Detecting Chinese Sexism Text in Social Media Using Hybrid Deep Learning Model with Sarcasm Masking

Lei Wang[1], Nur Atiqah Sia Abdullah[2], Syaripah Ruzaini Syed Aris[3]
College of Information Engineering & Computer Science, Hebei Finance University, Baoding, Hebei, China[1]
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1, 2, 3]
Knowledge and Software Engineering Research Group (KASERG), Research Nexus UiTM (ReNeU), Office of Deputy Vice Chancellor (Research and Innovation), Universiti Teknologi MARA, Shah Alam, Malaysia[2]
Hebei Provincial Key Laboratory of Financial Technology Application, Baoding, Hebei, China[1]

***Abstract*—Sexist content is prevalent in social media, which seriously affects the online environment and occasionally leads to offline disputes. For this reason, many scholars have researched how to automatically detect sexist content in social media. However, the presence of sarcasm complicates this task. Thus, recognizing sarcasm to improve the accuracy of sexism detection has become a crucial research focus. In this study, we adopt a deep learning approach by combining a sexism lexicon and a sarcasm lexicon to work on the detection of Chinese sexist content in social media. We innovatively propose a sarcasm-based masking mechanism, which achieves an accuracy of 82.65% and a macro F1 score of 80.49% on the Sina Weibo Sexism Review (SWSR) dataset, significantly outperforming the baseline model by 2.05% and 2.89%, respectively. This study combines the irony masking mechanism with sexism detection, and the experimental results demonstrate the effectiveness of the deep learning method based on the irony masking mechanism in Chinese sexism detection.**

***Keywords*—Sexism; chinese; deep learning; sarcasm; masking**

I. Introduction

There is a limitation on the current sexism detection. It is difficult to detect sexist text with a sarcastic sense [1]–[3]. Sarcasm uses irony to mock or express contempt, and there is a discrepancy between an utterance's literal meaning and intended meaning. It is challenging to detect automatically [4]. Sarcasm in long and short text remains a challenge in Natural Language Processing (NLP) and Sentiment Analysis (SA) [5]. In the research of [3], irony limits the performance of sexism classification. According to [1], numerous tweets are labeled under sexism categories due to their sarcastic meaning. However, their model is unable to detect the sarcastic intent. The researchers of [2] also argued that sexist posts with humor, irony, and sarcasm are challenging to spot and often contain no overt expressions of hatred. Therefore, it is necessary to identify sarcastic sexism in the detection of sexism.

For example, 如果她自己够优秀就不会在网络上怨天尤人了 (If she is excellent enough, she will not blame others on the internet) is misclassified sexism [2], which contains a sarcastic sense. It is a sarcastic remark that blames unsuccessful women while insulting others who support gender equality. When there is no explicit presence of harsh language, it might be challenging to recognize sexism in an automatic system.

In the research of [3][5], they discovered that sarcasm limits the performance of the sexism classification. In the research of [6], they reported that sarcasm knowledge is helpful for Spanish sexism classification. The researchers of [7] researched Arabic sexism and sarcasm text classification using deep learning methods and noted that AraBERT performs best for both sexism detection and sarcasm detection. It illustrates that AraBERT is useful for both sexism and sarcasm detection. However, the accuracy of sexism detection is 91.0%, higher than sarcasm detection, which is 88%. It can be concluded that the deep learning method is useful for both sexism and sarcasm detection, and the sarcasm knowledge is useful for improving the performance of sexism detection.

Various deep learning methods, such as Convolutional Neural Networks (CNN) [8], Long Short-Term Memory (LSTM) [9], Bidirectional Long Short-Term Memory (BiLSTM) [10], Bidirectional Gated Recurrent Unit (Bi-GRU) [11], Bidirectional Encoder Representations from Transformers (BERT) [12]-[14], Robustly Optimized BERT Pretraining Approach (RoBERTa) [2], Multilingual BERT (mBERT) [15], have been employed in sexism detection. In the research of [1][8][11][16], deep learning methods outperform traditional machine learning methods, such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT). Among the deep learning methods, transformers-based deep learning like BERT and RoBERTa usually perform the best [2][3][12][17]. The hybrid approach in the research [10][18]-[21] obtains the best result. Accordingly, deep learning methods, especially transformer-based methods and hybrid approaches, are considered in our research.

The researchers in [6] demonstrated that sarcasm aids in the identification of sexism. They adopted a multi-task learning strategy that simultaneously addresses numerous problems rather than treating them in isolation to enhance Spanish sexism detection. Nonetheless, the strategy is inappropriate for a singular-task situation. The research in [7] demonstrated that deep learning methods are good at detecting both sexism and sarcasm in Arabic. However, they employed the same deep learning model on both the misogyny dataset and the sarcasm

This work was funded by the Science Research Project of Hebei Education Department (QN2024149).

dataset to demonstrate the efficacy of their technology. Consequently, there is a lack of study on sarcasm-based sexism detection in a singular task context, while it is considered a prospective research avenue by [1]–[3][22].

This research aims to improve Chinese sexism detection using a deep learning approach fused with sarcasm. The dataset is the Sina Weibo Sexism Review (SWSR). To leverage sarcasm features in the model, sarcasm-detecting approaches are invested. There are text-only methods and extra information methods for sarcasm detection [23]. In particular, the text-only method concentrates on the utterance itself to detect the sarcasm text, such as exploiting linguistic features [24] and using inconsistent expressions [26]. This research uses sarcastic linguistic features in a sexist lexicon with a masking mechanism. Consequently, it combines them with a deep learning approach. The contributions of this paper are as follows:

1) Proposed Sarcasm-sexism-Nezha-mCNNs model. The model combines a sarcasm lexicon and a sexism lexicon with a transformer-based Nezha and three CNNs. The model's performance in detecting Chinese sexist text is significantly enhanced, achieving an accuracy improvement of 2.05% and a macro F1 score increase of 2.89% compared to the baseline.

2) Constructed a Chinese sarcasm lexicon. A Chinese sarcasm feature lexicon, consisting of 193 words, has been developed. This lexicon combines terms identified in previous studies with words generated through the chi-square test in this paper. It has potential applications in further sarcasm detection endeavors.

3) A sarcasm mask mechanism is proposed. The output from the sarcasm layer is masked with 1 or -1, and then the masked output is multiplied by the output from the sexism lexicon layer as 20% for the final output. This crucial design feature for the proposed model can be adapted for gender-based categorization of sexism in both Chinese and other languages.

The remainder of the paper is organized as follows: Section II reviews related works. Section III outlines the research methodology, including data processing, text representation, classification, evaluation, and experimental settings. Section IV presents the experimental results, while Section V interprets these outcomes and highlights the key findings. Finally, Section VI summarizes our contributions, acknowledges limitations, and suggests directions for future research.

II. Related Work

First, hybrid deep learning approaches for sexism detection are investigated. Next, the detection of sarcasm using Chinese sarcasm linguistic features is explored. The Chinese sarcasm dataset and feature extraction methods are also presented.

