information

MDPI

Article

Cyberbullying Detection in Social Networks Using Bi-GRU with Self-Attention Mechanism

Yong Fang [1], Shaoshuai Yang [1], Bin Zhao [2] and Cheng Huang [1,*]

[1] College of Cybersecurity, Sichuan University, Chengdu 610065, China; yfang@scu.edu.cn (Y.F.); yss@stu.scu.edu.cn (S.Y.)
[2] CETC Avionics Co., Ltd., Chengdu 611731, China; zhaob@cetca.net.cn
* Correspondence: codesec@scu.edu.cn

Abstract: With the propagation of cyberbullying in social networks as a trending subject, cyberbullying detection has become a social problem that researchers are concerned about. Developing intelligent models and systems helps detect cyberbullying automatically. This work focuses on text-based cyberbullying detection because it is the commonly used information carrier in social networks and is the widely used feature in this regard studies. Motivated by the documented success of neural networks, we propose a complete model combining the bidirectional gated recurrent unit (Bi-GRU) and the self-attention mechanism. In detail, we introduce the design of a GRU cell and Bi-GRU's advantage for learning the underlying relationships between words from both directions. Besides, we present the design of the self-attention mechanism and the benefit of this joining for achieving a greater performance of cyberbullying classification tasks. The proposed model could address the limitation of the vanishing and exploding gradient problems. We avoid using oversampling or downsampling on experimental data which could result in the overestimation of evaluation. We conduct a comparative assessment on two commonly used datasets, and the results show that our proposed method outperformed baselines in all evaluation metrics.

Keywords: cyberbullying detection; social network; neural networks; bidirectional gated recurrent unit; self-attention mechanism

check for updates

Citation: Fang, Y.; Yang, S.; Zhao, B.; Huang, C. Cyberbullying Detection in Social Networks Using Bi-GRU with Self-Attention Mechanism. *Information* **2021**, *12*, 171. https://doi.org/10.3390/info12040171

Received: 16 March 2021
Accepted: 13 April 2021
Published: 16 April 2021

1. Introduction

As one of the problematic phenomena along with the ubiquity of social networks, cyberbullying is defined as bullying that takes place over digital devices like cell phones, computers, and tablets [1]. Nowadays, young people are the main force of the Internet. For example, 95% of teens in the U.S. are online, and the vast majority access the social networks on their mobile device [2]. Compared with traditional bullying, cyberbullying is more likely to happen during the Web 2.0 era. 36.5% of people feel they have been cyberbullied in their lifetime, and 17.4% have reported it has happened at some point in the last 30 days [3], these numbers are more than double what they were in 2007, and both represent an increase from 2018 to 2019 [4]. Besides, teachers in the U.S. report that cyberbullying is their first safety concern in their classrooms, according to a Google survey in 2019 [5]. In the last few years, online communication tends to be more and more user-driven, and social network platforms have become the hardest hit area of cyberbullying due to millions of users worldwide, efficient communication and "24/7/365" services [6].

Cyberbullying can affect people's psychological state and other aspects of life. Especially for young people, cyberbullying even could lead them to do self-harm and suicide [7]. Researchers also studied an association between cyberbullying victimization and suicidal ideation risk [8–10].

It is an essential and challenging task to detect cyberbullying. As for importance, to minimize the harm of cyberbullying as much as possible, it is more important to developing an automatic model to predict cyberbullying events, rather than checking them manually or remedying them afterward.

The challenges of cyberbullying detection are as follows. First, from the perspective of manual recognition, the judgment of whether a particular behavior is cyberbullying varies from person to person. For example, it might be considered as a cyberbullying action when a sentence contains swear words, but sometimes that is not true when some adolescents communicate with their friends in social networks. Second, as places where cyberbullying occurs most frequently [11], social network platforms usually have the nature of expressing public and anonymous expression. Hence, posts are more independent and context-free, and easy to be ambiguous or misleading. Third, one major problem with cyberbullying research is the lack of standard data [12]. Although the data used in many previous studies are obtained from the same social networks (e.g., Twitter), they are created independently using public API or by scraping from websites. Therefore, single data can not be compared to each other and is not conducive to verifying the method's universality. Furthermore, towards text messages in social networks, the length of sentences are usually short and noisy, these messages can also be unstructured, i.e., messages might contain emoji, emoticons such as : and misspellings, which confuses models capturing knowledge from text messages [13].

In this paper, we design a model based on the bidirectional GRU and the self-attention mechanism to detect cyberbullying in textual form. The main contributions of our work could be summarized as follows:

(1) We design a model using the advantages of the bidirectional gated recurrent unit (Bi-GRU) optimized with dropout and pooling layers for the avoidance of the vanishing/exploding gradient problem, and appropriate elastic net regularization for better learning convergent.

(2) We leverage the advantages of the multi-head self-attention mechanism to distinguish the importance of each social network post, which helps improve the accuracy and the robustness of the whole model.

(3) Comparing with different methods including traditional machine learning and conventional deep learning, we conduct experiments on three datasets commonly used in recent years. Concerning the fairness and objectivity of performance evaluation, none of the sampling techniques is applied during the whole experiment process. The experiment results show that our proposed model has more advantages than other baselines in all metrics.

The remainder of this paper is structured as follows. Section 2 discusses related work in this regard, mainly from the computational aspect. Section 3 presents the proposed model architecture for cyberbullying detection in detail, including each layer in the whole architecture. The experiment-related setup and results are introduced in Section 4. Section 5 concludes the paper and puts forward a vision for future work.

2. Related Work

2.1. Definition of Terms

As quoted at the beginning of Section 1, the definition of cyberbullying in this paper follows a broad sense way. In detail, we regard it as cyberbullying when a post in a social network meets at least one of the following definitions:

Definition 1. *Posts that use swear words including racial discrimination, sexual suggestiveness, religious violence, etc.*

Definition 2. *Posts that use hateful or offensive speech to attack others without a well-founded argument.*

2.2. Cyberbullying Detection

Early Studies on cyberbullying are mainly based on statistics and investigation, which focus on the definition, statistical methods and the impacts of cyberbullying, these studies enhanced the factuality of cyberbullying and made researchers pay more attention to cyberbullying from the perspective of severity [14,15]. In the aspect of computational studies, machine learning and deep learning help researchers understand more about human behaviors [16]. Cyberbullying detection has been considered a natural language processing (NLP) task.

Traditional machine learning algorithms are widely used in detecting negative forms of human behavior [17], the most common classifier adopted by researchers is the support vector machine (SVM), and the most common feature extraction method is the Bag-of-words (BoW) model [6]. Xu et al. [18] presented a preliminary work in which they verified the possibility of automatic cyberbullying detection by using several NLP models such as BoW, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation(LDA). Narhar et al. [19] used the LDA to learn topics and then plus TF-IDF values as features for training an SVM classifier. Similarly, Zhao et al. [20] concatenated BoW features and latent semantic features, then trained with a linear SVM. In the work of Waseem et al. [21], a Twitter dataset of over 16k tweets was collected, and researchers performed a grid search over several n-gram features set combinations, then turned out that using character n-grams outperforms using word n-grams at least 5 F1-points. Lee et al. [22] designed an abusive text detection system based on abusive and non-abusive word lists. It was a hybrid model composed of part-of-speech (POS) tagging, skip-gram embedding, and blacklist filtering. However, manual tagging is needed in the process of tagging and filtering, and the prediction ability of the model relies on the quality of abusive and non-abusive word lists. Murnion et al. [23] put their eyes on cyberbullying in social games. An automatic data collection system was proposed in their work, and they adopted a simple naive classifier and a twinword sentiment classifier to detect cyberbullying. Collecting data in the in-game chat is a novel way to help to supplement cyberbullying datasets.

