CangjieToxi: A Chinese Offensive Language Detection Benchmark with Radical-Level Perturbations

Anonymous ACL submission

Abstract

In this paper, we introduce CangjieToxi, a novel benchmark dataset designed to address the challenges of detecting covert offensive language in Chinese social media. Existing detection systems are often ineffective against evasion techniques that manipulate character structure to bypass censorship. We focus on two key perturbation methods: character splitting and character substitution. Character splitting involves breaking down offensive words into visually similar but contextually distinct components, while character substitution replaces offensive characters with visually similar but non-offensive ones, thus concealing the original intent. Our dataset incorporates these techniques to create more complex forms of toxicity that are difficult for traditional models to detect. We conduct extensive experiments with state-of-the-art models, revealing their limitations in handling these perturbations and demonstrating the need for more robust systems. This work advances the field by providing a resource to improve the detection of cloaked offensive language and contributing to the development of censorship-resistant detection methods. Details can be found on GitHub repository [1].

Disclaimer: *This paper describes violent and discriminatory content that may be disturbing to some readers.*

1 Introduction

In China, while social media censorship is pervasive, it is somewhat less restrictive when it comes to gender and LGBTQ+ topics compared to other politically sensitive issues. Although certain boundaries remain, these discussions still manage to surface, particularly in "safe zones" such as international events, public health concerns (e.g., AIDS), and the arts, where censorship is more lenient. (Yu, 2024) This relatively relaxed approach has fostered a space where gender and LGBTQ+ topics can continue to be discussed, often in subtle ways, such as through the use of emojis or references to foreign contexts. (Gu and Heemsbergen, 2023) Despite these allowances, the digital space remains a battleground for gendered and LGBTQ+ hate speech, as harmful content targeting marginalized groups, like women and sexual minorities, thrives in covert forms. While censorship does not completely stifle feminist or LGBTQ+ discourse, it shapes the way these conversations unfold, contributing to both the visibility and the persistence of offensive language.

Researchers have developed machine learning and Natural Language Processing (NLP) systems, particularly large language models (LLMs), to detect offensive content across various languages. While these models show promise, they struggle against covert offensive language, which is designed to evade detection. Evasion tactics include homophonic substitutions, emoji replacements, and character splitting, techniques that obscure the harmful content from automated systems while remaining understandable to human readers. (Jiang et al., 2022) For example, the offensive phrase “操逼” (a vulgar insult) can be split using Chinese radicals into “扌辶,” (Chen, 2012) effectively disguising the original intent. Similarly, “操你妈逼” can be camouflaged as “澡称冯福” through radical substitution, making it difficult for automated models to flag as offensive while being easily comprehended by users familiar with the context.(Husain and Uzuner, 2021)

The Chinese language, in particular, is vulnerable to these evasion techniques due to lexicon-based censorship, which encourages users to creatively bypass detection. These covert methods often involve replacing offensive terms with homophones or emojis, techniques that can fool automated systems but are easily understood by human readers. As a result, offensive language continues

[1] https://anonymous.4open.science/r/CangjieToxi-6D02

to spread unchecked across social media platforms.

Current moderation systems are ill-equipped to detect these cloaked forms of offensive language, leaving harmful content to proliferate. This growing gap in detection capabilities highlights the urgent need for more robust and adaptable models that can recognize and interpret these subtle forms of toxicity.

To address this challenge, we introduce the dataset, which aims to push the boundaries of existing detection systems by incorporating innovative perturbations like radical-based character decomposition and radical substitution. These techniques create more complex forms of offensive language, challenging models to detect harmful content in ways that go beyond traditional methods.

This study offers several key contributions:

The introduction of the dataset, which serves as a benchmark for assessing the robustness of offensive language detection models. A comprehensive evaluation of state-of-the-art LLMs, demonstrating their limitations in detecting cloaked content.

