prevent secondary harm effectively. Hence, how to effectively detect multi-modal bullying information in time and prevent it from further discussion is a new challenge for cyberbullying detection.

To cope with the new forms of cyberbullying, we redefine cyberbullying as a process that combines textual, visual, and another meta-information to identify whether a post belongs is a bullying topic. To address the above challenges, we propose a novel Multi-Modal Cyberbuulying Detection (MMCD) framework. It can integrate textual, visual, and other meta-information uniformly to identify various cyberbullying instances in social networks. Explicitly, we assume that cyberbullying posts received offensive comments. We model all of the comments by Hierarchical Attention Networks (HAN) [25] to judge the feedback of comments and then encode visual and other meta-information. Finally, we integrate these features and text contents to improve cyberbullying detection performance. The main contributions of this work are:

1) We define a new form of cyberbullying detection and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

formulate the entire problem process. The definition focuses on fusion and unified processing of multi-modal data to deal with the multi-form of cyberbullying.

2) We propose a novel multi-modal cyberbullying detection framework to model textual, visual, and other content respectively, and then construct the multi-modal processing framework. The framework has several components: topic-oriented bidirectional long-short term memory (BiLSTM) model with self-attention; comment-based HAN model to focus on word-level and comments-level; visual embedding and other embeddings methods. We integrate these components to achieve the effect of information fusion, to face the multi-form of cyberbullying.
3) We use multi-modal data from two social networks to verify the effectiveness of this method. We also analyze the influence of multi-modal on cyberbullying.

II. Related Work

The existing cyberbullying detection work has paid attention to analyze the features of the text, identifying bullying behavior. Generally, the methods applied in text content were similar to text classification, emotional analysis, and other technologies [12].Some text feature extraction methods were also effective, such as N-gram models, bags-of-words models (Bow), TF-IDF, etc [2], [8], [29]. In response to these features, sclassification methods provided great effect, which include Random Forest, support vector machine(SVM), Logistic Regression,Naive Bayes, Random Forest, etc [3], [7], [8], [16]. Chavan et al. [4] extracted several features (e.g., TF-IDF, Bow, bullying vocabulary) and used SVM and Logistic Regression to identify bullying behavior. In addition to text characteries, network characteristics have received attention [1]. Other information on social networks has received the researchers' attention, such as the number of tweets, locations, and social relationships on Twitter. Chen et al. [5] constructed the social network topology structure to identify bullying. Algaradi et al. [1] integrated several available information. They used the network, activity, user, and tweet content to build an effective detection model. To get better results, Lu Cheng et al. [7] built a complex heterogeneous network by metadata, such as an image, video, user profile, time, location and comments. They learned the vector representation of posts by network embedding and then classified the post by SVM, Random Forest, etc. These models attempt to solve the problem that the lack of text characteries.

In addition, deep learning as an end-to-end method, has better representation capabilities in text content. Most of the text classification has improved by these methods, using convolutional neural network CNN [11], RNN [13] and the combination of CNN and RNN (RCNN) [27], etc. Miswriting often occurs in social media, especially using miswrite to avoid detection in bullying text. Park et al. [19] used a hybrid model to solve that. They applied convolutional neural networks at both the character-level and word-level, then connected them to classify. Ziqi [28] joined convolution layers and Gate Recurrent Unit (GRU) to encoded the text, the method combined structural features and sequence features. Attention mechanisms are introduced in cyberbullying detection to pay more attention to essential words. Zhang [26] proposed a bidirectional RNN (BiRNN) [31] model with the attention mechanism to identify bullying text. This model integrated the contextual feature by BiRNN, used attention mechanism change the words weigh. Similar to the above methods, some meta-information has played an important role in the deep model. Founta [10] trained the hybrid model by combining the latent representation from text and metadata.

With the diversity of social media, some studies not only focused on textual but also paid attention to visual data. More and more researches attempted to integrate multi-modal data. Soni [23] attempted to extra visual features to supplement the lack of textual elements. To understand the context, ziyi [17] used the parent-child relationship between comments to analyze child semantics and child comments on related topics. Moreover, the document classification methods were applied in comments, such work [6] built a time-dependent hierarchical attention network to capture comments features. These methods indicated that multi-modal data has a positive effect on cyberbullying detection. Feature extraction in multi-modal and feature fusion were research trend of cyberbullying detection.