Feature engineering in cyberbullying detection has developed a lot in the last few years. Researchers collected not only texts in social networks, but also user profiles, textual-based statistical features, etc. A typical study is the one by Davidson et al. [24], they collected a Twitter dataset using a crowd-sourced hate speech lexicon. In their research, multiple text-based features of tweets were extracted such as TF-IDF, POS tag and sentiment scores. They also counted indicators for hashtags, mentions, retweets and URLs, as well as the number of characters, words, and syllables. After testing with five machine learning algorithms (Logistic Regression, Naïve Bayes, Decision Tree, Random Forests, and linear SVM), they found that Logistic Regression and linear SVM classifier perform better than others. Finally, they used Logistic Regression with L2 regularization as the final model and reached the performance with precision, recall, and F1-score of 0.91, 0.90 and 0.90. Chatzakou et al. [25] did similar work. They extracted text, user and network-based attributes as features for training several classifiers, including J48, LADTree, NBTree, Random Forest (RF) and Functional Tree. Balakrishnan et al. [26] explored the roles of user personalities from a psychological perspective, they calculated the Big Five (extraversion, openness, conscientiousness, agreeableness, neuroticism) and Dark Triad (Machiavellianism, psychopathy, narcissism) scores, along with user-based and text-based features; different combinations of features were tested using Weka, i.e., an open-source software.

Deep learning and neural networks are widely applied in recent years. Zhao et al. [27] developed a semantic-enhanced marginalized denoising auto-encoder (smSDA) based on stacked denoising autoencoder (SDA), it improved semantic learning ability by adding dropout noise and sparsity constraints. Badjatiya et al. [28] performed extensive experiments with multiple deep learning methods including convolutional neural network (CNN), long short-term memory (LSTM) and FastText, combined with gradient boosted decision trees, the performance of deep learning model did slightly improved. Agrawal et al. [29] also verified the feasibility of neural networks for detecting cyberbullying. They experimented with four neural networks on three cyberbullying datasets from different social network platforms. However, two noteworthy studies [30,31] discussed the limitations of the above two works [28,29] in data processing. In these two works, the oversampling method was handled for data processing, which led to overfitting of data, in other words, performance claims of the models in these two works had become overestimated. Lu et al. [32] proposed a character-level convolutional neural network with shortcuts, their model turned out that char-level features could be learned to overcome spelling errors and intentional obfuscation in online posts. Zhang et al. [33] introduced a method combining the one-dimensional CNN and the single GRU network, they experimented on the dataset of the Twitter platform and obtained an increase between 1 and 13% in F1-score. Albadi et al. [34] tried to combine feature engineering with RNN to detect religious hate speech on Twitter platform, they collected 6000 Arabic posts using Twitter search API, and the experimental result turned out that the single GRU layer with pre-trained word embeddings provided best precision (0.76) and F1-score (0.77), while training the same neural network on additional time, user and content features can provide better recall (0.84).

Approaches in cyberbullying detection in recent years rely upon neural language models that map words onto an embedding space. Cheng et al. [35] proposed a hierarchical attention network, in order to distinguish cyberbullying sessions that are typically composed of multiple insulting comments. In their another study [36], network representation learning (NRL) was adopted to detect cyberbullying within a multi-model context. This is a novel and representative way to track down cyberbullying session since multi-elements in social networks like user profile, media and post popularity are embedding into a heterogeneous graph neural network. Various heterogeneous graph neural networks are often used to understand the social characters of the users, such as the users' emotions, personality types, and interests [37]. Dadvar et al. [38,39] performed a reproducibility study with deep learning based models that are conducted in [29], their reproduced experiment showed that deep learning based models with word embedding outperform the machine learning models, they also thought the effect of oversampling method should be assessed.

Given the above related works, we believe that the existing studies in cyberbullying detection have the following limitations. First, regardless of the type of methods, texts in social networks are still the key feature of cyberbullying detection. To capture more semantic information and context dependence in social network posts, a better and more suitable text representation method is needed. Second, without using any resampling method, few studies detect cyberbullying against imbalanced datasets of multiple social network platforms. Third, there have been studies using the attention mechanism to detect cyberbullying, but there have been no studies using self-attention to distinguish the sentiment polarity in social network posts.

2.3. *Recurrent Neural Networks with Attention Mechanism*

With the function of processing temporal sequence inputs, recurrent neural networks (RNN) have performed well in the NLP field. Since Google AI put forward the Transformer network [40], different recurrent neural networks with various attention mechanisms are created in recent years to improve the ability of features capturing, which is critical to the NLP. For example, Korovesis et al. [41,42] presented a novel bidirectional LSTM/CNN network with an attention layer in order to capture sentiment features of texts in social networks. Based on a Greek-language corpus with a number of 342,507, their experiment showed that their attention-based RNN model has a strong ability of aspect-based sentiment prediction. Besides, a number of document metadata including the number of repetitions, the existence of emojis and the presence of keywords were examined for encouraging sentiment prediction performance. In the work of Wang et al. [43], Attention Recurrent Neural Network (AttRNN) is proposed to predict the personality traits, they experimented on a dataset of 110,728 Facebook users with 1,580,284 digital footprints and showed that their model has a better performance compared with the original Bi-GRU. Chen et al. [44] combined LSTM with a deterministic soft attention mechanism to achieve a time-sensitive early rumor detection in social networks, their experiment on two datasets including Twitter and Sina Weibo showed that the LSTM network with a soft attention mechanism could pay particular attention to distinct features over time.

3. Proposed Method

In this section, we illustrate the design of the bidirectional gated recurrent unit model with the self-attention mechanism, the overall architecture of our model is shown in Figure 1. We elaborate the details of our model by hierarchical order in the rest of this section.

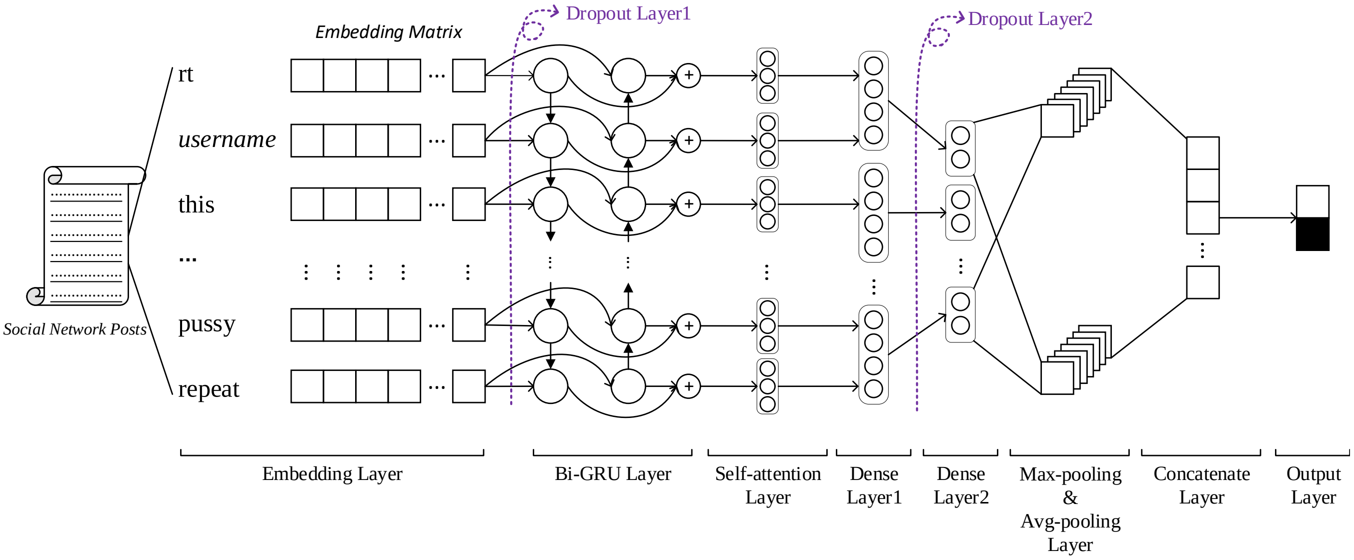


Figure 1. The architecture of proposed model.

The model accepts social network posts as the input. It is a necessary step to preprocess the input texts, which is described in Section 4.2. The first layer is the embedding layer, as the entrance to the whole model, it encodes words of sentences into word vectors. Instead of initializing the neural work with random embedding, we set the weights for the embedding layer by using pre-trained GloVe word embeddings [45]. The embedding weights are constructed by mapping words onto a fixed dimensional vector, and word embeddings have 50 dimensions. Besides, the sequence length of each sentence is fixed to the maximum in all input sequence lengths, and shorter texts are padded with zero values.

