Nepali Offensive Language Detection and Sentiment Analysis: A Comprehensive Study on Transfer Learning techniques

Merishna Singh Suwal, Sujan Bhusal, Drishtant Regmi
University of New Haven
West Haven, CT
msuwa1@unh.newhaven.edu, sbhus1@unh.newhaven.edu, dregm1@unh.newhaven.edu

Abstract

Addressing offensive content in the digital realm has become crucial with the widespread use of online platforms and social media. However, the challenges posed by linguistic diversity make pre-trained models, originally designed for English or other languages, less effective in multilingual contexts. In response to this, our research aims to comprehensively explore the limitations and challenges associated with applying pre-trained models to multilingual offensive language detection tasks.

This paper focuses on developing and evaluating a system specifically designed for identifying offensive entities in Nepali, aiming to improve content moderation tools and foster safer online spaces for the Nepali-speaking community. The research extensively investigates the effectiveness of various pre-trained BERT architectures for Named Entity Recognition (NER), with a specific emphasis on aspect term extraction to accurately categorize abusive entities in Nepali text. Additionally, the study extends its scope to Sentiment Analysis (SA), concentrating on sentiment classification in the Nepali language. Through this dual approach, our research strives to augment content moderation tools and create safer online spaces for the Nepali-speaking community. Ultimately, our goal is to pave the way for enhanced cross-linguistic understanding and improved model performance across diverse languages. **Warning: This paper contains offensive language.**

Introduction

Nepali, the official language of Nepal and a prominent language in the Indian state of Sikkim and the Himalayan regions, serves as a linguistic bridge uniting diverse communities. With over 30 million native speakers worldwide, Nepali plays a pivotal role in connecting the rich tapestry of cultures and traditions present in the South Asian region. In recent years, the explosive growth of digital communication has brought people together in unprecedented ways, transcending geographical and cultural boundaries. However, this interconnection has also exposed a darker side – the rise of hate speech in online spaces. Hate speech, characterized by offensive language and discriminatory expressions, poses a serious threat to harmonious discourse and, in extreme cases, can incite real-world violence.

Considering the growing digital landscape and the increasing prominence of online communication, the need for effective offensive language detection tools in Nepali has become more crucial than ever. Offensive language, spanning a spectrum from casual profanity to hate speech, poses significant challenges in maintaining a respectful and inclusive online environment. The vast Nepali-speaking population engages in various digital platforms, making it imperative to develop robust tools capable of discerning and mitigating offensive content. A particularly concerning aspect is the pervasive presence of offensive comments in Nepali on social media platforms, which often escape detection due to linguistic nuances and the platforms' limited focus on major languages. This oversight not only perpetuates the spread of hate speech but also marginalizes the Nepali-speaking community, leaving them vulnerable to online harassment and abuse. While significant progress has been made in hate speech detection for major languages, there remains a critical need to extend these advancements to languages with limited computational resources and less explored linguistic characteristics. The existing research landscape in offensive language detection often lacks a dedicated focus on the nuances of the Nepali language, necessitating a tailored bench-marking effort to address this gap.

This paper introduces a novel approach to address this gap by leveraging transfer learning techniques, specifically fine-tuning pre-trained language models, to create an effective hate speech detection system tailored for the Nepali language. Fine-

tuning involves adapting a pre-trained model on a general language understanding task to a specific domain or task, allowing the model to learn domain-specific features without extensive labeled data. The objectives of this research are twofold:

- First, to develop a fine-tuned model capable of identifying hate speech in Nepali text,
- Second, to contribute to the broader landscape of hate speech detection by showcasing the adaptability of transfer learning techniques to underrepresented languages.

We seek to lay the groundwork for future advancements in this critical domain. Through this benchmarking initiative, we aspire to refine and adapt offensive language detection tools to the unique linguistic characteristics of Nepali, ultimately serving the diverse and vibrant community of Nepali speakers across the globe.

Related work

The landscape of abusive content detection and Named Entity Recognition (NER) has seen extensive exploration across various languages, with a primary focus on major languages such as English. Numerous studies have addressed the challenge of identifying and mitigating offensive and abusive content in online platforms. Approaches range from traditional rule-based methods to state-of-the-art deep learning techniques. Early work often relied on lexical analysis, pattern matching, and rule-based systems to flag offensive language. However, these methods struggled to adapt to the dynamic and context-dependent nature of online conversations.