III. Problem Definition

Let $P = \{p_1, p_2, ..., p_N\}$ be a corpus of N social media posts. Each post includes the text content, consecutive comments, media object (i.e., image or video) and other information (e.g., time stamp, the user's profile, Likes, Shares). Thus, the post i is denoted as $p_i = \{t_i, c_i, m_i, o_i\}$, where t_i means text content; c_i as a set of comments, m_i means media object, o_i is other meta-information. Moreover, we use $c_i^{(1)}$ to represent the first comment in c_i, the length of $c_i^{(1)}$ is donated as $l_i^{(1)}$. In addition, we use a binary label $y_i = \{0, 1\}$ to identify the post i, where 1 represents bullying behavior, 0 is the otherwise.

Compared with simple text classification for cyberbullying detection, we define cyberbullying detection as the process of learning the bullying behavior by *context, comments, media, and other information*.

IV. Proposed Model

In this section, we introduce the proposed multi-modal cyberbullying detection (MMCD) framework in detail. The model is divided into two processes of encoding and decoding. For the encoder, we encode different data contents separately. It consists of several components: Topic-oriented encoder by BiLSTM, comment-based hierarchical attention, media embedding, and other meta-information embedding layer. In addition, we take into account the sequence of the comment set in the comments encoder. As for the decoder, we use the multilayer perceptron (MLP) to train the multi-mode data independently, and finally integrate multi-mode data for training.

The proposed framework of multi-modal cyberbullying detection is shown in Figure 1.

A. Multi-Modal encoder

In this subsection, we introduce several components for encoding multi-modal data, including BiLSTM, hierarchical attention mechanism for comments, and embedding of media information.

1) Bidirectional LSTM : Recurrent neural networks are a common method in natural language processing. Considering the word order of the sentence, we use BiLSTM to model the sentence. As a type of RNN, LSTM has three additional gated units which are the input gate i_t , the forget gate f_t , and the output gate o_t. Specifically, we use h_t represent the state at time t, then the state of the three gates are calculated by the previous state h_{t-1}. For the t-th word vector x_t:

$$i_t = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i), \tag{1}$$

$$f_t = \sigma(x_t W_{xf} + h_{t-1} W_{hf} + b_f), \tag{2}$$

$$o_t = \sigma(x_t W_{xo} + h_{t-1} W_{ho} + b_o), \tag{3}$$

where x_t is the word embedding vector at t step, the parameters W_{xi},W_{xf},W_{xo},W_{hi},W_{hf},W_{ho} are the relevant weight matrix, and bias b_i,b_f,b_o.

As for the hidden layer h_t, it depends on the current memory c_t and the candidate memory $\tilde{c_t}$.

$$\tilde{c_t} = tanh(x_t W_{xc} + h_{t-1} W_{hc} + b_c), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t}, \tag{5}$$

$$h_t = o_t \odot tanh(c_t), \tag{6}$$

where W_{xc}, W_{hc} are weight matrix, and b_c is bias.

We use BiLSTM to encode the text content, which captures the sentence features from both directions. For a sentence with n words $\{w_1, w_2, ...w_n\}$, and the embedding vectors $\{x_1, x_2....x_n\}$, we calculate the hidden state of $i_t h$ word:

$$\overrightarrow{h}_i = \overrightarrow{\mathrm{LSTM}}(x_i), i \in [1, n], \tag{7}$$

$$\overleftarrow{h}_i = \overleftarrow{\mathrm{LSTM}}(x_i), i \in [n, 1], \tag{8}$$

where $\overrightarrow{h}_i$ is the hidden state from word w_1 to w_i, $\overleftarrow{h}_i$ means the opposite direction. In addition, the word-level self-attention mechanism is introduced to improve awareness of negative words.

2) Comments Enbedding With Hierarchical Attention networks: We assume that each comment is affected by all previous comments. Thus, we regard all of the comments as a document generated in the order of publication. Based on the method of document classification, we apply the Hierarchical Attention Network (HAN) [25] to encode the comments. This study focuses on the attention mechanism at the word-level and comments-level, which makes the HAN model pay more attention to important content when encoding the document. Hence, we do the word-level attention in each comment and then apply attention mechanisms at comments level. We treat the comments by hierarchical attention architecture.

We use Bidirectional GRU to encode word level and comment level. Similar to LSTM, GRU is other type of RNN with two gates: the update gate and reset gate.

