CLTL@HarmPot-ID: Leveraging Transformer Models for Detecting Offline Harm Potential and Its Targets in Low-Resource Languages

Yeshan Wang, Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
y.wang11@student.vu.nl, i.markov@vu.nl

Abstract

We present the winning approach to the TRAC 2024 Shared Task on Offline Harm Potential Identification (HarmPot-ID). The task focused on low-resource Indian languages and consisted of two sub-tasks: 1a) predicting the offline harm potential and 1b) detecting the most likely target(s) of the offline harm. We explored low-source domain specific, cross-lingual, and monolingual transformer models and submitted the aggregate predictions from the MuRIL and BERT models. Our approach achieved 0.74 micro-averaged F1-score for sub-task 1a and 0.96 for sub-task 1b, securing the 1st rank for both sub-tasks in the competition.

1. Introduction

In the age of digital interconnectedness, social media platforms like Facebook, Instagram, and Twitter have become key places for billions of users worldwide to connect, share insights and perspectives easily and quickly. It has greatly enhanced communication between different cultures and helped online communities to grow. However, it has also led to the proliferation of content that contains violent language, potentially inciting real-world harm (Olteanu et al., 2018). This type of content, ranging from overt expressions of aggression to subtler forms of hate speech, not only violates platform community standards but poses a significant risk of leading to real-world violence (Millar, 2019). Recognizing the gravity of this issue, governments, research community, and social media companies are increasingly working on ways to limit the spread of such violence-inciting content.

However, the effort to detect and withstand online violence has mostly focused on widely spoken languages such as English, leaving behind many low-resource languages spoken in diverse countries like India, such as Meitei, Hindi, and Bangla, each with its own complex features and regional differences. This complexity makes it hard to identify violent content, a problem exacerbated by the lack of resources and limited research dedicated to these languages.

The TRAC 2024 Shared Task[1] introduced the task of predicting the offline harm potential of social media posts: whether a specific post is likely to initiate, incite or further exaggerate an offline harm event, as well as detecting the most affected target categories if an offline harm event was triggered.

The task focused on three low-resource Indian languages – Bangla, Hindi, Meitei - and for each of these languages the data was code-mixed with English or different varieties of English. The task consisted of two sub-tasks. Sub-task 1a focused on predicting the offline harm potential of social media posts, where the participants were required to predict the level of offline harm potential as a four-way multi-class classification task:

- 0: it will never lead to offline harm, in any context
- 1: it could lead to incite an offline harm event given specific conditions or context
- 2: it is most likely to incite in most contexts or probably initiate an offline harm event in specific contexts
- 3: it is certainly going to incite or initiate an offline harm event in any context

Sub-task 1b consisted in identifying the most likely target(s) of offline harm if an offline harm event was triggered, as a multi-label classification problem with the following five target categories:

- Gender
- Religion
- Descent
- Caste
- Political Ideology

While there have been numerous shared tasks on identifying different types of harmful content, including hate speech (Mandl et al., 2019), offensive language (Zampieri et al., 2019), and aggression (Kumar et al., 2018), amongst others, few have focused on predicting the offline harm potential of social media posts, especially in the context of low-resource languages. To the best of our knowledge,

[1] https://codalab.lisn.upsaclay.fr/competitions/17646

TRAC-2024 Workshop, pages 21–26

the most similar shared task related to this topic was the Shared Task of Violence Inciting Text Detection (Saha et al., 2023), which focused on the Bengali language.

From the machine learning perspective, various approaches have been explored to detect harmful content online and its targets, including lexicon-based approaches (Schouten et al., 2023), conventional machine learning approaches (Waseem and Hovy, 2016; Wiegand et al., 2018; Markov and Daelemans, 2021; Lemmens et al., 2021), neural networks (van Aken et al., 2018), and transformer-based pre-trained language models (Risch and Krestel, 2020; Markov and Daelemans, 2022; Ghosh and Senapati, 2022), with the latter usually outperforming the other strategies for detecting harmful content in social media posts (Zampieri et al., 2019, 2020). Therefore, we focus on exploring various transformer-based language models to tackle the tasks at hand.

