Benchmarking Language Models for Offensive Sentences Classification in Offensive Nepali Roman Multi-Label Dataset

Anonymous ACL submission

Abstract

This paper presents a comprehensive methodology for benchmarking and evaluating multiple language models to detect offensive language in Romanized Nepali text. Recognizing Nepali as a low-resource language, we introduce the Offensive Nepali Roman Multi-label Dataset (ONRMD), labeled for abuse, scam, sexual, and neutral content,specifically designed for this study. We employ various models, including BERT-base-multilingual-cased, RoBERTa-base, distilbert-base-nepali, FastText, and LASER + CNN, and compare their performance on the ONRMD. Our approach encompasses thorough preprocessing and tokenization of the dataset, followed by training and evaluation using standard metrics such as accuracy, precision, recall, and F1 score. Additionally, we conduct human evaluations with two distinct groups to further validate our findings, given the novelty of our dataset and the absence of a standard baseline. The results demonstrate the potential of these models in effectively handling the nuances of Romanized Nepali text for offensive language detection. This study serves as a foundation for future research involving other pre-trained language models and multilingual datasets.

1 Introduction

The proliferation of offensive language on social media platforms poses significant challenges, especially in low-resource languages like Nepali.(Magueresse et al., 2020) Offensive content, including abuse (Lahti et al., 2024), scams (Coluccia et al., 2020), and sexual messages (Alaggia and Wang, 2020), threatens the safety and well-being of users, necessitating effective detection mechanisms. While significant advancements have been made in offensive language detection for high-resource languages (Caselli et al., 2020) (Razavi et al., 2010). There were some good research done in some low resource South Asian language like Dravidian (Roy et al., 2022), Urdu (Akhter et al., 2020) and Also on Hindi (Mathur et al., 2018), with Devnagari Script of Hindi (Jha et al., 2020) which is same as Nepali.

Nepali, particularly in its Romanized form, remains underexplored. To address this gap, we introduce the Offensive Nepali Roman Multi-label Dataset (ONRMD), which is specifically designed to detect abuse, scam, sexual, and neutral content in Romanized Nepali text.

Previous research on natural language processing (NLP) for Nepali has been limited, particularly concerning the challenges posed by Romanized text. Romanization introduces variations and inconsistencies in spelling and syntax, complicating offensive language detection. (Singh et al., 2020) focused on aspect-based abusive sentiment detection in Nepali social media text, extracting comments from YouTube videos and benchmarking with classic and deep learning methods. Existing studies primarily focus on traditional Nepali scripts, leaving a significant gap in resources and methodologies for Romanized Nepali. (Shrestha and Bal, 2020) annotated an equal number of positive and negative sentences for sentiment analysis, addressing class imbalance. (Niraula et al., 2021)collected and annotated 7,462 comments for sentiment analysis, finding that Multilingual BERT (M-BERT) performed inadequately compared to traditional ML models due to the limited size of available content for low-resource languages like Nepali.Thus, we also mostly used transformer based models in our experiment.

In this study, we evaluate to benchmark various language models—BERT-base-multilingual-cased, RoBERTa-base, distilbert-base-nepali, FastText, and LASER + CNN—to identify offensive content. Using comprehensive preprocessing, tokenization, and training steps, we evaluate the models with metrics such as accuracy, precision, recall, and F1 score. Human evaluations with two groups validate our findings, given the novelty of

our dataset and lack of a standard baseline. The introduction of ONRMD and the evaluation of multiple models on this dataset establish a foundation for future research, contribute to robust online content moderation systems, and support efforts to control harmful language on social media. Our findings demonstrate the potential of these models to handle Romanized Nepali text effectively, highlighting the importance of developing specialized resources and methodologies for low-resource languages and paving the way for more inclusive NLP research.

2 Offensive Nepali Roman Multi-label Dataset (ONRMD)

We collected data from different social media platforms using specific keywords and searched for relevant tweets and Facebook posts. The list of offensive terms used for this search can be found at [1]. Additionally, we conducted a survey circulating Google Forms with volunteers to gather data on scam and sexual content, which were not readily available on public platforms.

2.1 Challenges in Dataset Collection

Language Filtering: Nepal has more than 125 ethnic groups and 123 spoken languages.[2] On public platforms, people often mix their local language with standard Nepali. Since we aimed to create a dataset solely in Nepali, we had to exclude scarcely used sentences that contained mixtures of Nepali and other local languages, such as *Nepali + Newari* or *Nepali + Bhojpuri.*

Unclear Sentence Labels: We encountered sentences that did not fit any of our labels. Some sentences contained slangs that could be interpreted differently based on context. Also, some romanized sentences' literal meaning would be neutral however some words would carry sarcastic offensive meaning.We had to clean such sentences to maintain the integrity of our dataset.

