LLMsAgainstHate @ NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification in Devanagari Languages via Parameter Efficient Fine-Tuning of LLMs

Rushendra Sidibomma[1], **Pransh Patwa**[2], **Parth Patwa**[3],
Aman Chadha[4,5*], **Vinija Jain**[4], **Amitava Das**[6],

[1]IIIT Sri City, India, [2]Aditya English Medium School, India, [3]UCLA, USA,
[4]Stanford University, USA, [5]Amazon GenAI, USA, [6]University of South Carolina, USA
rushendra.s20@iiits.in, pransh.patwa@aemspune.edu.in, parthpatwa@g.ucla.edu
hi@aman.ai, hi@vinija.ai, amitava@mailbox.sc.edu

Abstract

The detection of hate speech has become increasingly important in combating online hostility and its real-world consequences. Despite recent advancements, there is limited research addressing hate speech detection in Devanagari-scripted languages, where resources and tools are scarce. While large language models (LLMs) have shown promise in language-related tasks, traditional fine-tuning approaches are often infeasible given the size of the models. In this paper, we propose a Parameter Efficient Fine tuning (PEFT) based solution for hate speech detection and target identification. We evaluate multiple LLMs on the Devanagari dataset provided by Thapa et al. (2025), which contains annotated instances in 2 languages - Hindi and Nepali. The results demonstrate the efficacy of our approach in handling Devanagari-scripted content. Our code is available at https://github.com/Rushendra10/Hate-Speech-Detection-and-Target-Identification-in-Devanagari-Languages.

1 Introduction

In recent years, the rise in online hate speech has led to severe social consequences, often escalating into real-world violence and disproportionately affecting vulnerable communities (Laub, 2019). This issue is especially challenging for low-resource languages, where the lack of technological tools limits effective monitoring and mitigation of harmful content (Shen et al., 2024; Court and Elsner, 2024). Addressing hate speech in these languages is important to minimize societal harm and foster safer online environments.

Large Language Models (LLMs) have shown significant potential in handling various language-related tasks, including hate speech detection. However, techniques such as in-context learning (ICL) are increase the cost and latency of LLMs with the increase in data (Liu et al., 2022b). While fine-tuning can improve performance, it remains resource-intensive, given the billions of parameters of LLMs. To address these challenges, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a more adaptable and cost-effective solution, making it a compelling choice for this application (Patwa et al., 2024).

In this paper, we present our system for detection hate-speech in Devanagari-scripted languages. Our key contributions are:

- We introduce a PEFT-based system for detecting hate speech and identifying targeted individuals or groups.
- We evaluate the effectiveness of various LLMs in this context.
- We focus on Devanagari-scripted languages, but our system can be potentially applied to other languages as well.

2 Related Work

Detecting hate speech online has become a critical issue due to the potential for real-world consequences. Traditional research in this area focused primarily on high-resource languages like English, where robust datasets and NLP tools facilitated effective models (Davidson et al., 2017; Fortuna and Nunes, 2018). However, applying these methods to low-resource languages remains a significant challenge due to limited annotated datasets and language-specific resources. For instance, recent research on hate speech detection in Hindi, a low-resource language despite its global prevalence, has highlighted the importance of building dedicated

*Work does not relate to position at Amazon.

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 301–307
January 19, 2025.

datasets and methodologies tailored to these linguistic contexts (Kapil et al., 2023).

Efforts to address these challenges have led to new datasets such as IEHate (Jafri et al., 2023), which specifically captures hate speech in the political discourse of the Indian Assembly Elections. This dataset provides valuable insights and benchmarks for hate speech in low-resource languages, underscoring the need for refined algorithms and hybrid human-machine approaches. Similarly, the HHSD (Kapil et al., 2023) dataset offers a multilayer annotated dataset for hate speech detection in Hindi, structured hierarchically to categorize hate speech into explicit and implicit forms and target attributes. This dataset demonstrates how multitask learning (MTL) frameworks, which combine similar tasks across related languages, can improve performance, further advancing hate speech detection in resource-limited languages.

Researchers have attempted hate-speech detection in low resources languages using various deep learning techniques. Some of the languages explored include Bengali (Safi Samghabadi et al., 2020; Das et al., 2022), Hindi (Patwa et al., 2021b; Shukla et al., 2022; Velankar et al., 2021; Patwa et al., 2021a), Dravidian languages (Tula et al., 2021; Sreelakshmi et al., 2024; Tula et al., 2022) etc. Some researchers have also explored multimodal low resource hate-speech detection (Mishra et al., 2023b,a; Guo et al., 2023). For a detailed discussion, please refer to (Parihar et al., 2021).