Research on Offensive Language Detection (OLD) has predominantly focused on the English language, yet the need for effective detection extends across diverse linguistic landscapes. In acknowledging this, several studies have broadened the scope to include multiple languages, recognizing the importance of tailoring models to linguistic and cultural nuances. Deng et al. (2023) delved into the realm of offensive language detection in Chinese, acknowledging the linguistic challenges specific to this language. The research contributes to the growing understanding of offensive language dynamics in Chinese online communication. Similarly, Park et al. (2023) explores offensive language detection in Korean, recognizing the need for language-specific models to effectively address the intricacies of Korean online discourse. Offensive language detection in Danish is investigated by Sigurbergsson and Derczynski (2020) and Mridha et al. (2021) have contributed to the understanding of offensive language in Bengali.

In the context of Nepal, the research landscape for detecting offensive language in online platforms is still evolving. One significant contribution is from Niraula, Dulal, and Koirala (2021), who focus specifically on offensive language detection in the Nepali language. Their work is pivotal in understanding the unique challenges and nuances of Nepali text in the realm of digital communication. Although Niraula et al.'s research does not directly reference or utilize the findings of 'Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts' by Singh et al. (2020), it nonetheless plays a complementary role in the broader effort to address online hate speech in Nepali.

Separately, the foundational work presented by Singh et al. (2020) delves into sentiment analysis within Nepali social media, which is highly relevant to our project's aim of Offensive entity recognition. Their pioneering efforts in creating the NepSA dataset—a corpus of Nepali YouTube comments annotated for aspects such as profanity, violence, and target entities—offer valuable insights. The methodologies they employed, notably the use of advanced NLP models like BERT and BiLSTM, serve as a benchmark for our approach, especially in processing a low-resource language like Nepali. Their work provides crucial guidance in understanding the complexities of processing Nepali text, including challenges like code-switching and script intricacies. Although (Singh et al., 2020)'s research is not directly cited by Niraula, Dulal, and Koirala (2021), it independently contributes to the broader understanding of sentiment in Nepali social media and lays a foundation that is instrumental for our research in abusive named entity recognition within this linguistic context.

Language models utilized in offensive language detection often rely on training data distributions, reflecting temporal and cultural contexts. Ghosh, Maji, and Desarkar (2022) highlight the challenge of maintaining model effectiveness across diverse linguistic landscapes, given the influence of cultural and temporal factors on language use. Lwowski, Rad, and Rios (2022) conducted a comprehensive analysis focusing on geographically-related content and its impact on performance disparities in offensive language detection models. Their findings emphasize the necessity of accounting for geographical variations in language use to enhance the robustness of offensive language detection models. The recognition of these performance disparities raises crucial questions about the generalizability and fairness of NLP models, necessitating a conscientious approach to model development and evaluation.

As we navigate the landscape of offensive language detection, it becomes imperative to consider the diversity of languages, dialects, and cultural nuances to foster inclusivity and equitable model performance. Recent advancements in deep learning have led to the development of sophisticated models capable of capturing contextual nuances. Pre-

trained language models, such as BERT and GPT, have shown remarkable performance in understanding the semantics of language, enabling more accurate identification of abusive content. However, the majority of these studies have primarily focused on widely spoken languages, leaving a significant gap in languages with limited computational resources, like Nepali.

Despite the progress in abusive content detection and NER, applying these techniques to the Nepali language presents unique challenges. Nepali, with its distinct script and linguistic characteristics, lacks the extensive labeled datasets available for major languages. The scarcity of resources hinders the development of robust models, necessitating innovative approaches to overcome these challenges. As we extend our focus to Nepali in this paper, we build upon this foundation to address the unique challenges and opportunities presented by the Nepali language in the context of offensive language detection. We aim to focus our research on fine-tuning pre-trained models for hate speech detection in Nepali, exploring uncharted territories in the intersection of language-specific challenges and state-of-the-art techniques.