2.2 Dataset Composition

After scraping through social media and collection through survey , we got datasets on both Devanagari and Nepali Roman , we manually translated the Devanagari Scripts into commonly used Nepali Roman form .

[1] https://github.com/nowalab/offensive-nepali/blob/master/offensive-terms-in-nepali.csv

[2] https://mofa.gov.np/nepal-profile-updated/

2.2.1 Devanagari:

1. **Translation:** We manually translated the data. Issues arose such as variations in transliteration, e.g., "timi kaha chau?" versus "timi kaha xau?". We ensured that both variations were included by dividing the task such that one would write the same letter as "chha" and another as "xa". This approach was applied to other words as well.

2. **Dirgikaran:** This approach was first implemented on (Niraula et al., 2021), which normalizes words with different orthographic forms to a standard form, e.g., "pipal" and "peepal" were standardized to "peepal", while translating to Nepali Roman dataset.

2.2.2 Nepali Roman:

We translated Devanagari script sentences to Nepali Roman and performed preprocessing on existing Roman sentences:

1. **Normalization of Tense and Honorifics:** In Nepali, the same verb "to eat" can be expressed in different tenses and honorifics, such as "khanchu" for present tense, "khako thyo" for past tense, and "khane chu" for future tense. Additionally, verbs change based on honorifics, like "khalas" for low honor and "khaibaksinchha" for high honor when addressing elders. We normalized all sentences to the present tense and low honor form.

The final dataset sizes are shown in Table 1:

Label	Count
Abuse	1186
Neutral	2000
Scam	2000
Sexual	1000

Table 1: Dataset Composition

2.3 Dataset Labeling

We used fine-grained and coarse-grained labels. Initially, we labeled the dataset as neutral and offensive. For performance benchmarking, we further labeled them as scam, neutral, abuse, and sexual.

3 Experiment

3.1 Data Preprocessing

After our Offensive Nepali Roman Multi-Label Dataset (ONRMD) got finalized, we performed

label analysis, calculated sentence lengths, determined a maximum token length (95th percentile, 52 tokens), tokenized the data with truncation and padding, and split the dataset into 70% training, 15% validation, and 15% evaluation. The data was encoded using the tokenizer with truncation and padding to the maximum length.

3.2 Experimental Setup and Hyperparameters

For Models 1, 2, and 3, we utilized similar training code with adjustments for each model's requirements. We added a classifier layer on top of the pre-trained models, configured 500 warmup steps to stabilize initial training, and empirically determined other hyperparameters. Training parameters included 25 epochs, batch sizes of 16, and a weight decay of 0.01.

Models 4 and 5 involved different architectures but followed a consistent approach in data preprocessing and training setup. Models 6 and 7 involved human evaluation, which is discussed in more detail in the upcoming sections.

3.3 Computation and Resource

We used Google Colab for the experiments, GPU used was T4GPU with each model taking nearly 30 minutes to train.

3.4 Models Used in the Experiment

Model 1: We utilized the *bert-base-multilingual-cased* model from Hugging Face's Transformers library [3]. The tokenizer used for this model was *BertTokenizer*, and the model architecture was *BertForSequenceClassification* with the number of labels set to the length of the label dictionary. The tokenization process involved truncating and padding the input sequences to a maximum length determined by the 95th percentile of sentence lengths. Labels were converted to tensors, and custom dataset classes were defined to handle the data.

Model 2: The *roberta-base* model was employed, sourced from Hugging Face [4]. For tokenization, *RobertaTokenizer* was used, and *RobertaForSequenceClassification* served as the model architecture, with the number of labels set to the length of the label dictionary. The training setup was similar to that of Model 1, with the same preprocessing steps, dataset creation, and training arguments.

Model 3: Additionally, we used *distilbert-base-nepali* [5], another model from Hugging Face. The tokenizer and model used were *AutoTokenizer* and *AutoModelForSequenceClassification*, respectively, with the number of labels configured similarly.