Large Language Models (LLMs) have improved detection capabilities but require considerable resources for fine-tuning. Parameter-Efficient Fine-Tuning (PEFT) techniques allow for efficient adaptation by tuning only a subset of model parameters, making them practical for low-resource settings (Li and Liang, 2021; Lester et al., 2021). Language-agnostic models, leveraging machine translation to standardize inputs, also show promise in multilingual hate speech detection (Khan and Phillips, 2021).

In-context learning (ICL) has been explored for adapting LLMs without full retraining, though it incurs higher inference costs as examples scale (Brown et al., 2020). In contrast, PEFT methods offer scalable adaptation (Liu et al., 2022a; Patwa et al., 2024), supporting efficient hate speech detection across languages with fewer resources. In our work, we explore LoRA (Hu et al., 2021) for hate speech detection in Devanagari languages Hindi and Nepali.

3 Data

We use the dataset released as a part of the shared task (Thapa et al., 2025) in the CHiPSAL workshop (Sarveswaran et al., 2025). It contains two tasks - hate speech detection and hate speech target identification in two Devanagari scripted languages: Hindi (Jafri et al., 2024, 2023) and Nepali (Thapa et al., 2023; Rauniyar et al., 2023).

3.1 Hate Speech Detection

For hate speech detection, the data consists of devanagari-scripted text annotated into 2 classes - hate speech and not hate speech. The texts are diverse and collected from various sources including social media posts, news articles, and forums, reflecting a wide range of topics and styles. Table 1 shows the data distribution. We can see that there is a significant class imbalance towards the non-hate class. This imbalance poses a challenge for training the models, as they may tend to favor the majority class.

Class	Train	Valid	Test
Not Hate	16805	3602	3601
Hate	2214	474	475
Total	**19019**	**4076**	**4076**

Table 1: Data distribution of the hate speech detection dataset.

3.2 Hate Speech Target Identification

The second subtask focuses on identifying the targets of hate speech in Devanagari-scripted text. The goal is to classify whether hate speech is directed towards an individual, an organization, or a community. The dataset for this task contains text samples annotated with target labels. The distribution of targets, as indicated in Table 2, shows a more balanced representation for individual and organizational targets, with approximately equal numbers of samples for both classes. However, there is a notable scarcity of samples where the target is a community, resulting in a skewed distribution towards individual and organizational targets. This data limitation introduces a potential challenge in accurately predicting hate speech directed at communities.

Class	Train	Valid	Test
Individual	1074	230	230
Organizational	856	183	184
Community	284	61	61
Total	**2214**	**474**	**475**

Table 2: Data distribution of the hate speech target identification dataset.

4 Methodology

LLMs leverage the transformer (Vaswani et al., 2023) architecture to model linguistic patterns across vast corpora, utilizing multi-head self-attention mechanisms to capture both local and global dependencies in text. LLMs have billions of parameters and are pretrained on extensive general-purpose corpora. As a result they demonstrate great zero shot capabilities on many natural language tasks (Kojima et al., 2022). However, they struggle on low resource languages (Cassano et al., 2024).

ICL is a way to improve performance of LLMs. It refers to providing few labeled examples in the prompt to guide the LLM. However, as the number of examples increase, the cost and lantency of inference increases (Liu et al., 2022b).

Fully fine tuning (FFT) an LLM with billions of parameters is infeasible because of the costs and computational resources needed (Xu et al., 2023).

Parameter Efficient Fine Tuning (PEFT) is a method in which we only finetune a small number of parameters as compared to the size of the LLM. It is more effective than ICL while being more efficient than FFT (Xu et al., 2023).

For our system we use a PEFT method called Low Rank Adaptation (LoRA) (Hu et al., 2021). LoRA reduces the number of trainable parameters by decomposing weight updates into low-rank matrices, which are inserted into the model's attention layers. Specifically, for a weight matrix W, LoRA approximates the update as:

$$W' = W + \Delta W = W + AB^T \quad (1)$$

where A and B are low-rank matrices. By freezing the core parameters of the pretrained model and only updating the low-rank matrices during training, LoRA significantly decreases computational and memory requirement for training while being as effective as FFT (Hu et al., 2021). Furthermore, LoRA does not add to the inference latency, as after training, the weight update AB^T is added to the model weights, hence the total number of model weights remains the same.