Methodology

Dataset

The Nepali Sentiment Analysis (NepSA) dataset Singh et al. (2020) serves as the cornerstone of our project, offering a comprehensive exploration of abusive content in the Nepali language. Our choice to utilize this dataset stems from its distinctive characteristics, unravelling the diverse linguistic challenges associated with abusive language in Nepali.

The dataset was curated from comments originating from the most popular Nepali YouTube channels within the News & Politics category, specifically those with the highest subscriber counts. Comprising 3068 comments extracted from 37 distinct YouTube videos spanning 9 channels, this dataset employs a binary sentiment polarity schema. Comments are categorized into six aspects: General, Profanity, Violence, Feedback, Sarcasm, and Out-of-scope, providing comprehensive annotations. Notably, the annotations are tailored to the sentiment expressed towards a specific target entity, ensuring a nuanced understanding rather than a generalized interpretation of sentences. The target entities are further categorized into Person, Organization, Location, and Miscellaneous, contributing to the dataset's richness and applicability.

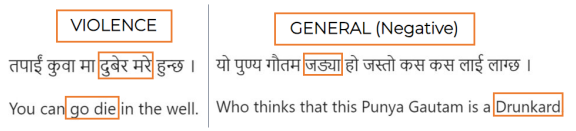

Figure 1: Named Entity Recognition dataset example

सुशील जि धन्यवाद् जन्ता को आवाज वि ले को मा । Thank you Sushil ji for being the voice of the people.	GENERAL POSITIVE (Positive Criticism)
येस्ता मानव अधिकार कर्मि को काम छइन ... । This human rights activist is of no use.	GENERAL NEGATIVE (Derogatory remarks, insults)
यो खाते अधिवक्ता दिनेश त्रिपाठी को अावज सुन्न पनि मन पर्दैन । I don't even like to hear the voice of this slum dweller advocate Dinesh Tripathi.	PROFANITY (disrespectful and inappropriate)
भ्रष्टचारी हरुलाइ टुंडिखेल मा ल्याइ झुन्ड्याएर गोलि ठोक्नु पर्छ । The corrupt should be brought to Tundikhel and shot.	VIOLENCE (Discrimination, abuse and hate speech)

Figure 2: Sentiment Classification dataset example

We have utilized the NepSA dataset for Named Entity Recognition (NER) for determining the offensive aspects within a text and for Sentiment Analysis (SA) to determine the Sentiment of the text based on the aspect word.

Dataset Analysis

From this dataset, we observe linguistic diversity on social media platforms, evident in the increasing use of code-mixed, code-switched, and transliterated comments in Nepali, based on user convenience. The ratio of English words to Nepali words is 0.0185 in this dataset.

Dataset statistics reveal that general negative sentiment prevails among all aspect categories. Mildly profane and positive feedback terms are also prominent, whereas comments related to violent sentiment are least frequent. On average, aspect terms consist of two words, with the majority having fewer than five words. Approximately 41% of aspect terms are unigrams, while bigrams and trigrams constitute 36% and 15%, respectively. Unigrams are more prevalent in general and profanity categories, while bigrams and trigrams are more common in violence and feedback categories.

Top words in each category of comments include:

- General: सलाम, झोले (Translation: salute, sycophant)
- Profanity: चोर, साला (Translation: thief, moron)
- Violence: ह□यारा, यातना (Translation: killer, torture)

The NER dataset incorporates diverse tags such as PROFANITY, VIOLENCE, GENERAL (NEGATIVE OR POSITIVE) and FEEDBACK, each representing distinct facets of offensive content, along with PERSON, LOCATION, ORGANIZATION and MISCELLANEOUS tags for defining the target for the offensive word or hate-speech.

On the other hand, the SA dataset contains 4 classes: GENERAL, PROFANITY, VIOLENCE, and FEEDBACK with 0 (Positive) and 1 (Negative) Polarity as shown in the Table 2.