Model 4: The *FastText* model was trained using supervised learning on the prepared dataset. The data was saved in FastText format and the model was trained with hyperparameter control settings: 25 epochs, learning rate of 0.1, and word n-grams of 2. The team (Modha et al., 2018) demonstrated the effectiveness of combining FastText embeddings with CNN, outperforming 18 other approaches, in a task organized by (Kumar et al., 2018). Given the similarity between Devanagari and Roman scripts for Nepali and Hindi, we adopted this approach as well for comparison.

Model 5: For this model, we used *LASER embeddings* followed by a *CNN classifier*, inspired by (Aluru et al., 2021). The embeddings were converted to PyTorch tensors and used as input to a simple CNN model. The CNN architecture consisted of a convolutional layer, max-pooling layer, and a fully connected layer. The model was optimized using the Adam optimizer with a learning rate of 0.001.

Model 6 & 7: In addition to machine learning models, we conducted human evaluation for Models 6 and 7 using two groups: Group A and Group B, randomized using a random number generator to ensure variation in gender, age group, and demographics. We summarized our work to our volunteers, some of whom contributed data. Proper consent was obtained, and they volunteered for the project. Personal identification was not collected. This step was crucial because our novel dataset lacks a standard baseline. Human evaluation provided an additional layer of validation, following the recommendations by (Schuff et al., 2023) on the importance of user studies in NLP and the insights by (Nguyen, 2018) on comparing automatic and human evaluation methods.

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

[4] https://huggingface.co/docs/transformers/en/model_doc/roberta

[5] https://huggingface.co/Sakonii/distilbert-base-nepali

	Model 1			Model 2			Model 3		
	precision	**recall**	**f1-score**	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
abuse	0.90	0.93	0.92	0.92	0.89	0.91	0.81	0.81	0.81
neutral	0.98	0.97	0.97	0.98	0.98	0.98	0.94	0.95	0.94
scam	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99
sexual	0.87	0.87	0.87	0.86	0.89	0.87	0.72	0.71	0.71
accuracy	0.96			0.96			0.90		
macro avg	0.94	0.94	0.94	0.94	0.94	0.94	0.86	0.86	0.86
weighted avg	0.96	0.96	0.96	0.96	0.96	0.96	0.90	0.90	0.90

Table 2: Performance Metrics for BERT-based Models

	Model 4 (FastText)			Model 5 (LASER)			Model 6 (A)	Model 7 (B)
	precision	**recall**	**f1-score**	**precision**	**recall**	**f1-score**	**accuracy**	**accuracy**
abuse	0.89	0.15	0.25	0.82	0.80	0.81	0.93	0.91
neutral	0.70	0.23	0.35	0.92	0.95	0.94	1.00	0.99
scam	0.99	0.56	0.72	1.00	1.00	1.00	1.00	1.00
sexual	0.21	0.97	0.34	0.75	0.73	0.74	0.95	0.91
accuracy		0.44			0.91		0.97	0.95
macro avg	0.70	0.48	0.41	0.87	0.87	0.87		
weighted avg	0.77	0.44	0.46	0.90	0.91	0.91		

Table 3: Performance Metrics for FastText, LASER Models and Human Groups

4 Results

The results highlight the effectiveness of various language models in detecting offensive content in Romanized Nepali text using the ONRMD dataset.

4.1 Performance of BERT-based Models

The BERT-based models demonstrated strong performance (Table 2). Model 1 (BERT-base-multilingual-cased) achieved 96% accuracy, with high precision, recall, and F1-scores across all categories, excelling in scam detection with a perfect F1-score of 1.00. Model 2 (RoBERTa-base) showed similar performance, with 96% accuracy and slightly lower performance in the abuse category (F1-score of 0.91). Model 3 (distilbert-base-nepali) had lower overall performance, particularly in sexual content detection (F1-score of 0.71) and 90% accuracy.

4.2 Performance of FastText and LASER Models

Model 4 (FastText) had significantly lower performance with 44% accuracy and struggled with recall, particularly in abuse and neutral categories (F1-scores of 0.25 and 0.35, respectively). Model 5 (LASER + CNN) showed improvement with 91% accuracy and high precision, recall, and F1-scores in scam and neutral content detection. The performance in sexual content was moderate (F1-score of 0.74).

4.3 Human Evaluation

Human evaluations with two groups (Models 6 and 7) further validated our findings. Group A achieved 97% accuracy, and Group B 95%, particularly excelling in neutral and scam content detection. Slightly lower performance in abuse and sexual categories reflected the complexities of offensive language detection.

