# Chinese Cyberbullying Detection: Dataset, Method, and Validation

Yi Zhu<sup>1,2</sup>, Xin Zou<sup>1</sup>, Xindong Wu<sup>2,3</sup>

<sup>1</sup>School of Information Engineering, Yangzhou University, Yangzhou 225009, China
 <sup>2</sup>Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230009, China

<sup>3</sup>School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

zhuyi@yzu.edu.cn, mz120241002@stu.yzu.edu.cn, xwu@hfut.edu.cn

#### **Abstract**

Existing cyberbullying detection benchmarks were organized by the polarity of speech, such as "offensive" and "non-offensive", which were essentially hate speech detection. However, in the real world, cyberbullying often attracted widespread social attention through incidents. To address this problem, we propose a novel annotation method to construct a cyberbullying dataset that organized by incidents. The constructed CHNCI is the first Chinese cyberbullying incident detection dataset. which consists of 220,676 comments in 91 incidents. Specifically, we first combine three cyberbullying detection methods based on explanations generation as an ensemble method to generate the pseudo labels, and then let human annotators judge these labels. Then we propose the evaluation criteria for validating whether it constitutes a cyberbullying incident. Experimental results demonstrate that the constructed dataset can be a benchmark for the tasks of cyberbullying detection and incident prediction. To the best of our knowledge, this is the first study for the Chinese cyberbullying incident detection task.

#### 1 Introduction

Cyberbullying has become a pervasive issue and attracted widespread social attention, which manifests in diverse online platforms and often causing severe emotional, psychological, and societal repercussions [Mahmud et al., 2023; dos Santos et al., 2024]. In recent decades, cyberbullying detection aims to identify and mitigate abusive content promptly, thereby safeguarding online environments and ensuring user well-being [Bozyiğit et al., 2021]. Despite extensive research conducted on cyberbullying detection in various languages, including English [Kim et al., 2021; Salawu et al., 2017], German [Fischer et al., 2020; Schultze-Krumbholz et al., 2013], Russian [Boronenko et al., 2013; Kintonova et al., 2021], and Arabic [ALBayari et al., 2021; Musleh et al., 2024], Chinese cyberbullying detection has received limited attention. In this paper, we address this gap by focusing on the Chinese cyberbullying detection task.

To enable the development and evaluation of effective Chinese cyberbullying detection methods, a large-scale dataset is intuitively important. Existing widely used English cyberbullying benchmarks, such as Cyberbullying\_Tweets [Wang et al., 2020a], Formspring [Reynolds et al., 2011], and MySpace [Kumar and Sachdeva, 2022], have provided valuable resources for developing and evaluating models aimed at detecting cyberbullying and related abusive behaviors online. However, these datasets were organized by the polarity of speech, which have the following two problems.

- (1) Low Coverage: The datasets classified by the polarity of speech may have limitations in addressing real practical problems, and these datasets are actually also applied for hate speech detection. However, just judging comments as "cyberbullying" and "non-cyberbullying" or "offensive" and "non-offensive" is far from enough in real-world cyberbullying related tasks, such classification does not capture the temporal dynamics or social amplification of cyberbullying incidents, which often escalate rapidly and cause widespread harm before interventions can be deployed. For example, a single offensive comment may not be problematic in isolation, but when thousands of similar comments appear in a short time frame targeting a specific individual or group, the cumulative effect can be severely damaging.
- (2) High Cost: Annotating the categories of sentences is a time-consuming and labor-intensive task. It requires human annotators to carefully consider suitable categories, taking into account the humanistic histories or cultural stories behind the text. For example, given the sentence as "The famous German landmark in 1941: The Louvre.", the true intention behind the statement is "Louvre is French, in 1941, the German Nazi regime occupied France in the World War II.", so the sentence is indeed an offensive comment. Due to the complexity of the task, annotating a large number of instances becomes challenging within a reasonable timeframe and budget.

To address these challenges, we propose a novel annotation method to construct a cyberbullying dataset that organized by incidents. Firstly, we propose an ensemble method that leverages three cyberbullying detection methods based on explanations generation to generate the pseudo labels. This automated methods can quickly generate potential labels and corresponding explanations, reducing the burden on human annotators. Then we let human annotators assess these pseudo

labels, this collaborative process harnesses the expertise of human annotators while leveraging the efficiency and scalability of machine-generated labels. Secondly, we propose the evaluation criteria for judging whether it constitutes a cyberbullying incident, and the following incident prediction can be validated. All these efficiencies allow for the creation of a larger dataset within a reasonable budget.

Our work is motivated by the following two findings:

- (1) Cyberbullying incidents are more specific and impactful than general cyberbullying behaviors. While general cyberbullying involves isolated abusive interactions, cyberbullying incidents refer to sustained, event-related abusive behaviors that escalate over a short period, often leading to serious consequences and widespread social attention. These cyberbullying incidents amplify harm due to their concentrated nature and the scale of participation, making their detection and prediction crucial for timely intervention. Notably, in the experiments, the collected incidents are the real-world events or trending topics that gain widespread public attention on Chinese social media. Cyberbullying incidents are identified within these events; however, not all such events necessarily involve cyberbullying behavior.
- (2) Machine-generated cyberbullying detection methods based on explanations can introduce interpretability for detection results. By leveraging detection techniques like language models, paraphrasing models, or multi-agents, the explanations for why it is detected as cyberbullying can be generated. These explanations enrich the dataset by identifying the humanistic histories and cultural stories behind the sentences, capturing the true intentions for cyberbullying detection. Assessing the rationality of these explanations is much simpler for the annotator compared to making decisions without prior knowledge.

In summary, our contributions are listed below:

- (1) We propose a novel annotation approach based on human and machine collaboration to construct a cyberbullying dataset that organized by incidents. Our approach provides a good idea for constructing large-scale, high-coverage datasets, especially for tasks involving distinguishing cyberbullying that are characterized by unclear decision boundaries. Based on our method, we construct the first large-scale Chinese Cyberbullying Incident dataset CHNCI that consists of 220,676 comments in 91 incidents, which cover five different text genres, namely Business, Entertainment, Sports, Society and Politics.
- (2) We present three cyberbullying detection methods based on explanations (paraphraser-based, Chain-of-Thought (CoT)-based, and multi-agents-based), and give an ensemble method that combines the three methods. Experimental results on CHNCI show that the ensemble method can be served as a strong baseline for future studies.
- (3) We provide evaluation criteria for determining whether the statements surrounding a certain event would constitute a cyberbullying incident, then the tasks of incident prediction can be further validated. Experimental results demonstrate that the constructed CHNCI can effectively reflect the development trend of cyberbullying incidents.

The dataset and code are available at https://github.com/zhuyiYZU/CHNCI.

#### 2 Related Work

# 2.1 Cyberbullying Detection Resources

Existing cyberbullying detection datasets are available for various languages, including English and other languages. In existing datasets, each instance is composed of a sentence, some attributes, and corresponding labels.

In English, the first cyberbullying detection dataset from Formspring [Reynolds et al., 2011], which has been subject to updates throughout the years. When Formspring was first created in 2011, it had nearly 4000 samples, but it has since tripled in size to 2018 [Rosa et al., 2018]. The ratio of cyberbullying instances it contains is 0.194, which refers to about 2,500 bullying text. By and large, even in English, there are very few standard datasets available for cyberbullying detection [Rosa et al., 2019]. Although most studies recur to the same social networks in order to obtain data (e.g., Twitter, YouTube), the datasets are independently created by using a publicly available API or scrapping the website for samples. For example, cyberbullying Tweets dataset [Wang et al., 2020a] comes from Twitter, comprises over 47,000 tweets annotated with cyberbullying labels, including non-cyberbullying comments and those categorized into five distinct bullying types: Age, Ethnicity, Gender, Religion, and Other. To maintain data balance, approximately 8,000 records are allocated to each category. Recently, the repeating datasets among different works have been from MySpace [Kumar and Sachdeva, 2022], which is one of the largest datasets used for cyberbullying detection, containing over 381,000 posts organized into approximately 16,000 threads. The dataset captures conversations from the MySpace platform, with posts authored by a diverse demographic, including 34% female and 64% male contributors. Each thread represents an interactive discussion, providing a rich source of contextual information for analyzing cyberbullying behaviors. The dataset is annotated with labels identifying instances of cyberbullying, offering a valuable resource for studying patterns, linguistic cues, and user interactions associated with cyberbullying. Due to its size and detailed annotations, the MySpace dataset is often employed to develop and evaluate models for cyberbullying detection and to explore gender-based differences in abusive online behavior.