A. Hybrid Deep Learning Approach for Sexism Detection

1) Combining more than one deep learning method: In the research of [18], they employed a CNN-LSTM model to detect sexual harassment in Hindi. They substituted all newline characters with a space, eliminated English letters, numerals, hyperlinks, emojis, and special characters, and employed a tokenizer with the Keras library. An accuracy of 93.53% is achieved with CNN-LSTM, surpassing RNN-LSTM.

The approach in a study by [19] utilized a BiLSTM combined with a Temporal Convolutional Network (TCN-BiLSTM) to analyze sexist speech in Arabic social media, surpassing BiLSTM, TCN, and XGBoost. This method benefits languages that provide morphological challenges, such as Arabic, Turkish, and Lithuanian.

The researchers in [20] combined customBERT with a CNN to detect sexism in social media on the dataset of SemEval Task 10. It combines output from BERT, XLM-RoBERTa, and DistilBERT and then fed into CNN. 'Random deletion' improves the model's ability to understand incomplete information. 'Synonym replacement' improves the model's understanding of context. Then, back translation is used for data augmentation. Shapley additive explanation values are used to increase model explainability.

In the research of [22], they employed a semi-supervised model to enhance the EXIST dataset through data augmentation. The pre-trained XLM-R and sentence BERT combined with BiLSTM are utilized for sexism detection, achieving an accuracy of 81.2% and an F1 of 81.1%. They indicated a prospective research direction involving the integration of sarcasm and irony detection into their system.

It illustrates that CNN, LSTM, BiLSTM, and transformer-based models such as BERT, mBERT, and XLM-R are commonly used in combinations of two or more deep learning methods.

2) Combining deep learning with linguistic features: The researchers in [10] combined lexicon and sentiment features with LSTM, obtaining the best accuracy of 87.2% and an F1 of 82.4% on the dataset of AMI EN-EVALITA. They discovered that the optimal feature type identified is Term Frequency-Inverse Document Frequency (TF-IDF) Ngrams. However, lexical features and word2vec embeddings can enhance the outcomes when integrated with TF-IDF. Their model outperforms BERT, XLNET, RoBERTa, and DistilBERT.

In the research of [2], they combined BERT, RoBERTa, CNN, SVM, and LR with the sexism lexicon, respectively. The sexism lexicon is expressed using TF-IDF. The researchers find that RoBERTa, with a sexism lexicon, obtains the best F1 score of 78% on the SWSR dataset. However, it has a lower accuracy value than BERT without a lexicon. It indicates that the sexism lexicon should be combined with the deep learning method with a proper method.

The researchers in [21] integrated LSTM with sentiment analysis, obtaining the best performance detecting sexism with an F1 of 83.01%, which outperforms LSTM-sexism and LSTM-RoBERTa. Thus, pre-trained models are used for word representation.

It can be noted that linguistic features such as sexism and sentiment have been combined into deep learning methods

such as LSTM, BERT, and RoBERTa. However, a proper combining method is required.

3) Sarcasm detection with Chinese sarcasm features. Sarcasm uses irony to mock or express contempt, and there is a discrepancy between the literal meaning and intended meaning of an utterance. It is challenging to detect automatically [4][25]. According to [26], there are text-only methods and extra information methods. The text-only method concentrates on the utterance itself to detect the sarcasm text, such as exploiting linguistic features [27]-[29] and using inconsistent expressions [23][30][31]. In contrast, the extra information method focuses on external knowledge, such as user features and common sense knowledge. In this research, we focus on sarcasm linguistic features.

In the studies in [24][27]-[29][32]-[34], linguistic features of Chinese sarcasm are employed as significant features for the detection of Chinese sarcasm. The chi-square test is employed to pick feature words more pertinent to the target category and obtain Chinese sarcastic feature words. The traits are categorized into more specific classifications, including interjection words, adverbs of degree, internet vocabulary, homophonic words, and punctuation. The Chinese sarcasm feature words presented in the studies referenced as [24][27][28][32][34] are integrated. Table I illustrates some examples of Chinese sarcastic linguistic feature words.

TABLE I. CHINESE SARCASM LINGUISTIC FEATURE WORDS

Word Type	Words
Interjection words	呵呵(hehe), 啊(ah), 哇(wow), 哟(yo), 哈哈(haha), 嘿嘿(heihei), 唉(sigh), 哼(hmm）
Adverbs of degree	真是(really), 非常(very), 极其(extremely), 牛(awesome), 牛逼(awesome), 不愧是(worthy of being）
Internet vocabulary	醉了(drunk), 凡尔赛(Versailles), 躺平(lie flat), 涨姿势(gaining knowledge), 囧(confused), 逗比(funny person),
Homophonic words	河蟹-和谐(harmony), 杯具-悲剧(tragedy), 灰常-非常(very), 餐具-惨剧(tragedy), 表-婊(bitch), 虾米-什么(what),
Punctuation	？，！，“”, ……，。。。

1) Interjection words: The inclusion of interjection words in a sentence imparts a sense of teasing and irony [28]. In sentence 哇？？？？？？？吐槽鬼居然是女性我真的觉得她可爱了哈哈哈哈哈 (Wow??????? The complaining ghost is actually a woman. I really think she is cute, hahahahahaha), both 哇 (Wow) and 哈哈哈哈哈 (hahahahahaha) are interjection words.

2) Adverbs of degree: Adverbs of degree will enhance the semantic intensity of the words in the text, illustrating an exaggerated effect and presenting irony [27]. In sentence 不过我经常碰到异常自信，自以为是的男性，也让人吃惊。(But it's also surprising how often I run into unusually confident, self-righteous males.) 异常(exceptional) expresses exaggeration, and it is ironic to somebody.

3) Internet vocabulary: Internet vocabularies are widely used on the internet, which may be helpful for sarcasm detection, such as 醉了(drunk), which expresses an attitude of finding someone's actions or something unexplainable

4) Homophonic words: Most homophonic words are deformed words with similar pronunciation to the original words or new words created on the internet to incorporate fun and humor or to express a specific mood [28]. Somebody may humorously write 杯具 (cupcake) instead of 悲剧 (tragedy) since they share the same pronunciation.

5) Punctuation: According to [35], some special punctuation can emphasize the effect of irony in the proper context and help readers understand the author's intent. Punctuation marks such as question marks, exclamation points, quotation marks, and apostrophes could emphasize irony [27]. If these symbols are used more than three times consecutively, they convey further emphasis and enhance irony [36].

B. Chinese Sarcasm Datasets

To extract more sarcasm linguistic feature words, a collection of Chinese sarcasm datasets is conducted. Table II illustrates the Chinese sarcasm datasets. All datasets are about the classification of sarcasm except for the first dataset, SMP ECISA, and the last dataset, Weibo Sentiment. The first dataset, SMP ECISA, focuses on implicit sentiment analysis. It contained implicit positive sentiment, implicit negative sentiment and no sentiment. The rationale for including the final dataset is that, although it pertains to sentiment analysis, the author regards sarcastic semantic recognition as a significant aspect of the research.