Target entities	Count
Person	2327
Location	432
Organization	300
Miscellaneous	1235

Table 1: Count of Target Entities in the NER Dataset

Preprocessing

Stemming In our project, we have implemented a NepaliStemmer oya163 (2020), a vital component for text processing and analysis. This NepaliStemmer is designed to simplify and standardize Nepali text by iteratively separating suffixes (postpositions) until further separation is not possible. The underlying algorithm draws inspiration from the Hindi stemmer, adapting its principles to suit the specific characteristics of the Nepali language. The stemmer employs an iterative approach to systematically extract postpositions, contributing to the simplification of words in Nepali text. For instance, when given the word "नेपाललाई," the stemmer performs iterative separation, resulting in "नेपाल लाई," thereby breaking down the word into its essential components. To ensure accuracy and linguistic fidelity, the stemmer incorporates cross-verification with a Nepali dictionary. This step enhances the reliability of the stemming process.

Class balancing Furthermore, the label distribution within the NepSA dataset brings to light the need for a more nuanced understanding of linguistic cues, particularly in distinguishing between general and abusive contexts. The tags PROFANITY and VIOLENCE highlight explicit forms of offensive language, while GENERAL (Polarity 1: Negative) and GENERAL (Polarity 0: Positive) encompass a broader spectrum of sentiments. To address class imbalances, we first combined the polarity of the labels Profanity and Violence and removed Feedback. We then divided the General category into General positive and General negative to define a broader spectrum of general comments which do not classify as an offensive comment.

To further address the class imbalances particularly in the GENERAL category, random undersampling was employed, ensuring a more equitable distribution among classes. This strategic balancing is crucial for training models that accurately capture the intricacies of offensive language across various categories. The resulting class distribution after preprocessing is shown in the table 3 and bar chart below 3.

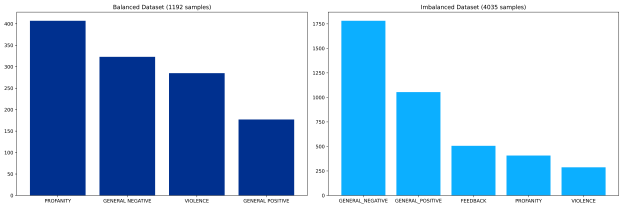

Figure 3: Balanced vs Imbalanced dataset

Experiments

Model Fine-tuning The first step in training the Offensive language detection model involves performing Named Entity Recognition (NER) to identify and tag different entities in a sentence. This process involves analyzing the sentence to detect and classify words or phrases into predefined categories like names of people, organizations, locations, etc.

Once the entities are identified and tagged, the second step involves performing Sentiment Analysis (SA) on the same sentence. Here, the goal is to determine the sentiment (General Positive/Negative, Profanity or Violence) associated with each identified entity or aspect. This step goes beyond basic sentiment analysis by not only detecting the sentiment of the entire sentence but also linking specific sentiments to the identified entities, thus providing a more nuanced understanding of the text. This two-step approach is particularly useful for detailed text analysis, where understanding the sentiment towards specific aspects or entities within the text is crucial.

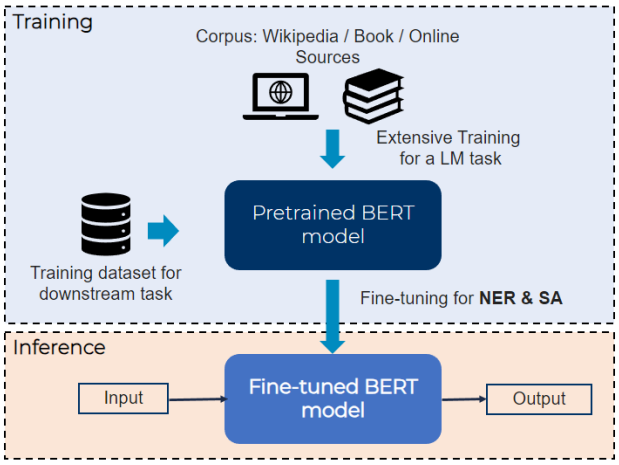

Figure 4: Fine-tuning process for Offensive Language detection (NER) and SA in Nepali

Transfer Learning on BERT Our methodology incorporates transfer learning, focusing on the fine-tuning stages for Named Entity Recognition

Polarity	General	Profanity	Violence	Feedback
0	1203	344	122	447
1	1971	120	190	84
Total	3174	464	312	531

Table 2: Count of classes in the SA Dataset

Class Balancing	General Positive	General Negative	Profanity	Violence
After	177	323	407	285
Before	1203	1971	464	312

Table 3: Class Balancing Comparison

(NER) and sentiment analysis (SA) tasks. In the initial phase, we experimented with different models, leveraging their pre-trained weights obtained from diverse linguistic backgrounds.