An in-depth analysis of context-dependent toxicity in single-character tokens, revealing that existing automated methods struggle to accurately distinguish between toxic and non-toxic usage. A critical assessment of lexicon-based filtering, highlighting its high false positive rate due to the misclassification of socially critical but non-toxic comments.

Recommendations for improving toxicity detection through context-aware modeling and hybrid approaches that integrate lexicon-based methods with machine learning.

2 Related Work

2.1 Chinese Offensive Content Dataset

Several datasets have been developed for detecting offensive content in Chinese, each addressing specific types of offensive language. The Chinese Offensive Language Dataset (COLD) categorizes content into attacks on individuals, groups, and anti-bias categories, although it is limited in diversity and lacks representation of the full spectrum of offensive language (Deng et al., 2022). The TOCP (Yang and Lin, 2020) and TOCAB (Chung and Lin, 2021) datasets , originating from Taiwan's PTT platform, focus on detecting profanity and abusive language, while Sina Weibo Sexism Review (SWSR) specifically targets sexism within Chinese social media, offering a lexicon for abusive and gender-related terms (Jiang et al., 2022). The ToxiCN dataset (Lu et al., 2023), which incorporates multi-level labeling for offensive language, hate speech, and other categories, serves as the foundation for the newly introduced ToxiCloakCN, which enhances detection by addressing the challenge of cloaked offensive content, such as homophonic substitutions and emoji transformations (Xiao et al., 2024). These datasets provide valuable resources but often fall short in capturing evolving tactics like cloaking or nuanced expressions of offense.

2.2 Chinese Offensive Content Detection

A range of models have been developed to detect offensive content in Chinese, leveraging techniques such as lexicon-based approaches, supervised learning, and fine-tuned pre-trained models. Lexicon-based models have been widely used but struggle to detect emerging offensive terms (Deng et al., 2022). Machine learning models, including supervised and adversarial learning, offer improved detection, but their performance is often limited by the evolution of language and the subjectivity of offensive content (Liu et al., 2023). Research on domain adaptation (Ying et al., 2024) and cross-cultural transfer learning (Zhou et al., 2023) has further shown that language models trained on other languages can be adapted to explicit detection of Chinese offensive languages with promising results. Recent research has highlighted the effectiveness of large language models (LLMs) in context-aware hate speech detection. Guo et al. showed that LLMs outperform traditional models by using specialized prompting strategies to better capture the context of hate speech. (Guo et al., 2023) Kumarage et al. also explored the strengths of LLMs in hate speech classification (Kumarage et al., 2024) Additionally, Nirmal et al. (2023) introduced an interpretable hate speech detection method using LLM-extracted rationales. (Nirmal et al., 2024)

Our proposed ABC dataset introduces new perturbations like radical-based decomposition and substitution to challenge existing models, aiming to improve the detection of more complex forms of offensive content.

2.3 Language Perturbation

Language perturbation techniques have been explored to examine vulnerabilities in NLP models, especially in adversarial settings. Techniques like emoji insertion (Kirk et al., 2022) and token replacement (Garg and Ramakrishnan, 2020) are

