System Report for CCL25-Eval Task 10: SRAG-MAV for Fine-Grained Chinese Hate Speech Recognition

Jiahao Wang, Ramen Liu, Longhui Zhang, Jing Li*
Harbin Institute of Technology, Shenzhen, China
wjh123king@gmail.com, jingli.phd@hotmail.com

Abstract

This paper presents our system for CCL25-Eval Task 10, addressing Fine-Grained Chinese Hate Speech Recognition (FGCHSR). We propose a novel SRAG-MAV framework that synergistically integrates task reformulation(TR), Self-Retrieval-Augmented Generation (SRAG), and Multi-Round Accumulative Voting (MAV). Our method reformulates the quadruplet extraction task into triplet extraction, uses dynamic retrieval from the training set to create contextual prompts, and applies multi-round inference with voting to improve output stability and performance. Our system, based on the Qwen2.5-7B model, achieves a Hard Score of 26.66, a Soft Score of 48.35, and an Average Score of 37.505 on the STATE ToxiCN dataset, significantly outperforming baselines such as GPT-4o (Average Score 15.63) and fine-tuned Qwen2.5-7B (Average Score 35.365). The code is available at https://github.com/king-wang123/CCL25-SRAG-MAV.

1 Introduction

The growth of social media has significantly amplified the spread of hate speech (Fortuna and Nunes, 2018), with malicious content targeting attributes such as race, region, and gender, causing substantial harm to individuals and society. Effective hate speech detection has become a critical focus in Natural Language Processing (NLP), aiming to mitigate these negative impacts (Davidson et al., 2017; Waseem and Hovy, 2016). Moreover, ensuring the fairness of detection models to avoid potential biases is essential for their practical deployment (Sap et al., 2019). Traditional methods often rely on binary classification to identify hateful content (Fortuna and Nunes, 2018), but these approaches lack the granularity to capture the internal structure of hate speech, limiting their interpretability and utility for downstream applications (Yin and Zubiaga, 2021). Consequently, Fine-Grained Chinese Hate Speech Recognition (FGCHSR), which extracts structured information such as specific targets or types of hate, has gained increasing attention (Basile et al., 2019; Mathew et al., 2021; Ren et al., 2021).

CCL25-Eval Task 10 focuses on extracting quadruplets (Target, Argument, Targeted Group, Hateful) from Chinese social media texts. This task is particularly challenging due to the subtle, context-dependent nature of Chinese hate speech (Pavlopoulos et al., 2020), the interdependence of quadruplet elements, and the limited availability of high-quality annotated data (Yin and Zubiaga, 2021). The STATE ToxiCN study (Bai et al., 2025) highlights these difficulties, showing that even the most advanced models like GPT-4o achieve an Average Score of only 15.63, while fine-tuned open-source models like Qwen2.5-7B reach 35.365, but still require further optimization.

To address these challenges, we propose a novel SRAG-MAV framework synergistically combining Task Reformulation (TR), Self-Retrieval-Augmented Generation (SRAG), and Multi-Round Accumulative Voting (MAV). Our approach simplifies quadruplet extraction into triplet extraction, enhances contextual understanding through dynamic retrieval inspired by Retrieval-Augmented Generation (RAG)

*Corresponding author: Jing Li (jingli.phd@hotmail.com).

Proceedings of the 24th China National Conference on Computational Linguistics, pages 395-402, Jinan, China, August 11-14, 2025.

(Lewis et al., 2020), and ensures stable outputs via multi-round inference based on the principles of Parallel Scaling Law (PARSCALE) (Chen et al., 2025). Our contributions include:

- Proposed a novel SRAG-MAV framework that integrates TR, SRAG, and MAV, demonstrates superior performance for FGCHSR and is adaptable to other structured NLP tasks.
- Conducted comprehensive experiments, validating the effectiveness and robustness of our approach and assessing the performance contributions of individual components.
- Released code at `https://github.com/king-wang123/CCL25-SRAG-MAV`, promoting reproducibility and facilitating further research in hate speech detection and other related NLP domains.

2 Methodology

2.1 System Overview

Our system combines TR, SRAG, and MAV to extract fine-grained hate speech quadruplets from Chinese social media texts, as shown in Figure 1. The workflow simplifies the task into triplet extraction, enhances contextual understanding through retrieval, and stabilizes outputs via iteratively voting.