Besides English, the German cyberbullying detection dataset originated from the GermEval 2018 and contains around 8,000 instances collected from Twitter, with a proportion of approximately 0.34 instances of cyberbullying. The Russian dataset 2019 is the collection of annotated comments from Russian online communication platforms, which consists of a total of 14,412 comments, the ratio of cyberbullying instances is about 0.334. The annotations were validated with the help of Russian language speakers (laypeople) using a crowd sourcing application. All the above datasets in all languages are constructed by human annotators and organized by the polarity of text. Due to their relatively small size, all of these datasets can only be used for evaluation and not for training.

Unfortunately, research on Chinese cyberbullying detection is still scarce, although there are already some existing datasets, they are either not specifically designed for Chinese

cyberbullying detection [Deng et al., 2022] or are not publicly available and accessible [Yang et al., 2025]. More importantly, there is currently no publicly available cyberbullying datasets organized by incidents, nor any benchmark dataset suitable for evaluating the performance of cyberbullying detection models in incident-based scenarios.

#### 2.2 Cyberbullying Detection Methods

Based on the different techniques to feature learning, traditional cyberbullying detection methods can roughly be categorized into four groups: machine learning-based methods [Balakrisnan and Kaity, 2023], dictionary-based methods [Mahbub et al., 2021], rule-based methods [Chong et al., 2022], and hybrid methods [Raj et al., 2021]. (1) Machine learning-based methods typically employ classification models such as Support Vector Machines (SVM) for detecting cyberbullying. For example, Raisi et al. proposed a weakly supervised machine learning approach that relies on a limited set of bullying-related vocabulary provided by experts [Raisi and Huang, 2017], which automatically infers users' bullying tendencies and language through the analysis of extensive social media interaction data. (2) Dictionary-based detection methods use predefined word lists to identify cyberbullying. For example, Wang et al. introduced a cyberbullying detection algorithm based on FastText and word similarity metrics [Wang et al., 2020b]. They constructed a list of cyberbullying-related words from the dataset and assessed whether a text contained cyberviolence risks based on text similarity. (3) Rule-based matching methods involve applying predefined rules to match texts with bullying behaviors. For example, Chong et al. compare zero-shot text classification and rule-based matching approaches for identifying cyberbullying behaviors on social media, demonstrating that zero-shot models, particularly using BART-based architectures, outperform rule-based methods in recognizing behaviors like flaming, though both methods struggle with broader behavior detection accuracy [Chong et al., 2022]. (4) Hybrid frameworks combine multiple methods based on the characteristics of the data to achieve better classification results. For instance, Almomani et al. proposes a hybrid framework combining deep learning and traditional machine learning to detect cyberbullying in images on social media platforms [Almomani et al., 2024]. By utilizing pre-trained CNN models (e.g., InceptionV3, ResNet50, VGG16) as feature extractors and feeding these features into classifiers such as Logistic Regression and SVM, the approach achieves improved detection accuracy.

Recently, deep learning methods have been increasingly employed in cyberbullying detection, owing to their ability to learn abstract features by disentangling underlying explanatory factors in the data. For instance, Iwendi et al. conducted an empirical study to evaluate the performance of traditional deep learning models, including LSTM, Bidirectional LSTM (BiLSTM), and RNN [Iwendi et al., 2023]. The study involved applying data pre-processing steps, and the results highlighted the effectiveness of these deep learning approaches. Additionally, Maity et al. proposed a graph neural network-based multitask framework for sentiment-aided cyberbullying detection [Maity et al., 2022]. The GNN frame-

work accurately identifies unlabeled or noisily labeled nodes (sentences) by aggregating information from similarly labeled nodes.

Despite the significant advantages of these deep methods, they are still limited by the effects of labeled data. More recently, with the widespread use of pre-trained language (PLMs) and Large Language Models (LLMs) [Qiang et al., 2023], these methods have shown to be extremely helpful in cyberbullying detection. For example, Yadav et al. proposed to a PLMs-based cyberbullying detection method [Yadav et al., 2020], which introduced pre-trained BERT model with a single linear neural network layer on top as a classifier. Kaddoura et al. evaluated the efficacy of open-source LLMs, e.g., Mistral 7B and Llama3, against the transformer-based model on cyberbullying detection [Kaddoura and Nassar, 2025].

# 3 Creating CHNCI

In this section, we describe our method to build a cyberbullying detection dataset organized by incidents, and the overall architecture for constructing this Chinese dataset is illustrated in Figure 1.

# 3.1 Data Preparation

In this step, we first extract the text for data preparation. To ensure diversity and complexity in our dataset, we utilize five distinct text genres: Business, Entertainment, Sports, Society and Politics. Here, genre does not refer to the linguistic style or literary form of the cyberbullying comments themselves, but rather to the topical category or domain of the incident that the comments revolve around. That is, the genre is based on the nature of the event that gave rise to the online discussion. For example, if the cyberbullying occurs in response to the "Tesla Owner Rights Defense Incident," the genre is categorized as "Business". This genre-based organization allows us to analyze how cyberbullying behavior varies across different real-world contexts. By incorporating multiple categories, we aim to capture the richness and intricacy of the Chinese language.

During the data preparation phase, we extracted relevant information about incidents, including the content of each online comment, its timestamp, and the platform of origin. The incidents here refer to the real-world events or trending topics that gain widespread public attention on Chinese social media. Considering the rapid spread and fermentation of public opinion surrounding incidents, we selected multiple mainstream Chinese social media platforms as data sources, including Douyin<sup>1</sup>, Weibo<sup>2</sup>, Xiaohongshu<sup>3</sup>, and Bilibili<sup>4</sup>. These platforms represent the primary arenas of Chinese social media, encompassing a broad user base and diverse forms of discussion. To discover such incidents, we monitored the top-trending topics on these platforms, which provide event-based or topic-tagged aggregations of posts and discussions. For an identified incident, the relevant contents

<sup>1</sup>https://www.douyin.com/

<sup>&</sup>lt;sup>2</sup>https://weibo.com/

<sup>&</sup>lt;sup>3</sup>https://www.xiaohongshu.com/

<sup>4</sup>https://www.bilibili.com/

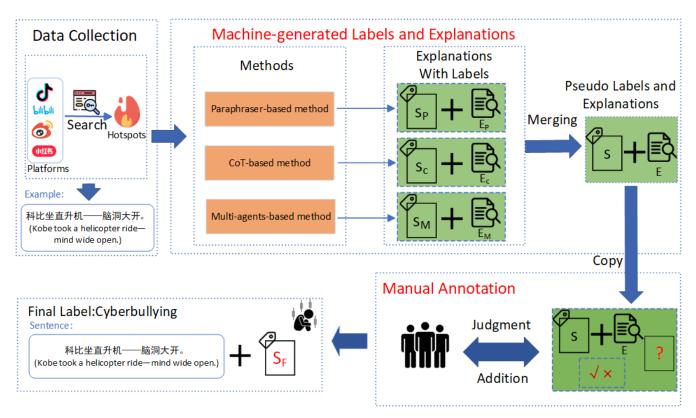


Figure 1: The overview of our method for building Chinese cyberbullying detection dataset organized by incidents. The data are collected from multiple mainstream Chinese social media platforms. Our method is composed of two phrases: machine-generated pseudo labels and manual annotation. The first phase combines three cyberbullying detection methods based on explanations as an ensemble method to generate the pseudo labels. The second phase utilizes native Chinese annotators to judge the pseudo labels with generated explanations.

of user comments and posts are directly collected from the associated topic pages or event tags, ensuring that all retrieved content was contextually aligned with the selected incident. This method guarantees that the messages in our dataset are highly relevant to the real-world event being studied.