2. Data

The dataset used in the TRAC 2024 Shared Task is composed of social media texts collected from different social media platforms such as YouTube, Twitter, and Telegram. It was manually annotated by multiple annotators for the level of offline harm potential (sub-task 1a) and the likely target(s) of offline harm (sub-task 1b) (Kumar et al., 2024). The data covers three Indian languages: Meitei, Bangla (Indian variety), and Hindi, where each of the languages is code-mixed with English or English varieties (i.e., English used in the context of these languages).

The dataset statistics in terms of the number of instances per class, as well as the class distribution is provided in Tables 1 and 2 for sub-tasks 1a and 1b, respectively.

Label	Train		Dev	
	# posts	%	# posts	%
0	16,135	31.77	2,017	31.77
1	21,554	42.44	2,695	42.44
2	12,211	24.04	1,526	24.04
3	888	1.75	111	1.75
Total	**50,788**	**100**	**6,349**	**100**

Table 1: Sub-task 1a: statistics of the dataset in terms of the number of posts and their distribution per class.

It can be observed that the dataset is highly imbalanced in terms of represented classes, with the majority class constituting more than 42% of the entire dataset for sub-task 1a and more than 55% for sub-task 1b.

Label	Train		Dev	
	# posts	%	# posts	%
Gender	9,599	56.80	1,180	55.90
Religion	4,876	28.85	645	30.55
Descent	1,456	8.62	180	8.53
Caste	561	3.32	58	2.75
Political Ideology	407	2.41	48	2.27
Total	**16,899**	**100**	**2,111**	**100**

Table 2: Sub-task 1b: statistics of the dataset in terms of the number of posts and their distribution per class.

3. Methodology

3.1. Preprocessing steps

In the text preprocessing phase, we used a python module for text normalization (Hasan et al., 2020). It is intended to be used for normalizing / cleaning Bengali and English texts. Considering certain similarity of Bengali to the other Indian languages covered in this shared task, we used this module to perform text preprocessing. We conducted an ablation study of two commonly used text preprocessing strategies when dealing with social media texts (converting emojis to text and removing URLs) using the BERT-base model[2], observing the effectiveness of these two steps when used in combination (see Table 3).

Converting emojis to text	Removing URLs from texts	Micro-F1
✓	✓	70.66%
✓	×	70.56%
×	×	70.26%
×	✓	70.23%

Table 3: Ablation study of the text preprocessing strategies on sub-task 1a.

3.2. Transformer models

After determining the usefulness of the examined preprocessing steps, we conducted a comparative experiment using the currently publicly available transformer-based language models, which we fine-tuned on the shared task training data and evaluated on the development set. Specifically, we examined the following categories of language models:

1. **Low-source domain specific language model**: Low-source language models are pre-trained on extensive datasets comprising one or more low-resource languages. We used

[2] https://huggingface.co/google-bert/bert-base-uncased

the MuRIL model[3], which is based on a BERT large architecture with 24 layers, pre-trained on 17 Indian languages and their transliterated counterparts (Khanuja et al., 2021).

2. **Cross-lingual language models**: These models leverage large multilingual datasets for pre-training, supporting over 100 languages for cross-lingual classification tasks. Our experimentation included XLM-RoBERTa-Large[4] and its two derivatives: XLM-T[5] and Multilingual E5[6]. XLM-RoBERTa-Large was introduced by Facebook AI in 2019, which is a multilingual adaptation of RoBERTa (Liu et al., 2019) pre-trained on 2.5TB of CommonCrawl data spanning 100 languages (Conneau et al., 2020). XLM-T, built upon XLM-RoBERTa-Large framework, was re-trained on more than 1 billion tweets in diverse languages up to December 2022 (Barbieri et al., 2022). Multilingual E5, released by Microsoft in 2023, is the newest derivative of XLM-RoBERTa-Large, incorporating additional training on a variety of multilingual datasets to enhance its versatility across languages and tasks (Wang et al., 2024).