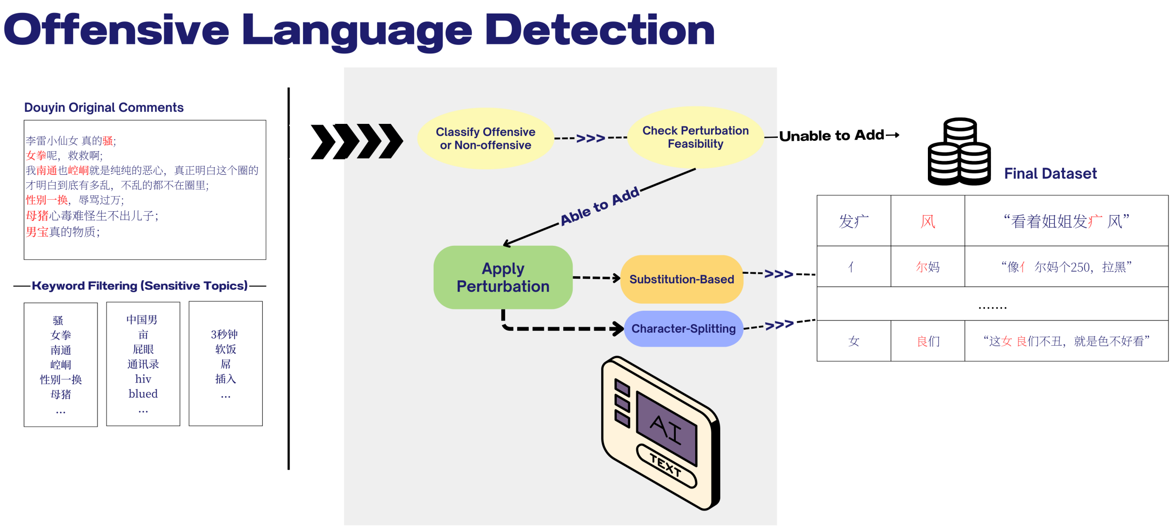

Figure 1: Offensive Langurage Detection Flowchart

commonly used to test the robustness of models against subtler forms of offensive content. In Chinese, language perturbation faces additional challenges due to the language's character-based structure, where meaning can shift dramatically with slight modifications in characters or word order. Previous work on Chinese offensive language detection has addressed perturbations such as word perturbation and synonym usage (Su et al., 2022), while the introduction of ToxiCloakCN demonstrates the impact of homophonic substitutions and emoji transformations on model performance (Xiao et al., 2024).

Our dataset expands on these perturbation techniques by incorporating radical splitting and substitution of character components, adding a new layer of complexity to model testing and addressing emerging evasion tactics in Chinese offensive language detection.

3 Dataset Construction

In this section, we describe the process of constructing the dataset used for offensive language detection, including data collection, preprocessing, offensive keyword extraction, and annotation, as well as the techniques used to introduce meaningful perturbations to the dataset for training purposes. The visualization of the comprehensive process is shown in 2.

3.1 Data Source and Preprocessing

We collect comments from Douyin, a major short video platform in China. Due to the site's filtering system, posts containing offensive language are relatively rare. To address this, we focus our data collection on several sensitive topics, such as marriage, gender, fertility, LGBTQ issues, and race, which are frequently discussed online. We then compile a list of keywords for each topic and use them to gather 45484 comments that do not have replies. We exclude texts that are too short to convey meaningful content, such as those consisting only of auxiliary words or inflections. Additionally, we remove irrelevant data, such as duplicate entries and advertisements. Ultimately, 28080 comments are retained. During the data cleaning process, we standardize the unique web text formats as outlined by Ahn et al. (2020), removing unnecessary newlines and spaces. To protect privacy, we anonymize the data by filtering out usernames, links, emails and stickers. Since emojis may contain valuable emotional cues, we retain them for the purpose of offensive language detection.

3.2 Offensive Keywords Extraction

In order to enrich our dataset with meaningful perturbations, we applied a multi-step approach for offensive keyword extraction. First, we utilized the BERTopic model for topic modeling on our dataset, identifying offensive terms from the representative words of each topic. Additionally, we leveraged existing lexicons, such as the SexHate Lexicon from the SWSR dataset and the gender and LGBTQ+ lexicon from the ToxiCN dataset, to filter relevant offensive keywords. After filtering, we merged these external lexicons with the