By integrating data from multiple platforms, we aimed to comprehensively capture the diversity, complexity, and rapid dissemination characteristics of cyberbullying incidents. Using a custom-designed web crawler, we successfully collected online data for nearly 100 incidents, with about 2,000 comments gathered per incident from various platforms.

#### 3.2 Machine-generated Pseudo Labels

Considering the sentence x, we employ cyberbullying detection methods based on explanations to generate the pseudo labels and corresponding explanations. For a more accurate detection result, we adopt an ensemble approach that combines three distinct methods: paraphraser-based, CoT-based, and multi-agents-based method. By leveraging these diverse methods, each of which taps into different semantic knowledge, we aim to enhance the overall diversity of explanations available for consideration.

**Pseudo Labels Generation.** We present three baseline approaches by adapting existing cyberbullying detection methods based on explanations:

(1) Paraphraser-based: The paraphraser-based method uti-

lized the open-source LLMs to paraphrase the given input to generate explanation and pseudo label. The predictions are made through a conversational approach as "Please provide an explanation of the text and predict whether it is cyberbullying". In the experiments, the Llama3 [Dubey *et al.*, 2024] is introduced as LLM and the text is input with a 5-shot prompt.

- (2) CoT-based: The CoT-based method [Huang et al., 2023] employs a Chain-of-Thought prompt to generate high-quality explanations for cyberbullying detection. The CoT prompt templates used in this study are shown in Table 1, which are designed to guide the model through a step-by-step reasoning process to evaluate online comments for cyberbullying detection. Each template directs the model to analyze different aspects of the comment, including offensive language, targeting of individuals or groups, tone, emotional impact, and the broader context of trending events. By breaking down the analysis into distinct steps, the model is able to generate more accurate and detailed explanations for its decisions.
- (3) Multi-agents-based: The multi-agent-based method generates explanations and introduces a two-layer multi-agent voting strategy for cyberbullying detection. In this approach, a multi-agent system is employed where multiple independent agents collaborate to process the task. Each agent operates based on a distinct prompt template, enhancing the model's robustness and accuracy.

Table 1: Chain-of-Thought Prompt Templates for Cyberbullying Detection

No.	Prompt Template
1	Please analyze the following comment step-by-step. First, identify if there are any offensive or harmful words. Second, determine if the comment targets an individual or group. Finally, decide if it constitutes cyberbullying.
2	Evaluate the following online comment. Start by examining the tone and wording. Then, assess the intent and potential emotional impact on the recipient. Conclude whether the comment is cyberbullying.
3	This comment was posted in the context of a trending event. Please break down the statement, identify any signs of aggression or personal attack, and make a final judgment on whether it's cyberbullying.
4	Given the text below, follow these steps: (1) Look for negative or abusive language (2) Determine if there is a target (3) Assess the severity and impact (4) Provide a decision on cyberbullying.
5	Analyze the user's comment carefully. Begin by detecting any offensive language. Then check whether it's directed at someone. Explain your reasoning before classifying the comment as bullying or not.

Specifically, five different agents are designed, each corresponding to a prompt templates that generates a decision for each comment, the details of these agents are presented in Table 2. The voting among these agents constitutes the external voting. Each agent processes the input using its designated prompt template and produces both an "Explanation" and a "Label". The final decision is determined by majority voting across all five agents. If the majority of agents classify the comment as cyberbullying (y = 1), the comment is labeled as cyberbullying. Additionally, each agent performs an internal voting process. That is, each agent executes its prompt three times for the same input and votes based on the results. The purpose of internal voting is to reduce the potential bias of a single execution, ensuring the agent's final decision is more robust. If the number of cyberbullying predictions (y = 1)exceeds the number of non-cyberbullying predictions (y = 0)in the three runs, the agent's final decision is set to cyberbullying.

Through this double-layer voting strategy, the model first stabilizes each agent's output through internal voting and then aggregates all agents' final decisions via external voting. This method significantly improves the accuracy and reliability of cyberbullying detection.

An ensemble Method. Considering that each classification method produces different detection results, and to avoid overwhelming annotators with excessive workload or causing fatigue that might affect label quality, we decided to combine the results from these three methods. Specifically, we assigned voting weights of 1 to paraphraser-based, CoT-based, and multi-agents-based methods individually. Using these three approaches, we selected all comments labeled as either "cyberbullying" and "non-cyberbullying" as pseudo-labels. This selection process ensures that the labels generated by multiple methods are more likely to reflect potential classification outcomes, enhancing the reliability and diversity of the labels.

#### 3.3 Manual Annotation

To ensure the accuracy of the annotations, we engage multiple annotators for annotation. It is worth mentioning that all the annotators involved in this process are native Chinese undergraduates, and all of whom are native Chinese speakers with extensive experience on social media platforms. These

students are experienced users of major Chinese social media platforms, with over five years of account registration history and more than 20 hours of weekly activity. This extensive usage enables them to develop a deep understanding of offensive comments commonly found on these platforms. The information of these annotators is provided in Table 3.

Before starting the annotation process, we provided them with a detailed definition of cyberbullying to ensure they had a clear and comprehensive understanding of the annotation task. In the context of this study, cyberbullying refers to language or behavior in online platforms that intentionally insults, threatens, humiliates, or maligns individuals or groups, often leading to emotional distress or reputational damage. Unlike general hate speech, cyberbullying is typically personalized, targeted, and often context-dependent, emerging in the comments and discussions related to specific incidents. It can manifest in both direct and implicit expressions, such as sarcasm, derogatory innuendo, or culturally coded language. Annotators were trained to identify both explicit and subtle indicators of abuse, guided by generated explanations. To clarify, cyberbullying differs from general hate speech in that it typically involves personal attacks or targeted harassment rather than generic negative sentiment toward a group. For instance, a comment such as "All old people shouldn't be allowed to drive" would be categorized as hate speech, as it targets a demographic group (age) but lacks personalization. In contrast, "You're too old to drive, just stay home next time, Grandpa Li" constitutes cyberbullying, as it involves direct, targeted, and demeaning language toward a specific individual. In our annotation, only abusive content that is directed at identifiable individuals or groups in the context of a public incident is labeled as cyberbullying.

During the annotation process, we first performed a joint annotation of 1,000 samples by all annotators to establish a set of reference labels, hereafter referred to as "annotated labels". These 1,000 samples were then randomly inserted into the original dataset, without the annotators being aware of it. After completing the annotations, we compared the results of these 1,000 samples with the "annotated labels". If significant discrepancies were found, indicating that the annotations might be unreliable, we replaced them with annotations from different annotators. To ensure that each comment was accurately classified, we required at least three annotators to reach

Table 2: Five Agents Used in the Multi-Agents-based Approach

Agent	Prompt Template			
Agent 1	Read the following comment and determine whether it contains any form of cyberbullying.			
Agent 1	Provide your reasoning and output a final label: "Cyberbullying" or "Non-Cyberbullying".			
Agent 2	Analyze the text below and decide if it exhibits cyberbullying behavior. Explain your			
Agent 2	reasoning and give a label ("1" for cyberbullying, "0" for non-cyberbullying).			
Agent 3	You are an expert in online safety. Review the comment and judge whether it should be			
Agent 3	classified as cyberbullying. Justify your answer and output the classification result.			
Agent 4	Evaluate the given comment and assess whether it constitutes cyberbullying. Include your			
Agent 4	reasoning process and conclude with a label: cyberbullying or not.			
Agent 5	Determine if the following comment is an instance of cyberbullying. Write a brief			
Agent 3	explanation and assign a label (Cyberbullying/Non-Cyberbullying).			