Initially, we transform the quadruplet dataset into a triplet dataset to reduce task complexity, aligning with the principles of TR. Then, we encode all training inputs into a vector database using a retrieval model, enabling efficient retrieval for both training and inference phases.

Training Phase: For each input, we retrieve the most similar sample from the training set excluding the input itself, concatenate the retrieved sample with the input to form a prompt. These prompts along with their corresponding triplet outputs, form new training samples for fine-tuning.

Inference Phase: For each test input, we retrieve the top-k similar training samples to construct k prompts. The model iteratively generates triplets for these k prompts across multiple rounds until the frequency of the most frequent triplet exceeds the threshold τ, at which point MAV selects it as the final triplet answer. This selected triplet is then converted into a quadruplet by inferring hatefulness from the target group.

This pipeline, visualized in Figure 1, balances simplicity, contextual richness (Lewis et al., 2020), and output stability, with k and τ as key hyperparameters.

2.2 Task Reformulation (TR)

Analysis of the training data reveals a strong correlation between the target group and hatefulness label: "no-hate" hatefulness occurs only when the target group is "no-hate"; otherwise, the hatefulness is labeled as "hate." Leveraging this pattern, we reformulate the original quadruplet extraction task into a triplet extraction task. This simplification reduces the complexity of structured generation, as the hatefulness label can be deterministically inferred from the target group, thereby improving the efficiency and accuracy of large language models (LLMs).

Figure 2 provides a concrete data example to illustrate how TR simplifies the extraction process while maintaining the integrity of the structured output.

2.3 Self-Retrieval-Augmented Generation (SRAG)

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm in NLP, widely applied to tasks such as question answering, dialogue systems, and knowledge-intensive text generation (Lewis et al., 2020; Gao et al., 2023). By integrating external knowledge through retrieval, RAG enhances the contextual relevance and factual accuracy of generated outputs (Ram et al., 2023). Recent advancements have further refined RAG to handle structured data and improve robustness in low-resource settings (Zhang et al., 2023). However, FGCHSR poses unique challenges, including the lack of high-quality external corpora and the complexity of structured quadruplet generation.

To address these challenges, we propose the Self-Retrieval-Augmented Generation (SRAG) framework, which adapts the RAG paradigm by using the training set itself as the retrieval corpus. SRAG

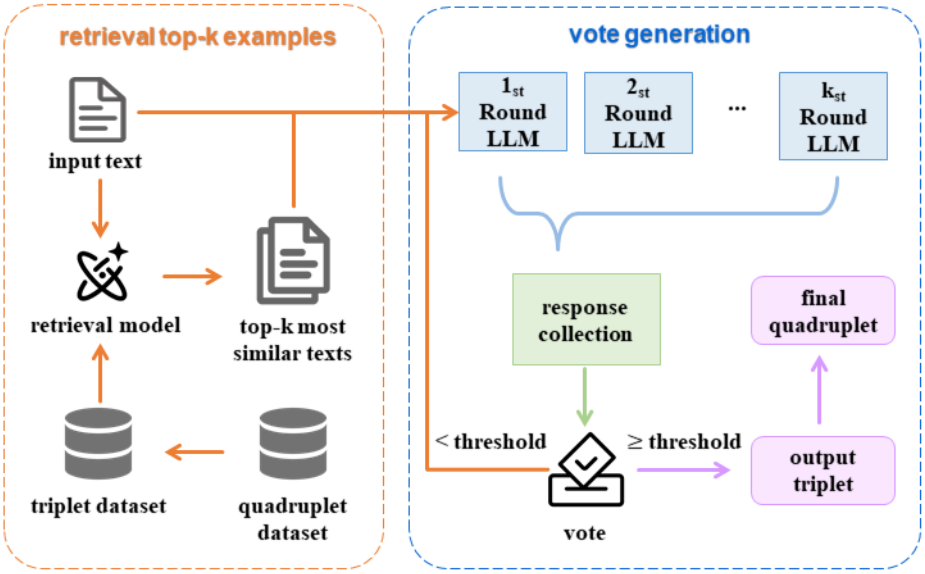

Figure 1: System architecture of SRAG-MAV, depicting the workflow from input text to final quadruplet output. The process includes: (1) transforming the quadruplet dataset into a triplet dataset, (2) retrieving top-k similar samples and concatenating each with the input text to construct prompts, (3) voting to select triplet answer with frequencies exceeding the threshold τ, otherwise continuing iterative inference until the threshold is met, and (4) converting the selected triplet into the final quadruplet output.

leverages semantically similar annotated examples to guide triplet generation, ensuring contextually relevant outputs without requiring external resources. The SRAG pipeline includes:

1. **Corpus Construction**: We use the bge-large-zh-v1.5 model (Xiao et al., 2023) to generate embeddings for training set texts, building a retrieval corpus based on cosine similarity.