TABLE II. CHINESE SARCASM DATASET

Name	Size	Used by	Model	Performance
SMP ECISA	Neutral: 10012 Positive: 5925 Negative: 5973	[37]	KG-MPOA	Macro F1 0.777
		[38]	BERT-BiLSTM	Macro F1 0.775
Ciron	Not ironic: 4343 Unlikely ironic: 3391 Insufficient evidence: 64 Weakly ironic: 838 Strongly ironic: 130	[39]	BERT	F1: 0.572 A: 0.603
Ciron + Chinese Sarcasm	Not ironic: 5343 Ironic: 5425	[27]	IDIHR	F1:0.8124 A: 0.8149
ToSarcasm	Not ironic: 2435 Ironic: 2436	[40]	TOSPrompt	F1:0.7320 A: 0.7176
Weibo Sarcasm	Not ironic: 2000 Ironic: 2000	[24]	ISR	F1: 0.7793
		[33]	NB	A: 0.6945
Multimodal Sarcasm	Not ironic: 1179 Ironic: 1009	[41]	FCAM	F1: 0.8778 A: 0.8813

C. Chi-Square Test for Sarcasm Linguistic Feature Extracting

The chi-square test can be used to extract sarcastic linguistic features that are closely related to sarcasm. The chi-square test posited that the characteristics and categories operate independently, subsequently assessing the correlation through deviation analysis. A low chi-square test value indicated that the correlation between the two variables might be coincidental and lacked significance. Conversely, a high chi-square test value suggested a stronger correlation that could be utilized as a categorical feature. Consequently, the chi-

square test stood out as the efficient and suitable approach for feature selection [24].

III. Research Methodology

This section will introduce data description and preprocessing, text representation, text classification, evaluation of sexism detection, and experiment settings.

A. Data Description and Preprocessing

The dataset utilized in this study is a Chinese sexism dataset named Sina Weibo Sexism Review (SWSR). It was gathered from the Sina Weibo platform between June 2015 and June 2020 by [2]. The comments are labeled as sexism or not sexism. Among the 8,969 initial comment texts, there are six instances of duplication in contents but with different user information. In this research, user information is not used. Hence, the duplicated comments are removed, resulting in 8,963 comments.

To eliminate the noise in the text, data preprocessing is conducted. The original dataset was converted from traditional Chinese to simple Chinese by [2]. We conducted the following processes: converting full-corner text content to half-corner, converting punctuation symbols from English to Chinese, and converting uppercase English to lowercase English.

B. Text Representation

Three text representations are conducted in the final model.

1) Pre-trained Nezha tokenizer: Nezha is a transformer-based pre-trained model, like BERT and RoBERTa. It has been trained on large-scale Chinese corpus and has obtained outstanding performance in many Chinese NLP tasks [42]. Then, we expressed the comments with the pre-trained Nezha tokenizer. In this study, the sentence length is set to 150. Sentence lengths exceeding 150 are truncated, and sentences less than 150 are filled with [PAD]. After the Nezha tokenizer, the sentence is expressed with input ids, token type ids, and attention masks.

2) Text representation with sexist lexicon: A sexism lexicon with 3,016 words created by [2] is utilized for comments representation using Bag of Word Vector (BoWV) on the sexism lexicon. The frequency of each word in the sexism lexicon for each sentence phrase is calculated to a vector as a linguistic attribute.

3) Text representation with sarcasm lexicon: Firstly, the sarcasm lexicon is constructed. To construct the sarcasm lexicon, the sarcasm-related feature words are collected. Irony-related feature words are obtained by the chi-square test. The chi-square test is conducted on the SMP ECISA dataset which is for evaluating Chinese implicit sentiment analysis. Non-implicit content and negative implicit content are used, while positive implicit content is excluded. The sarcastic feature words identified using the chi-square test are subsequently picked manually, obtaining 85 words. The same operation is conducted on the Ciron dataset. The content is relabelled before the chi-square test. After relabeling the content, items labeled 1 and 2 are classified as non-sarcasm, while those labeled 4 and 5 are categorized as sarcasm. Meanwhile, entries labelled with 3 are disregarded because of insufficient evidence of sarcasm. Subsequent to the chi-square test and manual selection, 28 words remained as feature words. Accordingly, no prominent sarcasm features are discovered in the ToSarcasm and BSC datasets using the chi-square test. The words selected from SMP ECISA 2021 and Ciron are combined with the words selected from other related research, and 193 sarcastic linguistic feature words are finally obtained.

Then, the sarcastic lexicon with 193 words is utilized for comment representation using BoWV. Following text representation approach with sexism lexicon, the frequency of each word in the sarcasm lexicon for each comment is calculated as a vector to represent sarcasm linguistic features.

C. Text Classification

To enhance the classification accuracy of sexism detection, sarcastic features and sexism features are considered to be combined with the deep learning method. Fig. 1 illustrates the architecture of the Sarcasm-sexism-CNNs-Nezha model. This model has four layers: mCNNs-Nezha layer, sexism lexicon layer, sarcasm lexicon layer, and output layer. The output of the sexism layer is multiplied by the output of the sarcasm lexicon layer, resulting in a combined feature output that contributes 20% to the final output. In contrast, the output from the mCNNs-Nezha layer accounts for 80% of the final output. Upon applying SoftMax to the result, the text is categorized as either sexist or non-sexist.

1) mCNNs-Nezha layer: The mCNN-Nezha layer accounts for 80% of the final result. It is a deep learning-based model consisting of the Nezha layer, transposition layer, mCNNs layer, and linear layer.

a) Nezha layer: Through the Nezha tokenizer, it can obtain input ids, token type ids and masks. As a pre-trained transformer-based model, Nezha contains the Nezha embeddings layer, Nezha encoder layer, and Nezha pooler layer. There are 12 hidden layers in the Nezha encoder layer. The last hidden layer is used as input for the transposition layer. The whole information, both the [CLS] information, which contained the whole information of the sentence and word embedding for characters in the sentence, is used. The data size is [b, l, h]. b is the batch size, l is the length of sentences, and h is the hidden size in the hidden layer. In our research, it is [32, 150, 768].

b) Transposition layer: A transposition is conducted on the dimensions of 2 and 3 in (1) and obtains data with the size of [b, h, l], which is [32, 768, 150] in this research.