Table 3: Demographic and Activity Profile of Annotators

<b>Total Annotators</b>	Gender (M/F)	Age Range	Avg. Registration Time	Avg. Weekly Active Hours
3	2/1	22 – 25 yrs	>5 years	>20 hours

a consensus on the label. Based on different events, we constructed annotated datasets, and the detailed statistics of the final dataset are shown in Table 4.

As shown in Figure 2, we have created a specialized website for annotating data. On each page of the website, a sentence is presented with the generated explanations. For each pseudo label, there are two radio buttons labeled "noncyberbullying" and "cyberbullying". The task of annotators is to select "cyberbullying" if they considered the given sentence to be a offensive text. Conversely, they were to choose "non-cyberbullying" if they determined that the sentence would be non-offensive.

# 4 Dataset Analysis

The statistical information of the constructed Chinese cyberbullying detection dataset organized by incidents, CHNCI, is presented in Table 4. The dataset consists of a total of 220,676 sentences with a ratio of cyberbullying instances is about 19%. On average, each sentence contains nearly 20 words. To validate the quality of constructed dataset, we conducted the following data analysis. **The Analysis of Dataset Construction** We conducted a pilot test with a single annotator, who was able to annotate approximately 360 instances within one hour when using explanation-based prompts. Notably, the average time spent per task was about 10 seconds, which is quite remarkable in terms of efficiency. In contrast, the same annotator processed only about 120 instances per hour without providing explanations, indicating that explanation-guided annotation significantly improved speed. This high

efficiency can be attributed to two main factors: first, native speakers are able to quickly make binary judgments regarding malicious language detection; second, annotators only need to read each target comment once to make a judgment on malicious language detection within a single task.

The Analysis of Dataset Coverage We show that CHNCI achieves high coverage. The dataset consists of 91 incidents, and the proportion of cyberbullying ones is about 50%. CHNCI covers five text genres: Business, Entertainment, Sports, Society and Politics. The "Cyber Violence" and "Normal Events" in the last two columns on Table 4 refers to the proportion of offensive sentences in cyberbullying and noncyberbullying incidents, respectively. The high coverage of CHNCI can be a a benchmark for evaluating incident prediction, which has shown in Section Validation.

The Analysis of Dataset Quality To validate the quality of constructed CHNCI dataset, we randomly selected 300 instances of online comments for evaluation. An annotator, proficient in Chinese and familiar with social media, was assigned to assess the accuracy of detecting malicious comments within the selected instances. The annotator carefully examined each instance to determine whether the comment was offensive or non-offensive, classifying them accordingly into cyberbullying or non-cyberbullying categories. The accuracy was calculated as 281/300, which corresponds to 93.7%. This accuracy rate, exceeding 90%, demonstrates the high quality of the dataset.

The Analysis of Dataset Agreement To measure the consistency of annotators, we further calculate common agree-

Table 4: Dataset Statistics Overview. Proportion: The average proportion of offensive comments within incidents.

Total Comments (220,676)	Cyberbullying Comments	19.20%
Total Comments (220,070)	Non-Cyberbullying Reviews	80.80%
Hotspot Events (91)	Cyber Violence	44
Tiouspot Evenus (91)	Normal Events	47
Average	Length Of Reviews	19.9
riverage	Reviews Per Event	2425
Proportion	Cyber Violence	25.76%
Troportion	Normal Events	8.85%



Figure 2: Screenshot of an annotation example on the annotation website. The red text indicates the English translation.

ment metrics such as Cohen's Kappa [Cohen, 1960] and Fleiss's Kappa [Fleiss, 1971]. Cohen's Kappa measures agreement between two raters and Fleiss's Kappa is used to assess the degree of agreement among multiple raters. The Kappa result be interpreted as follows: values ≤0 as indicating no agreement and 0.01-0.20 as none to slight, 0.21?0.40 as fair, 0.41?0.60 as moderate, 0.61?0.80 as substantial, and 0.81?1.00 as almost perfect agreement.

The inter-annotator agreement scores for the three annotators are presented in Table 5. Specifically, Fleiss' Kappa is calculated for the dataset, yielding a score of 0.609. This result indicates a substantial level of agreement among annotators, supporting the consistency and reliability of the annotation process.

# 5 Experiments

# 5.1 Experimental Setup

**Dataset.** We split the whole dataset CHNCI into train (80%), valid (10%), test (10%) set. The distribution of the text genres in CHNCI dataset is shown in Fig 3. The experimental results are validated on test sets.

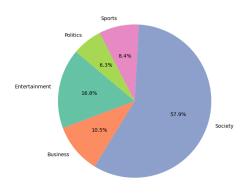


Figure 3: Category distribution of the CHNCI dataset.

**Baselines.** To evaluate the performance of different methods on our dataset, we implemented two representative baselines: HateBERT [Caselli *et al.*, 2020] and ConPrompt [Kim *et al.*, 2023], using their publicly available source code and default configurations as reported in the original studies. Both methods have been widely adopted in cyberbullying detection tasks and have demonstrated promising performance. The details are as follows:

- **HateBERT** [Caselli *et al.*, 2020]: A BERT-based model pre-trained on a large corpus of banned Reddit comments, including content flagged for offensiveness, abuse, and hate speech.
- ConPrompt [Kim et al., 2023]: A contrastive learningbased method that generates machine-inferred statements from prompts and compares them with the original prompts, which aims to pre-train a BERT model for detecting implicit hate speech.

In addition, to investigate potential biases introduced by using LLMs for explanation generation, we conducted a

Table 5: Cohen's Kappa agreement scores for pairs of annotators

Cohen's Kappa (A1-A2)	Cohen's Kappa (A1-A3)	Cohen's Kappa (A2-A3)	Fleiss's Kappa (A1-A2-A3)
0.436	0.712	0.690	0.609

comparative experiment between explanation-based annotation (involving LLM-generated reasoning) and direct labeling without explanations.

**Metrics.** We employ the designated official metrics, namely "Acc" and "F1-score" as outlined in the SemEval 2007 task. These metrics provide a comprehensive and detailed evaluation from multiple perspectives.

**Implementation Details.** Paraphraser-based (Para) and CoT-based (CoT): detections were made directly through a conversational approach with 5-shot training data. For multiagents-based (M-Agent) method, the number of agents is 5, and the threshold of internal and external voting are 3 and 5, respectively. It is worth mentioning that Llama 3-Chinese-8B is introduced as the LLMs for all the methods.

#### 5.2 Experimental Results

The results of all methods across various metrics are summarized in Figure 4. Each experiment was conducted at least three times, and the average results were calculated to ensure a fair comparison. The key findings are as follows:

- (1) Among the individual methods, M-Agent outperforms the baselines Para and CoT. This is attributed to M-Agent's incorporation of a double-layer multi-agent voting strategy, which effectively mitigates the hallucinations of LLMs, thereby enhancing cyberbullying detection performance. Without the M-Agent strategy, CoT performs better than Para, indicating that the chain-of-thought prompting mechanism in CoT better leverages the distributed knowledge within pre-trained models. When compared to HateBERT and ConPrompt, the results demonstrate that Para, CoT, and M-Agent all achieve superior performance. This validates the effectiveness of generating explanations, as these explanations capture the nuanced humanistic histories and cultural contexts underlying posts.
- (2) The experimental results further show that the methods incorporating generated explanations significantly outperform their counterparts without explanations. The improvement can be attributed to the ability of explanations, which identify the humanistic histories and cultural stories behind the sentences, thereby capturing the true intentions for cyberbullying detection. These findings indicate that prompting models to generate explanations can indeed improve prediction accuracy. Without such explanations, even large language model (LLM)-based approaches may struggle to accurately infer the true intent behind the content.
- (3) The proposed ensemble method significantly outperforms all individual models across all evaluation metrics, with statistically significant improvements. Ensemble method achieves superior results by expanding the coverage of possible explanations through the integration of multiple methods. While individual methods are often limited by specific biases and constraints, their integration within an ensemble frame-

work mitigates these shortcomings, leading to broader coverage. Furthermore, individual cyberbullying detection methods often display varying sensitivities to different linguistic contexts, word senses, or syntactic structures. By leveraging the strengths of diverse methods, the ensemble approach demonstrates enhanced robustness in handling various linguistic scenarios, thereby contributing to its superior performance.