2. **Dynamic Retrieval**: For each input text, we retrieve the most similar sample except the input itself during training, while retrieving the top-k most similar samples during inference.

3. **Prompt Generation**: Retrieved samples are combined with the input text to create structured prompts, guiding the model to produce task-compliant triplets, exemplified in Figure 2.

SRAG innovatively leverages the training set as a dynamic retrieval corpus, enabling few-shot learning through similar annotated examples to enhance task understanding and output accuracy. Unlike traditional RAG, SRAG eliminates the need for external data, making it particularly suited for resource-constrained environments and domain-specific tasks like FGCHSR.

2.4 Multi-Round Accumulative Voting (MAV)

Parallel Scaling Law (PARSCALE) (Chen et al., 2025) illustrates that applying diverse transformations to an input to generate multiple variants, followed by parallel inference and result aggregation with learnable parameters, can significantly enhance LLMs performance. This approach improves robustness and accuracy without requiring model retraining, making it efficient for complex tasks in resource-constrained settings.

Inspired by PARSCALE's emphasis on parallel processing of diverse inputs, we introduce Multi-Round Accumulative Voting (MAV) as an innovative adaptation for FGCHSR, which generates diverse prompts with SRAG-retrieved examples and selects the optimal triplet output through a voting mechanism. The MAV pipeline includes:

1. **Diverse Prompts**: Through SRAG, retrieve the top-k most similar samples from the triplet dataset for each input text, and concatenate each retrieved sample with the input to construct k distinct prompts.

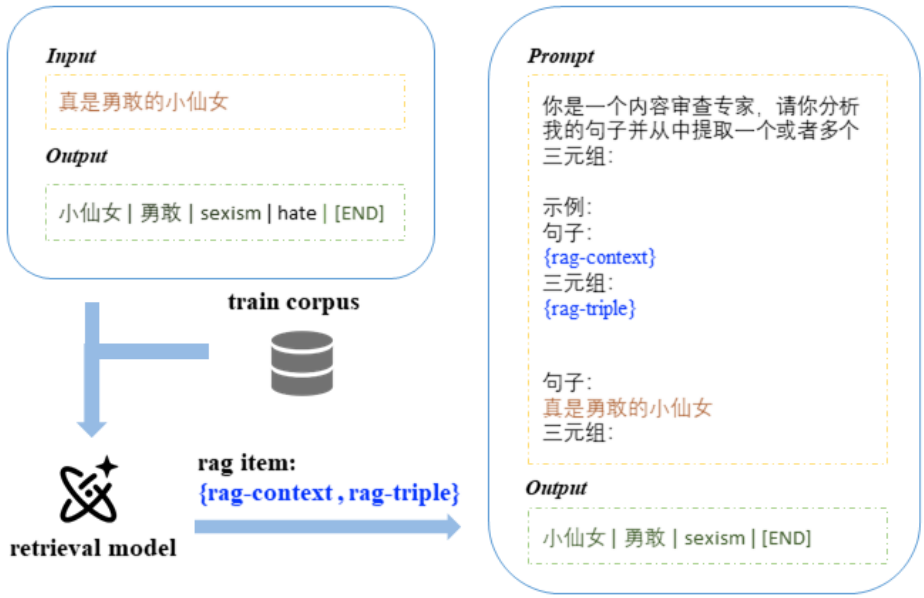

Figure 2: Illustration of TR and SRAG: the retrieval model retrieves similar texts from the training corpus, concatenates them into a prompt, and generates the corresponding triplet.

2. **Multi-Round Inference**: Iteratively perform inference on each prompt, generating and accumulating the frequencies of triplet results across iterations until the most frequent triplet exceeds the threshold τ.

3. **Voting Mechanism**: Select the triplet output reaching a frequency threshold τ and convert it to a quadruplet as the final result.