Specifically, we chose to use pre-trained BERT models for the task Devlin et al. (2019). This decision was driven by a careful balance between model complexity and the limited labeled data available for Nepali. The BERT architecture comprises transformer encoder blocks, each featuring multi-head self-attention mechanisms and feedforward neural networks. This design enables the model to capture bidirectional contextual information, crucial for understanding intricate linguistic relationships within a sequence.

Two of the models we experimented with (XLM-RoBERTa Conneau et al. (2019) and bert-base-multilingual-cased Devlin et al. (2018), were pre-trained in high-resource languages and multilingual text, providing a broader understanding of generalized language representations. Additionally, three models were experimented for fine-tuning which were pre-trained specifically in Nepali (NepBERTa Timilsina, Gautam, and Bhattarai (2022), deberta-base-nepali, distilbert-base-nepali Maskey et al. (2022) and NepaliBERT **?**, tailoring their knowledge to the intricacies of the low-resource Nepali language.

To adapt BERT for NER and sentiment analysis in Nepali, we introduced a task-specific output layer as illustrated in 4. This layer includes a dense neural network with a softmax activation function, facilitating the model's ability to predict the probability distribution of named entity labels(O', 'PER', 'ORG', 'LOC', 'MISC', 'GENERAL', 'PROFANITY', 'VIOLENCE', 'FEEDBACK') and sentiment labels ('GENERAL POSITIVE', 'GENERAL NEGATIVE', 'PROFANITY','VIOLENCE') at the token level. This layer acts as a bridge between BERT's contextualized representations, and the specific outputs needed for each task.

For the training and evaluation process, we adhered to an 80-10-10 percentage split, dividing the dataset into training, validation, and test sets, respectively. This partitioning scheme ensures robust model training, thorough validation, and unbiased evaluation. We trained the models using Huggingface Wolf et al. (2019) framework for both the tasks for 6 epochs using an Adam Optimizer with a learning rate of 5e-5 and a weight decay of 0.01. A comparative study of the model performances on the test set after fine-tuning is given in tables 4 and 5.

The evaluation of Named Entity Recognition (NER) models after 6 epochs reveals diverse performance characteristics. In Table 4, the xlm-roberta-large model achieves a balanced F1-score of 52.22% and an accuracy of 87.48%, demonstrating proficiency in identifying named entities. In contrast, bert-base-multilingual-cased, while achieving a high accuracy of 84.73%, shows comparatively lower precision, recall, and F1-score, indicating a potential trade-off between these metrics. NepBERTa exhibits a decent balance, although there is room for improvement in the overall F1-score. The deberta-base-nepali model emerges as the top performer, achieving the highest F1-score of 40.14% and an accuracy of 86.6%. The distilbert-base-nepali model demonstrates reasonable performance with an F1-score of 36.77% and an accuracy of 85.67%.

Moving on to sentiment classification, as shown in Table 5, xlm-roberta-large displays competitive scores across all metrics, indicating its effectiveness in classifying sentiment with an accuracy of 71.1%. The bert-base-multilingual-cased model exhibits a balanced performance with a high accuracy of 69.41%, yet there is potential for improvement in precision, recall, and F1-score. NepaliBERT performs consistently well, achieving an accuracy of 69.26% and balanced F1-score. The distilbert-base-nepali model emerges as the top performer, boasting the highest accuracy (72.24%) and F1-score (71.8%), showcasing its proficiency in sentiment classification. The deberta-base-nepali model also demonstrates strong precision, recall, and F1-score, achieving an accuracy of 71.67%.