5 Experiments

We conduct experiments on 4 different LLMs to address challenges in processing Devanagari-scripted languages. The considered models include the LLama-3.1-8B (Dubey et al., 2024), Nemo-Instruct-2407 (AI and NVIDIA, 2023), Qwen2.5-7B-Instruct (Yang et al., 2024), and Phi3-medium-4k-Instruct (Abdin et al., 2024). Each model is fine-tuned using task-specific datasets. Quantization of the models to 4-bit precision was employed to reduce memory consumption and to speed up training and inference. All fine-tuning models used LoRA with $rank = 16$, $alpha = 16$ and no dropout.

All fine-tuning experiments are performed using a 16GB NVIDIA T4 GPU. For the hate speech detection task, all models were fine-tuned for 2 epochs. For the target identification task, models were fine-tuned for 4 epochs in order to accommodate a relatively small training set. The code is implemented using the Unsloth (Daniel Han and team, 2023) library, which helps accelerate training. Our code is available at https://github.com/Rushendra10/Hate-Speech-Detection-and-Target-Identification-in-Devanagari-Languages.

6 Results and Analysis

The test performance of the models for the hate speech detection and target identification tasks are provided in Tables 3 and 4 respectively. We can see that for both the tasks Nemo has the best performance (F1 scores 90.05% and 71.47% respectively). Notably, Nemo performs better than Llama despite having smaller size. Furthermore, we can see that the overall performance is better on hate speech detection as compared to target identification. This is because the latter task has 3 classes whereas the former task has only 2 classes.

6.1 Class-wise Analysis

Table 5 shows the class-wise results of Nemo for hate speech detection task. The F1 score on the hate class is much lower than on the non-hate class. The Confusion Matrix (Figure 1) shows that the instances of hate class are often mis-predicted as Non Hate. These observations can be attributed to the class imbalance in the training dataset.

Model	Size	Acc.	F1
Llama-3.1	8.03B	88.71%	88.02%
Phi-3-medium	7.36B	90.06%	88.91%
Qwen2.5	4.46B	88.62%	87.90%
Nemo	6.97B	**90.75%**	**90.05%**

Table 3: Performance of various models for hate speech detection task on the test set, along with the quantized model size. Acc. refers to accuracy. F1 refers to weighted average F1 score.

Model	Size	Acc.	F1
Lama-3.1	8.03B	67.37%	66.58%
Phi-3-medium	7.36B	68.21%	67.80%
Qwen2.5	4.46B	70.32%	70.41%
Nemo	6.97B	**72.00%**	**71.47%**

Table 4: Performance of various models for target identification task on the test set along with the quantized model size. Acc. refers to accuracy. F1 refers to weighted average F1 score.

Table 6 shows the class-wise results of Nemo for hate target identification task. The F1 on Individual class is comparable to that in Organization class, whereas it is significantly lower for the Community class. From the Confusion Matrix (Figure 2), we can see that instances of hate directed towards community are frequenty mis-predicted into one of the other 2 classes. Similar to the hate speech detection task, these observation are also a result of the imbalanced training dataset.

	P	R	F1
Non Hate	93.10%	96.70%	94.86%
Hate	64.58%	45.68%	53.51%

Table 5: Class-wise performance of Nemo on test set of the hate speech detection task. P = Precision, R= Recall, F1 = F1 score.

	P	R	F1
Individual	76.57%	79.56%	78.04%
Organization	72.49%	74.46%	73.46%
Community	46.81%	36.07%	40.74%

Table 6: Class-wise performance of Nemo on test set of the target identification task. P = Precision, R= Recall, F1 = F1 score.

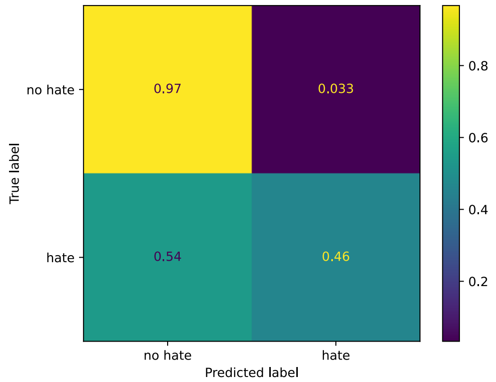

Figure 1: Confusion matrix of Nemo on the test set for hate speech detection.

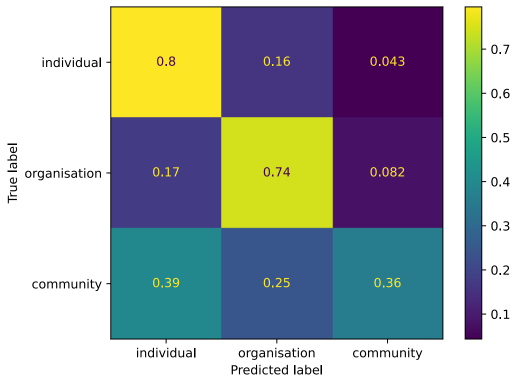

Figure 2: Confusion matrix of Nemo on the test set for hate speech target identification.