$$D' = D^{T_{2,3}} \quad (1)$$

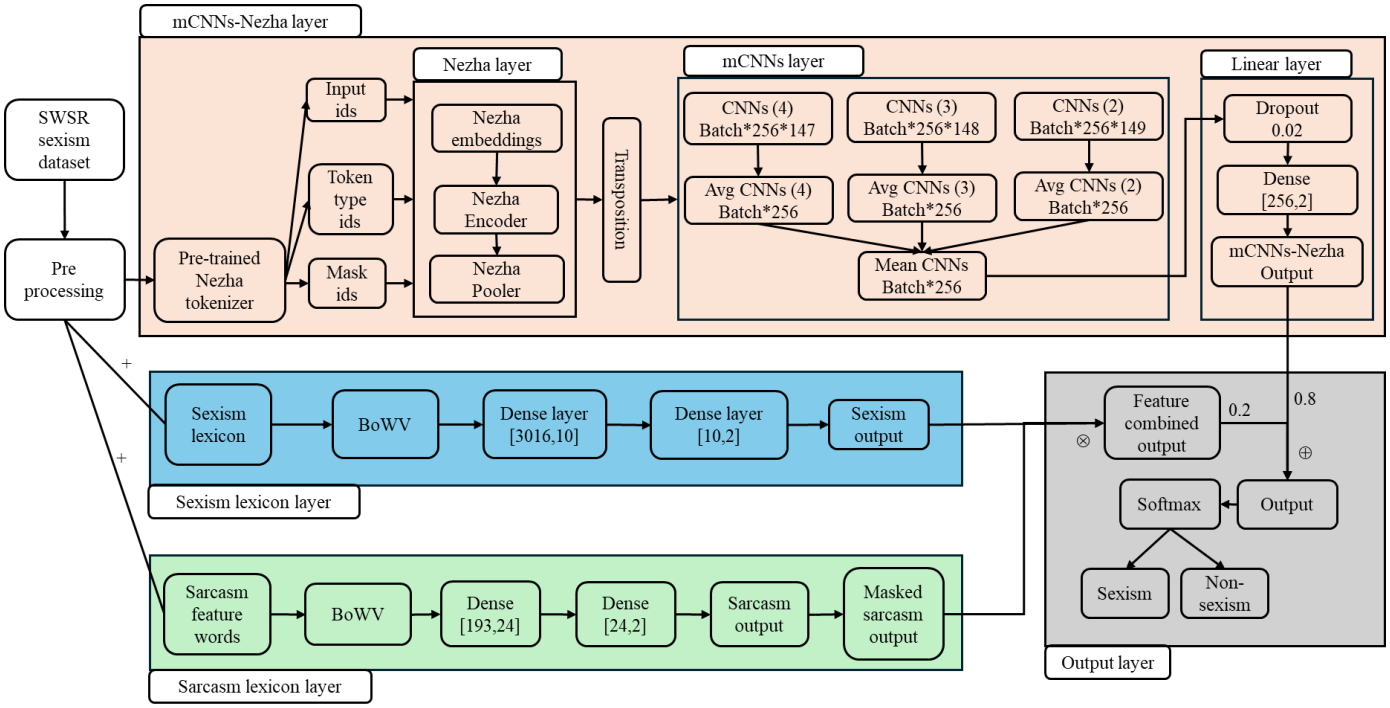

Fig. 1. The architecture of the sarcasm-sexism-mCNNs-Nezha model

c) mCNNs layer: The mCNNs layer contains three CNNs with kernels of 2, 3, and 4. After the CNN in (2), the data size is [b, o, l-k+1]. k is kernel size. o is the output channel with a value of 256. Hence, the feature is reduced.

$$CNN_k = CNN(D', k), \quad k=2,3,4 \tag{2}$$

To align features, an average operation is conducted on the third dimension in (3), and each CNN obtains data with the size of [b,o].

$$mCNN_k = Average(CNN_k^3), \quad k=2,3,4 \tag{3}$$

Then, an average operation is conducted on the three $mCNN_k$ on dimension 2 in (4). The final output of this layer is [b,o]. It can better capture words with two, three, or four characters, which is consistent with the Chinese word expressing a meaningful thing with two, three, or four characters.

$$mCNN = \frac{\sum_2^4 mCNN_k}{3} \tag{4}$$

d) Linear layer: The output from the mCNNs layer is fed into a linear layer. After a dropout of 0.02, a fully connected neural network is conducted and obtains an output with two values. The output of this layer is [b,2].

2) Sexism lexicon layer: The comments have been expressed as vectors using BoWV with the sexism lexicon. Then, the vector is fed into two densely connected neural networks in order in (5) (6) and results in an output with a two-dimensional vector. The first dense layer integrates the characteristics of sexism and executes feature mapping, encapsulating the intricate relationships among the features. The subsequent layer additionally conducts feature mapping and utilizes the extracted features to correlate with the probabilities of the two predictive classifications.

$$z^{(1)} = W^{(1)T} X_{sex} + b^{(1)} \tag{5}$$

$$O_{sex} = z^{(2)} = W^{(2)T} z^{(1)} + b^{(2)} \tag{6}$$

3) Sarcasm lexicon layer: The comments have also been expressed as vectors using BoWV with a sarcastic lexicon. Then, it is fed into two densely connected neural networks in order in (7) (8) and obtains a two-dimensional output. A masked operation is adopted on the output, where the first value is bigger than the second value and masked with -1, otherwise with 1, as indicated in (9). The masked output would multiply with the output of the sexism lexicon layer. This sarcasm masking mechanism has an important impact on the final output.

$$z^{(1)} = W^{(1)T} X_{sar} + b^{(1)} \tag{7}$$

$$O(l) = z^{(2)} = W^{(2)T} z^{(1)} + b^{(2)} \tag{8}$$

$$O_{sar} = \begin{cases} -1 & O(l)_0 > O(l)_1 \\ 1 & O(l)_0 \leq O(l)_1 \end{cases} \tag{9}$$

4) Output layer: The influence of the sexism lexicon layer's output O_{sex} on the final result, whether positive or negative, is dependent upon the sarcasm lexicon layer's output O_{sar}. If the value of O_{sar} is -1, it indicates a negative effect. Conversely, a value of O_{sar} equal to 1 signifies a positive effect, as presented in (10). The combined output, O_{lex}, constitutes 20%, while the output from O_{mn} mCNNs-Nezha layer comprises 80% of the output O_v in (11). A SoftMax is performed on O_v in (12).

$$O_{lex} = O_{sex} \otimes O_{sar} \tag{10}$$

$$O_v=0.8*O_m+0.2*O_{lex} \quad (11)$$

$$Let\ O(v)=(v_0,v_1), P(y=i)=SoftMax(v_i)=\frac{e^{v_i}}{\sum_{j=0}^{1} e^{v_j}},\ i=0,1 \quad (12)$$

The sarcasm masking mechanism algorithm is described as follows:

Algorithm 1: Sarcasm masking mechanism

```
Input: Output from mCNNs-Nezha layer (O1)
       Output from sexism lexicon layer (O2)
Output from sarcasm lexicon layer without masking (O3)
While (data in batch) do
    For (each comment) do
        If (O3[0] > O3[1]) then
            O2[0] = O2[0] * (-1)
            O2[1] = O2[1] * (-1)
        End
        O[0] = O1[0] * 0.8+O2[0] * 0.2
        O[1] = O1[1] * 0.8+O2[1] * 0.2
    End
```

D. Evaluation

As a classification task, confusion metric, accuracy, and F1 score are used for evaluation.