5 Conclusion and Future Work

Our findings highlighted the superior performance of BERT-based models in handling the nuances of Romanized Nepali text for offensive language detection, validating the importance of specialized resources and methodologies for low-resource languages. In the future, we aim to include other regional Nepali sentences and expand the labels in our dataset. We also plan to benchmark additional pre-trained language models and techniques like ensemble learning to further improve detection accuracy. Additionally, we will investigate the impact of different transliteration schemes on model performance and develop methods to handle mixed-language content more effectively. By addressing these areas, we hope to contribute to more inclusive and effective natural language processing research for low-resource languages.

6 Limitation

We present a multi-label dataset in this study for the categorization of offensive statements in romanized Nepali. However, our work must acknowledge a number of constraints. First of all,for scam and sexual labels, rather than directly scraping tweets or posts made by the offenders themselves, we had to rely on volunteers who anonymously provided us with texts and remarks they've received or what they believe to be commonly sent out scams and sexually abusive sentences. This is because those sentences are typically received via texts or private messages and aren't usually posted publicly. Second, we came across a few statements that didn't fit any of our labels but still contained profane and offensive language. Depending on the nature of the relationship between the parties involved, these sentences could be construed in several ways. What might be obvious abuse to a stranger could be seen as friendly banter between friends.Third, we also discovered a few sentences that combined local language with romanized Nepali, such as Nepali and Newari, Nepali and Maithili, etc. However, we decided not to include them in the dataset because they were far less common and scarcer than sentences written entirely in romanized Nepali or in a combination of English and romanized Nepali.

7 Ethical Consideration

We gave ethical issues first priority when performing our study and used anonymous data collection from volunteers. This strategy protected participants' privacy by making sure their identities could not be connected to the information they submitted. The replies provided by each volunteer were de-identified, which means that no personal information like their email address or such was linked to them and that there was no way to trace the data back to any specific person. In addition to safeguarding the participants' privacy, this anonymity protocol promoted a feeling of confidence that encouraged candid and unreserved comments. In order to optimize participation and guarantee diversity, the URL for anonymous data entry was shared among multiple social media groups and Discord servers. Our reach was increased and a wide range of viewpoints from a large audience were gathered thanks to this tactic.

This is our first submission, and we are considering open-sourcing the scripts and data utilities used in our study. However, we face several ethical and procedural concerns:

Anonymity Compliance: We are unsure if releasing the code and data during the review process aligns with anonymity guidelines. We will consult the conference policies on this matter.

Licensing: We need guidance on selecting an appropriate license for the datasets and scripts to ensure responsible use by the community.

Content Sensitivity: Our dataset includes offensive language, which raises concerns about potential misuse. We are considering implementing access controls or usage agreements to manage this risk.

Balancing open research with ethical responsibility is crucial. We welcome feedback on how to address these challenges while contributing positively to the field.

References

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. 2020. Automatic detection of offensive language for urdu and roman urdu. *IEEE Access*, 8:91213–91226.

Ramona Alaggia and Susan Wang. 2020. "i never told anyone until the# metoo movement": What can we learn from sexual abuse and sexual assault disclosures made through social media? *Child abuse & neglect*, 103:104312.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, pages 423–439. Springer.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Anna Coluccia, Andrea Pozza, Fabio Ferretti, Fulvio Carabellese, Alessandra Masti, and Giacomo Gualtieri. 2020. Online romance scams: relational dynamics and psychological characteristics of the victims and scammers. a scoping review. *Clinical practice and epidemiology in mental health: CP & EMH*, 16:24.

Vikas Kumar Jha, Pa Hrudya, PN Vinu, Vishnu Vijayan, and Pa Prabaharan. 2020. Dhot-repository and classification of offensive tweets in the hindi language. *Procedia Computer Science*, 171:2324–2333.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

Henri Lahti, Marja Kokkonen, Lauri Hietajärvi, Nelli Lyyra, and Leena Paakkari. 2024. Social media threats and health among adolescents: evidence from the health behaviour in school-aged children study. *Child and Adolescent Psychiatry and Mental Health*, 18(1):62.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the sixth international workshop on natural language processing for social media*, pages 18–26.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 199–207.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1069–1078. Association for Computational Linguistics.

Nobal B Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive language detection in nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23*, pages 16–27. Springer.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Birat Bade Shrestha and Bal Krishna Bal. 2020. Named-entity based sentiment analysis of nepali news media texts. In *Proceedings of the 6th workshop on natural language processing techniques for educational applications*, pages 114–120.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308. IEEE.