7 Conclusion and Future Work

In this study, we present our approach for hate speech detection in Devanagari-scripted languages using LLMs fine-tuned with LoRA. Our methodology demonstrates good performance, as evidenced by accuracy and F1 score metrics. By leveraging the CHiPSal dataset, we effectively address the challenges posed by low-resource languages. We notice that the performance is lower on the the classes with fewer data instances.

Future research could involved enhancing the model's capabilities by developing data generation techniques to address class imbalance, ensuring robust performance across all classes. Additionally, investigating the integration of more sophisticated techniques, such as ensemble methods, can further boost detection accuracy and robustness.

8 Limitation

We assume the existence of a decently sized train dataset to fine-tune our model. Further, we assume that the LLMs will have some knowledge of devanagari languages for PEFT to work.

9 Ethical Statement

Hate speech detection is a sensitive topic and can be subjective. LLMs are known to have inherent biases. Any censoring decisions based on the LLMs predictions should involve comprehensive human reviews.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mistral AI and NVIDIA. 2023. Mistral nemo-instruct-2407: A 12b parameter transformer model. `https://mistral.ai/news/mistral-nemo/`. Accessed: 2024-11-10.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Anders Freeman, Carolyn Jane Anderson, Molly Q Feldman, Michael Greenberg, Abhinav Jangda, and Arjun Guha. 2024. Knowledge transfer from high-resource to low-resource programming languages for code llms. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA2):677–708.

Sara Court and Micha Elsner. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem. *Preprint*, arXiv:2406.15625.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online only. Association for Computational Linguistics.

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*, pages 512–515.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Xiaoyu Guo, Jing Ma, and Arkaitz Zubiaga. 2023. Nuaa-qmul-aiit at memotion 3: Multi-modal fusion with squeeze-and-excitation for internet meme emotion analysis. *Preprint*, arXiv:2302.08326.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines.

Prashant Kapil, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and B. N. Vinutha. 2023. Hhsd: Hindi hate speech detection leveraging multi-task learning. *IEEE Access*, 11:101460–101473.

Heena Khan and Joshua L. Phillips. 2021. Language agnostic model: detecting islamophobic content on social media. In *Proceedings of the 2021 ACM Southeast Conference*, ACMSE '21, page 229–233, New York, NY, USA. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Zachary Laub. 2019. Hate speech on social media: Global comparisons. *Council on foreign relations*, 7.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Preprint*, arXiv:2101.00190.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Preprint*, arXiv:2205.05638.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022b. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Shreyash Mishra, S Suryavardan, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023a. Overview of memotion 3: Sentiment and emotion analysis of codemixed hinglish memes. *arXiv preprint arXiv:2309.06517*.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. 2023b. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas Pykl, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021a. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 42–53. Springer.

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, and Shervin Malmasi. 2024. Enhancing low-resource llms classification with peft and synthetic data. *Preprint*, arXiv:2404.02422.

Parth Patwa, Srinivas Pykl, Amitava Das, Prerana Mukherjee, and Viswanath Pulabaigari. 2021b. Hater-o-genius aggression classification using capsule networks. *arXiv preprint arXiv:2105.11219*.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Kengatharaiyer Sarveswaran, Bal Krishna Bal, Surendrabikram Thapa, Ashwini Vaidya, and Sana Shams. 2025. A brief overview of the first workshop on challenges in processing south asian languages (chipsal). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *Preprint*, arXiv:2401.13136.

Shubham Shukla, Sushama Nagpal, and Sangeeta Sabharwal. 2022. Hate speech detection in hindi language using bert and convolution neural network. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 642–647.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Debapriya Tula, Prathyush Potluri, Shreyas Ms, Sumanth Doddapaneni, Pranjal Sahu, Rohan Sukumaran, and Parth Patwa. 2021. Bitions@ dravidianlangtech-eacl2021: Ensemble of multilingual language models with pseudo labeling for offence detection in dravidian languages. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 291–299.

Debapriya Tula, MS Shreyas, Viswanatha Reddy, Pranjal Sahu, Sumanth Doddapaneni, Prathyush Potluri, Rohan Sukumaran, and Parth Patwa. 2022. Offence detection in dravidian languages using code-mixing index-based focal loss. *SN Computer Science*, 3(5):330.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. *Preprint*, arXiv:2110.12200.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.