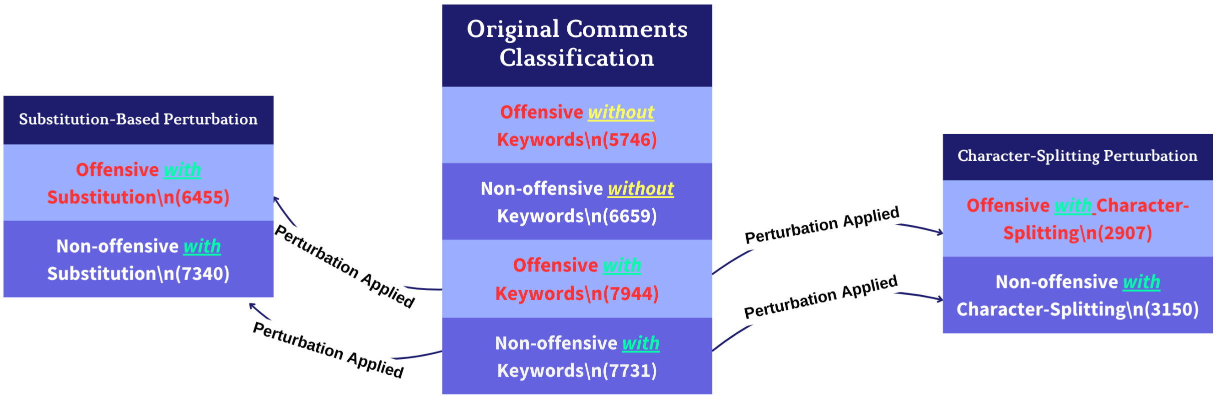

Figure 2: Offensive Langurage Detection Flowchart

offensive terms we defined ourselves, creating a comprehensive keyword list, consisting of 300 offensive keywords. This lexicon was then used to screen the entire dataset for offensive content.

3.3 Human Annotation

For the annotation process, we conducted a manual review of the filtered dataset. A total of four native Chinese annotators with social science backgrounds were involved, ensuring gender balance in the team. To assess the reliability of the annotations, we calculated the interannotator agreement using Fleiss's Kappa, which yielded a value of 0.829, indicating a high level of agreement among the annotators. This robust agreement suggests the reliability and consistency of the offensive labels applied to the dataset.

3.4 Character-Level Perturbation

To better simulate the process of character substitution and splitting used by people to evade censorship on social media, our approach follows key principles grounded in visual recognition studies. Research has shown that substitutions or variations in character structure, as long as the distribution of information within the character remains consistent—such as maintaining the relative positions of phonetic and semantic radicals—do not significantly affect a reader's ability to recognize meaning or pronunciation (Hsiao and Cheng, 2013). This aligns with findings that visual recognition advantages in the right visual field (RVF) persist when phonetic components appear on the right and semantic components on the left, a structure commonly observed in Chinese characters (wen Hsiao, 2011). Additionally, studies on radical combinability indicate that position-specific radical combinability (SRC) is a stronger predictor of neural activation in character recognition than position-general radical combinability (GRC), suggesting that radical position matters more than sheer frequency (Liu et al., 2022). By preserving these positional relationships—especially in left-right and up-down structures—our modifications ensure that the altered characters remain easily interpretable by human readers while disrupting automated detection systems.

Our perturbation strategy differs for offensive and non-offensive text:

1. *Perturbation of offensive Text:* We only perturb words that appear in a predefined list of specific offensive keywords. This selective perturbation ensures that modifications are concentrated on words strongly associated with toxicity while avoiding unnecessary changes to unrelated words. For example, in the phrase "妈逼" (a profane expression), the character "妈" will be perturbed, whereas in "妈妈" (mother), no perturbation will occur.

2. *Perturbation of Non-offensive Text:* We perturb all individual characters that appear in the keyword list, even if they are not part of offensive words. While these perturbations are unrelated to toxicity, this design prevents the model from learning incorrect associations during training—such as mistakenly linking rare characters or structural variations with toxicity. For instance, in the word "妈妈" (mother), the character "妈" will be perturbed.

Our approach to character perturbation adheresit to three main principles:

1. *Character Structure:* We selected characters whose structure could be further split, avoiding non-split characters such as "广" (which

cannot be split further). We primarily chose left-right and top-bottom structured Chinese characters, as they are the most frequently used formations in written Chinese.

