NepX-Hate: A Nepali Hate Speech Corpus with Fine-Grained Sociocultural Annotations

Anonymous ACL submission

Abstract

Hate speech detection in low-resource languages remains a significant challenge due to the scarcity of annotated datasets. We introduce NepX-Hate, a new benchmark dataset for hate speech detection in low-resource languages, centered on Nepali with an auxiliary Hindi subset for cross-lingual experiments. The dataset comprises 10,000 annotated tweets labeled across multiple dimensions: hate speech presence, fine-grained category (e.g., casteism, xenophobia), offensiveness, target type, and sentiment. NepX-Hate is the first publicly available hate-speech dataset with multi-aspect sociocultural annotations, covering general social media discourse beyond prior domain-specific efforts. We provide benchmarks across traditional classifiers and multilingual transformer models, revealing challenges in detecting implicit hate and highlighting how fine-grained labels aid model interpretability. NepX-Hate provides a comprehensive testbed for hate speech research in underrepresented languages, enabling both sociocultural analysis and multilingual transfer. We release the dataset and code publicly, aiming to support robust, explainable hate speech detection in the Global South.

Content Warning: This paper contains examples and discussions of hate speech, including potentially offensive language and discriminatory content, used solely for academic research purposes. Reader discretion is advised.

1 Introduction

Hate speech is expressions that spread hatred, incite violence, or discriminate against individuals or groups based on their protected attributes or identity. Hate speech is subjective, and many define it differently. According to the United Nations (UN, 2019), Hate Speech is any kind of communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor. Hate speech has become a growing concern on social media due to the lack of moderation and anonymity. It also provides a platform for people to collectively incite hatred against others (Walther, 2022).

To tackle the increasing cyberbullying and hate speech, researchers have opted to use Natural Language Processing (NLP), and machine learning to develop tools for the automatic detection of hate speech (Jahan and Oussalah, 2023). All such tools depend on annotated datasets, mostly collected from social media websites such as Facebook, X (formerly known as Twitter), Reddit, etc. However, most of the annotated hate speech datasets are only available for the English language (Poletto et al., 2021), with limited resources for some other major languages (Vidgen and Derczynski, 2021). Distinguishing hate speech in a particular language from offensive or abusive speech is difficult due to the subjective nature of interpretation and cultural context, suggesting dataset requirement of the same language (Schmidt and Wiegand, 2017).

South Asian languages—particularly Nepali and Hindi—face distinct challenges: complex morphology, and culturally rooted hate expressions. Yet, resources for hate speech detection in these languages remain limited or domain-specific, with most efforts constrained to political contexts (Thapa et al., 2023). Furthermore, most datasets either lack nuance (e.g., binary labels only) or do not distinguish between hate, offensiveness, and targeted speech—dimensions critical to understanding and moderating online discourse in multilingual societies.

To bridge this gap, we present **NepX-Hate**, a novel

Figure 1: Word Cloud depicting the most frequent occuring words in the True cases of hate speech in the Nepali split of the dataset

dataset of 10,621 Nepali tweets annotated across five sociolinguistic dimensions: hate_label, fine-grained hate_category (e.g., casteism, racism), offensiveness, sentiment, and target type. Data was collected via keyword-driven crawling of publicly available tweets, followed by multi-round preprocessing and native speaker annotation with strong inter-annotator agreement. Our core contributions are:

- A high-quality, multi-annotated hate speech dataset in Nepali, with a parallel Hindi subset for cross-lingual experiments.
- Annotation guidelines and sociolinguistic labels that support nuanced detection, interpretability, and cultural analysis.
- Benchmarks using traditional ML and multilingual transformer models for binary classification and category prediction.
- Interpretability analysis between traditional, single task and multitask baselines

NepX-Hate is the first publicly available dataset of its kind for any language and contributes toward equitable NLP resources for underrepresented languages. All data, annotation guidelines, and models are made available to support reproducible and extensible research in hate speech detection.