After the embedding layer, the first dropout layer is applied. The next layer is the Bi-GRU layer. Both GRU [46] and long short-term memory (LSTM) [47] are special variants of RNN with logic gates. LSTM has three gates in its structure including forget gates, input gates and output gates, while GRU has only two gates: update gates and reset gates. GRU has no output gate, so it always outputs the complete state. In most cases, GRU and LSTM have similar performance in model prediction capabilities [48,49]. From the perspective of the recurrent unit structure, GRU uses fewer connections and fewer parameters in all places of networks, so model training and generalizing are easier and faster, especially on small sequence data [50]. Given the above case, considering that the sequence length of social network posts is usually not so long, GRU is more suitable for cyberbullying detection in social networks.

The architecture of a GRU cell is shown in Figure 2, for a given time step t, a GRU cell includes one reset gate $\mathbf{R}_t$ and one update gate $\mathbf{Z}_t$. Both the reset gate and the update gate take the current time step input $\mathbf{X}_t$ and the hidden state of the previous time step $\mathbf{H}_{t-1}$ as gate input, and a sigmoid function, respectively, activates these two gates. The output of the GRU cell is the hidden state $\mathbf{H}_t$, which can be calculated by the following equations [46,51]:

$$\begin{aligned}
\mathbf{R}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xr} + \mathbf{H}_{t-1}\mathbf{W}_{hr} + \mathbf{b}_r) \\
\mathbf{Z}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xz} + \mathbf{H}_{t-1}\mathbf{W}_{hz} + \mathbf{b}_z) \\
\widetilde{\mathbf{H}}_t &= \tanh(\mathbf{X}_t\mathbf{W}_{xh} + (\mathbf{R}_t \odot \mathbf{H}_{t-1})\mathbf{W}_{hh} + \mathbf{b}_h) \\
\mathbf{H}_t &= \mathbf{Z}_t \odot \mathbf{H}_{t-1} + (1 - \mathbf{Z}_t) \odot \widetilde{\mathbf{H}}_t
\end{aligned} \tag{1}$$

where tanh represents hyperbolic tangent function, $\mathbf{W}_{xr}, \mathbf{W}_{xz}, \mathbf{W}_{xh} \in \mathbb{R}^{d\times h}$ and $\mathbf{W}_{hr}, \mathbf{W}_{hz}, \mathbf{W}_{hh} \in \mathbb{R}^{h\times h}$ are weight matrices, d and h denote to the number of inputs and hidden states, respectively, $\widetilde{\mathbf{H}}_t$ is the candidate activation variable, $\mathbf{b}_r, \mathbf{b}_z, \mathbf{b}_h \in \mathbb{R}^{1\times h}$ are bias parameters of GRU, $\odot$ stands for elementwise multiplication.

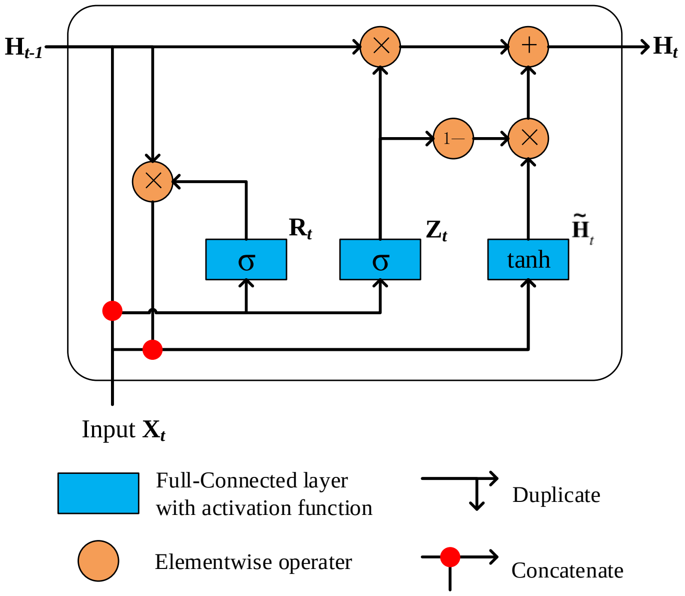


Figure 2. The architecture of a gated recurrent unit (GRU) cell.

The existence of logic gates has a crucial effect on GRU, two logical gates act like two switches in GRU. For one GRU layer, reset gates control the combination of the current input and historical information, while update gates adjust how much historical information is retained in their GRU outputs. Conventional RNNs commonly have the disadvantage of the vanishing gradient problem, the design of a full gated version of GRU cell helps our model avoid the vanishing gradient problem as much as possible.

The Bi-GRU layer consists of two GRU layers in opposite directions, which can obtain the semantic representation from different sequence order. Specifically, as shown in the Equations (2)–(4), at each time step t, the GRU cell in the forward sub-layer learns the sentence information in the order of the input sequence $\mathbf{X}_t$, the $\overrightarrow{\mathbf{H}_{t-1}}$ and the output $\overrightarrow{\mathbf{H}_t}$ have the forward direction. The backward sub-layer does the same work except that it has a reverse sequence order, the previous hidden state becomes $\overleftarrow{\mathbf{H}_{t-1}}$ and each GRU cell gets the hidden state $\overleftarrow{\mathbf{H}_t}$. At the end of Bi-GRU layer, a new hidden state is calculated by elementwise sum between the hidden state in two sub-layers. This design of the Bi-GRU layer can extract more latent relationships between words than the single GRU layer [52].

$$\overrightarrow{\mathbf{H}_t} = \overrightarrow{GRU}(\mathbf{X}_t, \overrightarrow{\mathbf{H}_{t-1}}) \tag{2}$$

$$\overleftarrow{\mathbf{H}_t} = \overleftarrow{GRU}(\mathbf{X}_t, \overleftarrow{\mathbf{H}_{t-1}}) \tag{3}$$

$$\mathbf{H}_t = \overrightarrow{\mathbf{H}_t} \oplus \overleftarrow{\mathbf{H}_t} \tag{4}$$

The outputs of the Bi-GRU layer are feed into the self-attention layer. Since the words in social network posts differ in their importance to the sentence, the self-attention layer has the function of capturing not only the semantics of individual words from both directions but also importance weights from posts. This layer includes a multi-head self-attention mechanism, which is developed by the original attention mechanism [40] that is defined as Equation (5). The original attention mechanism finds the important words from the keys for the query word, this is done by passing the key **K** and value **V** as the encoder, the output is the multiplication of the probability vector obtained by the softmax function with the value **V**. The new output vector reflects the similarity between the query **Q** and the value **V**.

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{5}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times d_k}$, $\mathbf{K} \in \mathbb{R}^{m\times d_k}$ and $\mathbf{V} \in \mathbb{R}^{m\times d_v}$, the factor $\sqrt{d_k}$ plays a regulatory role so that the inner product of **Q** and $\mathbf{K}^T$ is not too large.

The multi-head attention mechanism runs the original attention computing multiple times. Equations (6) and (7) describe the multi-head attention mechanism, the **Q**, **K**, **V** map through their own weight matrix **W** before being passed into attention function, the attention function is performed *h* times. The independent attention outputs are simply concatenated and linearly transformed into the expected dimensions. Especially, for the multi-head self-attention, as shown in Equation (8), the input parameter **Q**, **K**, **V** are all generated from the same input sequence **X**.

$$head_i = Attention\left(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}\right) \tag{6}$$

$$Multi\_head\_attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_h) \tag{7}$$

$$Multi\_head\_self\_attention = Multi\_head\_attention(\mathbf{X}, \mathbf{X}, \mathbf{X}) \tag{8}$$

Two fully-connected layers are applied after the self-attention layer in order to alleviate the vanishing gradient problem. The first fully-connected layer has 32 units, while another has 16 units. Both layers have a rectified linear unit (ReLU) as the activation function and a kernel regularizer with an L2 regularization penalty. Between two fully-connected layers, there is another dropout layer.