(4) In summary, the advantages of ensemble arise from its ability to harness the diversity, expanded coverage, and robustness of individual detection methods. These factors collectively explain the significant performance improvements of the ensemble approach over individual methods across all evaluation metrics.

### 6 Validation

To validate the constructed CNHCI, we further design the validation experiments in two parts. The first part focuses on cyberbullying language detection, where various baseline methods are employed to evaluate the performance. The second part centers on time series trend forecasting, utilizing different approaches for cyberbullying incidents prediction. The process of validation is illustrated as Figure 5.

# **6.1** Baseline Methods for Validation Cyberbullying Language Detection.

The following nine baselines are selected to validate cyberbullying language detection on CHNCI, including fine-tuning PLMs, prompt-tuning methods, and the SOTA LLMs.

**Fine-tuning PLMs:** We include BERT-base-Chinese and HateBERT as classic PLM baselines. BERT represents a widely used generic language model, while HateBERT is specifically pre-trained on offensive and hate-related content, making it a strong domain-specific baseline. The details are presented as follows.

- Bert-base-Chinese [Kenton and Toutanova, 2019]:A
  pre-trained Chinese language model specifically designed for processing tasks in Chinese. By fine-tuning
  the model, it can effectively detect cyberbullying comments or non-cyberbullying language.
- HateBERT [Caselli et al., 2020]: A model tailored for hate speech detection, fine-tuned from the BERT framework. It excels in identifying offensive, racist, or hateful content in text.

**Prompt-tuning Methods:** We adopt ConPrompt, P-Tuning, KPT, and KPT++ to test the effectiveness of lightweight adaptation techniques in few-shot scenarios. These methods have shown promising performance in cyberbullying detection tasks, and allow models to leverage task-relevant knowledge with minimal label information. The details are presented as follows.

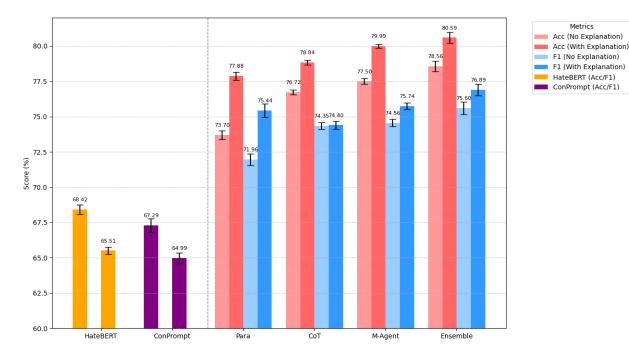


Figure 4: Performance Comparison with Baseline Methods

- ConPrompt [Kim et al., 2023]:A contrastive learningbased prompt technique designed to enhance performance in few-shot learning tasks. It remains effective in detecting cyberbullying language even in scenarios with limited data.
- P-Tuing [Liu et al., 2024]: A parameterized promptbased fine-tuning method that improves performance in few-shot tasks by optimizing model parameters. It is particularly advantageous in detecting cyberbullying language in imbalanced or small datasets.
- KPT [Hu et al., 2021]:A method that integrates external knowledge with prompt techniques to enhance model performance by leveraging background information. It is suitable for detecting complex or implicit cyberbullying language.
- KPT++ [Ni and Kao, 2023]:An enhanced version of KPT that further optimizes the combination of knowledge injection and prompt learning methods. It performs exceptionally well in cross-cultural or cross-domain cyberbullying language detection.

Large Language Models (LLMs): We evaluate Llama3, Qwen, and DeepSeek-V3 in zero-shot settings. These open-source LLMs represent the current state-of-the-art in general-purpose language understanding and generation. Including these models allows us to assess the upper bounds of performance without task-specific fine-tuning, which is crucial for scalable real-world deployment. The details are presented as follows.

 Llama3: A large-scale multilingual open-source language model developed by Meta, excelling in handling

- long texts and multilingual tasks. It is ideal for largescale and diverse cyberbullying language detection.
- Qwen: A large-scale language model developed by Alibaba, featuring multimodal capabilities and strong language comprehension. It excels at detecting cyberbullying comments and hate speech on Chinese and multilingual platforms.
- DeepSeek-V3: A large language model developed by DeepSeek, supporting both Chinese and English tasks with 128K context length capability, demonstrating excellent performance in detecting cyberbullying content in complex contexts.

# **Cyberbullying Incidents Prediction.**

The following nine baselines are selected to validate cyberbullying incidents prediction on CHNCI, including the deep learning and transformer-based methods. These models are selected for their strong performance in time series forecasting and event trend modeling.

**Deep Learning Methods:** The deep learning methods including GRU, TCN, and LSTM provide a strong reference point for evaluating temporal dynamics of incident-level data. DLinear and NLinear are recently proposed time series models and have shown excellent performance with lower computational cost, which help assess the efficacy of lightweight forecasting models.

GRU [Cho, 2014]:An optimized type of Recurrent Neural Network (RNN) that improves computational efficiency by simplifying the structure of traditional RNNs. GRU is particularly suitable for tasks involving the processing of time series data and excels in short-term time series forecasting.

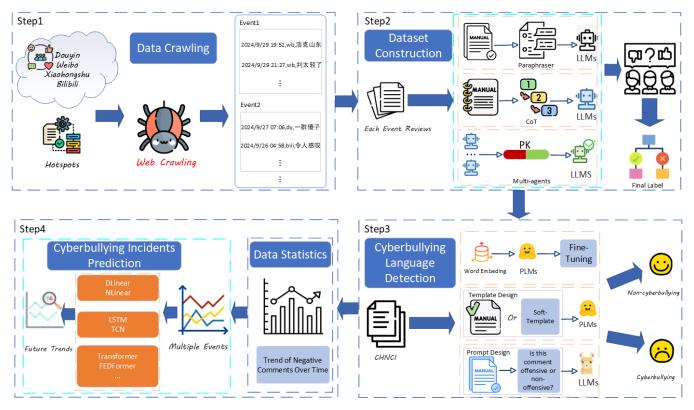


Figure 5: The process of dataset validation. Step 1: Scraping comments related to cyberbullying incidents, which may include offensive expressions such as "????" (Hulk Shandong), "????" (The judgment is too lenient), and "????" (A bunch of idiots). Step 2: Data Preparation. Step 3: Cyberbullying language detection. Step 4: Data statistics and Cyberbullying incidents prediction.

- TCN [Bai et al., 2018]: A time series modeling method based on Convolutional Neural Networks (CNNs), capturing long-term dependencies through causal convolution and dilated convolution. It is widely used in time series regression tasks, such as trend forecasting.
- LSTM [Hochreiter, 1997]:A classic type of Recurrent Neural Network that effectively addresses the problem of long-term dependencies with specially designed memory cells. It performs exceptionally well in tasks such as weather forecasting and stock trend prediction that require processing of long-term dependencies.
- DLinear [Zeng *et al.*, 2023]:A model that utilizes a decomposition mechanism to divide time series into trend and seasonal components, models them separately, and then merges the outputs. It demonstrates outstanding performance in time series forecasting tasks that exhibit trends and seasonal patterns.
- NLinear [Zeng et al., 2023]:An improved version of DLinear, which enhances the model's forecasting stability and generalization ability by introducing a normalization mechanism. It is suitable for dealing with nonstationary time series scenarios.