2 Related Work

The rise of hate speech on social media has prompted substantial initiatives to create annotated datasets for automated hate speech identification. Most existing datasets focus on high-resource languages like English, with popular examples including the Stormfront dataset (de Gibert et al., 2018), HateEval (Basile et al., 2019), and Large Scale Crowdsourcing and Characterization of Twitter Abusive corpus (Founta et al., 2018). These datasets have supported progress in NLP for automatic hate speech detection by providing annotated examples that differentiate hate speech, offensive speech, and neutral speech. Similar efforts have been made in other languages, including Hindi (Bohra et al., 2018), Arabic (Mubarak et al., 2017), and Spanish (Romim et al., 2020). However, most existing datasets are domain-specific or limited in diversity, often neglecting linguistic and cultural nuances. Additionally, there is a growing recognition of the need for multilingual datasets to address hate speech in diverse linguistic contexts (Chhabra and Vishwakarma, 2023). For instance, the Multilingual HateCheck (MHC) (Röttger et al., 2022) covers functionalities across ten languages, providing a benchmark for evaluating hate speech detection models. Additionally, the LAHM (Yadav et al., 2023) dataset offers a large annotated resource for multi-domain and multilingual hate speech identification, encompassing languages such as En-

Keyword	English Code	English Translation/Meaning
साले धोती	Saale Dhoti	Bloody Dhoti (derogatory term for a particular ethnic group of Terai)
धोती को काम छैन	Dhoti ko kaam chaina	Dhoti has no work
मधिसे	Madhise	Madhise (derogatory version of Madhesi)
वेश्या	Waisya	Prostitute
छक्का	Chakka	Derogatory version of Gay

Figure 2: Sample keywords with their English code and meaning used for data collection. The English Code and English Translation are not part of the dataset and are provided as examples only.

glish, Hindi, Arabic, French, German, and Spanish. Yet, most of these benchmarks don't include low-resource languages as creating hate speech datasets in low-resource languages remains a significant challenge due to the lack of language resources, annotator expertise, and linguistic diversity.

In the context of Nepali, there has been limited work on building annotated datasets for hate speech detection. A notable effort is the dataset by Thapa et al. (2023), which comprises over 13,000 tweets collected during elections, focusing on political discourse. Most existing work for automated Nepali hate speech detection utilizes this dataset (Purbey et al., 2024). While this dataset provides a large-scale resource, its scope is restricted to political hate speech, making it less suitable for general hate speech detection. Another resource is the list of offensive keywords compiled by Niraula et al. (2022), which serves as a lexicon for identifying hate speech but lacks the annotated contextual data required for supervised learning. These limitations highlight the need for a diverse, richly annotated dataset that captures various forms of hate speech beyond political contexts. Our dataset addresses this gap by focusing on offensive and hateful slurs, providing annotations for hate speech, its categories, targets, and offensive speech, enabling more nuanced hate speech detection in Nepali.

3 Dataset: NepX-Hate

We introduce **NepX-Hate**, a multilingual benchmark dataset for hate speech detection in low-resource South Asian languages, primarily Nepali and Hindi. The dataset consists of 10,621 tweets, each annotated across multiple dimensions to support fine-grained hate speech classification, sociolinguistic analysis, and cross-lingual modeling.

3.1 Data Collection and Preprocessing

We crawled over 20,000 tweets using X's advanced search functionality with a manually curated list of 145 offensive and hate-indicative keywords sourced from online discourse. To ensure linguistic coverage and diversity, keywords targeted expressions related to caste, religion, gender, ethnicity, and nationality. Data was collected over a three-month period, filtered for duplicates, cleaned of metadata, and deduplicated post-annotation for quality control. The data was collected from publicly available tweets using a keyword-based search approach with the help of a web driver (Selenium, 2023). A small sample of such keywords with their English code and translation is shown in Figure 2. Embedded information in tweets, such as links, URLs, hashtags, mentions, and user-specific information, were cleaned. Records with null values as tweets, were removed. Identical tweets were removed after data cleaning, resulting in unique records only. Hindi and Nepali languages both utilize the Devanagari script and share many words. Language filtering was performed using the fastText classifier (Joulin et al., 2016), and manual verification. Subsequent analysis and experiments in this paper primarily focus on the Nepali subset due to the completeness of annotations and as a primary contribution.

3.2 Annotation Schema

Annotations were conducted by native Nepali-speaking linguists using a detailed guideline manual. A subset of 500 samples was jointly annotated for agreement analysis (results in-progress). The hindi subset was annotated by native hindi speakers. Each tweet is labeled across six dimensions:

- **Hate_Label (Binary)**: Whether the tweet constitutes hate speech (yes/no), guided by the United Nations' definition (UN, 2019).
- **Hate_Category (Multilabel)**: One or more of ten predefined categories: *casteism, sexism, racism, xenophobia, religious intolerance, body shaming, ableism, homophobia, others, none*. *others* include personal attacks, violent speech, attacks based on class, etc.

- **Target (Multiclass)**: The group targeted: *individual, community, organization, country, none*.
- **Offensiveness (Binary)**: Whether the tweet contains offensive or abusive language.
- **Sentiment (Single-label)**: General sentiment as perceived from the tweet: *positive, negative, neutral*.
- **Language (ne/hi)**: Identified language for each tweet.