In summary, the models present diverse strengths and areas for improvement. The xlm-roberta-

Model	P	R	F1	Acc
xlm-roberta-large	43.92	51.76	52.22	87.48
bert-base-multilingual-cased	32.6	38.76	35.41	84.73
NepBERTa	34.91	42.34	38.26	84.19
deberta-base-nepali	37.95	42.58	40.14	86.6
distilbert-base-nepali	34.59	39.23	36.77	85.67

Table 4: Results for Named Entity Recognition after 6 epochs

Model	P	R	F1	Acc
xlm-roberta-large	68.99	71.1	68.97	71.1
bert-base-multilingual-cased	67.92	69.41	66.53	69.41
NepaliBERT	68.0	69.26	66.25	69.26
distilbert-base-nepali	72.05	72.24	71.8	72.24
deberta-base-nepali	69.67	71.67	67.56	71.67

Table 5: Sentiment Classification Results after 6 epochs

large, distilbert-base-nepali and deberta-base-nepali models stand out as top performers, excelling in sentiment analysis and named entity recognition tasks. Further fine-tuning and optimization could enhance their overall performance, contributing to more accurate and nuanced analysis of sentiment and named entities in the Nepali language.

Early Stopping with F1 Score Optimization We implemented an early stopping mechanism grounded in F1 score for further optimization of our fine-tuned model. A patience parameter, set to 2, was employed to monitor F1 score improvements on the validation set during training. If consecutive epochs failed to exhibit enhancements beyond the specified patience threshold, the training process was gracefully halted. This precautionary step aimed to prevent overfitting, ensuring that the model retained its best-performing weights.

The model was then fine-tuned with carefully chosen hyperparameters: a learning rate in the range of 2e-5 to 5e-5, a batch size of 16 or 32, and an appropriate number of training epochs to balance learning and overfitting risks. This resulted in a good improvement of classification metrics on our test dataset as shown in Table 6. These steps were critical in ensuring that the model was not only well-trained on the available data but also robust enough to handle the intricacies and variances inherent in the Nepali language.

Out-of-domain testing For assessing the robustness and generalization capabilities of our model, we curated an out-of-domain test set comprising 30 manually selected YouTube comments. These comments, sourced from diverse content on the platform, introduce variability in language usage and thematic context that may not have been fully covered during training. Evaluating the model's performance on this out-of-domain test set provides insights into its ability to handle real-world comments from different contexts and linguistic variations. This rigorous evaluation ensures a more comprehensive understanding of the model's effectiveness beyond the specific domains encountered during the training phase.

The SA model performance in out-of-domain test set is shown in the classification report 8 and the confusion matrix 5. The model achieved high precision (0.90) for Class 0 (General positive), indicating that when it predicted instances as belonging to Class 0, it was correct 90% of the time. The recall for Class 1 (General Negative) is relatively high (0.79), suggesting that the model effectively captured a significant portion of true instances for this class. Class 2 (Profanity) shows a balanced precision-recall trade-off (0.62 and 0.83, respectively). However, Class 3 (Violence) has lower precision and recall (0.33 each), indicating challenges in correctly identifying instances of this class. The overall accuracy of the model is 72%, with a macro-average F1-score of 0.65, suggesting a reasonable balance between precision and recall across classes. The weighted average F1-score is 0.72, providing an aggregate measure of the model's overall performance across different classes. is well in the out-of-domain dataset with an overall accuracy of 67% and F1 score of 69%.

Results and Discussion

After an exhaustive exploration of various pre-trained model architectures and meticulous fine-tuning processes, our study has yielded insightful results in the domain of Named Entity Recognition (NER) and Sentiment Analysis (SA) tasks for the Nepali language. This section discusses the performance of different pre-trained models, shedding

Task (Model)	Before Optimization				After Optimization			
Metrics	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
NER (xlm-roberta-large)	56.95	61.98	59.36	82.85	58.32	63.8	60.60	**83.25**
SA (distilbert-base-nepali)	72.05	72.24	71.8	72.24	73.51	74.12	72.3	**74.52**

Table 6: Test Metrics for Offensive Entity Recognition and Sentiment Classification

Label	Count
0 (General Positive)	14
1 (General Negative)	13
2 (Profanity)	6
3 (Violence)	3