Confusion matrix is an essential instrument for evaluating the performance of classification models in classification tasks. As indicated in Table III, True Positive (TP) refers to the count of TP instances; True Negative (TN) denotes the count of TN instances; False Positive (FP) indicates the count of FP instances; False Negative (FN) represents the count of FN instances.

TABLE III. CONFUSION MATRIX FOR CLASSIFICATION EVALUATION

	Predict Positive	**Predict Negative**
Actual Positive	True Positive (TP)	False Positive (FP)
Actual Negative	True Negative (FN)	True Negative (TN)

Accuracy denotes the proportion of samples accurately identified by the model relative to the total number of samples (13).

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

F1 score is the harmonic average of precision and recall, designed to address the trade-off between these two metrics in classification problems, as indicated in (14). Meanwhile, precision denotes the ratio of correctly predicted positive instances to the total number of instances predicted as positive by the model, as presented in (15). At the same time, recall denotes the ratio of accurately predicted positive instances by the model to the total number of actual positive instances, as given in (16).

$$F1\ score=\frac{2*Precision*Recall}{Precision+Recall} \quad (14)$$

$$Precision=\frac{TP}{TP+FP} \quad (15)$$

$$Recall=\frac{TP}{TP+FN} \quad (16)$$

The 5-fold cross-validation method is employed to assess the model's performance and stability by partitioning the dataset into five subsets. By systematically employing several subsets as training and validation sets, the bias in model evaluation can be significantly mitigated.

E. Experimental Settings

Table IV illustrates the experimental settings.

TABLE IV. EXPERIMENTAL SETTINGS

Items	**Value**
GPU	NVIDIA RTX 4090 24G
Memory	64G
Cuda	v12.1.105
PyTorch	v2.2.1
Max Length of Sentence	150
Batch size	32
Epochs	4
Loss	Cross Entropy Loss
Optimizer	AdamW

IV. RESULT

A range of models are conducted to improve the model performance, including BERT-Fc, RoBERTa-Fc, Nezha-Fc, mCNNs-BERT, mCNNs-RoBERTa, mCNNs-Nezha, BERT-BiLSTM, RoBERTa-BiLSTM, and Nezha-BiLSTM, excluding the baseline and the proposed Sarcasm-sexism-mCNNs-Nezha. In the model of BERT/RoBERTa/Nezha-Fc, a fully connected neural network is concatenated after BERT/RoBERTa/Nezha. In the model of mCNNs-BERT/RoBERTa/Nezha, the mCNNs layer is concatenated after BERT/RoBERTa/Nezha. In the model of BERT/RoBERTa/Nezha-BiLSTM, the BiLSTM is concatenated after BERT/RoBERTa/Nezha.

The experiment result is illustrated in Table V. The proposed Sarcasm-sexism-mCNNs-Nezha model achieves the following performance metrics: 82.65% accuracy, 80.49% macro F1 score, 82.49% weighted F1 score, 86.94% non-sexism F1 score, and 74.05% sexism F1 score. It can be reported that the proposed model obtains the best performance on all five metrics except for sexism F1. Although the sexism F1 score in the proposed model is 0.39% lower than that of the RoBERTa-BiLSTM model, its accuracy is 0.55% higher. The mCNNs-Nezha model demonstrates the second-best performance.

As the best-performing model, Sarcasm-sexism-mCNNs-Nezha integrates sexism and sarcasm features, leveraging a deep learning framework with a masking mechanism. It outperforms all other deep learning approaches in Table V that do not incorporate sarcasm detection. It demonstrates significant improvements over the baseline model (BERT [2]) with an increase of 2.05% in accuracy and a 2.89% rise in macro F1. Specifically, the model achieves a 3.65% improvement in sexism F1, outperforming the 1.14% improvement in non-sexism F1. This highlights its superior capability in addressing sexism compared to non-sexism.

TABLE V. EXPERIMENTAL RESULT (%)

Model	Accuracy	Macro F1	Weight F1	Non-sexism F1	Sexism F1
BERT [2]	80.6	77.6	-	85.8	69.4
RoBERTa-lexicon [2]	80.4	78.0	-	85.3	70.7
BERT-Fc	81.86	79.77	81.78	86.25	73.28
RoBERTa-Fc	82.07	80.10	82.03	86.32	73.87
Nezha-Fc	81.84	79.61	81.69	86.31	72.91
mCNNs-BERT	81.55	79.45	81.48	85.99	72.90
mCNNs-RoBERTa	82.45	80.32	82.32	86.78	73.86
mCNNs-Nezha	82.53	80.24	82.32	**86.94**	73.55
BERT-BiLSTM	81.69	79.52	81.58	86.15	72.90
RoBERTa-BiLSTM	82.10	80.33	82.16	86.22	**74.44**
Nezha-BiLSTM	82.06	79.88	81.93	86.48	73.29
Sarcasm-sexism-mCNNs-Nezha	**82.65**	**80.49**	**82.49**	**86.94**	74.05

V. DISCUSSION

A. Confusion Matrix of Proposed Model

The confusion matrix of the proposed model is displayed in Fig. 2. There are 2,229 true predicted as sexisms, 863 false anticipated as sexisms, 5,179 accurately predicted as non-sexisms, and 692 false predicted as non-sexisms.

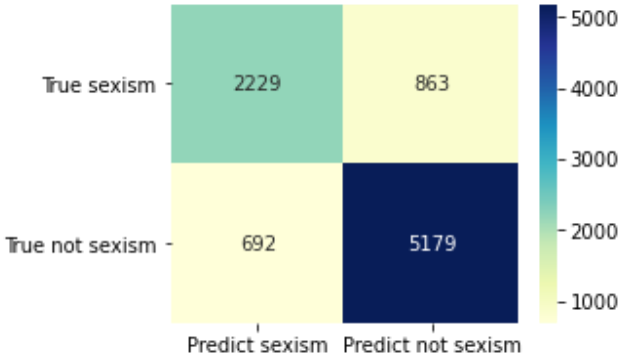

Fig. 2. Confusion matrix of proposed sarcasm-sexism-mCNNs-Nezha model

B. The Stability of Proposed Sarcasm-sexism-mCNNs-Nezha Model

The stability of the proposed model is demonstrated in Fig. 3. Fold 2 to fold 5 exhibits a higher concentration across all metrics. Fold 1 demonstrates a focus on non-sexism F1, suggesting that the capacity to identify non-sexism has not significantly altered. Nonetheless, on the other four metrics, fold 1 exhibits inferior performance compared to other folds, particularly for the sexism F1 score.