2. *Position Consistency:* For both substitution and splitting, we ensured that the components retained their relative positions within the character. This structural stability minimizes disruptions in visual recognition, allowing readers to process the modified text with minimal effort.

3. *Radical Frequency:* We focused on structural components (radicals) frequently employed in character variations, ensuring that the substitutions remained consistent with real-world linguistic modifications and had minimal impact on readability.

By following these principles, our character perturbation strategy effectively mimics real-world tactics used by social media users to bypass censorship while preserving readability for human readers.

3.4.1 Character Splitting

In the Character Splitting step, we used the splitting dictionary provided by the `funnlp` library[2] to match characters in our offensive word list. The library offers multiple splitting methods for each character, and we selected the most optimal splitting method based on our principles.

The splitting rules were as follows:

1. We only split characters into two components. If a character's components exceeded two, they were placed in non-typical positions, negatively affecting recognition. For example, the character "搏" (bó) splits into '手' (hand) + '甫' (fu) + '寸' (inch), but '寸' is expected to be at the bottom of "甫," making the split unnatural.

2. When multiple splitting methods were available, we chose the method where the components' positions most closely resembled those of the original character. For instance, the character "擦" (wipe) has three splitting methods:

   - "擦" → "手" (hand) + "察" (inspect)
   - "擦" → "扌" (hand radical) + "察" (inspect)
   - "擦" → "才" (only) + "察" (inspect)

   We chose the second method because "扌" (hand radical) is most frequently seen on the left side of a character, making it the most natural and recognizable modification.[3]

3.4.2 Character Substitution

In the Character Substitution step, we relied on the library of the *Chinese Text Project* (中国哲学书电子化计划) to substitute the radical of characters from 101 offensive words, selected from a total of 300 offensive terms. These substitutions involved modifying 427 Chinese characters using different radicals.[4]

Since a single Chinese character can be substituted with multiple radicals, we followed the principle of radical frequency to determine the most suitable replacements. Specifically, we used the *Xiandai Hanyu Changyong Zibiao* (List of Frequently Used Characters in Modern Chinese) provided by the Ministry of Education [5]. Based on the individual character frequencies, we selected the most frequent substitute character with the highest frequency of occurrence as the replacement. For example, the character "猥琐" (lewd) was substituted with "偎唢" following this approach, as these substitutions closely align with commonly used radicals in modern Chinese.

This method ensures that the substitutions reflect both linguistic frequency and the intended meaning while avoiding arbitrary or non-standard replacements, helping to maintain the readability of the altered text.

4 Experiments

To evaluate the effectiveness of existing models and methods on our proposed benchmark, we employed the following experimental setup and methodologies. This systematic approach ensures a comprehensive assessment of model performance and robustness in detecting offensive language under various perturbations.

4.1 Baseline

The evaluation of three state-of-the-art models—DeepSeek-V3, GPT-4o, and Qwen-Max—revealed notable trends in their performance

[2] https://github.com/fighting41love/funNLP

[3] https://lingua.mtsu.edu/chinese-computing/statistics/index.html

[4] https://ctext.org/dictionary.pl?if=gb

[5] https://lingua.mtsu.edu/chinese-computing/statistics/index.html

Model	Accuracy	Macro F1 Score
DeepSeek-V3	0.7286	0.7255
GPT-4o	0.7329	0.7309
Qwen-Max	0.7447	0.7432

Table 1: Performance of Models on Full Dataset

under character decomposition (拆字) and character substitution (换字) perturbations. On the original data, Qwen-Max achieved the highest accuracy (0.7868) and Macro F1 score (0.7858), followed by DeepSeek-V3 and GPT-4o. After applying character decomposition, all models experienced a performance decline, with DeepSeek-V3 dropping to an accuracy of 0.7165 and a Macro F1 score of 0.7150, GPT-4o dropping to an accuracy of 0.6875 and a Macro F1 score of 0.6839, and Qwen-Max dropping to an accuracy of 0.7281 and a Macro F1 score of 0.7267.