**Transformer-based Methods:** Four Transformer-based methods including Transformer, Informer, Autoformer, and FEDformer incorporate innovations like sparse attention, decomposition mechanisms, and frequency-domain analysis.

These methods are particularly suited to handling complex and long-term temporal patterns, making them ideal for incident-level forecasting.

- Transformer [Vaswani, 2017]:A deep learning model based on attention mechanisms that has gained widespread attention due to its powerful global context modeling capabilities. It is suitable for complex tasks such as multi-variable time series forecasting and financial analysis.
- Informer [Zhou et al., 2021]:An optimized variant of Transformer for long-term time series forecasting, significantly improving computational efficiency through sparse attention mechanisms. It is particularly suitable for time series forecasting tasks with long time spans.
- Autoformer [Wu et al., 2021]: A variant of Transformer that focuses on time series trend modeling and seasonal pattern capturing. Its automated design reduces the need for parameter tuning and performs particularly well in tasks with strong trends and cycles.
- FEDformer [Zhou et al., 2022]: A variant of Transformer that combines Fourier transformation and decomposition techniques, enhancing the model's performance by strengthening frequency domain analysis. It is suitable for complex time series forecasting tasks, such as traffic flow and meteorological data analysis.

# **6.2** Implementation Details and Evaluation Metrics

Criteria for validating cyberbullying incident. Based on the data statistics of CHNCI, we propose the evaluation criteria for validating whether it constitutes a cyberbullying incident, including two rules: (1) Offensive Comments Peak Phenomenon: If the number of offensive comments within a certain time interval exceeds 5% of the current total number of comments, it is considered that this may constitute a cyberbullying incident. (2) Multiple Clusters of Offensive Sentiments: When the proportion of offensive comments in multiple time intervals (threshold is set to 5 in the experiments) exceeds 50% of that time interval, it is considered as a cyberbullying incident.

**Experimental Setup.** In the task of cyberbullying language detection, we adopted different training strategies for various model types. When fine-tuning pre-trained language models (PLMs), we randomly selected training sets consisting of 600, 1,000, and 1,600 samples, with the remaining data reserved for testing. For prompt-tuning-based methods, we simulated few-shot learning scenarios by employing 30-shot, 40-shot, and 50-shot settings, using the rest of the samples for evaluation. In contrast, large language models (LLMs) were evaluated in a zero-shot setting without any task-specific fine-tuning.

In the task of cyberbullying incidents prediction, we first calculated the number of offensive comments for each event on an hourly basis, constructing complete time series data for each event (i.e., the number of offensive comments per hour). We then selected data from five representative events as the training set, including three cyberbullying incidents and two normal events, while using data from the remaining events as the test set. We employed a sliding window method, using the number of offensive comments from the past five time intervals as input to predict the number of offensive comments for the next time interval.

**Parameter settings** In the cyberbullying language detection, to ensure fairness, we set the following parameters for methods of the same type: for prompt-tuning methods (Ptuing, KPT, KPT++), the learning rate is set to 4e-5, batch size to 32, and the number of training epochs to 20. For finetuning PLMs (BERT, HateBERT, and Conprompt), the learning rate is set to 2e-5, batch size to 16, and the number of training epochs to 20.

Additionally, in the cyberbullying incidents prediction, to ensure fairness, we set the learning rate to 1e-3, the number of training epochs for each event to 10, and the sliding window size to 5 for all the methods.

**Evaluation metrics** In the cyberbullying language detection, the same metrics, Acc and F1-socre, are used as the evaluation metrics. In the cyberbullying incidents prediction, MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) are used as evaluation metrics. Notably, the smaller value of MAE and RMSE indicates better results.

#### **6.3** Validation Results

The results of the cyberbullying language detection and cyberbullying incidents prediction on CHNCI are shown in Table 6 and Table 7, respectively. It is worth noting that, each

experiment was conducted three times, and the average and standard deviation were computed.

Cyberbullying Language Detection. In PLMs, the data reveals that HateBERT achieved 62.59% ( $\pm 0.07$ ) accuracy with 600 samples, slightly lower than standard BERT's 63.00% ( $\pm 0.14$ ). However, it demonstrated better scalability at 1000 and 1600 sample sizes, reaching 68.22% ( $\pm 0.31$ ) and 70.49% ( $\pm 0.31$ ) accuracy respectively. This suggests that the advantages of domain-specific pretraining become more apparent as data volume increases. Conprompt showed more outstanding performance, achieving 64.97% ( $\pm 0.12$ ), 72.33% ( $\pm 0.44$ ), and 73.89% ( $\pm 0.45$ ) accuracy across the three data scales, significantly outperforming the baseline BERT model.

In few-shot learning scenarios, KPT++ demonstrated outstanding performance. With only 30 samples, it achieved an accuracy of 69.60% ( $\pm 1.00$ ) and an F1-score of 72.22% ( $\pm 0.71$ ), significantly outperforming both P-tuning and the standard KPT method. In particular, KPT++ improved the F1-score by approximately 6.5 percentage points compared to KPT, fully validating its strong generalization ability and robustness under limited data conditions.

The performance of large language models (LLMs) is particularly noteworthy. Qwen-7B achieved comprehensive leadership with 75.86% ( $\pm 0.09$ ) accuracy and 77.74% ( $\pm 0.02$ ) F1-score, maintaining its performance advantage consistently across all data scales. Llama3-8B and DeepseekV3 attained 73.04% ( $\pm 0.09$ ) and 73.43% ( $\pm 0.06$ ) accuracy respectively. While not matching Qwen-7B, they still outperformed most specialized models and prompttuning approaches. This phenomenon strongly demonstrates the powerful generalization capabilities that LLMs acquire through massive pretraining, enabling them to maintain stable high performance across different data scales.

As training data size increases, the performance of all PLMs and prompt-tuning methods improves significantly. This indicates that data quantity plays a crucial role in determining model performance in cyberbullying language detection tasks. However, prompt-tuning methods (especially KPT) show more pronounced growth on small datasets, demonstrating their adaptability in few-shot scenarios. LLMs (such as Llama3-8B, Qwen-7B, and DeepseekV3) already achieve high performance under zero-shot conditions, with Qwen-7B even surpassing all prompt-tuning methods. This suggested that LLMs, through extensive pre-training corpora and scale effects, can achieve outstanding performance even in the absence of task-specific data, making them well-suited for unsupervised or zero-shot applications.

Cyberbullying Incidents Prediction FEDformer outperforms all other methods, significantly excelling in terms of performance. Its adaptive feature selection and dynamic modeling capabilities allow it to capture essential features in time series data, making it especially suitable for tasks involving long time series predictions or tasks with long-term dependencies. FEDformer is ideal for scenarios that require high-precision forecasting, especially in the context of complex and dynamic time series data. NLinear follows closely behind, demonstrating strong capabilities in modeling non-linear relationships. This method is well-suited for

Table 6: Results (%) of cyberbullying language detection. The best values are bolded.