Given the comments C with L comments, and the i-th comment c_i has L_i words $w_{it}, t \in [0, L_i]$.

Word-Level Attention. We use the embedding matrix to W_e to embed the words, $x_{ij} = W_e w_{ij}$. Then, we put the words of comment i into bidirectional GRU model for encoding comment.

$$\overrightarrow{h}_{it} = \overrightarrow{\mathrm{GRU}}(x_{it}), t \in [1, L_i], i \in [1, L], \tag{9}$$

$$\overleftarrow{h}_{it} = \overleftarrow{\mathrm{GRU}}(x_{it}), t \in [L_i, 1], i \in [1, L], \tag{10}$$

where $\overrightarrow{h}_{it}$ means the hidden state from word w_{i1} to w_{iL_i}, and $\overleftarrow{h}_{it}$ means the backward hidden state. Then we concatenate hidden vector in both direction, $h_i = [\overrightarrow{h}_{it}; \overleftarrow{h}_{it}]$

Because each word has the different influence on the comment, we attempt to reconstruct the vector of these words to form comments for these important words through the attention mechanism. Specifically, we use a multi-layer perceptron with a hidden layer to extract higher-level hidden layer representation u_{it}:

$$u_{it} = \tanh(W_w h_{it} + b_w), \tag{11}$$

where W_w is weight matrix, b_w is a bias of word-level. We measure the similarity u_{it} and a vector u_w that is regarded as a word-level context vector. Then normalizing the weight matrix,

$$\alpha_{it} = \frac{\exp\left(u_{it}^\top u_w\right)}{\sum_t \exp\left(u_{it}^\top u_w\right)}. \tag{12}$$

Finally, we reconstruct the vector representation of this comment.

$$c_i = \sum_t \alpha_{it} h_{it}. \tag{13}$$

Similar to the word-level approach, We user bidirectional GRU encode each comment c_i by time stap:

$$\overrightarrow{h}_i = \overrightarrow{\mathrm{GRU}}(c_i), i \in [1, L], \tag{14}$$

$$\overleftarrow{h}_i = \overleftarrow{\mathrm{GRU}}(c_i), i \in [L, 1]. \tag{15}$$

Then, we calculate the comments attention and rebuild comments vector representation,

$$u_i = \tanh(W_c h_i + b_c), \tag{16}$$

$$\alpha_i = \frac{\exp\left(u_i^\top u_c\right)}{\sum_i \exp\left(u_i^\top u_c\right)}, \tag{17}$$

$$v = \sum_i \alpha_i h_i, \tag{18}$$

where W_c is weight matrix, b_c is a bias of comment-level and u_c is the comments level context vector. The vector v summarizes all information of the comments C.