All tweets were stripped of personally identifiable information (PII), usernames, and links. Sensitive tokens like names and countries were replaced with neutral placeholders.

3.3 Annotation Agreement

To evaluate label consistency, we conducted inter-annotator agreement (IAA) analysis on a subset of 500 tweets, evenly split between Nepali and Hindi, annotated by six native speakers for Nepali and three native speaker for Hindi. For single-label tasks, we report Fleiss' Kappa and for multilabel category annotations, we report average pairwise Jaccard similarity.

Label Type	Nepali (K / Jaccard)	Hindi (K / Jaccard)
Hate Speech (Binary)	0.845	0.822
Offensive (Binary)	0.919	0.879
Target (Multiclass)	0.897	0.887
Sentiment (Multiclass)	0.919	0.918
Category (Multilabel)	0.228	0.279

Table 1: Inter-annotator agreement scores by language. K denotes Fleiss' Kappa for single-label tasks and Jaccard denotes average pairwise agreement for multilabel annotations.

The high agreement across tasks suggests strong label reliability, especially in offensive and sentiment categories. While category-level agreement is lower, this reflects the inherent subjectivity and intersectionality of hate speech, which often spans multiple overlapping categories. These findings support the dataset's robustness and suitability for both classification and sociolinguistic analysis.

3.4 Dataset Composition

The dataset underwent a final round of preprocessing to ensure consistency:

- Removal of newline characters, excessive white spaces, and special characters.
- Removal of residual non-Nepali words and numbers that added no context to tweets
- Standardization of punctuation to maintain uniformity
- Manual check correction of spellings and any privacy issues across the entire dataset

Removal of PII resulted in many tweets with identical speech. Tweets with duplicate meanings were removed, resulting in a refined dataset of 10621 high-quality tweets. Out of the full dataset, 3,448 tweets are labeled as hate speech, with 7,173 labeled as non-hate. Nepali tweets constitute the majority of hate content (2,784 of 3,448).

Language	Total Tweets	Hate Tweets (%)
Nepali (ne)	7,428	2,784 (37.5%)
Hindi (hi)	3,193	664 (20.8%)

Table 2: Language-wise hate speech distribution in NepX-Hate.

The Nepali subset is split into **train (5,942)**, **validation (743)**, and **test (743)** sets for supervised experiments. All experiments in this paper are conducted on this partition unless otherwise noted.

3.5 Potential Applications

NepX-Hate provides a critical resource for NLP research in low-resource languages. It is designed to support a wide range of potential usages and applications, including but not limited to:

- **Hate Speech Detection Models**: The dataset serves as a foundational resource for training and evaluating machine learning and deep learning models for automatic hate speech and offensive language detection in Nepali.
- **Linguistic and Cultural Analysis**: Researchers can use the dataset to study linguistic patterns, cultural nuances, and the prevalence of hate speech in Nepali social media discourse, providing insights into societal attitudes and behavior.
- **Hate Speech Category Detection**: The dataset can be utilized for training and evaluating deep learning models and fine-tuning language models for detecting sub-categories of hate speech in the Nepali language.