For character substitution, Qwen-Max again led with an accuracy of 0.8132 and a Macro F1 score of 0.8122, while DeepSeek-V3 and GPT-4o achieved accuracies of 0.7752 and 0.7818, respectively. The performance drop following character decomposition highlights the increased difficulty posed by this perturbation type. Notably, Qwen-Max exhibited the smallest performance degradation, suggesting stronger robustness to adversarial transformations compared to the other models. These results underscore the challenges of character-level perturbations and the varying resilience of models in handling such modifications. Detailed model performance can be seen in Table 1 and Table 2.

4.2 Experiment Settings

To ensure standardized and reproducible outputs from large language models (LLMs) in our experiments, we utilized the dspy framework. This framework provides a structured approach to prompt engineering and output generation, enabling consistent evaluation across different models and settings.

Prior research has demonstrated that using Chinese prompts yields marginally better performance in detecting offensive language in Chinese text compared to English prompts (Xiao et al., 2024). To align with these findings and maintain consistency, we adopted a uniform Chinese prompt across all experiments. The prompt used for offensive language detection is as follows:

Prompt: "下面是一条来自于社交媒体的评论，请判断这条评论是否是冒犯性的(offensive)。若你认为评论是’冒犯性的(offensive)’的，请输出1，否则输出0。" (Translation: "Below is a comment from social media. Please determine whether this comment is offensive. If you consider the comment ’offensive,’ output 1; otherwise, output 0.")

This prompt was designed to explicitly instruct the model to classify the input text as either offensive or non-offensive, ensuring clarity and consistency in model responses.

Model Configuration For all experiments involving LLMs, we set the temperature parameter to 0 to minimize randomness in model outputs and ensure deterministic behavior. All other hyperparameters were kept at their default values to maintain a fair and controlled evaluation environment.

This standardized setup allows for a rigorous comparison of model performance on our benchmark, while also providing a foundation for future reproducibility and extension of our work.

5 Results and Discussion

5.1 Single Char

During the construction of the dataset, we observed that certain single-character tokens (e.g., “鸡” often used as a sexualized insult, and “艾”, “梅”, “淋” commonly appear as the first character in sexually transmitted disease names) could potentially indicate toxic content. However, a significant portion of comments containing these tokens were found to be non-toxic upon manual inspection, highlighting that the toxicity of such tokens is highly context-dependent.

In our toxic comment dataset, we included comments containing these specific tokens and manually annotated them to determine their toxicity. Despite this effort, we identified a critical limitation: current automated methods struggle to accurately distinguish whether comments containing these tokens are toxic or not. This underscores the need for more sophisticated context-aware approaches to improve the precision of toxicity detection. Future work should focus on developing models capable of capturing nuanced contextual cues to address this challenge effectively.

5.2 Lexicon and False Positive

The lexicon-based filtering approach exhibited a high false positive rate, where non-toxic content was frequently misclassified as toxic. A primary reason for this is the prevalence of com-

Model	Before Split		After Split		Before Substitution		After Substitution	
	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DeepSeek-V3	0.7665	0.7661	0.7165	0.7150	0.7752	0.7737	0.7107	0.7101
GPT-4o	0.7629	0.7619	0.6875	0.6839	0.7818	0.7807	0.6793	0.6792
Qwen-Max	0.7868	0.7858	0.7281	0.7267	0.8132	0.8122	0.7157	0.7157

Table 2: Model Performance in Different Conditions

ments criticizing socially undesirable behaviors (e.g., fraud, promiscuity), which, despite their harsh tone, do not constitute offensive language. This phenomenon poses a significant challenge for offensive language detection systems, as it blurs the line between legitimate criticism and actual toxicity.

To mitigate this issue, future research should prioritize the development of more advanced semantic understanding and context-aware models. Incorporating domain-specific knowledge and leveraging larger, more diverse datasets could help reduce false positives. Additionally, exploring hybrid approaches that combine lexicon-based methods with machine learning models may offer a more robust solution for distinguishing between toxic content and socially critical discourse.