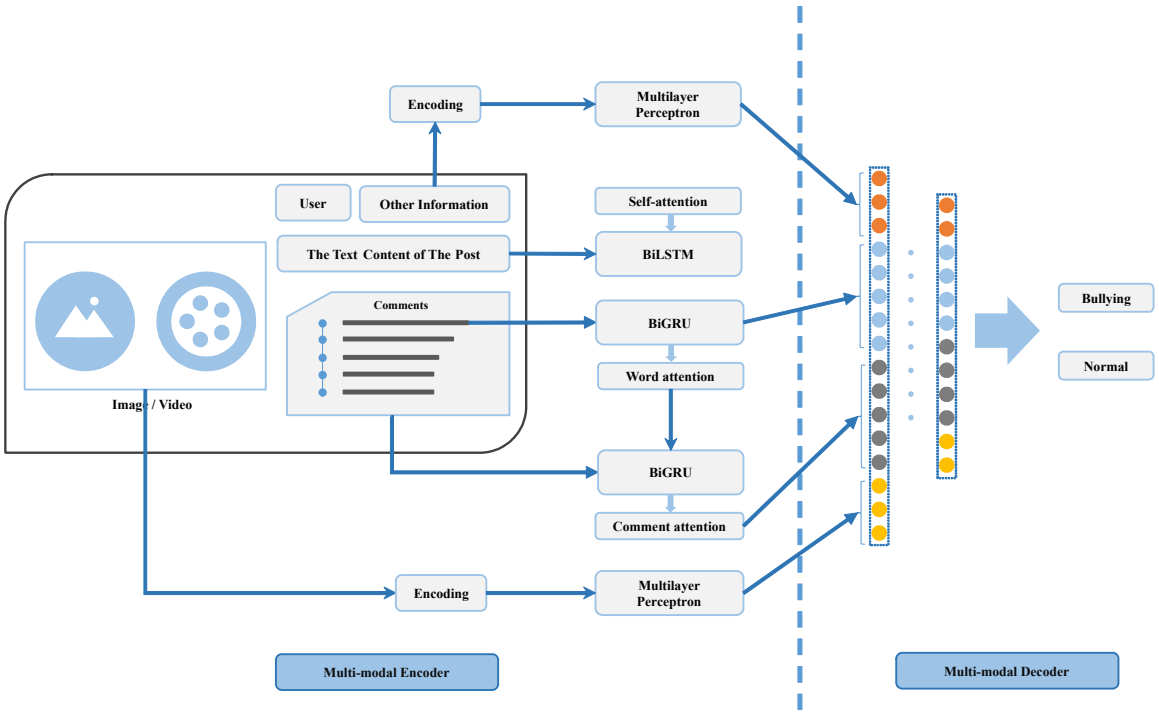

Fig. 1. The proposed multi-modal cyberbullying detection framework. Given a post in social media, we first analyze various content in this post, model different modalities, and get the latent vector for each modality data (multi-modal encoder). We integrate all the latent vectors from multi-modal data and train classifiers(multi-modal decoder).

3) Other Embedding : As for the encoding of media information, we first construct the one-hot encoding by the tags(e.g.,text, scenery, portraits, etc.) of media. To reduce the dimension of one-hot, we adopt multi-layer perceptron for feature extraction.

B. Multi-Modal decoder

We assume that the information of different modalities make different contributions to bullying behavior detection. Based on this assumption, we propose a new method to adjust the weights of modalities' vector in the decoding step.

In the encoding step, we extract the different modalities information, and vectors of different sizes represent the information. We use v_t, v_c, v_m and v_o to represent the encode vector form comments, text, media, and other meta-information, respectively. In the decoding step, we use fully connected units for each modal independently and calculate between each layer:

$$h_{dt} = tanh(W_{dt}v_t + b_{dt}), \tag{19}$$

$$h_{dc} = tanh(W_{dc}v_t + b_{dc}), \tag{20}$$

$$h_{dm} = tanh(W_{dm}v_t + b_{dm}), \tag{21}$$

$$h_{do} = tanh(W_{do}v_t + b_{do}), \tag{22}$$

where h_{dt}, h_{dc}, h_{dm}, h_{do} are the hidden layer of full connected units, the parameters W_{dt}, W_{dc}, W_{dm}, W_{do} are the weight matrix, and the bias b_{dt}, b_{dc}, b_{dm}, b_{do}.

Then, we adjust the size of each hidden vector to fit the importance of the modalities information. We compute the output of each hidden vector:

$$\tilde{out} = tanh(Wh + b), \tag{23}$$

where W, b are the weight matrix and bias for each hidden vecto, and get the output vector out_t, out_c, out_m, out_o. We concatenate the each output vector, $out = [out_t; out_c; out_m; out_o]$. We complete the final decoding operation based on the vector out.

V. EXPERIMENTS

In this section, we evaluate the effectiveness of our methods by using two real-world dataset collected from Instagram and Vine. For each dataset, we use 80% of data for tranning, the remaining for testing and evaluated with ACC scores and F1 scores. As for word embedding, we use pre-trained Glove model to train words and embed into 300-dimensions vectors.

A. Datasets

In our experiments, we use two datasets[1] from social media network; Instagram(photo and video sharing) and Vine (Short video sharing). Both of these datasets are publicly available and contains multi-modal data (e.g., text, image, etc.).

[1]Available from the site: https://sites.google.com/site/cucybersafety/home/cyberbullying-detection-project/dataset

TABLE I
DATASET STATISTICS.

Dataset	Posts	Bullying	Normal	Comments	Average
Instagram	2,218	678	1540	15260	71
Vine	970	304	666	78250	80

Instagram. [14] A photo and video-sharing social network. The dataset contains 2,218 posts, among which 678 are labeled as *bullying* and 1540 are labeled *Normal*. Besides,there are 155,260 discussion comments for the posts. Some other information is also available, such as user comments, the tags of images, time stamp and user profiles, etc.

Vine. [21] A short video sharing social network that allows users to edit and share six-second-long, looping video clips. There are 970 posts in the dataset, of which 666 refer to Normal behavior and 304 refer to *Bullying behavior*. Each post of Vine is associated with post content, user comments, the tag of video, etc.