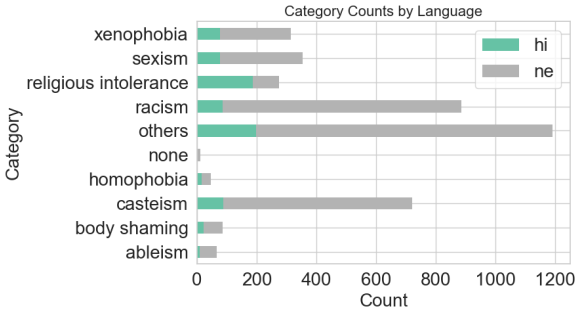

Figure 3: category frequency per language for hate speech true cases only

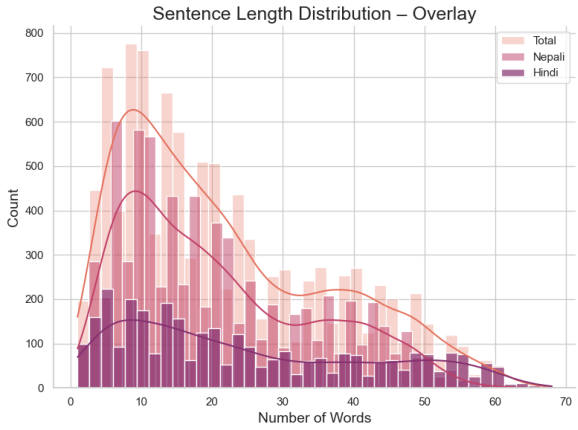

Figure 4: Sample count per word length in the dataset

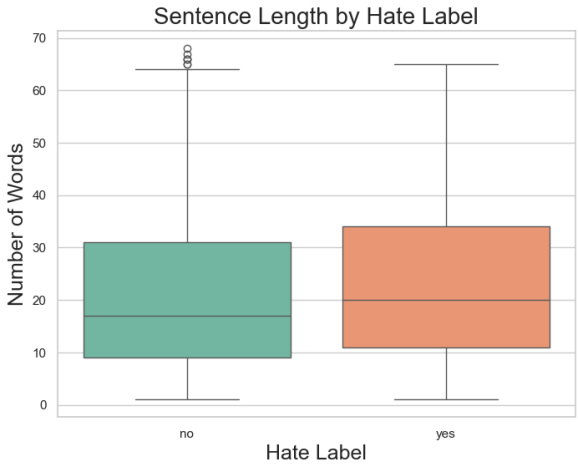

Figure 5: Sentence length per hate label in the dataset

- **Cross-Lingual and Multilingual Studies**: NepX-Hate can be integrated with hate speech datasets from other languages to develop cross-lingual models, enabling hate speech detection in multilingual or code-mixed environments.

- **Policy and Content Moderation**: The dataset can aid policymakers, NGOs, and social media platforms in understanding and addressing hate speech, contributing to the development of localized content moderation tools and strategies.

- **Creation of Dataset for Other Languages**: The methodologies employed in collecting, annotating, and preprocessing NepX-Hate can serve as a blueprint for creating similar high-quality hate speech datasets in other low-resource languages. Researchers can adapt these approaches to address linguistic and cultural nuances in different contexts.

4 Exploratory Data Analysis

To better understand the linguistic and structural properties of **NepX-Hate**, we conduct exploratory analysis across the dataset's key dimensions: label distribution, lexical characteristics, sentence structure, and category co-occurrence. These insights offer context for the modeling challenges and inform downstream experimental design.

4.1 Label Distribution

The dataset is moderately imbalanced across the hate and offensive labels. Of the 10,621 tweets, 3,448 (32.5%) are labeled as hate speech, while 7,173 are non-hate. Offensive speech is more prevalent, covering 75.74% of the data, with a strong overlap observed between hate and offensive labels.

As shown in Figure 3, category-wise distributions are skewed: *others*, *casteism*, *racism*, and *sexism* dominate, while categories such as *ableism* and *homophobia* are rare. This suggests severe class imbalance, posing challenges for categorical and multi-label modeling. Caste-based hate dominates Nepali tweets, while communal hate is more common in Hindi.

4.2 Lexical and Length Analysis

We observe notable structural differences in sentence length by hate label. Hate tweets are generally longer (median 20 tokens) than non-hate tweets (median 16.8), as shown in Figure 5. Figure 4, showing the distribution of texts by word counts for both Hindi and Nepali language. Word clouds show stark lexical distinctions between hate and non-hate tweets, particularly for caste-based and gendered slurs.