5.3 Future Works

Addressing offensive language that evades censorship mechanisms through techniques such as character splitting or using visually similar characters may involve two potential approaches. One approach is to employ computer vision (CV) methods to identify and associate similar characters and split characters. However, this method is costly and complicated, as the flexible structure of Chinese characters makes the problem more challenging. An alternative approach is to use "masking" techniques, which obscure key offensive terms while still allowing offensive language to be understood and recognized through contextual semantic clues—essentially enabling the system to infer meaning even when specific words are not explicitly stated (i.e., "although nothing was directly said, the intent is still understood"). The dataset we propose, which introduces perturbations only to offensive terms, is adaptable to both of these strategies.

6 Limitations

Despite the contributions made by CangjieToxi, there are several limitations in this study that should be acknowledged. First, while the dataset introduces novel perturbations such as character splitting and character substitution, it remains limited to Chinese language contexts, and the effectiveness of these evasion techniques may vary in other languages with different writing systems or character structures. Second, the perturbation methods used in this work, although effective in creating subtle forms of offensive language, are still constrained by the manual construction of these transformations, and there may be additional, unforeseen evasion tactics that were not covered. Third, the performance of state-of-the-art models on our dataset demonstrates clear limitations, but further research is needed to explore new model architectures and training methodologies that can better adapt to these types of perturbations. Finally, while we have focused on offensive language detection within social media contexts, the dataset's applicability to other domains, such as formal text or legal documents, remains to be evaluated. Future work will aim to expand these methods, explore additional types of perturbations, and assess the robustness of models across different languages and content domains.

References

Wangdao Chen. 2012. *Rhetoric introduction*. Fudan University Press. Publication date: January 1, 2008.

I Chung and Chuan-Jie Lin. 2021. Tocab: A dataset for chinese abusive language processing. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 445–452.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), pages 6174–6181, Online. Association for Computational Linguistics.

Yijia Gu and Luke Heemsbergen. 2023. The ambivalent governance of platformed chinese feminism under censorship: Weibo, xianzi, and her friends. *International Journal of Communication*, 17(0).

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Janet H. Hsiao and Liao Cheng. 2013. The modulation of stimulus structure on visual field asymmetry effects: The case of chinese character recognition. *The Quarterly Journal of Experimental Psychology*, 66(9):1739–1755. PMID: 23391072.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *Preprint*, arXiv:2403.08035.

Hanyu Liu, Chengyuan Cai, and Yanjun Qi. 2023. Expanding scope: Adapting English adversarial attacks to Chinese. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 276–286, Toronto, Canada. Association for Computational Linguistics.

Xiaodong Liu, David Wisniewski, Luc Vermeylen, Ana F. Palenciano, Wenjie Liu, and Marc Brysbaert. 2022. The representations of chinese characters: Evidence from sublexical components. *Journal of Neuroscience*, 42(1):135–144.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. RoCBert: Robust Chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, Dublin, Ireland. Association for Computational Linguistics.

Janet Hui wen Hsiao. 2011. Visual field differences in visual word recognition can emerge purely from perceptual learning: Evidence from modeling chinese character pronunciation. *Brain and Language*, 119(2):89–98. Neurocognitive Processing of the Chinese Language.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Hsu Yang and Chuan-Jie Lin. 2020. TOCP: A dataset for Chinese profanity processing. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 6–12, Marseille, France. European Language Resources Association (ELRA).

Hao Ying, Qiongrong Ou, Chengjun Fan, Lin Mei, Shuyu Zhang, and Xu Xu. 2024. Domain adaptation fornbsp;chinese offensive language detection. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part IV*, page 146–158, Berlin, Heidelberg. Springer-Verlag.

Jinyang Yu. 2024. *Shifting shadows: media attention and censorship of gay people in China (1949-2023)*. Ph.D. thesis, University of British Columbia.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.