Method	600/30/0		1000/40/0		1600/50/0	
Wichiod	Acc	F1s	Acc	F1s	Acc	F1s
Bert	63.00 (±0.14)	66.69 (±0.36)	67.14 (±0.10)	69.86 (±0.20)	70.10 (±0.08)	72.48 (±0.17)
HateBert	62.59 (±0.07)	66.56 (±0.18)	68.22 (±0.31)	70.29 (±0.51)	70.49 (±0.31)	71.70 (±0.52)
Conprompt	64.97 (±0.12)	68.20 (±0.21)	72.33 $(\pm 0.44)$	72.23 $(\pm 0.65)$	73.89 (±0.45)	73.86 (±0.63)
P-tuning	$62.90 \ (\pm 0.79)$	66.83 $(\pm 0.70)$	$68.66 \ (\pm 1.30)$	71.49 (±0.94)	71.36 $(\pm 1.21)$	73.61 $(\pm 0.95)$
KPT	61.61 (±1.14)	65.65 ( $\pm 0.99$ )	$70.70 \ (\pm 0.03)$	$73.18$ ( $\pm 0.02$ )	73.97 (±1.71)	$75.63$ ( $\pm 0.79$ )
KPT++	$69.60 \\ (\pm 1.00)$	72.22 (±0.71)	70.79 ( $\pm 0.47$ )	$73.29$ ( $\pm 0.24$ )	$72.63$ ( $\pm 2.16$ )	74.56 (±1.59)
Llama3-8B	$73.04$ ( $\pm 0.09$ )	$73.96$ ( $\pm 0.09$ )	$73.04$ ( $\pm 0.09$ )	$73.96$ ( $\pm 0.09$ )	$73.04$ ( $\pm 0.09$ )	$73.96$ ( $\pm 0.09$ )
Qwen-7B	75.86 $(\pm 0.09)$	77.74 $(\pm 0.02)$	75.86 $(\pm 0.09)$	77.74 $(\pm 0.02)$	75.86 (±0.09)	77.74 $(\pm 0.02)$
DeepseekV3	$73.43$ ( $\pm 0.06$ )	$76.09$ ( $\pm 0.08$ )	$73.43$ ( $\pm 0.06$ )	$76.09$ ( $\pm 0.08$ )	$73.43$ ( $\pm 0.06$ )	$76.09$ ( $\pm 0.08$ )

handling time series data with significant nonlinear characteristics, which performs excellently in scenarios where data exhibits complex trends or fluctuations. DLinear also shows stable performance, with its advantage lying in the simplified linear model that provides efficient and robust predictions for simpler linear time series tasks. DLinear is suitable for scenarios with limited computational resources or where data variation is relatively small.

Table 7: Results of cyberbullying incidents prediction. The best values are bolded.

MAE	RMSE
$1.4407 (\pm 0.0611)$	$2.8429 (\pm 0.1434)$
$1.4169 (\pm 0.0649)$	$2.7558 (\pm 0.0695)$
$1.4039 \ (\pm \ 0.0468)$	$2.6394 (\pm 0.1155)$
$1.2746~(\pm~0.0096)$	$2.6000 (\pm 0.0194)$
$1.2512~(\pm~0.0181)$	$2.5527 (\pm 0.0517)$
$1.3952 \ (\pm \ 0.0732)$	$2.7860 (\pm 0.0703)$
$1.3996 (\pm 0.0847)$	$2.7931 (\pm 0.1533)$
$1.4463 \ (\pm \ 0.1002)$	$2.9284 (\pm 0.2083)$
$\boldsymbol{1.2077}\ (\pm\ \boldsymbol{0.0303})$	$\pmb{2.5259}\ (\pm\ \pmb{0.0166})$
	$1.4407 (\pm 0.0611)$ $1.4169 (\pm 0.0649)$ $1.4039 (\pm 0.0468)$ $1.2746 (\pm 0.0096)$ $1.2512 (\pm 0.0181)$ $1.3952 (\pm 0.0732)$ $1.3996 (\pm 0.0847)$ $1.4463 (\pm 0.1002)$

LSTM and GRU provide stable results, but their performance in more complex tasks is inferior to newer methods. These traditional architectures can capture long-term dependencies in time series data, but their adaptability to dynamic changes is limited, making them more appropriate for scenarios with smaller datasets and relatively stable variations. Transformer and Autoformer, despite their powerful global dependency modeling capabilities, show weaker performance

in the experiments, possibly due to poor adaptability of the model architecture to the dataset with overfitting. Transformer is more suitable for large-scale data scenarios where global information capture is essential, while Autoformer is better suited for data with strong seasonal components.

To visually express the differences between Cyberbullying and Non-Cyberbullying Incidents, we present two incidents in CHNCI, along with their corresponding word clouds, as illustrated in Figure 6 and Figure 7.

For Figure 6, the y-axis represents the ratio of cyberbullying comments within each hour. We can see that in Figure 6(a), there is a clear peak in the number of cyberbullying comments (the peak point is about 7%), which account for a large proportion of the total number of comments. In contrast, the number of cyberbullying comments in Figure 6(b) is significantly more evenly distributed (the peak point is just 0.35%), and accounts for a smaller proportion of the total number of comments. It is evident that these figures are consistent with our cyberbullying incidents prediction.

Figure 7 illustrates the differences in word clouds between cyberbullying and non-cyberbullying incidents. In Figure 7(a), high-frequency words such as "Qingdao prawns," "Who do you think you are," and "temporary worker" carry strong sarcastic and aggressive tones, reflecting the negative emotions and accusatory attitudes often expressed during cyberbullying events. In contrast, Figure 7(b) features more positive words like "Great," "Support," and "WeChat Pay," indicating approval and anticipation from users. This contrast in word frequency clearly highlights the emotional divergence in public opinion between the two types of incidents.

#### 7 Conclusions

This study presents the first comprehensive exploration of the Chinese cyberbullying detection task. We propose a novel an-

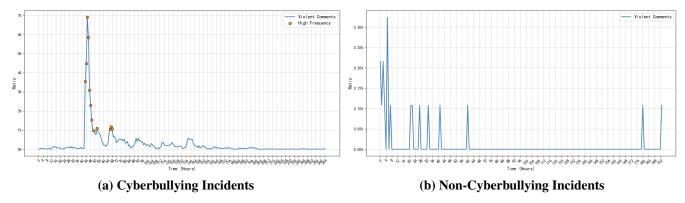
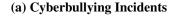


Figure 6: Hourly Trend of Comments: Comparison of Cyberbullying Incidents and Normal Events. The x-axis represents the hours elapsed since the event began, while the y-axis represents the ratio of offensive comments within each hour.







(b) Non-Cyberbullying Incidents

Figure 7: Word clouds of online comments during the event. (a) shows cyberbullying content, and (b) shows non-cyberbullying content. The figure highlights differences in high-frequency words between the two.

notation method to construct a large-scale Chinese dataset organized by incidents through a collaborative human-machine approach. The constructed dataset consists of 220,676 instances in 91 incidents, with a ratio of cyberbullying instances is about 19%. Our proposed ensemble method by leveraging the strengths of each method while mitigating their weaknesses, our ensemble approach significantly outperforms the individual cyberbullying detection methods with generated explanations across all evaluation metrics. The constructed dataset can not only be a benchmark for evaluating cyberbullying detection, but also more important, for the task of cyberbullying incident prediction.

In the future, we plan to extend our work in two directions. Firstly, we aim to explore automatic criteria for validating cyberbullying incident, feature representation learning methods can be introduced to identify cyberbullying incidents. Secondly, the constructed CHNCI dataset will be used as a benchmark for future research.

In conclusion, our study fills the research gap on how to construct a large-scale cyberbullying detection dataset with high coverage and low cost, providing a solid foundation for further development. The construction of a high-quality dataset and the development of an effective ensemble method showcase the potential for improved cyberbullying detection

in the Chinese language.

# Acknowledgments

This research is partially supported by the National Natural Science Foundation of China under grants (62076217), National Language Commission of China under grants (ZDI145-71), Key Research and Development Program of Jiangsu Province in China (BE2023315), Yangzhou Science and Technology Plan Project City School Cooperation Special Project (YZ2023199), Open Project Program of Key Laboratory of Knowledge Engineering with BigData (the Ministry of Education of China, NO. BigKEOpen2025-06).

#### References

[ALBayari *et al.*, 2021] Reem ALBayari, Sharif Abdullah, and Said A Salloum. Cyberbullying classification methods for arabic: A systematic review. In *The International Conference on Artificial Intelligence and Computer Vision*, pages 375–385. Springer, 2021.

