Profanity and Offensiveness Detection in Nepali Language Using Bi-directional LSTM Models

Abiral Adhikari[1], Prashant Manandhar[1], Reewaj Khanal[1], Samir Wagle[1], Praveen Acharya[2], and Bal Krishna Bal[1]

[1] Information and Language Processing Research Lab (ILPRL)
Department of Computer Science and Engineering
Kathmandu University, Nepal
[2] School of Computing, Dublin City University, Ireland
abiraladhikari1222@gmail.com, prashantmanandhar2002@gmail.com, emailreewaj@gmail.com, waglesameer5@gmail.com, acharyaprvn@gmail.com, bal@ku.edu.np

Abstract

Offensive and profane content has been on the rise in Nepali Social Media, which is very disturbing to users. This is partly due to the absence of proper tools and mechanisms for the Nepali language to deal with profanity and offensive texts. In this work, we attempt to develop a deep learning-based profanity and offensive comments detection tool. We develop a Bi-LSTM (Bidirectional Long Short Term Memory) based model for the classification of Profane and Offensive comments and study different variations of the task. Furthermore, Multilingual BERT embedding and vocab embedding were used among others for an accurate understanding of the intent and decency of the posts. While previous related studies in the Nepali language are more focused on sentiment and offensiveness detection only, our study explores profanity and offensiveness detection as two distinct tasks. Our Bi-LSTM model outputs 87.8% accuracy for "profane" and "non-profane" content classification and 78.2% accuracy for "offensive" and "non-offensive" content classification. Similarly, it outputs 65% accuracy with the multilingual BERT embedding and 66.3% accuracy with the vocab embedding for the "none", "offensive" and "profane" content classification.

Keywords: Profanity detection, Bidirectional LSTM Model, Transliteration, NLP, Vocab Embedding, Multilingual BERT Embedding, Nepali Language.

1 Introduction

With the development of technology and affordable access to the internet, the number of users has surged across social media sites like Facebook, YouTube, X (formerly Twitter), Reddit, etc. Users interact on social media by commenting and posting in natural language. Unfortunately, the interaction is not always social and pleasant. Internet content is getting more toxic by the day with increasingly profane, offensive, and hateful content flooding in (Cheng et al., 2021). Thus, detecting and moderating profane and offensive language content is crucial in social media to keep the discussion decent and publicly acceptable. Manual detection of foul language in social media platforms with a large user base is costly and time-consuming. Therefore, it is necessary to develop automatic profane and offensive content detection systems (Niraula et al., 2021).

There are multiple studies conducted in this problem domain, particularly for high-resource languages like English (Teh and Cheng, 2020), German (Assenmacher et al., 2021; Risch et al., 2019), and Spanish (Díaz-Torres et al., 2020). However, the problem is not adequately studied for Nepali - a low-resource language, albeit a few works of (Niraula et al., 2021; Singh et al., 2020). The previous works have focused more on abusive sentiment and offensive language detection whereas profanity detection and the gender associations with the problem are underexplored. Building on these previous studies, we attempt to detect profanity and offensiveness in Nepali, primarily observed in controversial Facebook posts. The profanity and offensiveness of text are differentiated with the notion that the presence of taboo words in the abusive text makes it profane while its absence makes it offensive.

In this study, we contribute to dataset development by developing an annotated dataset for profane and offensive comments in the Nepali language, collected from various controversial Facebook posts. Besides the dataset development, we also propose a Bi-LSTM (Bidirectional Long Short Term Memory) model-based working prototype to handle both binary (profane vs non-profane, offensive vs non-offensive) and multi-class (offensive vs profane vs none) classification of the comments. Furthermore, we study the multi-class classification problem using multilingual BERT embedding and vocabulary embedding. In this regard, the key contributions of this study can be summarized as follows:

1. We developed an 11681 manually labeled dataset for the detection of profanity and offensiveness in the Nepali language.

2. We proposed a baseline Bi-LSTM model tailored for multi-class and binary-class classification of Nepali Social Media comments and demonstrated its effectiveness.

2 Related Work

Profanity and offensiveness detection systems have been developed for high-resource languages like English (Teh and Cheng, 2020), Hindi (Chopra et al., 2020), German (Risch et al., 2019), and Spanish (Díaz-Torres et al., 2020) but such systems are still under development for low-resource languages like Nepali (Niraula et al., 2021), Sinhala (Ranasinghe et al., 2024) etc. The study conducted in the Mexican Spanish language focuses on different categories of abusive texts such as vulgarity, aggressiveness, and offensiveness based on intent and context (Díaz-Torres et al., 2020). Furthermore, the classification of offensive and vulgar language requires a fine-grained approach to differentiate the aspects including disability, sexism, racism, etc. (Teh and Cheng, 2020). Profanity is one of such classifications of abusive texts, detection of which is more effective if we associate hatred with profane words (Teh and Cheng, 2020).