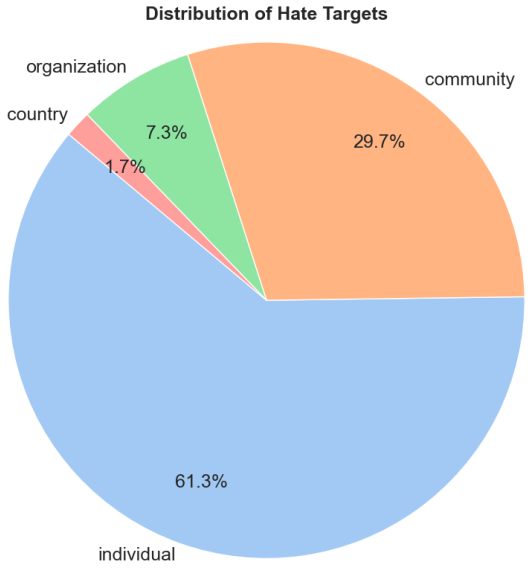

Figure 6: Target distribution in the dataset

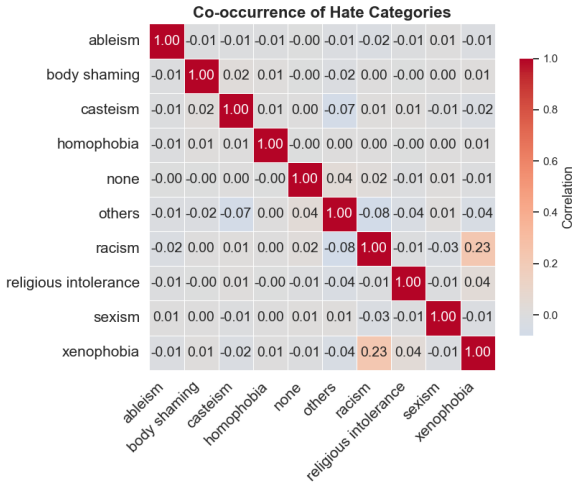

Figure 7: Co-occurrence matrix for hate speech categories

4.3 Target and Sentiment Analysis

Most hateful tweets target *individuals* (61.3%), followed by *communities* (29.7%), as visualized in Figure 6. Organization and country-level hate is rare, but often politically charged. Sentiment patterns suggest a strong alignment between hate speech and negative sentiment, although sarcasm and implicit hate complicate this correlation.

4.4 Hate Category Co-occurrence

We observe substantial co-occurrence between certain hate categories. Figure 7 shows strong co-occurrence between certain categories, especially *xenophobia*, and *racism*. This complexity highlights the importance of multilabel modeling over simplistic single-label classification frameworks and motivates the inclusion of multitask setups in our experiments.

5 Experiments and Baselines

We conduct a series of experiments to evaluate the effectiveness of NepX-Hate for hate speech detection under various modeling paradigms. Our evaluation includes binary classification and a multitask setup, both of which serve as strong baselines for future research. We use the Nepali subset of NepX-Hate (7,428 tweets) for all experiments. The data is split into training (5,942), validation (743), and test (743) sets using stratified sampling on the binary hate label. All models are trained and evaluated on the same splits for comparability. We report Accuracy, Precision, Recall, and F1-Score for all tasks. For multitask settings, evaluation is performed per task (e.g., hate detection, category classification), and macro-averaged metrics are reported for multi-label outputs.

5.1 Binary Hate Speech Classification

We benchmark a range of models for binary hate speech detection, including traditional classifiers, neural architectures, and multilingual transformers. Traditional models use TF-IDF vectorization (word-level, unigrams and bigrams), followed by classification using Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), Random Forests (RF), and XGBoost. GRU is trained with XLM Roberta Embeddings. Transformer models (mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) , MuRIL (Khanuja et al., 2021), IndicBERTv2 (Doddapaneni et al., 2023)) are fine-tuned using Pytorch custom trainer. GPT-4o-mini (OpenAI et al., 2024) and Gemini (Team, 2025) are evaluated via zero-shot prompting on the test set. Full hyperparameter details are provided in the Appendix.

Table 3 shows the results for all the models in the binary hate detection task. Transformer models outperform traditional baselines, with mBERT and IndicBERT V2 achieving the highest F1 scores. GPT-4o-mini and Gemini, while zero-shot, perform competitively but underperform compared to fine-tuned models. Surprisingly, simpler models like Linear Regression, Naive Bayes and SVM also perform competitively. However, none of the baseline crosses a score of 80% for any metric, showing the challenge of detecting implicit hate speech in

Mode	Accuracy	Precision	Recall	F1
Linear Regression	0.705	0.685	0.686	0.686
Naive Bayes	0.720	0.702	0.682	0.687
Support Vector Machine	0.708	0.693	0.702	0.696
Random Forest	0.692	0.702	0.611	0.603
XGBoost	0.703	0.685	0.651	0.656
Gated Recurrent Unit (GRU)	0.717	0.697	0.686	0.690
mBERT	0.758	0.747	0.721	0.729
XLM Roberta	0.732	0.715	0.718	0.716
Muril	0.734	0.716	0.718	0.717
IndicBERT V2	0.740	0.722	0.717	0.720
Gemma3-1b-it	0.598	0.583	0.587	0.583
Gemma3-4b-it	0.659	0.650	0.659	0.649
Gemma3-12b-it	0.701	0.682	0.652	0.657
GPT4o-mini	0.693	0.680	0.688	0.681
Gemini-2.5-flash-preview	0.703	0.714	0.727	0.700

Table 3: Performance of different models for binary classification task of hate speech detection

Nepali language.

5.2 Multiclass Category Classification

We evaluate traditional machine learning models on the task of multilabel hate category classification using TF-IDF features and One-vs-Rest classifiers. As shown in Table 4, Naive Bayes achieves the highest micro-F1 (0.605), while Logistic Regression leads on macro-F1 (0.410), indicating better performance across imbalanced classes. However, all models struggle with class sparsity and low support categories, highlighting the challenge of fine-grained hate classification in low-resource settings.