Detailed statistics for the Vine and Instagram datasets are shown in Table 1.

B. Baseline Methods

To verify the effectiveness of the framework, we compare the framework with several baseline methods, including SVM, Naive Bayesian, Logistic Regression, Random Forest methods with different textual features. We use several textual features for these classification models, including word-level TF-IDF vectors, character-level TF-IDF vectors, and psychological features from Linguistic Inquiry Word Count (LIWC) [20].

We also use some deep learning models as baseline, including LSTM [30], LSTM with attention, Text-CNN [15]. These models are commonly used in cyberbullying detection. We compare our methods with HAN to verify the effectiveness of other information (such as media information).

In addition, we compare our methods with some existing cyberbullying detection models, i.e., Xu et al. [24] and Lu et al. [6]. Next, we briefly introduce the two models.

Xu et al. The method extracted several text features to understand the text content, and use some models to detect cyberbullying behavior, including text classification, role labeling, sentiment analysis, and topic modeling.

Lu et al. This method regarded texts that with a timestamp as a media session. It modeled used hierarchical attention texts, and attention mechanisms applied to the word-level and comment-level. In addition, the method took into account the time interval between two comments and time features.

C. Result

We use ACC-scores and F1-scores to evaluate the performance of these models on Instagram and Vine dataset. Because these datasets are imbalanced, we pay more attention to the F1-scores. The results are shown in Table 2-3.

As the results are shown in Table 2-3, MMCD has the best performance in F1-scores and ACC-scores among all of the

TABLE II
COMPARISON OF THE F1-SCORES AND ACC-SOCRES FOR THE BASELINE MODELS IN INSTAGRAM DATASET.

Methods		ACC	F1
SVM	Char TF-IDF	0.576	0.583
	Word TF-IDF	0.556	0.562
	LIWC	0.623	0.597
Naive Bayesian	Char TF-IDF	0.625	0.676
	Word TF-IDF	0.653	0.668
	LIWC	0.592	0.504
Logistic Regression	Char TF-IDF	0.594	0.583
	Word TF-IDF	0.605	0.573
	LIWC	0.73	0.653
Random Forest	Char TF-IDF	0.619	0.669
	Word TF-IDF	0.695	0.637
	LIWC	0.758	0.604
LSTM		0.791	0.613
LSTM with Attention		0.813	0.692
Text-CNN		0.781	0.643
HAN		0.804	0.708
Xu et al.		0.513	0.502
Lu et al.		0.851	0.783
MMCD		**0.864**	**0.86**

TABLE III
COMPARISON OF THE F1-SCORES AND ACC-SOCRES FOR THE BASELINE MODELS IN VINE DATASET.

Methods		ACC	F1
SVM	Char TF-IDF	0.529	0.622
	Word TF-IDF	0.571	0.587
	LIWC	0.638	0.672
Naive Bayesian	Char TF-IDF	0.631	0.658
	Word TF-IDF	0.662	0.697
	LIWC	0.638	0.559
Logistic Regression	Char TF-IDF	0.612	0.596
	Word TF-IDF	0.641	0.595
	LIWC	0.726	0.684
Random Forest	Char TF-IDF	0.625	0.658
	Word TF-IDF	0.746	0.781
	LIWC	0.761	0.729
LSTM		0.783	0.641
LSTM with Attention		0.813	0.692
Text-CNN		0.761	0.674
HAN		0.817	0.797
Xu et al.		0.684	0.697
Lu et al.		0.817	0.797
MMCD		**0.838**	**0.841**

models. For the Instagram dataset, MMCD outperforms the best baseline model Lu et al. by 7.7% and 1.3% in F1-scores and ACC-scores. Although the score has little promoted in ACC-scores, the F1-scores is more useful in cyberbullying detection. For the Vine dataset, MMCD outperforms the best baseline model Lu et al. by 5.4% and 2.1% in F1-scores and ACC-scores. It shows that our model not only has high accuracy on cyberbullying detection, but also has higher stability than other models. Whereas the Lu et al. model uses comments, temporal and the structure of social media sessions, it neither considers the media information such as video and image. The Xu et al. model based several manually extracted textual feature and takes into account the keyword in the comments, but these features have a lot of limitations. The results prove that the effectiveness of media information in cyberbullying detection and the ability of deep learning to extract features is stronger than traditional methods.