In the next layer, both global max-pooling and global avg-pooling are performed simultaneously to keep main features, reduce the deviation of the estimated average value, and improve the robustness of the model. Two pooling results are concatenated as one.

The final output layer has two units since our model performs a binary classification. With sigmoid as its activation function, two units represent the prediction of cyberbullying or not.

4. Experiment

4.1. Datasets

Three datasets are used in our experiment, including two Twitter datasets and one Wikipedia dataset. Twitter is a popular microblogging and social networking service on which users post and interact with messages known as "tweets" [53], as of the end of 2020, Twitter had more than 340 million monthly active users and users posted 500 million tweets a day [54]. Wikipedia talk pages (also known as discussion pages) are administration pages where editors can discuss improvements to articles or other Wikipedia pages [55].

The first dataset was kindly provided by Waseem and Hovy [21], it includes over 16,000 annotated tweets. In their work, they collected the data by manually searching tweets that contain religious, sexual, gender, and ethnic minorities using the public Twitter search API. These tweets were collected over the course of two months and annotated manually by following the criteria of McIntosh [56] and DeAngelis [57], and finally got the inter-annotator agreement score of 0.84.

The second dataset was retrieved by Davidson [24], it includes over 24,000 annotated tweets. In their work, they used Twitter API to search tweets containing terms from the lexicon compiled by *Hatebase.org* and got 85.4 million tweets. After that, they took a random sample of 24,783 tweets. These tweets were annotated manually by CrowdFlower (CF) workers, who were provided with the label definition along with a detailed explanation, and finally got the inter-annotator agreement score of 0.92.

The third dataset was created by Wulczyn [58], it consists of over 110,000 labeled discussion comments from English Wikipedia. In their work, they used MediaWiki (https://www.mediawiki.org/wiki/MediaWiki, accessed on 12 April 2021) to generate a corpus from Wikipedia's talk pages that include the history of discussion among users. With 63 M generated comments, they also conducted manual annotation via CrowdFlower annotators and finally got Krippendorff's alpha score of 0.45.

The class distribution of two datasets is shown in Table 1.

Table 1. The statistics of original datasets.

Dataset	Label	Counts
Tweets [21]	Racism	1954
	Sexism	3122
	Neither	11,014
	Total	16,090
Tweets [24]	Hate	1430
	Offensive	19,190
	Neither	4163
	Total	24,783
Wikipedia [58]	Attack	13,590
	Normal	102,275
	Total	115,865

Following the definition of cyberbullying in Section 2.1, in our experiment, we regard all of racism, sexism, hate, and offensive speech as cyberbullying behavior, so we target binary classification task. The statistics of experiment datasets are shown in Table 2.

Table 2. The statistics of experiment datasets.

Dataset	Cyberbullying	Non-Cberbullying	Total
Tweets [21]	5076	11,014	16,090
Tweets [24]	20,620	4163	24,783
Wikipedia [58]	13,590	102,275	115,865

4.2. Data Preprocessing

To preprocess the input data, several text processing methods are applied. First, we convert url strings like www.* and http(s)://* to URL. Second, we convert the mention word @username to USERNAME. Third, special characters include &amp;, &gt; and &lt; are converted to &, > and <. Fourth, we apply lowercase to all strings. Finally, we remove any tokens with a word frequency less than 2. A worth noting point is that, as discussed in Section 2, we do not adopt any sampling measure during the whole experiment process to avoid the performance bias of all models.

4.3. Comparison of Methods

In order to evaluate the performance of the proposed model comprehensively and objectively, we introduce several baseline methods including traditional machine learning models and deep learning models in this regard.

TF-IDF + SVM: As elaborated in Section 2, TF-IDF is a classic feature in feature engineering for text classification, this kind of machine learning combination is the most popular method for cyberbullying detection even in recent years. In fact, after the experiment, we could see that the performance of this machine learning model is impressive.

The implementation of this method in our experiment is mainly powered by Scikit-learn (https://scikit-learn.org, accessed on 12 April 2021).

Char N-grams + LR [21]: This model uses character-level N-gram for hate speech detection. They perform a grid search over several n-gram features set combinations, and use logistic regression as the classifier. The implementation of this method in our experiment is also mainly powered by Scikit-learn, and the parameter settings are the same as in their paper.

Multi-features + LR [24]: This model adopts multi-features in feature engineering (https://github.com/t-davidson/hate-speech-and-offensive-language, accessed on 12 April 2021). For text surface features, it uses word n-grams that are weighted by TF-IDF. For the syntactic structure feature, it uses the Natural Language Toolkit (NLTK) (https://www.nltk.org/, accessed on 12 April 2021) to construct Penn POS tags. For the quality feature of each sentence, Flesch-Kincaid readability tests (including the Flesch reading-ease test and the Flesch—Kincaid grade level test) are applied to score the sentence readability. For sentiment features, it uses an open-source tool vaderSentiment (https://github.com/cjhutto/vaderSentiment, accessed on 12 April 2021) to assign sentiment scores for each post. There are several statistical features including the number of hashtags, mentions, retweets, and URLs. The model also use Logistic Regression with L2 regularization as the classifier.

KimCNN [59]: This is a classic CNN model proposed by Kim [59], it has been used a lot in text classification. The implementation of this method in our experiment is powered by Keras (https://keras.io, accessed on 12 April 2021), and the parameter settings are the same as in their paper.

Bi-LSTM + Attention [29]: LSTM is also a typical neural network which consists of logic gates like GRU, it is commonly used to train classification model for word embeddings. This model also has two LSTM sub-layers forward and backward with an additive attention mechanism (https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs, accessed on 12 April 2021). Note that our work differs from the above in that, in our experiment, oversampling is not performed.

4.4. Experimental Environment

The experimental environment is concluded as Table 3, we chose Keras with TensorFlow backend to construct our proposed model. For 3 machine learning models, we adopt 10-fold cross-validation, the other neural network models are performed 5-fold cross-validation due to the training time considering.

Table 3. Experiment Configuration.

Items	Configuration
System hardware	Intel® Xeon® CPU E5-2680 v4 @ 2.40 GHz 128 GB RAM NVIDIA® TITAN RTX™ GPU * 1
Operating system	Ubuntu 18.04.5 LTS
Main python libraries	Keras, TensorFlow, NumPy, scikit-learn, Pytorch

4.5. Hyperparameters Setting

For our proposed model, in the Bi-GRU layer, each GRU sub-layer has 64 hidden units, each GRU sub-layer is set to return the hidden states in all time steps and does not return the last hidden state in addition to the output, as for the bias parameter, the bias vector is enabled and initialized by zero values. In the self-attention layer, the number of 'heads' is 8 and the dimension of each 'heads' is 16. In the next two fully-connected layers, we set L2 regularizer with a regularization factor of 0.01 for both of them. For model compiling, we use binary_accuracy as metrics, use Adam with initial learning rate 1×10^{-3} as optimizer and use focal loss [60] as the loss function. Validation loss monitor is applied, learning rate will decay 10% after patience 2. Focal loss is designed for overcoming the data imbalance problem, it takes two parameters α and γ to reduce the weight of easy-to-classify samples so that the model is trained to focus more on difficult-to-classify ones. The α parameter in the focal loss function controls the weight of positive samples, while the γ parameter adjusts the difficulty level of learning samples. In our proposed model, the value of α and γ is set to 0.25 and 2, respectively.

For other compared methods, the hyperparameters go with the settings in their papers.

4.6. Evaluation Method

The output result of the proposed model may either be cyberbullying or non-cyberbullying due to we target a binary classification task. Evaluation metrics in this study including precision, recall and F1-score. The precision describes the ratio of predicting positive observations, it distinguishes the actual cyberbullying posts in all predicted positive ones. The recall tells the ratio of correctly predicted positive observations in the actual positive observations, it reflects how many cyberbullying posts have been predicted in all actual positive samples. The F1-score calculates the weighted combination of precision and recall, it is an important metric when the data class distribution is unbalanced that reflects the robustness of the model. Furthermore, we introduce the receiver operating characteristic curve (ROC) to evaluate all models comprehensively. Within the ROC curve, it is our wish that one model has a larger area under the curve (AUC), which means the model has a better performance in all metrics.