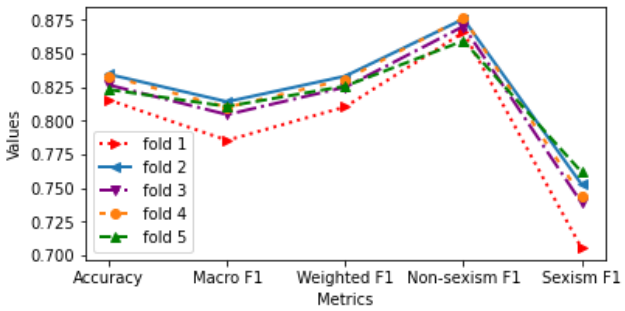

Fig. 3. The stability of the proposed Sarcasm-sexism-mCNNs-Nezha model

To examine the performance of this model on fold 1, a comparison between the mCNNs-Nezha model with the second-highest accuracy and the proposed model is conducted. The result is depicted in Fig. 4. It illustrates that the non-sexism F1 score is improved in this model, while the sexism F1 score is decreased. However, the accuracy is increased. This indicates that even though the performance of this model is worse on fold 1, it also improved the accuracy compared with the mCNNs-Nezha model. It illustrates the effectiveness of the proposed model.

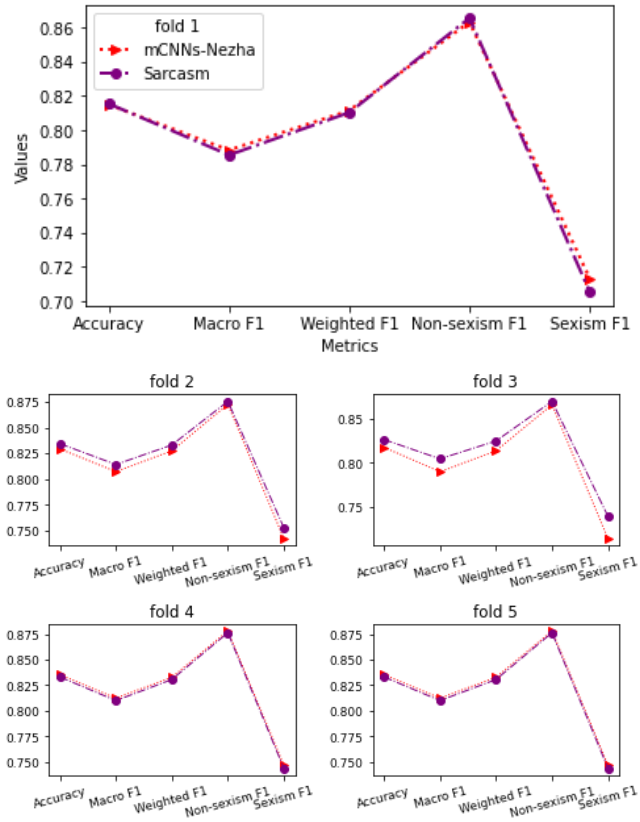

Fig. 4. Compared with the mCNNs-Nezha model on five folds

Nevertheless, the comparison of fold 2 to fold 5 demonstrates similar results in fold 4 and fold 5, while exhibiting a relatively greater enhancement in fold 2 and fold

3, as displayed in Fig. 4. It indicated that the proposed model is beneficial in enhancing performance in most cases.

C. Manually Analysis

A manual evaluation of sexist texts incorrectly identified by the mCNNs-Nezha model yet accurately by the proposed model is conducted. Table VI presents identical examples of sexist remarks that are misclassified by the mCNNs-Nezha model but accurately classified by the proposed approach. Sentences 2 and 3 contain metaphors, such as 田园狗 (a rural dog), which means pastoral feminist, and 鸡 (chicken), which means prostitute. Sentence 3 contains irony, which means, 'I do not admire you at all.'

TABLE VI. EXAMPLES OF SEXISM TEXT WRONGLY CLASSIFIED BY BASIC MODEL BUT RIGHTLY BY SARCASM-SEXISM-MCNNS-NEZHA MODEL

Id	Comments
1	明显田园狗做法。哈哈哈。这婚结的都不如去找鸡 This is obviously a rural dog behavior. Hahaha. This kind of marriage is worse than finding a prostitute.
2	说人家坦克真的好笑吗？那你们金针菇我们也可以笑吧？不会这么开不起玩笑吧，不会吧不会吧？ Is it really funny to say that others are tanks? Then can we, the Enoki mushrooms, also laugh? You can't be so incapable of taking a joke, right? No way, no way?
3	你真贴心，身为女人这么照顾男人的各种情绪，佩服你呢 You are so considerate. As a woman, you take care of a man's emotions. I admire you.

D. Ablation Experiment

To evaluate the significance of each component, an ablation experiment is conducted. There are three components except for Nezha. Table VII illustrates the result of the ablation experiment.

TABLE VII. ABLATION EXPERIMENT RESULT (%)

Id	Sarcasm	Sexism	mCNNs	accuracy	Macro F1	Weight F1	Non-sexism F1	Sexism F1
1	√	√	√	**82.65**	**80.49**	**82.49**	86.94	**74.05**
2		√	√	82.56	80.32	82.37	86.93	73.71
3	√		√	82.55	80.23	82.32	**86.98**	73.48
4	√	√		82.26	80.05	82.10	86.67	73.43
5	√			82.07	79.65	81.81	86.63	72.67
6		√		82.14	79.81	81.93	86.63	73.00
7			√	82.53	80.24	82.32	86.94	73.55
8				81.84	79.61	81.69	86.31	72.91

1) All three components have a positive impact: As summarized in Table VII, no matter eliminating one, two, or three components like sarcasm layer, sexism layer, and mCNNs layer from this model, the performance of the model on each metric such as accuracy, macro F1 score, weighted F1 score, sexism F1, non-sexism F1, is decreased except that the non-sexism F1 is improved 0.04% while eliminating sexism component. This indicates that every component has a positive effect on the proposed model.

2) mCNNs contributed more than sarcasm or sexism: Fig. 5 illustrates the model performance degradation while eliminating one component from the proposed model. It is evident that the model's performance declines when one component is removed and that mCNNs are the most crucial component of the proposed model, followed by sarcasm and sexism. It is noteworthy that the non-sexism F1 improves when the sexism component is eliminated, even though overall performance declines. The reason may attributed to the fact that the sexism lexicon layer tends to classify text containing sexist words as sexism. When the sexism component is deleted, some text containing sexist words may be classified as non-sexism, leading to improvement in non-sexism F1 score.

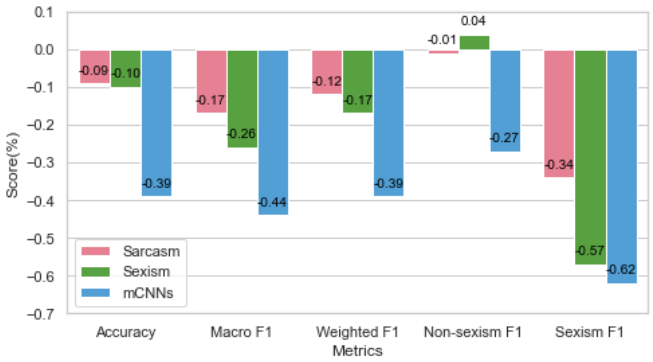

Fig. 5. Eliminating one component from Sarcasm-sexism-mCNNsNezha

3) Impact of sarcasm component: Fig. 6 illustrates the degradation of eliminating the sarcasm component and eliminating both the sarcasm component and the sexism component. If the sarcasm component is removed from the model, there will be a decrease of 0.09%, 0.17%, 0.12%, 0.01%, and 0.34% on the accuracy, macro F1, weighted F1, non-sexism F1, and sexism F1, respectively. It indicates that the sarcasm component has a positive effect on sexism detection, and it has affected more the sexism content than the non-sexism content.