MAV stands out for its cost-effectiveness, requiring only additional inference-time resources rather than retraining or parameter adjustments. Its flexibility is demonstrated in Section 3.2, where increasing the threshold progressively improves results, allowing dynamic adjustment based on available computational resources. Additionally, its straightforward implementation enhances reliability, making MAV particularly effective under constrained conditions.

3 Experiments

3.1 Experimental Setup

We evaluated our approach on the STATE ToxiCN dataset (Bai et al., 2025), comprising 4,000 training samples and 1,602 test samples, using 4×NVIDIA L40S 40GB GPUs. The base model, Qwen2.5-7B (Team, 2024), was trained with the LLaMA-Factory framework (Zheng et al., 2024) and deployed for inference using vLLM (Kwon et al., 2023). For retrieval, we employed the bge-large-zh-v1.5 model (Xiao et al., 2023), with generation parameters configured at a temperature of 0.7 during fine-tuning and 0.1 for MAV inference. MAV configuration utilized a top-k value of 10 and voting threshold τ of 200. The evaluation metrics included:

- **Hard Score**: F1 score for precise quadruplet matches.
- **Soft Score**: F1 score for partial matches, requiring identical target group and hatefulness, with Target and Argument similarity exceeding 50%.
- **Average Score**: Mean of Hard and Soft Scores.

Baselines were drawn from the STATE ToxiCN benchmarks (Bai et al., 2025), and our approach was compared to validate its effectiveness.

3.2 Experimental Results

We conducted three experiments, including a model comparison to benchmark our system against baselines, an MAV parameter sensitivity analysis to assess the impact of threshold variations on performance, and ablation studies to evaluate the contribution of each component.

3.2.1 Model Comparison

Model	Hard Score	Soft Score	Average Score
mT5-base	16.60	38.61	27.605
Mistral-7B	23.72	45.62	34.670
LLaMA3-8B	24.27	46.08	35.175
Qwen2.5-7B	23.70	47.03	35.365
ShieldLM-14B-Qwen	23.59	45.58	34.585
ShieldGemma-9B	23.49	47.14	35.315
Ours	**26.66**	**48.35**	**37.505**

Table 1: Performance comparison on the STATE ToxiCN test set. Results for baseline models are directly cited from the STATE ToxiCN paper (Bai et al., 2025), representing vanilla Supervised Fine-Tuning (SFT) outcomes. Our approach significantly outperforms these baselines across all metrics.

Table 1 demonstrates that our system achieves a Hard Score of 26.66, a Soft Score of 48.35, and an Average Score of 37.505 on the STATE ToxiCN test set, significantly surpassing all baseline models trained with vanilla Supervised Fine-Tuning (SFT). Our approach yields substantial improvements, particularly in the Hard Score, which reflects precise quadruplet matches and indicates robust performance in FGCHSR.

3.2.2 MAV Parameter Sensitivity Analysis

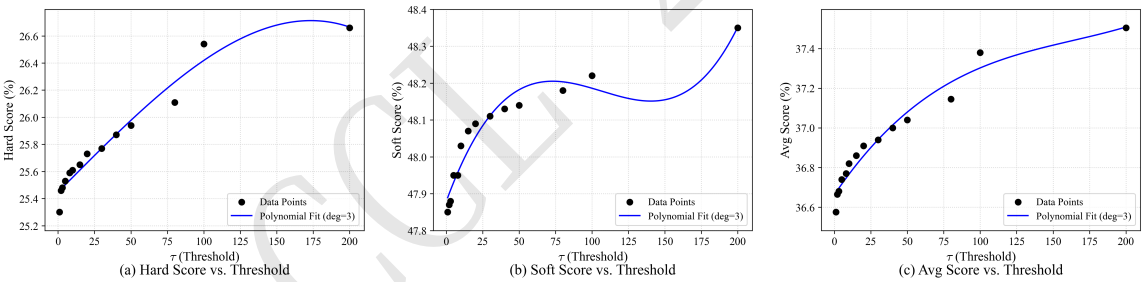

Figure 3: Impact of the MAV threshold parameter (top-k=10) on the STATE ToxiCN test set. The figure illustrates the relationship between the threshold (τ) values and the resulting Hard Score (a), Soft Score (b), and Average Score (c).