4.7. Results and Discussion

Table 4 shows all models' results under each class, and all evaluation metrics are applied to evaluate each model. It could tell that the proposed model always has the highest F1-score in each class of all datasets, which shows the best robustness among all models.

Table 4. Experiment Result for Each Class.

Dataset	**Method**	**Non-Cberbullying**			**Cyberbullying**		
		Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Tweets [21]	TF-IDF + SVM	0.868	0.849	0.859	0.686	0.719	0.702
	Char N-grams + LR [21]	0.874	0.844	0.859	0.683	0.735	0.708
	Multi-features + LR [24]	0.869	0.844	0.857	0.680	0.723	0.701
	KimCNN [59]	0.850	**0.889**	0.869	0.730	0.659	0.693
	Bi-LSTM + Attention [29]	0.857	0.864	0.861	0.698	0.686	0.692
	Proposed Method	**0.893**	0.885	**0.889**	**0.753**	**0.769**	**0.761**
Tweets [24]	TF-IDF + SVM	0.784	0.931	0.851	0.986	0.948	0.967
	Char N-grams + LR [21]	0.799	0.886	0.840	0.976	0.955	0.966
	Multi-features + LR [24]	0.785	**0.948**	0.859	**0.989**	0.948	0.968
	KimCNN [59]	0.844	0.833	0.838	0.966	0.969	0.968
	Bi-LSTM + Attention [29]	0.848	0.761	0.802	0.953	0.970	0.962
	Proposed Method	**0.862**	0.905	**0.883**	0.981	**0.971**	**0.976**
Wikipedia [58]	TF-IDF + SVM	0.971	0.917	0.944	0.562	0.797	0.659
	Char N-grams + LR [21]	**0.977**	0.912	0.944	0.560	**0.838**	0.671
	Multi-features + LR [24]	0.975	0.880	0.925	0.480	0.832	0.609
	KimCNN [59]	0.960	0.970	0.967	**0.809**	0.692	0.746
	Bi-LSTM + Attention [29]	0.958	**0.973**	0.965	0.766	0.676	0.718
	Proposed Method	0.965	**0.973**	**0.969**	0.780	0.734	**0.756**

In the first dataset, the proposed model achieves the best score in 5 metrics out of 6. In the second dataset, the proposed model achieves the best score in four metrics except for the recall score of the non-cyberbullying class and the precision score of the Cyberbullying class. It could tell that the model of Multi-features + LR [24] is reasonable for this dataset. In the third dataset, traditional machine learning methods all have a lousy performance on the precision score and F1-score when predicting Cyberbullying class. We think it is because the average sequence length of the third dataset (Wikipedia forum posts) is much longer than the tweets' length. Under the condition that none of the sampling methods are applied on the datasets, traditional machine learning models have a worse capability of capturing relations in long sequences. Besides, the performance of the convolutional neural network in the third dataset is impressive, especially when it comes to predicting Cyberbullying class, multiple convolution kernels had a strong ability to capture specific words in long sequences.

Table 5 shows the results of all models under the total weighted average. We can draw the observation from the comparison results that the proposed model outperformed the other compared models on all datasets.

Table 5. Experiment Result for Total Weighted Average.

Dataset	Method	Precision	Recall	F1-Score
Tweets [21]	TF-IDF + SVM	0.811	0.808	0.810
	Char N-grams + LR [21]	0.814	0.810	0.811
	Multi-features + LR [24]	0.811	0.807	0.809
	KimCNN [59]	0.813	0.816	0.814
	Bi-LSTM + Attention [29]	0.807	0.808	0.808
	Proposed Method	**0.849**	**0.848**	**0.849**
Tweets [24]	TF-IDF + SVM	0.952	0.945	0.947
	Char N-grams + LR [21]	0.947	0.943	0.944
	Multi-features + LR [24]	0.955	0.948	0.949
	KimCNN [59]	0.946	0.946	0.946
	Bi-LSTM + Attention [29]	0.935	0.937	0.936
	Proposed Method	**0.961**	**0.960**	**0.960**
Wikipedia [58]	TF-IDF + SVM	0.923	0.903	0.910
	Char N-grams + LR [21]	0.928	0.904	0.912
	Multi-features + LR [24]	0.917	0.875	0.888
	KimCNN [59]	0.939	0.940	0.939
	Bi-LSTM + Attention [29]	0.935	0.938	0.936
	Proposed Method	**0.943**	**0.945**	**0.944**

The ROC curves of all models on all datasets are shown in Figures 3–5. Our proposed model achieves the best AUC score on all datasets.

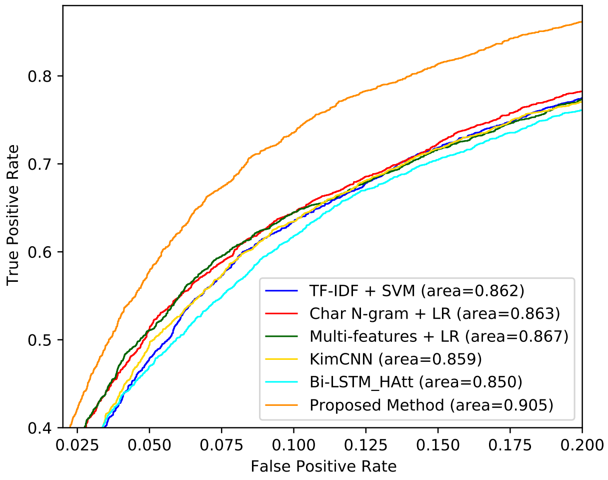


Figure 3. The ROC curve of dataset Tweets [21].

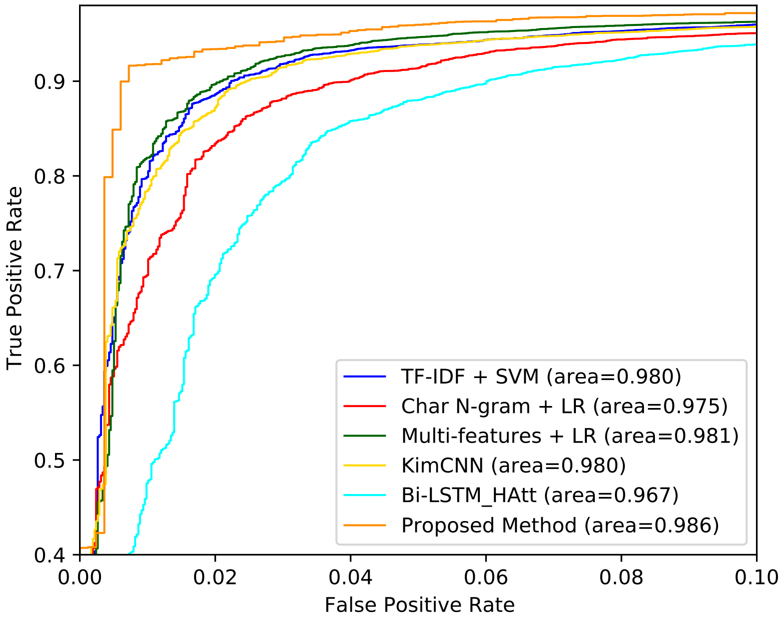


Figure 4. The ROC curve of dataset Tweets [24].

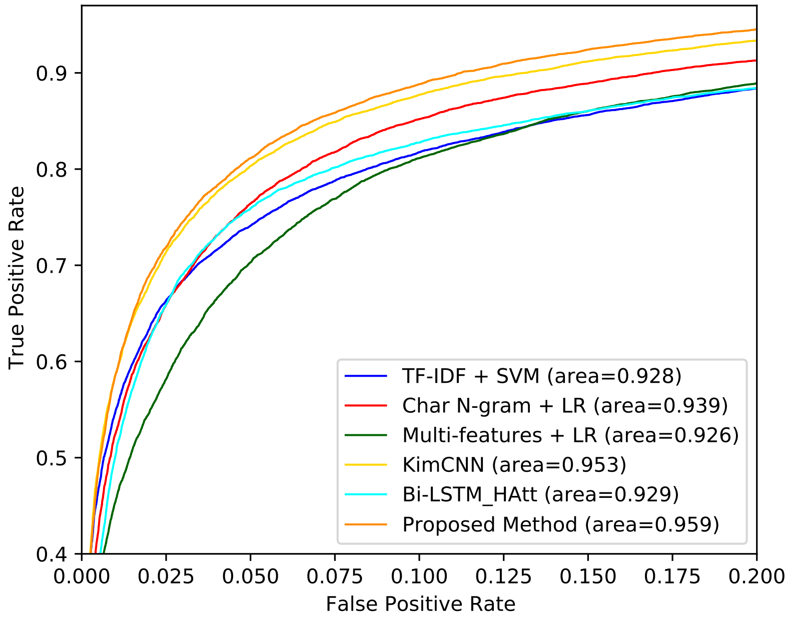


Figure 5. The ROC curve of dataset Wikiapedia [58].