Model	Micro-F1	Macro-F1
Naive Bayes	0.605	0.151
Logistic Regression	0.587	0.410
Linear SVM	0.567	0.369

Table 4: Multilabel hate category classification results using traditional models. Scores are reported on the test set using One-vs-Rest classification and macro/micro-averaged F1 metrics.

5.3 Multitask Classification

To capture the interdependence between hate presence and its categorical expression, we adopt a multitask learning setup using BERT-based models. Hate speech is treated as a single-label task, while category prediction is modeled as multilabel classification. As shown in Table 5, multitask models consistently outperform their single-task counterparts in binary hate detection. IndicBERTv2 shows the largest improvement, with F1 rising from 0.720 to 0.752 and accuracy from 0.740 to 0.764. mBERT also benefits modestly, with F1 increasing from 0.729 to 0.740. These gains suggest that auxiliary category supervision enhances the model's ability to identify implicit or coded hate. Our findings echo broader multitask NLP research, reinforcing that structured, semantically aligned labels can guide learning in resource-scarce settings, where each additional signal can meaningfully improve generalization and robustness. The performance gains observed across most models underscore the importance of designing multi-objective benchmarks in under-resourced language contexts, where each additional signal can play a disproportionate role in improving robustness.

5.4 Interpretability

To assess how multitask learning influences model interpretability and token attribution, we apply SHAP (Lundberg and Lee, 2017) to compare token-level importance in binary hate classification for both the multitask and single-task versions of IndicBERTv2. Figure 8 shows force plots for the same Nepali sentence. This example contains implicit hate expressions and was correctly classified as hateful by both models. The multitask IndicBERTv2 assigns a high probability of 0.934 to the hate class. It attributes the decision to a combination of hateful terms, while still factoring in structural tokens and conjunctions. This shows a more balanced attribution, where both harmful

Model	Binary F1 (macro)	Binary Acc.	Category F1 (macro)	Category F1 (weighted)	Category Acc.
IndicBERTv2-MLM-only	0.752	0.764	0.256	0.670	0.700
mBERT (bert-base-multilingual-cased)	0.740	0.756	0.251	0.674	0.704
MuRIL (google/muril-base-cased)	0.698	0.746	0.145	0.562	0.646
XLM-RoBERTa (xlm-roberta-base)	0.728	0.743	0.126	0.551	0.637

Table 5: Multitask classification performance across models. Binary classification is evaluated using macro F1 and accuracy. Hate category classification is evaluated using macro/weighted F1 and accuracy on single-label prediction (most prominent category).

(a) Multitask IndicBERTv2 (confidence = 0.934)

(b) Single-task IndicBERTv2 (confidence = 0.784)

Figure 8: SHAP interpretability comparison of multitask vs. single-task IndicBERTv2 on the same Nepali hate speech example.

and contextual tokens are used in decision-making. In contrast, the single-task IndicBERTv2 yields a lower probability of 0.784 for the same example. While it identifies some of the same toxic tokens, it shows a narrower focus on fewer tokens, potentially indicating lower contextual sensitivity.

6 Discussion

Our findings underscore the multifaceted challenge of hate speech detection in Nepali, a morphologically rich and under-resourced language. Even advanced transformer models struggle with implicit or culturally embedded hate, particularly when tweets rely on sarcasm, indirect language, or political euphemisms rather than overtly hateful expressions. These challenges are exacerbated by the use of regional slurs, code-mixed phrasing, and culturally specific idioms, all of which make it difficult for models trained solely on lexical features to generalize effectively. NepX-Hate's inclusion of sentiment, offensiveness, and target type provides valuable auxiliary signals, helping models to contextualize hate more effectively.

Multitask learning further enhances model performance by enabling shared representations across interrelated labels such as hate category and target group. This leads not only to quantitative gains—especially in recall—but also to improved interpretability, allowing systems to answer not just what was said, but who was targeted and why. The dataset's fine-grained annotations expose the intersectional nature of hate, with frequent co-occurrence of categories like casteism and xenophobia, motivating the use of multilabel frameworks over binary ones. Beyond classification, NepX-Hate opens up avenues for sociocultural analysis, fairness testing, and cross-lingual adaptation, particularly between Nepali and Hindi. Its design offers a replicable blueprint for creating ethically grounded, multi-aspect datasets in other low-resource contexts.