3. **Monolingual language model**: Monolingual models are pre-trained on vast datasets specific to a single language, facilitating extension and customization for domain-specific tasks. We explored the capabilities of BERT-Large[7], a transformer model pre-trained on a comprehensive corpus of English data through self-supervised learning methods (Devlin et al., 2019).

3.3. Experimental settings

We used the PyTorch framework (Paszke et al., 2019) and AutoGluon library (Shi et al., 2021) for models' implementation. We fine-tuned the transformer models on the training data provided by the organizers, without using any additional data for training. The models were fine-tuned with the following hyperparameters: a base learning rate of 1e-4, decay rate of 0.9 using cosine decay scheduling, batch size of 8, and a manual seed of 0 for reproducibility. The models were optimized using the AdamW optimizer for up to 4 epochs or until an early stopping criterion was met to prevent overfitting. All experiments were conducted on the Google Colaboratory platform with an NVIDIA A100 GPU.

[3] https://huggingface.co/google/muril-large-cased

[4] https://huggingface.co/FacebookAI/xlm-roberta-large

[5] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-large-2022

[6] https://huggingface.co/intfloat/multilingual-e5-large

[7] https://huggingface.co/google-bert/bert-large-uncased

4. Results

We present the results obtained on the development set in terms of the official evaluation metric: micro-averaged F1 score. The results for sub-task 1a are provided in Table 4.

Set	Language model	micro-F1
Dev	MuRIL	**73.89%**
	Multilingual E5	73.21%
	XLM-T	73.04%
	XLM-RoBERTa-Large	72.50%
	BERT-Large	72.00%
Test	MuRIL	0.74

Table 4: Results for sub-task 1a on the development and test sets.

As one can see, the MuRIL model outperformed the other examined models by a small margin in terms of micro-F1 score. The confusion matrix for the best-performing MuRIL model on the development set is shown in Figure 1.[8]

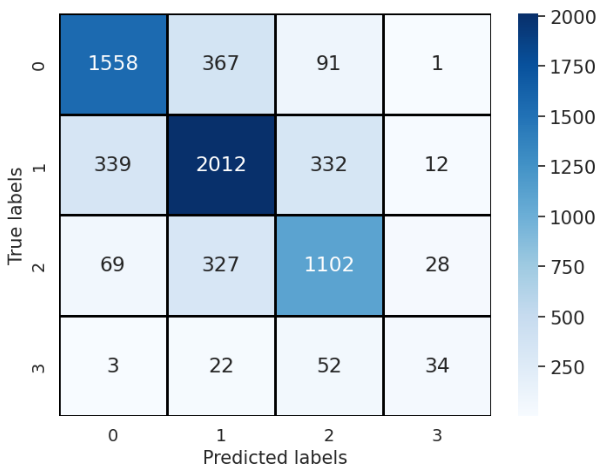

Figure 1: Confusion matrix for the MuRIL model on the development set.

As expected, we observe a high degree of confusion between the categories with less pronounced differences, i.e., 0 and 1, 1 and 2, 2 and 3.

We submitted the final predictions obtained with the MuRIL model for the official evaluation on the test set, achieving 74% micro-F1 score, as shown in Table 4.

[8] At the time of writing, the test labels were not made available by the organizers.