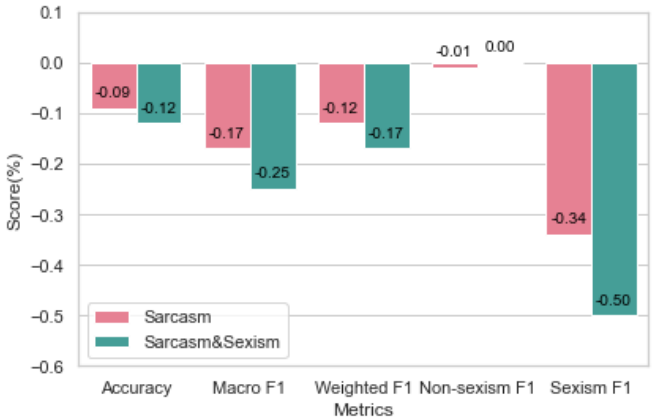

Fig. 6. Impact of sarcasm

It also illustrates that it decreases more by deleting both sarcasm and sexism from this model. It indicates that the combination of sarcasm and sexism has more influence than sarcasm. The effectiveness of sarcasm masking mechanisms is proven.

Both strategies demonstrate more impact on sexist content than non-sexism content. This indicates that they can improve the detection of sexism more than non-sexism.

4) Impact of sexism component: Fig. 7 illustrates the degradation of eliminating the sexism component and eliminating both sarcasm and sexism. It suggests that no matter whether the sexism *component* or a combination of sarcasm and sexism has a positive impact on the model. From the perspective of accuracy, the combination of sarcasm and sexism has a higher impact. Moreover, they have the same weighted F1 score.

Nevertheless, from the perspective of the macro F1 score, the combination of sarcasm and sexism has more influence than sarcasm (Fig. 6) but less influence than sexism. In addition, it is evident that the sexism *component* has a positive impact on detecting sexism while a negative impact on detecting non-sexism. Meanwhile, the combination of sexism and sarcasm has a positive impact on sexism but has no impact on non-sexism. It indicates a complex interaction between sexism and sarcasm.

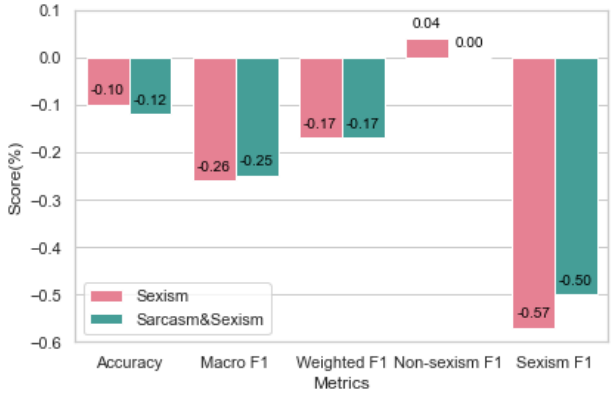

Fig. 7. Impact of sexism

5) Decision for coefficients: The output of the combination of sarcasm and sexism accounts for 20% of the final output, and the output of the mCNNs accounts for 80% of the final output. Different coefficients are adopted.

TABLE VIII. MODEL WITH DIFFERENT COEFFICIENTS (%)

Coefficient For mCNNs-Nezha	Coefficient for Combination of Sexism and Sarcasm	Accuracy	Macro F1
0.75	0.25	82.56	80.40
0.8	0.2	82.65	80.49
0.85	0.15	82.60	80.44
0.9	0.1	82.60	80.55

Table VIII illustrates the performance of the model with different coefficients. It can be discovered that the coefficient of (0.8, 0.2) contributes the best performance on accuracy. However, the coefficient of (0.9, 0.1) contributes the best performance on macro F1. Since (0.8, 0.2) has the highest accuracy, it is selected for the final model.

VI. CONCLUSION

An effective way to detect Chinese sexism text is proposed. The model combines a sarcasm lexicon, sexism lexicon, transformer-based Nezha, and three CNNs. The accuracy and macro F1 score reached 82.65% and 80.49%, outperforming the baseline with 2.05% and 2.89%, respectively. A sarcasm lexicon is constructed, and a sarcasm masking mechanic is proposed. In the ablation experiment, all evaluation metrics are decreased when sarcasm is removed, proving the effectiveness of the sarcasm masking mechanic. The sarcasm masking mechanic has the potential to generate the detection of sexism in other languages. However, this research only considers sarcasm feature words and does not account for inconsistent expressions in text or other sarcasm detection techniques that may be useful for detecting sexism, which contains sarcasm. Additionally, data augmentation methods, particularly back translation, can be considered, as some implicit sexist comments become explicit sexist content after being translated into English.

REFERENCES

[1] H. Mulki and B. Ghanem, “Let-Mi: an Arabic levantine twitter dataset for misogynistic language,” In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 154–163, Apr 2021.

[2] A. Jiang, X. Yang, L. Yang, and A. Zubiaga, “SWSR: a Chinese dataset and lexicon for online sexism detection,” Online Social Networks and Media, vol. 27, pp. 100182, Jan 2022.

[3] F. Rodriguez-Sanchez, J. Carrillo-de-Albornoz, and L. Plaza, “Automatic classification of sexism in social networks: an empirical study on twitter data,” IEEE Access, vol. 8, pp. 219563–219576, Dec 2020.

[4] S. Khotijah, J. Tirtawangsa, and A. A. Suryani, “Using LSTM for context based approach of sarcasm detection in twitter,” ACM Int. Conf. Proceeding Ser., no. 19, pp. 1–7, Jul 2020.

[5] T. Abdullah, A. Ahmet, and U. Kingdom, “Deep learning in sentiment analysis : a survey of recent architectures,” ACM Computing Surveys, vol. 55, no. 159, pp. 1–37, Dec 2022.

[6] F. M. Plaza-Del-Arco, M.-D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia, “Exploring the use of different linguistic phenomena for sexism identification in social networks,” in CEUR Workshop Proceedings, vol. 3202, Sept 2022.

[7] A. Y. Muaad et al., “Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts,” Comput. Intell. Neurosci., vol. 2022, pp. 7937667, March 2022.

[8] S. Sharifirad and A. Jacovi, “Learning and understanding different categories of sexism using convolutional neural network’s filters,” in Proceedings of the 2019 Workshop on Widening NLP, pp. 21–23, Aug 2019.