Among deep learning models, HAN has higher performance in both F1-scores and ACC-scores. It shows that it is useful to model all of the text continuously and hierarchical attention mechanisms. It also suggests the importance of comments and the social session in cyberbullying detection. Meanwhile, we notice that the attention mechanism has a positive effect on cyberbullying detection. In the Instagram dataset, the LSTM with attention model outperforms the LSTM model by 0.79% and 2.2% in F1-scores and ACC-scores. It shows that the attention mechanism improves the stability and accuracy of the model. Compared to the HAN model, our model has better performance. It shows that the posts and comments have different weights in cyberbullying detection on social media. In addition, another meta-information is also important for cyberbullying detection. As for the Text-CNN model, it does not work well with missing sequence information in the text.

For the traditional cyberbullying methods, no matter what kind of text features can not achieve high performance. Compared to other classifiers, the random forest model has higher performance. For the features, each feature plays a different role in different classifiers. It shows that the fusion of multiple features maybe gets better results.

D. Parameter Analysis

To explore the impact of different word embedding methods on the model, we use different embedding method to train the framework. At first, we select several different size pre-train model, including en-glove-6b-300d, en-glove-42b-300d, en-glove-840b-300d and en-word2vec-300[2]. The result are shown in Fig 2-3.

From Fig 2-3, we found that as the corpus grows, the pre-trained Glove model has better performance in F1-scores and ACC-scores. In the Instagram dataset, the en-glove-840b-300d model outperforms en-glove-42b-300d by 0.16% and 2.71% in F1-scores and ACC-scores. In the Vine dataset, the en-glove-840b-300d model outperforms en-glove-42b-300d by 0.66% and 1.5% in F1-scores and ACC-scores. The reason for the better outperforms is that, the larger corpus contains more words. The 840b model has nearly 10% more pre-trained words in both datasets than the 42b model. Compare with the word2vec model, the Glove model has better performance because the word2vec uses the news as the corpus and only includes 37% words of datasets. From the above result, we can conclude that a better word embedding model can improve the overall effect of the model.

To study the sensitivity and effect of the learning rate, We vary the values of learning rate and evaluate how it affects the overall performance (F1-scores). We summarize the effect of learning rate results in Fig 4.

[2]Available all of embedding model from the site: https://docs.qq.com/sheet/DVnpkTnF6VW9UeXdh?tab=BB08J2&c=B31A0AA0

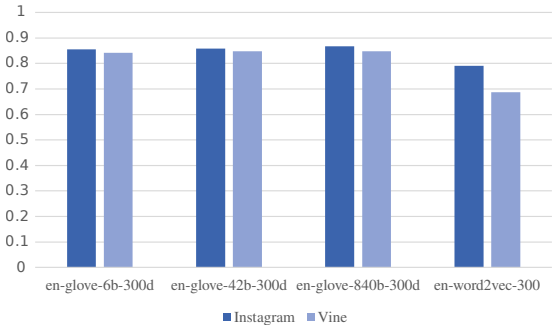

Fig. 2. F1 scores for different word embeddings on Instagram and Vine datasets.

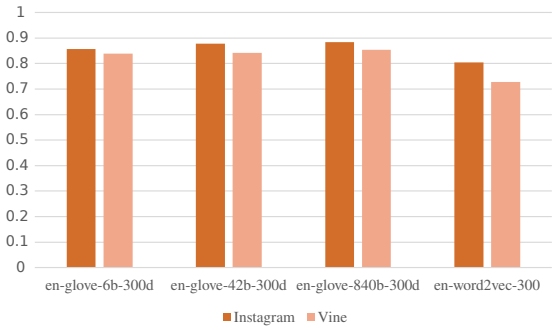

Fig. 3. ACC scores for different word embeddings on Instagram and Vine datasets.

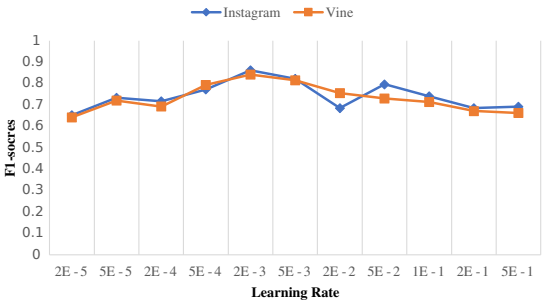

Fig. 4. The effect of learning rate on model.