7 Conclusion

The Nepali X Hate Speech Dataset (NepX-Hate) represents a significant step forward in enabling automated hate speech detection for low-resource languages. With 7,000 annotated Nepali tweets, the dataset provides a diverse and high-quality resource that captures hate speech, offensive speech, and targets, along with granular subcategories for hate speech classification. Our work highlights the potential for advancing natural language processing in Nepali for societal computation and contributes to broader efforts in developing multilingual and culturally aware AI systems.

By addressing key challenges such as linguistic nuances, privacy concerns, and bias mitigation, NepX-Hate sets a foundation for future research in hate speech detection for Nepali and similar low-resource languages. We anticipate that this dataset will not only support academic advancements but also foster the development of practical applications to counter online hate speech effectively.

Ethical Considerations

Curating datasets from publicly available data or user-posted data should always adhere to ethical guidelines and principles. We prioritized ethical principles throughout the dataset creation process to ensure compliance with privacy, fairness, and transparency standards. All personally identifiable information (PII), such as usernames, handles, and profile links, as well as personal names, contacts, and identities embedded in tweets, were either replaced or removed entirely. This ensures that the dataset cannot be used to trace back to any individuals, maintaining the anonymity of the users whose tweets were collected. All collected tweets were available in the public space, and no personal or followers-only tweets were collected. PII of targets in hate speech were also replaced with placeholders.

To create a comprehensive dataset that addresses general hate speech prevalent in the Nepalese online community, efforts were made to mitigate biases and address multiple facets of hate speech during data collection and annotation. Keywords used for data collection spanned various categories of hate speech, reducing the risk of dataset skewness and narrow specificity. To ensure fairness during annotation, annotators followed a structured set of instructions to ensure consistent labeling across diverse types of content.

The dataset is intended for academic and research purposes only, with the goal of improving hate speech detection systems and analyzing the trends of social media usage in the Nepali language. While we hope that this effort can be a foundation for future developments in content moderation systems for Nepali and other low-resource languages, the dataset is not a strict basis for surveillance, punitive measures, or other potentially harmful applications. Researchers using the dataset or collection steps advised here are encouraged to adhere to ethical AI practices, including transparency and accountability in their work.

Potential Risk: Due to the subjective and context-sensitive nature of hate speech, there is a risk that models trained on this dataset may produce biased or culturally insensitive predictions if deployed without proper oversight. Misuse could lead to over-censorship, mislabeling of critical speech, or marginalization of certain dialects or communities. Therefore, researchers and practitioners are strongly encouraged to adhere to ethical AI practices, including fairness auditing, transparency, and human-in-the-loop validation when using this dataset or derivative tools.

Limitations

Despite the extensive efforts made to ensure the quality and utility of NepX-Hate, several limitations remain, which should be considered when using the dataset for research or practical applications:

Datase Size

While the dataset contains 7,000 high-quality annotated tweets, this size may or may not be sufficient for training large-scale deep learning models or generalizing to all forms of hate speech encountered in real-world scenarios. We believe the dataset to be capable of fine-tuning rather than training from scratch. The dataset, while diverse, does not yet encompass every possible context or category of hate speech in Nepali.

Domain-Specific Bias

The dataset is primarily derived from social media, specifically X. As a result, it may reflect the biases, linguistic patterns, and user demographics of this platform, which could differ from other platforms or offline contexts. This may limit the generalizability of models trained on this dataset to other media of communication.

Amibiguity in Annotations

Despite providing detailed guidelines to annotators, the subjective nature of hate speech and offensive content classification introduces some level of ambiguity. Certain tweets that fall into a gray area may have been inconsistently labeled, which could impact the performance of models trained on this data for generalization.

Limited Context

Since the dataset is tweet-based, each sample is limited to a maximum of 280 characters, which may not provide sufficient context to fully understand the intent behind certain statements. Additionally, replies and broader conversational threads are not included, which could affect the accurate classification of hate speech that depends on context.

Continued Refinement

The definitions of hate speech are ever-changing and with the person. While we have made sure

to follow strict guidelines to provide objective annotation for the data, it may be possible that the definitions might change in the future, and views may differ.

Same Modality

The dataset exclusively focuses on written content in Nepali and does not include other modalities such as images, videos, or audio that may carry hate speech or offensive content. Multimodal hate speech detection is becoming increasingly important, and this dataset does not address such use cases.

Label Imbalance

Certain hate categories — such as *ableism* and *homophobia* — are underrepresented, making them challenging for supervised learning. Category co-occurrence and multi-label overlaps add further complexity, requiring specialized loss functions or resampling strategies.

References

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Anusha Chhabra and Dinesh Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29.