The number of GRU layers matters when applying the GRU network. Theoretically, more complex features can be captured by stacking GRU layers, but the training cost of the model should also be traded off. We compared the proposed model with multiple Bi-GRU layers in different aspects to verify the impact of the number of layers in the proposed model. Table 6 shows the total weighted average performance of three different numbers of Bi-GRU layers on three datasets. It can be seen that as the number of layers increases, the metrics of all datasets are remained or slightly decreased. Figure 6 shows the average training time required for the model to reach convergence in cross-validation. As the number of layers increases, the model costs more time to converge. Figure 7 shows the average GPU memory required for the model to train the network in cross-validation. It can be seen that the 2-Layer and the 3-Layer model require the same memory usage and are larger than the 1-Layer model (models are implemented by Keras with Tensorflow as backend).

Table 6. Experiment Result for Proposed Method under Various Number of Bi-GRU Layers.

Dataset	Number of Layers	Precision	Recall	F1-Score
Tweets [21]	1	0.849	0.848	0.849
	2	0.833	0.836	0.834
	3	0.830	0.831	0.830
Tweets [24]	1	0.961	0.960	0.960
	2	0.958	0.957	0.957
	3	0.958	0.957	0.957
Wikipedia [58]	1	0.943	0.945	0.944
	2	0.943	0.944	0.944
	3	0.940	0.942	0.941

GRU solves the problem of RNN long-distance dependence and gradient disappearance mainly within each layer rather than between layers. As the number of GRU layers increases, the gradient disappearance between layers becomes obvious, further bringing information loss in information compression. Coupled with the time series processing model, for models with multiple GRU layers, the update and iteration of the GRU layer close to the input layer slow down, which makes the effect and efficiency of model convergence drop significantly, and even causes to enter the local minimum dilemma. Compared to the conventional RNN model, the GRU layer based on time series costs more time and memory to train since the neurons increase. More GRU layers mean more training costs. Given the above case, the proposed model with a single Bi-GRU layer would be proper.

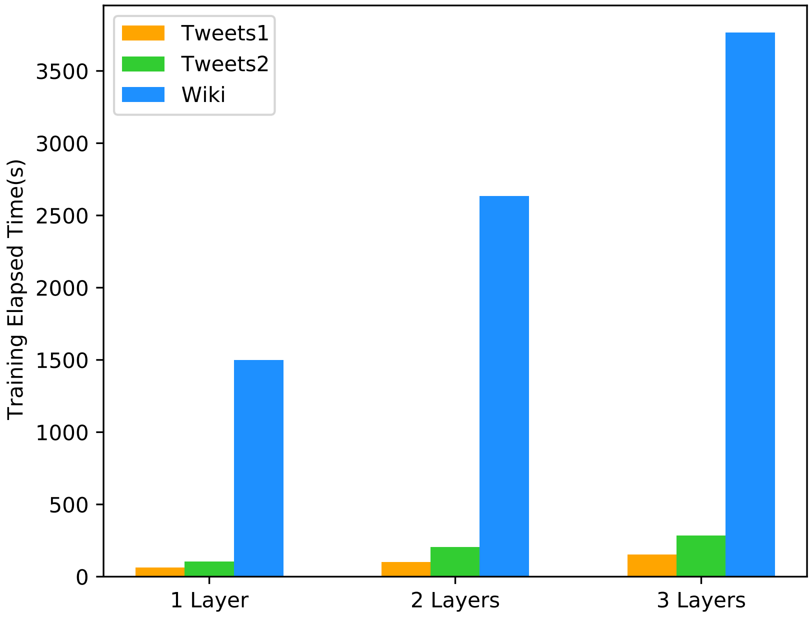


Figure 6. Average Training Time with Various Number of Bi-GRU Layers.

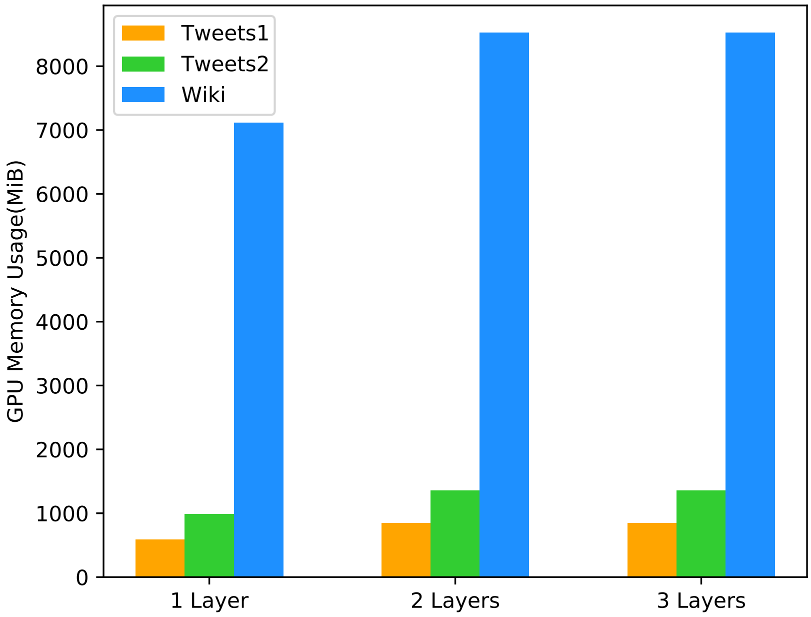


Figure 7. Average GPU Usage with Various Number of Bi-GRU Layers.

The proposed model without a multi-head self-attention layer was trained to explore the multi-head self-attention mechanism's role in the proposed model. Table 7 shows the F1-score of each class and total weighted average metrics of the proposed model with and without a multi-head self-attention layer. From the perspective of each class, the self-attention layer significantly helps the model improve the performance of the class with fewer samples on all three unbalanced datasets. From the total performance perspective, the former is 1–2% higher than the latter in all metrics.

Figure 8 is a heatmap that visualizes the multi-head self-attention weights in a sentence example, and it is clear that words in a sentence differ in weights from each other. Figure 9 shows some posts colored with their words' self-attention weights. The more the attention weights, the deeper the color. By observing the more profound colored words, the proposed model's self-attention layer mainly focuses on two types of terms: First, particular nouns and verbs related to swearing. Second, adjectives and adverbs contain emotional expressions. In view of the above case, we think the multi-head self-attention layer is necessary for the model.

Table 7. Experiment Result for Proposed Method with/without Self-attention.