[9] M. A. Bashar, R. Nayak, and N. Suzor, “Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set,” Knowl. Inf. Syst., vol. 62, no. 10, pp. 4029–4054, Jun 2020.

[10] A. Rahali, M. A. Akhloufi, A.-M. Therien-Daniel, and E. Brassard-Gourdeau, “Automatic misogyny detection in social media platforms using attention-based Bidirectional-LSTM,” in 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2706–2711, 2021.

[11] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, and D. Ging, "A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary," 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), pp. 1-8, Jun 2019.

[12] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An annotated corpus for sexism detection in French tweets," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., pp. 1397–1403, May 2020.

[13] N. Safi Samghabadi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio, "Aggression and misogyny detection using BERT: a multi-task approach," Proc. Second Work. Trolling, Aggress. Cyberbullying, pp. 126–131, May 2020.

[14] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, and M. A. Márquez-Vera, "Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases," IEEE Access, vol. 11, pp. 13179–13190, Feb 2023.

[15] A. Singh, D. Sharma, and V. K. Singh, "Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models," Comput. Speech Lang., vol. 89, pp. 101682, Jan 2025.

[16] A. Y. Muaad, H. J. Davanagere, M. A. Al-antari, J. V. B. Benifa, and C. Chola, "AI-based misogyny detection from Arabic levantine twitter tweets," vol. 2. no. 1, pp. 15, Sept 2021.

[17] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in twitter: a multilingual and cross-domain study," Inf. Process. Manag., vol. 57, no. 6, pp. 102360, Nov 2020.

[18] T. Jain et al., "Detection of sexually harassing tweets in Hindi using deep learning methods," Int. J. Softw. Innov., vol. 10, no. 1, pp. 1-15, 2022.

[19] N. A. Hamzah and B. N. Dhannoon, "Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network," Egypt. Informatics J., vol. 24, no. 2, pp. 365–373, Jul 2023.

[20] H. Mohammadi, A. Giachanou, and A. Bagheri, "A transparent pipeline for identifying sexism in social media: combining explainability with model prediction," Appl. Sci., vol. 14, no. 19, 2024, doi: 10.3390/app14198620.

[21] F. Belbachir, T. Roustan, and A. Soukane, "Detecting online sexism: integrating sentiment analysis with contextual language models," AI, vol. 5, no. 4, pp. 2852 – 2863, Dec 2024.

[22] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, and L. Plaza, "Leveraging unsupervised task adaptation and semi-supervised learning with semantic-enriched representations for online sexism detection," Expert Syst., vol. 42:e13763, no. 2, 2025.

[23] R. Wang, Q. Wang, B. Liang, Y. Chen, and B. Qin, "Masking and generation : an unsupervised method for sarcasm detection," In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, vol. 1, no. 1, pp. 2172-2177, Jul 2022.

[24] S. Wei, G. Zhu, G. Tan, and S. Zhang, "Ironic sentence recognition model integrating ironic language features," CAAI Trans. Intell. Syst., vol. 19, pp. 689–696, May 2024. http://tis.hrbeu.edu.cn/Upload/PaperUpLoad/f612a74d-dd2c-49b3-81b1-28422add80a4.pdf.

[25] E. Lunando and A. Purwarianti, "Indonesian social media sentiment analysis with sarcasm detection," 2013 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2013, pp. 195–198, 2013.

[26] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: a survey," ACM Comput. Surv., vol. 50, no. 5, pp. 1-22, Sept 2017.

[27] S. Li, G. Zhu, and J. Li, "A method of Chinese ironic detection integrated with hyperbolic representation," Data Anal. Knowl. Discov., no. 3, pp. 1–10, 2024. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CAPJ&filename=ZZDZ20241030001.

[28] W. Hu, L. Chen, T. Han, Z. Zhong, and C. Ma, "A multimodal Chinese sarcasm detection model integrated with linguistic features," J. Zhengzhou Univ. Sci. Ed., pp. 1–8, 2024. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CAPJ&filename=ZZDZ20241030001.

[29] J. Zhang and L. Jiang, "Research on irony recognition of travel reviews based on multi-modal deep learning," Inf. Stud. Theory Appl., vol. 45, no. 7, pp. 158–164, Jul 2022. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=QBLL202207022.

[30] A. Joshi, S. Agrawal, P. Bhattacharyya, and M. J. Carman, "Expect the unexpected: harnessing sentence completion for sarcasm detection," Commun. Comput. Inf. Sci., vol. 781, pp. 275–287, Mar 2018.

[31] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," Conference on Empirical Methods in Natural Language Processing, pp. 704–714, Oct 2013.

[32] X. Bai and R. Huo, "Research on recognition model of ironic text integrating language characteristics," Commun. Technol., vol. 54, no. 5, pp. 1126–1130, May 2021 https://cstj.cqvip.com/Qikan/Article/Detail?id=7104582899.

[33] X. Bai and R. Huo, "Ironic text recognition model incorporating language characteristics," Commun. Technol., vol. 54, no. 1, pp. 73–76, Jan 2021. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=TXJS202101012.

[34] X. Lu, Y. Li, and Wang , Suge, " Linguistic features enhanced convolutional neural networks for irony recognition," J. Chinese Inf. Process., vol. 33, no. 5, pp. 31 – 38, May 2019. https://www.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&filename=MESS201905004.

[35] V. Govindan and V. Balakrishnan, "A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 5110–5120, Sept 2022.

[36] X. Jia, Z. Deng, F. Min, and D. Liu, "Three-way decisions based feature fusion for Chinese irony detection," Int. J. Approx. Reason., vol. 113, pp. 324–335, Oct 2019.

[37] J. Liao, M. Wang, X. Chen, S. Wang, and K. Zhang, "Dynamic commonsense knowledge fused method for Chinese implicit sentiment analysis," Inf. Process. Manag., vol. 59, no. 3, pp. 102934, May 2022.

[38] J. Wei, J. Liao, Z. Yang, S. Wang, and Q. Zhao, "BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis," Neurocomputing, vol. 383, pp. 165–173, 2020.

[39] R. Xiang et al., "Ciron: A new benchmark dataset for Chinese irony detection," Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., no. May, pp. 5714–5720, 2020.

[40] B. Liang, Z. Lin, R. Xu, and B. Qin, "Topic-oriented sarcasm detection: new task, new dataset and new method," J. Chinese Inf. Process., vol. 37, no. 2, pp. 138–157, Oct 2023.

[41] W. Hu, L. Chen, X. Huang, C. Chen, and Z. Zhong, "A multimodal Chinese sarcasm detection model for emergencies based on cross attention," CAAI Trans. Intell. Syst., vol. 19, no. 2, pp. 392–400, 2024.

[42] J. Wei, X. Ren, X. Li, W. Huang, Y. Liao, and Y. Wang, et al. "NEZHA: neural contextualized representation for Chinese language understanding," arXiv, pp. 1–9, Nov 2021. https://doi.org/10.48550/arXiv.1909.00204%0A.