Figure 3 reveals a detailed analysis of the MAV threshold (τ) parameter's impact on model performance, with thresholds tested at [1, 2, 3, 5, 8, 10, 15, 20, 30, 40, 50, 80, 100, 200]. The Hard Score exhibits a pronounced increase from 25.30 to 26.66, reflecting a 1.36-point gain, with notable jumps at higher thresholds (e.g., from 26.11 at $\tau = 80$ to 26.66 at $\tau = 200$), underscoring the method's effectiveness in achieving precise quadruplet matches. The Soft Score also improves steadily from 47.85 to 48.35, a 0.50-point rise, indicating enhanced partial match accuracy. The Average Score rises from 36.575 to 37.505, a 0.93-point improvement, reflecting a balanced enhancement across both metrics. These results validate MAV's role in stabilizing outputs through accumulative voting, with the Hard Score's significant growth highlighting its superiority in precise fine-grained hate speech recognition.

Model training provides foundational capabilities, while inference strategies further unleash the model's potential by introducing moderate computational overhead, with the two complementing each

other.

3.2.3 Ablation Study

Configuration	Hard Score	Soft Score	Average Score
Base Model	23.70	47.03	35.365
+ TR	24.33	47.35	35.840
+ TR + SRAG	25.30	47.85	36.575
+ TR + SRAG + MAV	**26.66**	**48.35**	**37.505**

Table 2: Ablation study results, demonstrating the incremental contributions of Task Reformulation (TR), Self-Retrieval-Augmented Generation (SRAG), and Multi-Round Accumulative Voting (MAV) to the overall performance.

As shown in Table 2, the ablation study offers a high-level perspective on the effectiveness of each component in enhancing the model's performance. Beginning with the base model (Qwen2.5-7B trained via vanilla SFT), the introduction of TR simplifies the model's output structure, leading to a noticeable performance uplift that underscores its effectiveness in streamlining the task. Building on this, the addition of SRAG further strengthens the model by leveraging contextual retrieval, resulting in a clear improvement that highlights its role in refining predictions. The final incorporation of MAV delivers the most significant enhancement, markedly boosting the model's stability and accuracy through iterative inference, which emphasizes MAV's pivotal contribution to overall performance.

4 Conclusion

For CCL25-Eval Task 10, we developed a novel SRAG-MAV framework, for the purpose of effectively detecting and mitigating the spread of harmful content on social media. Our approach achieves a Hard Score of 26.66, a Soft Score of 48.35, and an Average Score of 37.505 on the STATE ToxiCN test set (Bai et al., 2025), significantly outperforming baselines such as GPT-4o (Average Score 15.63) and fine-tuned Qwen2.5-7B (Average Score 35.365). TR simplifies the quadruplet extraction task into triplet extraction, reducing complexity; SRAG enhances contextual understanding by leveraging the training set as a retrieval corpus; and MAV ensures output stability through iterative prompt generation and voting. These components work synergistically, as demonstrated by our ablation study, which highlights incremental performance gains from each module.

Our system's open-source implementation (https://github.com/king-wang123/CCL25-SRAG-MAV) fosters reproducibility and further research. However, limitations include the model's domain-specific performance, reliance on text-only data, and MAV's high voting thresholds increase computational costs. Future work will explore cross-domain transfer learning to enhance generalizability (Toraman et al., 2022), multimodal approaches integrating text and images for richer context (Gomez et al., 2020; Das et al., 2020), and optimization of MAV's computational efficiency to broaden its applicability.

Acknowledgements

This work was supported in part by National Science Foundation of China (62476070), Shenzhen Science and Technology Program (JCYJ20241202123503005, GXWD20231128103232001, ZDSYS20230626091203008, KQTD20240729102154066) and Department of Science and Technology of Guangdong (2024A1515011540).

We thank the CCL25-Eval organizers for their platform, the STATE ToxiCN dataset providers for supporting our experiments, and the reviewers for their valuable feedback.

References

Zewen Bai, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, Liang Yang, and Hongfei Lin. 2025. State toxicn: A benchmark for span-level target-aware toxicity extraction in chinese hate speech detection. *arXiv preprint arXiv:2501.15451*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Mouxiang Chen, Binyuan Hui, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Jianling Sun, Junyang Lin, and Zhongxin Liu. 2025. Parallel scaling law for language models. *arXiv preprint arXiv:2505.10475*.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2012.14891*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–628.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. *arXiv preprint arXiv:2109.06705*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 2215–2225.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. A two-stage adaptation of large language models for text ranking. *arXiv preprint arXiv:2311.16720*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llama-factory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410.