Dataset	Self-Attention	Non-Cberbullying	Cyberbullying	Total Weighted		
		F1-Score	F1-Score	Precision	Recall	F1-Score
Tweets [21]	Y	0.889	0.761	0.849	0.848	0.849
	N	0.879	0.720	0.828	0.831	0.829
Tweets [24]	Y	0.883	0.976	0.961	0.960	0.960
	N	0.846	0.969	0.948	0.948	0.948
Wikipedia [58]	Y	0.969	0.756	0.943	0.945	0.944
	N	0.960	0.727	0.924	0.926	0.925

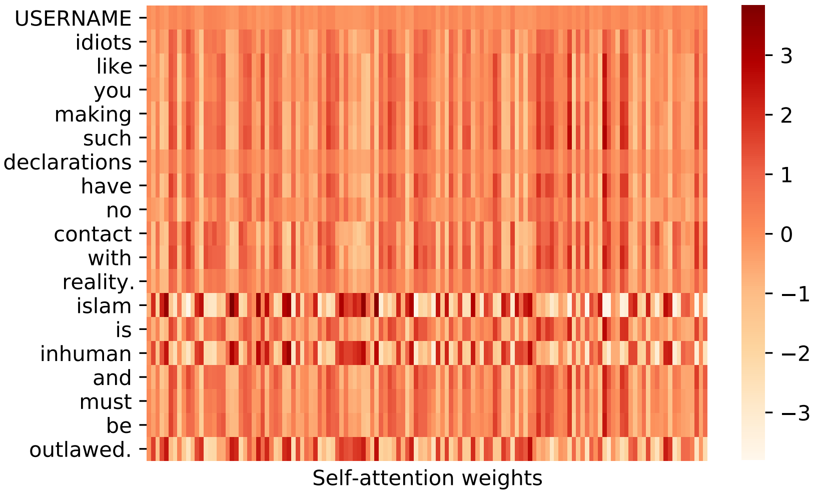
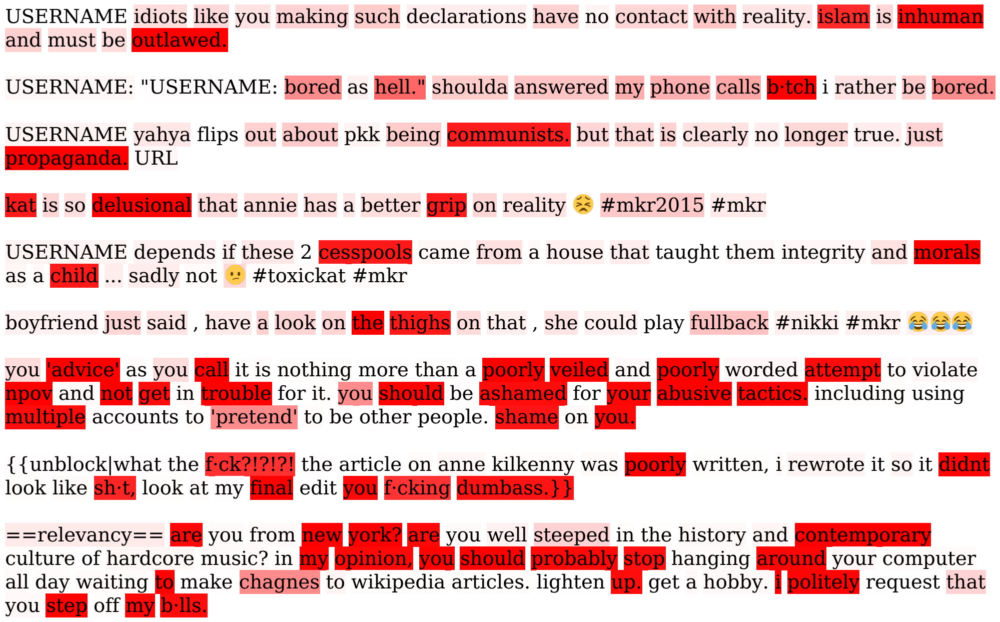


Figure 8. The Heatmap of the Self-attention Weights in an Example.


USERNAME idiots like you making such declarations have no contact with reality. islam is inhuman and must be outlawed.

USERNAME: "USERNAME: bored as hell." shoulda answered my phone calls b·tch i rather be bored.

USERNAME yahya flips out about pkk being communists. but that is clearly no longer true. just propaganda. URL

kat is so delusional that annie has a better grip on reality 😣 #mkr2015 #mkr

USERNAME depends if these 2 cesspools came from a house that taught them integrity and morals as a child ... sadly not 😕 #toxickat #mkr

boyfriend just said , have a look on the thighs on that , she could play fullback #nikki #mkr 😂😂😂

you 'advice' as you call it is nothing more than a poorly veiled and poorly worded attempt to violate npov and not get in trouble for it. you should be ashamed for your abusive tactics. including using multiple accounts to 'pretend' to be other people. shame on you.

{{unblock|what the f·ck?!?!?! the article on anne kilkenny was poorly written, i rewrote it so it didnt look like sh·t, look at my final edit you f·cking dumbass.}}

==relevancy== are you from new york? are you well steeped in the history and contemporary culture of hardcore music? in my opinion, you should probably stop hanging around your computer all day waiting to make chagnes to wikipedia articles. lighten up. get a hobby. i politely request that you step off my b·lls.

Figure 9. The Examples of Posts with Self-attention Weights Colored.

The proposed model benefits from the advantage of Bi-GRU on sequential modeling. The Bi-GRU layer learns relationships between words from both directions of a social network post, in the meantime, it has more ability to capture underlying semantics that might be cyberbullying. The multi-head self-attention mechanism helps improve the model's performance by focusing on those words or combinations important inside a post itself. Inspecting the proposed model, the model is more robust in avoiding the vanishing gradient problem compared to other neural network baselines.

5. Conclusions

Following the success of neural networks in the NLP field, we studied the use of the deep learning architecture for the detection of cyberbullying. In this paper, we designed a sequence model to automatically classify cyberbullying posts in social networks based on deep neural networks. Inside the proposed model architecture, we implemented the Bi-GRU that learns the word embedding sequence from two directions. It can further improve the representation quality of social network posts. The self-attention mechanism is adopted to help extract the attention weight of words from each sentence, which also can further improve classification accuracy. The proposed model architecture is not prone to cause the gradient vanishing and the gradient exploding problem. Often, the datasets for cyberbullying detection contain very imbalanced class distribution. We do not simply cover this issue by oversampling the training data, which could result in the overestimation of the model performance. The experimental results on two public Twitter datasets demonstrate our advantage over the compared baselines.

Concerning future work, we could explore a number of ways of cyberbullying detection. It is argued that deep neural networks can also benefit from features of other dimensions, and we will explore features such as user profile features and textual statistical features. Besides, we will study some novel word embedding ways to measure the effects on model performance. We also will try to apply graph neural networks for the model architecture to learn more relationships between posts from a broader perspective.

Author Contributions: Conceptualization, Y.F., S.Y. and C.H.; Data curation, S.Y.; Funding acquisition, Y.F. and C.H.; Investigation, S.Y., B.Z. and C.H.; Methodology, S.Y. and C.H.; Software, S.Y. and B.Z.; Supervision, Y.F. and C.H.; Validation, B.Z.; Writing—original draft, S.Y.; Writing—review & editing, Y.F. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Natural Science Foundation of China under Grant (No. 61902265), National Key Research and Development Program of China under Grant (No. 2018YFB0804503), Sichuan Science and Technology Program under Grant (No. 2020YFG0076).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in reference number [21,24,58].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stopbullying.gov. *What Is Cyberbullying*; Stopbullying.gov: Washington, DC, USA, 2020.
2. Hinduja, S.; Patchin, J.W. *Cyberbullying: Identification, Prevention, and Response*; Cyberbullying.org: Boca Raton, FL, USA, 2018.
3. Cyberbullying Research Center. *2019 Cyberbullying Data*; Cyberbullying.org: Boca Raton, FL, USA, 2020.
4. Cyberbullying Research Center. *Summary of Our Cyberbullying Research (2007–2019)*; Cyberbullying.org: Boca Raton, FL, USA, 2020.
5. Google. *Be Internet Awesome: Online Safety & Parents*; Google: Mountain View, CA, USA, 2019.
6. Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access* **2019**, *7*, 70701–70718. [CrossRef]
7. John, A.; Glendenning, A.C.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; Hawton, K. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *J. Med. Internet Res.* **2018**, *20*, e129. [CrossRef]
8. Hinduja, S.; Patchin, J.W. Bullying, cyberbullying, and suicide. *Arch. Suicide Res.* **2010**, *14*, 206–221. [CrossRef]
9. Sampasa-Kanyinga, H.; Roumeliotis, P.; Xu, H. Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans and Attempts among Canadian Schoolchildren. *PLoS ONE* **2014**, *9*, e102145. [CrossRef]
10. Zaborskis, A.; Ilionsky, G.; Tesler, R.; Heinz, A. The Association Between Cyberbullying, School Bullying, and Suicidality among Adolescents. *Crisis* **2018**, *40*, 100–114. [CrossRef]
11. Whittaker, E.; Kowalski, R.M. Cyberbullying via social media. *J. Sch. Violence* **2015**, *14*, 11–29. [CrossRef]
12. Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic cyberbullying detection: A systematic review. *Comput. Hum. Behav.* **2019**, *93*, 333–345. [CrossRef]
13. Baldwin, T.; Cook, P.; Lui, M.; MacKinlay, A.; Wang, L. How noisy social media text, how diffrnt social media sources? In Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–19 October 2013; Asian Federation of Natural Language Processing: Nagoya, Japan, 2013; pp. 356–364.
14. Patchin, J.W.; Hinduja, S. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence Juv. Justice* **2006**, *4*, 148–169. [CrossRef]
15. Smith, P.K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; Tippett, N. Cyberbullying: Its nature and impact in secondary school pupils. *J. Child Psychol. Psychiatry* **2008**, *49*, 376–385. [CrossRef]
16. Subrahmanian, V.; Kumar, S. Predicting human behavior: The next frontiers. *Science* **2017**, *355*, 489. [CrossRef] [PubMed]

17. Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Comput. Hum. Behav.* **2016**, *63*, 433–443. [CrossRef]
18. Xu, J.M.; Jun, K.S.; Zhu, X.; Bellmore, A. 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, Montreal, QC, Canada, 3–8 June 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 656–666.
19. Nahar, V.; Li, X.; Pang, C. An effective approach for cyberbullying detection. *Commun. Inf. Sci. Manag. Eng.* **2013**, *3*, 238–247.
20. Zhao, R.; Zhou, A.; Mao, K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In Proceedings of the 17th International Conference on Distributed Computing and Networking, Singapore, 4–7 January 2016; ACM: New York, NY, USA, 2016; pp. 1–6.
21. Waseem, Z.; Hovy, D. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 88–93.
22. Lee, H.S.; Lee, H.R.; Park, J.U.; Han, Y.S. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decis. Support Syst.* **2018**, *113*, 22–31. [CrossRef]
23. Murnion, S.; Buchanan, W.J.; Smales, A.; Russell, G. Machine learning and semantic analysis of in-game chat for cyberbullying. *Comput. Secur.* **2018**, *76*, 197–213. [CrossRef]
24. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; AAAI Press: Montreal, QC, Canada, 2017; pp. 512–515.
25. Chatzakou, D.; Kourtellis, N.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Vakali, A. Mean Birds: Detecting Aggression and Bullying on Twitter. In Proceedings of the 2017 ACM on Web Science Conference, Troy, NY, USA, 25–28 June 2017; ACM: New York, NY, USA, 2017; pp. 13–22.
26. Balakrishnan, V.; Khan, S.; Fernandez, T.; Arabnia, H.R. Cyberbullying detection on twitter using Big Five and Dark Triad features. *Personal. Individ. Differ.* **2019**, *141*, 252–257. [CrossRef]
27. Zhao, R.; Mao, K. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Trans. Affect. Comput.* **2016**, *8*, 328–339. [CrossRef]
28. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 759–760.
29. Agrawal, S.; Awekar, A. Deep learning for detecting cyberbullying across multiple social media platforms. In Proceedings of the European Conference on Information Retrieval, Grenoble, France, 26–29 March 2018; Springer International Publishing: Cham, Switzerland, 2018; pp. 141–153.
30. Arango, A.; Pérez, J.; Poblete, B. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 45–54.
31. Emmery, C.; Verhoeven, B.; De Pauw, G.; Jacobs, G.; Van Hee, C.; Lefever, E.; Desmet, B.; Hoste, V.; Daelemans, W. Current Limitations in Cyberbullying Detection: On Evaluation Criteria, Reproducibility, and Data Scarcity. 2019. Available online: https://arxiv.org/abs/1910.11922 (accessed on 12 April 2021).
32. Lu, N.; Wu, G.; Zhang, Z.; Zheng, Y.; Ren, Y.; Choo, K.K.R. Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurr. Comput. Pract. Exp.* **2020**, *32*, e5627. [CrossRef]
33. Zhang, Z.; Robinson, D.; Tepper, J. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*; Springer International Publishing: Cham, Switzerland, 2018; pp. 745–760.
34. Albadi, N.; Kurdi, M.; Mishra, S. Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. *Soc. Netw. Anal. Min.* **2019**, *9*, 1–19. [CrossRef]
35. Cheng, L.; Guo, R.; Silva, Y.; Hall, D.; Liu, H. Hierarchical attention networks for cyberbullying detection on the instagram social network. In Proceedings of the 2019 SIAM International Conference on Data Mining, Calgary, AB, Canada, 2–4 May 2019; pp. 235–243.
36. Cheng, L.; Li, J.; Silva, Y.N.; Hall, D.L.; Liu, H. Xbully: Cyberbullying detection within a multi-modal context. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; pp. 339–347.
37. Dhelim, S.; Ning, H.; Aung, N. ComPath: User Interest Mining in Heterogeneous Signed Social Networks for Internet of People. *IEEE Internet Things J.* **2020**, *8*, 7024–7035. [CrossRef]
38. Dadvar, M.; Eckert, K. Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *arXiv* **2018**, arXiv:1812.08046.
39. Dadvar, M.; Eckert, K. Cyberbullying detection in social networks using deep learning based models. In Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery, Bratislava, Slovakia, 14–17 September 2020; Springer: Cham, Switzerland, 2020; pp. 245–255.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 5998–6008.

41. Alexandridis, G.; Michalakis, K.; Aliprantis, J.; Polydoras, P.; Tsantilas, P.; Caridakis, G. A Deep Learning Approach to Aspect-Based Sentiment Prediction. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5–7 June 2020; Springer: Cham, Switzerland, 2020; pp. 397–408.
42. Korovesis, K.; Alexandridis, G.; Caridakis, G.; Polydoras, P.; Tsantilas, P. Leveraging aspect-based sentiment prediction with textual features and document metadata. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 168–174.
43. Wang, S.; Cui, L.; Liu, L.; Lu, X.; Li, Q. Personality Traits Prediction Based on Users' Digital Footprints in Social Networks via Attention RNN. In Proceedings of the 2020 IEEE International Conference on Services Computing (SCC), Beijing, China, 7–11 July 2020; pp. 54–56.
44. Chen, T.; Li, X.; Yin, H.; Zhang, J. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne, VIC, Australia, 3–6 June 2018; Springer: Cham, Switzerland, 2018; pp. 40–52.
45. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543.
46. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1724–1734.
47. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
48. Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of recurrent network architectures. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; PMLR: Lille, France, 2015; pp. 2342–2350.
49. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative Study of CNN and RNN for Natural Language Processing. 2017. Available online: https://arxiv.org/abs/1702.01923 (accessed on 12 April 2021).
50. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef] [PubMed]
51. Zhang, A.; Lipton, Z.C.; Li, M.; Smola, A.J. Dive into Deep Learning. 2020. Available online: https://d2l.ai (accessed on 12 April 2021).
52. Ma, C.; Yang, C.; Yang, F.; Zhuang, Y.; Zhang, Z.; Jia, H.; Xie, X. Trajectory factory: Tracklet cleaving and re-connection by deep siamese bi-gru for multiple object tracking. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; IEEE: San Diego, CA, USA, 2018; pp. 1–6.
53. Wikipedia. *Twitter–Wikipedia*; Wikipedia: Washington, DC, USA, 2020.
54. Omnicore. *Twitter by the Numbers: Stats, Demographics & Fun Facts*; Omnicore: London, UK, 2020.
55. Wikipedia. *Talk Pages*; Wikipedia: Washington, DC, USA, 2021.
56. McIntosh, P. *White Privilege: Unpacking the Invisible Knapsack*; ERIC: Norfolk County, MA, USA, 1988.
57. DeAngelis, T. Unmasking racial micro aggressions. *Monit. Psychol.* **2009**, *40*, 42.
58. Wulczyn, E.; Thain, N.; Dixon, L. Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 1391–1399.
59. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1746–1751.
60. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: Venice, Italy, 2017; pp. 2999–3007.