From Figure 4, we can know that our model is robust to learning rate a broad range but not perform well when the learning rate extremely large and small. When the learning rate is large, the model cannot accurately update the parameters, resulting in poor results. When the learning rate is low, it cannot be learned entirely in a limited epoch, resulting in poor performance. Overall, our model performs well for a wide range of learning rates. Therefore, the learning rate can be adjusted according to various purposes.

VI. Conclusions

In this paper, we propose a novel multi-modal cyberbullying detection framework that uses three modules to extract modality features in a social network and fuse multiple data types. The first module uses bidirectional LSTM with attention to extracting the post's characteristics. As for the comments of each post, We introduce hierarchical attention networks to apply at word and comment level. Then we use MLP to encode other meta information, such as video and image. By processing different modal data, we construct a multi-modal cyberbullying detection framework and utilize the framework to two real social platform datasets for validation. Experiments demonstrate that our model can make better use of multi-modal data to deal with new cyberbullying methods.

In the future, another vital direction may be multi-modality information fusion in cyberbullying detection, which will consider the associations of modal data in social media, and model the new type of cyberbullying behavior. Mini the characteristics of bullying behavior in social networks, and Efforts for more accurate and useful cyberbullying detection models.

Acknowledgment

This work was supported by the Major Science & Technology Program of Guangxi (Grant No.GKAA17129002), The Graduate Scientific Research and Innovation Foundation of Chongqing (No.CYS19028), The Key Research Program of Chongqing (Grant No.CSTC2017jcyjBX0025 and cstc2019jscx-zdztzx0031), and NSFC (61771077).

References

[1] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[2] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[3] ——, "Us and them: identifying cyber hate on twitter across multiple protected characteristics," *EPJ Data Science*, vol. 5, no. 1, p. 11, 2016.

[4] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2015, pp. 2354–2358.

[5] C. Chelmis, D.-S. Zois, and M. Yao, "Mining patterns of cyberbullying on twitter," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 126–133.

[6] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 235–243.

[7] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 339–347.

[8] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Canadian Conference on Artificial Intelligence*. Springer, 2014, pp. 275–281.

[9] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 52–67.

[10] A.-M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," *arXiv preprint arXiv:1802.00385*, 2018.

[11] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 85–90.

[12] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in *2016 European Modelling Symposium (EMS)*. IEEE, 2016, pp. 165–171.

[13] H. Hosseinmardi, S. Li, Z. Yang, Q. Lv, R. I. Rafiq, R. Han, and S. Mishra, "A comparison of common users across instagram and ask. fm to better understand cyberbullying," in *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*. IEEE, 2014, pp. 355–362.

[14] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *International conference on social informatics*. Springer, 2015, pp. 49–66.

[15] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[16] Y. Li and J. Ye, "Learning adversarial networks for semi-supervised text classification via policy gradient," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1715–1723.

[17] Z. Li, J. Kawamoto, Y. Feng, and K. Sakurai, "Cyberbullying detection using parent-child relationship between comments," in *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*. ACM, 2016, pp. 325–334.

[18] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in twitter data," in *2015 IEEE international conference on electro/information technology (EIT)*. IEEE, 2015, pp. 611–616.

[19] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," *arXiv preprint arXiv:1706.01206*, 2017.

[20] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[21] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ser. ASONAM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 617–622. [Online]. Available: https://doi.org/10.1145/2808797.2809381

[22] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, 2017.

[23] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 164, 2018.

[24] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, 2012, pp. 656–666.

[25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[26] A. Zhang, B. Li, S. Wan, and K. Wang, "Cyberbullying detection with birnn and attention mechanism," in *International Conference on Machine Learning and Intelligent Communications*. Springer, 2019, pp. 623–635.

[27] R. Zhang, H. Lee, and D. Radev, "Dependency sensitive convolutional neural networks for modeling sentences and documents," *arXiv preprint arXiv:1611.02361*, 2016.

[28] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European Semantic Web Conference*. Springer, 2018, pp. 745–760.

[29] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th international conference on distributed computing and networking*. ACM, 2016, p. 43.

[30] C. Zhou, C. Sun, Z. Liu, and F. Lau, "A c-lstm neural network for text classification," *arXiv preprint arXiv:1511.08630*, 2015.

[31] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2016, pp. 207–212.