[Almomani *et al.*, 2024] Ammar Almomani, Khalid Nahar, Mohammad Alauthman, Mohammed Azmi Al-Betar, Qussai Yaseen, and Brij B Gupta. Image cyberbullying detection and recognition using transfer deep machine learn-

- ing. International Journal of Cognitive Computing in Engineering, 5:14–26, 2024.
- [Bai et al., 2018] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.
- [Balakrisnan and Kaity, 2023] Vimala Balakrisnan and Mohammed Kaity. Cyberbullying detection and machine learning: a systematic literature review. *Artificial Intelligence Review*, 56(Suppl 1):1375–1416, 2023.
- [Boronenko *et al.*, 2013] Vera Boronenko, Vladimir Menshikov, and Gilberto Marzano. Topicality of cyberbullying among teenagers in russia and latvia. *Social Sciences Bulletin*, 1(16):84–104, 2013.
- [Bozyiğit *et al.*, 2021] Alican Bozyiğit, Semih Utku, and Efendi Nasibov. Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179:115001, 2021.
- [Caselli *et al.*, 2020] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv* preprint arXiv:2010.12472, 2020.
- [Cho, 2014] Kyunghyun Cho. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [Chong et al., 2022] Wei Jiek Chong, Hui Na Chua, and May Fen Gan. Comparing zero-shot text classification and rule-based matching in identifying cyberbullying behaviors on social media. In 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pages 1–5. IEEE, 2022.
- [Cohen, 1960] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46, 1960.
- [Deng et al., 2022] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. Cold: A benchmark for chinese offensive language detection. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11580– 11599. Association for Computational Linguistics, 2022.
- [dos Santos *et al.*, 2024] Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. Is this a violation? learning and understanding norm violations in online communities. *Artificial Intelligence*, 327:104058, 2024.
- [Dubey et al., 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.
- [Fischer *et al.*, 2020] Saskia M Fischer, Nancy John, Wolfgang Melzer, Anne Kaman, Kristina Winter, and Ludwig Bilz. Traditional bullying and cyberbullying among children and adolescents in germany–cross-sectional results of the 2017/18 hbsc study and trends. *Journal of health monitoring*, 5(3):53, 2020.

- [Fleiss, 1971] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [Hochreiter, 1997] S Hochreiter. Long short-term memory. Neural Computation MIT-Press, 1997.
- [Hu et al., 2021] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035, 2021.
- [Huang et al., 2023] Fan Huang, Haewoon Kwak, and Jisun An. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference* 2023, pages 90–93, 2023.
- [Iwendi et al., 2023] Celestine Iwendi, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*, 29(3):1839–1852, 2023.
- [Kaddoura and Nassar, 2025] Sanaa Kaddoura and Reem Nassar. Language model-based approach for multiclass cyberbullying detection. In *International Conference on Web Information Systems Engineering*, pages 78–89. Springer, 2025.
- [Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
- [Kim et al., 2021] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. A human-centered systematic literature review of cyberbullying detection algorithms. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–34, 2021.
- [Kim et al., 2023] Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. Conprompt: Pretraining a language model with machine-generated data for implicit hate speech detection. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10964–10980, 2023.
- [Kintonova *et al.*, 2021] Aliya Kintonova, Alexander Vasyaev, and Viktor Shestak. Cyberbullying and cyber-mobbing in developing countries. *Information & Computer Security*, 29(3):435–456, 2021.
- [Kumar and Sachdeva, 2022] Akshi Kumar and Nitin Sachdeva. A bi-gru with attention and capsnet hybrid model for cyberbullying detection on social media. *World Wide Web*, 25(4):1537–1550, 2022.
- [Liu *et al.*, 2024] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024.
- [Mahbub *et al.*, 2021] Syed Mahbub, Eric Pardede, and ASM Kayes. Detection of harassment type of cyberbullying: A dictionary of approach words and its impact.

- Security and Communication Networks, 2021(1):5594175, 2021.
- [Mahmud et al., 2023] Tanjim Mahmud, Michal Ptaszynski, Juuso Eronen, and Fumito Masui. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing & Manage*ment, 60(5):103454, 2023.
- [Maity et al., 2022] Krishanu Maity, Tanmay Sen, Sriparna Saha, and Pushpak Bhattacharyya. Mtbullygnn: a graph neural network-based multitask framework for cyberbullying detection. *IEEE Transactions on Computational Social Systems*, 11(1):849–858, 2022.
- [Musleh et al., 2024] Dhiaa Musleh, Atta Rahman, Mohammed Abbas Alkherallah, Menhal Kamel Al-Bohassan, Mustafa Mohammed Alawami, Hayder Ali Alsebaa, Jawad Ali Alnemer, Ghazi Fayez Al-Mutairi, May Issa Aldossary, Dalal A Aldowaihi, et al. A machine learning approach to cyberbullying detection in arabic tweets. Computers, Materials & Continua, 80(1), 2024.
- [Ni and Kao, 2023] Shiwen Ni and Hung-Yu Kao. Kpt++: Refined knowledgeable prompt tuning for few-shot text classification. *Knowledge-Based Systems*, 274:110647, 2023.
- [Qiang et al., 2023] Jipeng Qiang, Shiyu Zhu, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. Natural language watermarking via paraphraser-based lexical substitution. *Artificial Intelligence*, 317:103859, 2023.
- [Raisi and Huang, 2017] Elaheh Raisi and Bert Huang. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 409–416, 2017.
- [Raj et al., 2021] Chahat Raj, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad. Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics*, 10(22):2810, 2021.
- [Reynolds et al., 2011] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine learning and applications and workshops, volume 2, pages 241–244. IEEE, 2011.
- [Rosa et al., 2018] Hugo Rosa, David Matos, Ricardo Ribeiro, Luisa Coheur, and João P Carvalho. A "deeper" look at detecting cyberbullying in social networks. In 2018 international joint conference on neural networks (IJCNN), pages 1–8. IEEE, 2018.
- [Rosa et al., 2019] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93:333–345, 2019.
- [Salawu *et al.*, 2017] Semiu Salawu, Yulan He, and Joanna Lumsden. Approaches to automated detection of cyber-

- bullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24, 2017.
- [Schultze-Krumbholz et al., 2013] Anja Schultze-Krumbholz, Anne Jäkel, Martin Schultze, and Herbert Scheithauer. Emotional and behavioural problems in the context of cyberbullying: A longitudinal study among german adolescents. In Emotional and Behavioural Difficulties Associated with Bullying and Cyberbullying, pages 102–118. Routledge, 2013.
- [Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [Wang et al., 2020a] Jason Wang, Kaiqun Fu, and Chang-Tien Lu. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In 2020 IEEE International Conference on Big Data (Big Data), pages 1699–1708. IEEE, 2020.
- [Wang et al., 2020b] Kun Wang, Yanpeng Cui, Jianwei Hu, Yu Zhang, Wei Zhao, and Luming Feng. Cyberbullying detection, based on the fasttext and word similarity schemes. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1):1–15, 2020.
- [Wu et al., 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in neural information processing systems, 34:22419–22430, 2021.
- [Yadav et al., 2020] Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. Cyberbullying detection using pretrained bert model. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pages 1096–1100. IEEE, 2020.
- [Yang et al., 2025] Qingpo Yang, Yakai Chen, Zihui Xu, Yuming Shang, Sanchuan Guo, and Xi Zhang. Sccd: A session-based dataset for chinese cyberbullying detection. In Proceedings of the 31st International Conference on Computational Linguistics, pages 9533–9545. Association for Computational Linguistics, 2025.
- [Zeng et al., 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.
- [Zhou et al., 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.
- [Zhou et al., 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine* learning, pages 27268–27286. PMLR, 2022.