Table 7: Out-of-Domain Curated Dataset Label Distribution

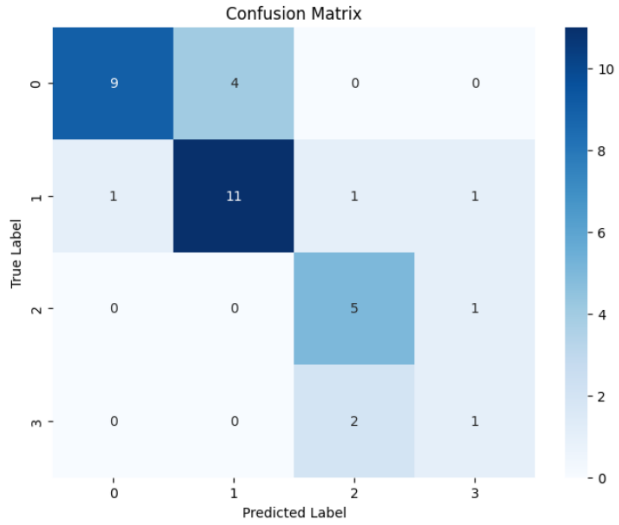

Figure 5: Confusion matrix on out-of-domain test set

light on their strengths, limitations, and overall suitability for the task at hand.

Fine-tuning BERT a transformer-based model pre-trained on a large corpus of general language understanding tasks, demonstrated a strong performance in capturing contextual information for Nepali Offensive language detection (NER) task. The bidirectional attention mechanism enabled the model to understand the dependencies between words effectively. XLM-R Conneau et al. (2019), designed for cross-lingual applications, demonstrated remarkable adaptability to Nepali NER tasks specifically for offensive language. Pretrained on a vast multilingual corpus, XLM-R exhibited a keen understanding of linguistic variations across different languages and largely contributed to its ability to handle Nepali's linguistic idiosyncrasies. Fine-tuning mBERT Devlin et al. (2018), another multilingual transformer-based model, showcased competitive performance in Nepali Sentiment Analysis to detect offensive contexts and sentences. Its versatility in handling multiple languages positioned it as a viable candidate for cross-lingual applications.

The adaptability of transformer-based models like XLM-R and mBERT across different languages is notable, especially when fine-tuned for specific tasks such as Offensive language entity recognition (NER) in languages like Nepali. While these models are pre-trained on vast, multilingual corpora, which facilitates their cross-lingual transfer capabilities, the performance of such models can be significantly enhanced with language-specific optimizations. This includes additional training epochs and fine-tuning on language-specific datasets, which enable the models to better capture the unique linguistic features and contextual nuances of a target language, as demonstrated by the performance improvements seen in the Nepali offensive language detection task.

In addition to the success of multilingual models, the performance of Nepali pretrained models stands out, demonstrating their effectiveness in handling tasks like abusive named entity recognition (NER) specific to the Nepali language. Models such as 'NepBERTa,' 'deberta-base-nepali,' and 'distilbert-base-nepali' have exhibited competitive results, showcasing the significance of pre-training on language-specific corpora. These models, fine-tuned for abusive NER in Nepali, demonstrate their proficiency in capturing the intricacies of the language, outperforming their multilingual counterparts in some aspects. This emphasizes the importance of leveraging pre-trained models tailored to the linguistic nuances of the target language for optimal performance in offensive content detection tasks. [1]

Practical Implications

The developed Named Entity Recognition (NER) model for Offensive Language Detection in Nepali offers versatile applications, particularly in the integration into content moderation systems and social media platforms. By automatically identifying and categorizing abusive named entities in textual content, the model becomes a powerful tool

[1]The code and dataset is available at https://github.com/merishnaSuwal/nep-off-langdetect

Class	Precision	Recall	F1-Score	Support
0	0.90	0.69	0.78	13
1	0.73	0.79	0.76	14
2	0.62	0.83	0.71	6
3	0.33	0.33	0.33	3
Accuracy			0.72	36
Macro Avg	0.65	0.66	0.65	36
Weighted Avg	0.74	0.72	0.72	36

Table 8: Classification Report for the Model

for enhancing the efficiency of content moderation processes. Its integration into social media platforms and other content-sharing platforms enables proactive detection and filtering of abusive entities, contributing significantly to the creation of a safer online environment. The model's capabilities hold promise for mitigating the spread of harmful and offensive content, making it an invaluable asset in the ongoing efforts to combat online abuse.