Set	Model	Overall micro-F1	Gender	Religion	Descent	Caste Bias	Political Ideology
Dev	MuRIL	**96.42%**	**90.41%**	**94.99%**	**97.86%**	**99.35%**	99.48%
	XLM-T	96.31%	90.25%	94.96%	97.61%	99.20%	99.53%
	Multilingual E5	96.24%	89.90%	94.79%	97.76%	99.21%	99.53%
	XLM-RoBERTa-Large	96.13%	89.84%	94.76%	97.70%	99.09%	99.24%
	BERT-Large	95.97%	89.13%	94.22%	97.72%	99.23%	**99.57%**
Test	MuRIL & BERT-Large	0.96	0.90	0.95	0.98	0.99	0.99

Table 5: Results for sub-task 1b on the development and test sets.

For sub-task 1b, we convert the multi-label classification task into five binary classification tasks, with each focusing on predicting the target of the offline harm (Gender, Religion, Descent, Caste, and Political Ideology). The results obtained by each model for sub-task 1b on the development set are provided in Table 5.

We observe a similar performance of the examined models within each target category covered in sub-task 1b. Surprisingly, the monolingual model: BERT-Large achieved similar results to the low-source domain specific and cross-lingual models, slightly outperforming the other models for the Political Ideology class. Furthermore, we observe overall high performance for this task and that Gender is the most difficult target category to predict, with the results on average 7.5% lower than for the other categories.

For the final evaluation, we submitted the aggregate predictions of the best-performing models for each target category based on the evaluation results on the development set, which contained predictions from the MURIL model for the first four targets (Gender, Religion, Descent, Caste) and predictions from the BERT model for the last target category (Political Ideology). The official results on the test set are provided in Table 5.

5. Conclusion

We presented the description of the CLTL approach to the TRAC 2024 Shared Task on Offline Harm Potential Identification. We explored low-source domain specific, cross-lingual, and monolingual transformer models: MuRIL, Multilingual E5, XLM-T, XLM-RoBERTa-Large, and BERT-Large. It was found during the preliminary experiments on the training and development sets that the low-source domain specific MuRIL model slightly outperforms the other examined transformer models for detecting the offline harm potential. For identifying the likely target(s) of offline harm, the examined models achieved similar results, with the MuRIL model outperforming the other models by a small margin in the vast majority of cases, while BERT-large performed best for predicting the Political Ideology target category. On the test set, our team achieved 0.74 micro-averaged F1-score for sub-task 1a and 0.96 for sub-task 1b, ranking 1st in both sub-tasks in the competition.

6. Bibliographical References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Koyel Ghosh and Dr. Apurbalal Senapati. 2022. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 853–865, Manila, Philippines. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. MuRIL: Multilingual representations for indian languages. *arXiv/2103.10730*.

Ritesh Kumar, Ojaswee Bhalla, Shehlat Maknoon Vanthi, Madhu Wani, and Siddharth Singh. 2024. Harmpot: An annotation framework for evaluating offline harm potential of social media text. In *Proceedings of the the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018. TRAC-1 shared task on aggression identification: IIT(ISM)@COLING'18. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 58–65, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv/1907.11692*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, page 14–17, New York, NY, USA. ACM.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

Ilia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Sharon Millar. 2019. Hate speech: Conceptualisations, interpretations and reactions. In *The Routledge handbook of language in conflict*, pages 145–162. Routledge.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 221–230. AAAI Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Julian Risch and Ralf Krestel. 2020. Bagging BERT models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61, Marseille, France. European Language Resources Association (ELRA).

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. BLP-2023 task 1: Violence inciting text detection (VITD). In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 255–265, Singapore. Association for Computational Linguistics.

Stefan F. Schouten, Baran Barbarestani, Wondimagegnhue Tufa, Piek Vossen, and Ilia Markov. 2023. Cross-domain toxic spans detection. In

Natural Language Processing and Information Systems: 28th International Conference on Applications of Natural Language to Information Systems, NLDB 2023, Derby, UK, June 21–23, 2023, Proceedings, page 533–545, Berlin, Heidelberg. Springer-Verlag.

Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. 2021. Multimodal AutoML on structured tables with text fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv/2402.05672*.

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, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.