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. *Preprint*, arXiv:1911.02116.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. 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. *Preprint*, arXiv:1810.04805.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Preprint*, arXiv:1802.00393.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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 Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Preprint*, arXiv:1705.07874.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2022. Linguistic taboos and euphemisms in nepali. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(6).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Jebish Purbey, Siddartha Pullakhandam, Kanwal Mehreen, Muhammad Arham, Drishti Sharma, Ashay Srivastava, and Ram Mohan Rao Kadiyala. 2024. 1-800-shared-tasks @ nlu of devanagari script languages: Detection of language, hate speech, and targets using llms. *Preprint*, arXiv:2411.06850.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *Preprint*, arXiv:2012.09686.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Selenium. 2023. Selenium: Web browser automation. `https://www.selenium.dev/`. Accessed: 2025-01-16.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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*.

United Nations UN. 2019. United nations strategy and plan of action on hate speech. `https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech`. [Accessed 15-01-2025].

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Joseph B. Walther. 2022. Social media and online hate. *Current Opinion in Psychology*, 45:101298.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. Lahm : Large annotated dataset for multi-domain and multilingual hate speech identification. *Preprint*, arXiv:2304.00913.

A Detailed Results

The tables list the results of the baselines:

Model	Class	Precision	Recall	F1-score
Logistic Regression	non-hate	0.85	0.88	0.86
Logistic Regression	hate	0.78	0.74	0.76
Logistic Regression	Macro Avg	–	–	0.81
Linear SVM	non-hate	0.87	0.89	0.88
Linear SVM	hate	0.80	0.76	0.78
Linear SVM	Macro Avg	–	–	0.83
Naive Bayes	non-hate	0.82	0.86	0.84
Naive Bayes	hate	0.71	0.65	0.68
Naive Bayes	Macro Avg	–	–	0.76

Table 6: Evaluation metrics for traditional machine learning models using TF-IDF features on hate speech classification.

Class	Precision	Recall	F1-score	Support
Non-hate (0)	0.756	0.811	0.782	465
Hate (1)	0.639	0.561	0.598	278
Accuracy	0.717			
Macro Avg	0.697	0.686	0.690	743
Weighted Avg	0.712	0.717	0.713	743

Table 7: Performance metrics of the GRU model on the hate speech dataset. Best threshold used: 0.65.

Model	Accuracy	Macro F1	Non-hate F1	Hate F1	Weighted F1
GRU	0.717	0.690	0.782	0.598	0.713
IndicBERTv2	0.764	0.752	0.808	0.696	0.766
BERT-multilingual-cased	0.756	0.740	0.805	0.676	0.757
MuRIL-base-cased	0.746	0.698	0.818	0.579	0.728
XLM-Roberta-base	0.743	0.728	0.791	0.665	0.744

Table 8: Performance comparison for binary hate speech classification across different models.

Hate Category	IndicBERTv2	BERT-multilingual-cased	MuRIL-base-cased	XLM-Roberta-base
Racism	0.607	0.662	0.385	0.443
Sexism	0.154	0.000	0.000	0.000
Casteism	0.608	0.627	0.260	0.000
Religious Intolerance	0.000	0.000	0.000	0.000
Body Shaming	0.000	0.000	0.000	0.000
Xenophobia	0.000	0.000	0.000	0.000
Ableism	0.000	0.000	0.000	0.000
Homophobia	0.000	0.000	0.000	0.000
Others	0.368	0.404	0.000	0.000

Table 9: Model-wise F1 scores for individual hate speech categories.

Table 10: Comparison of Model Performance Metrics

2***Model**	**Class "no"**				**Class "yes"**			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
bert-base-multilingual-cased	0.774	0.867	0.817	465	0.721	0.576	0.640	278
xlm-roberta-base	0.793	0.774	0.783	465	0.637	0.662	0.649	278
google/muril-base-cased	0.791	0.781	0.786	465	0.641	0.655	0.648	278
ai4bharat/IndicBERTv2-MLM-only	0.783	0.809	0.796	465	0.662	0.626	0.643	278

Table 11: Overall Model Performance Metrics

Model	**Accuracy**	**Macro Avg**		**Weighted Avg**		
		Precision	Recall	Precision	Recall	F1
bert-base-multilingual-cased	0.758	0.747	0.721	0.754	0.758	0.751
xlm-roberta-base	0.732	0.715	0.718	0.734	0.732	0.733
google/muril-base-cased	0.734	0.716	0.718	0.735	0.734	0.734
ai4bharat/IndicBERTv2-MLM-only	0.740	0.722	0.717	0.738	0.740	0.739