The potential impact of integrating the model into content moderation systems is profound, with a direct emphasis on user safety, online discourse, and community building. Firstly, it significantly enhances user safety by swiftly identifying and addressing abusive named entities, thereby reducing the risk of users encountering harmful content. Secondly, the model contributes to a higher standard of online discourse by flagging or removing abusive entities, fostering a more respectful and constructive digital environment. Lastly, the positive impact extends to community building, as the NER model helps create a safer and more inclusive online space, encouraging users to engage actively in discussions and fostering the development of supportive and meaningful online communities. In essence, the integration of the NER model has the potential to transform the digital landscape, promoting positive interactions and nurturing a sense of community among users.

Conclusion

This paper mainly focuses on fine-tuning a language-specific models in Nepali tailored for detecting offensive entities and sentiment in Nepali, while addressing linguistic nuances. The study emphasizes the crucial role of language-specific optimizations in promoting accurate abusive, offensive or hate-speech detection and fostering safer digital spaces.

The success of both Multilingual and Nepali pre-trained models suggests that language-specific adaptations play a crucial role in enhancing the models' understanding of the Nepali language's unique characteristics. As such, future endeavors in the field of offensive language detection could benefit from exploring more specialized, language-specific models and further fine-tuning strategies to maximize their efficacy. The positive performance of these models also opens avenues for creating similar pre-trained models for other low-resource languages, contributing to the development of effective solutions for offensive language detection in diverse linguistic landscapes. Looking forward, future research could explore more advanced model architectures, inclusion of romanized Nepali text datasets, and domain-specific adaptations to further enhance the effectiveness of these models in addressing the challenges of offensive language detection in Nepali.)

Acknowledgments

We extend our deepest gratitude to Professor Dr. Sayed, whose guidance and insightful feedback were instrumental in the success of this research. Special thanks are due to all the team members and collaborators who contributed tirelessly to this project. We are also immensely grateful to Singh et al. (2020) for providing the NepSA dataset, a crucial resource that laid the foundation for our work. Their effort in creating and maintaining this dataset has significantly propelled research in the field of Nepali language processing. We also wish to acknowledge the support of all other research papers we studied through, whose support and resources were vital in facilitating various aspects of our research.

This research would not have been possible without the collective effort and support of each individual and organization mentioned, and for this, we are profoundly thankful.

References

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116.

Deng, J.; Zhou, J.; Sun, H.; Zheng, C.; Mi, F.; Meng, H.; and Huang, M. 2023. Cold: A benchmark for chinese offensive language detection.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ghosh, S.; Maji, S.; and Desarkar, M. S. 2022. Gnom: Graph neural network enhanced language models for disaster related multilingual text classification.

Lwowski, B.; Rad, P.; and Rios, A. 2022. Measuring geographic performance disparities of offensive language classifiers. volume 29.

Maskey, U.; Bhatta, M.; Bhatt, S.; Dhungel, S.; and Bal, B. K. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In Melero, M.; Sakti, S.; and Soria, C., eds., *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 106–111. Marseille, France: European Language Resources Association.

Mridha, M. F.; Wadud, M. A. H.; Hamid, M. A.; Monowar, M. M.; Abdullah-Al-Wadud, M.; and Alamri, A. 2021. L-boost: Identifying offensive texts from social media post in bengali. *IEEE Access* 9.

Niraula, N. B.; Dulal, S.; and Koirala, D. 2021. Offensive language detection in nepali social media.

oya163. 2020. Github - oya163/nepali-stemmer: Simple rule-based nepali stemmer. flask web app deployed on heroku platform. created pip package.

Park, S. H.; Kim, K. M.; Lee, O. J.; Kang, Y.; Lee, J.; Lee, S. M.; and Lee, S. K. 2023. "why do i feel offended?' korean dataset for offensive language identification.

Sigurbergsson, G. I., and Derczynski, L. 2020. Offensive language and hate speech detection for danish.

Singh, O. M.; Timilsina, S.; Bal, B. K.; and Joshi, A. 2020. Aspect based abusive sentiment detection in nepali social media texts.

Timilsina, S.; Gautam, M.; and Bhattarai, B. 2022. NepBERTa: Nepali language model trained in a large corpus. In He, Y.; Ji, H.; Li, S.; Liu, Y.; and Chang, C.-H., eds., *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 2: Short Papers)*, 273–284. Online only: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR* abs/1910.03771.