The use of curse words is found to have a contextual variable associated with it: when, where, and how, via the research on English language data from X (formerly Twitter) (Wang et al., 2014). The variations in how curse words are used negatively affect the profanity detection task resulting in low recall (Wang et al., 2014). Even Nepali social media texts come in different forms: native (pure Devanagari), code-mixed (Romanized), and code-switched (Roman + Devanagari) making it difficult to pre-process (Singh et al., 2020) and classify accurately. The study by Niraula et al. (2021) provides a way to handle the pre-processing of such Nepali text using techniques like Dirgikaran and Romanization (transliteration) for classification purposes.

Among the models used for profanity detection, LSTM was found to outperform other sequential models like CNN (Convolutional Neural Network), and GRU (Gated Recurrent Unit) by studies done in Korean (Yi et al., 2021) and Marathi (Velankar et al., 2021) languages. With these references as a basis for our research, we decided that LSTM architecture would be a better suit to solve our tasks given its capacity for handling sequential tasks and addressing long-term dependencies, in text classification across different linguistic contexts.

In the detection of profanity and offensiveness in the Nepali language, only a handful of resources and research studies have been conducted (Niraula et al., 2021) and (Singh et al., 2020). The work of (Singh et al., 2020) was conducted for the classification of abusive texts but included only 407 instances of profanity which seems to be insufficient data for the profanity detection task. The work of (Niraula et al., 2021) has a fine-grained classification of racist, sexist, and others under the broader umbrella of offensive comments but this study does not encompass the profanity-based classification. As a bid to address this research gap in previous studies, we attempt to build a larger dataset for the profanity detection tasks. There are also gender rank factors associated with profanity (Wang et al., 2014), which is not explored in the Nepali language. So, we also look into this aspect in our study.

3 Methodology

3.1 Data Collection

The raw dataset was compiled by scraping comments from controversial Facebook posts in Nepali language using the Apify[1] console, an online-based web scrapping tool that can extract data from various social media like Facebook, TikTok, etc. As there exists a rate limit on the use of APIs for social media data extraction and authorization protocol complexities in web scraping, Apify was selected for its simplicity. Most of the profanity and offensive comments were seen on the Facebook comment section of controversial people, celebrities, and political figures. After shortlisting the controversial comments on the posts of popular Nepali celebrities like Samir Ghising aka. VTEN, Jyoti Magar, Rabi Lamichhane, etc., a total of 11681 comments in Devanagari script (native), Latin script (Romanized Nepali), and mixed script (code-switched) were selected, which were then classified as "offensive", "profane", or "none".

The gender of the users posting the comments was distinguished by employing Nepali gender names collected from Kaggle[2], which was the scrapped data from the website of the Election Commission of Nepal[3].

3.2 Data Preprocessing and Annotation

As mentioned in section 3.1 earlier, the comments in the posts were either in pure Devanagari script, pure Romanized script, or mixed Roman and Devanagari script which are showcased in Table 1.

Script Type	Examples
Devanagari script	यो मुलाको नाटक ।
Pure Latin script	All credit goes to Sarban Yadav chor fatha gadha mayor
Mixed Latin and Devanagari script	good luck ravi dai भन्ने शब्द हरायो

Table 1: Various script types and their examples

The presence of code-mixed (texts in multiple scripts - Devanagari and Latin) and code-switched (texts coming in multiple scripts and being switched from one script to the other) texts in comments poses difficulty in pre-processing. Since a significant amount of texts come in transliterated form, i.e., Nepali texts written in the Latin script, we converted such texts into Devanagari Nepali using the "ai4bharat-transliteration"[4] tool. The use of the "ai4bharat-transliteration" package with an 80% performance score for English-to-Nepali transliteration on the Aksharantar (native words) benchmark enabled a consistent conversion of text from Roman to Devanagari script as shown in Table 2. But even with the use of the package, there are several issues in terms of accurate transliteration as different users write the same word differently, for example, Kando, Kanndo, आइमाइ, आइमाई, उल्लू, उल्लु making it difficult to pre-process the texts consistently.

We obtained 166 profane keywords and 720 offensive keywords which were written in Romanized Nepali as well as Devanagari Script from the "Data Set for Offensive Language Detection in Nepali"[5] and "NepSA"[6] datasets. In addition, new keywords were added to enrich our keyword dictionary, thereby making auto-labeling comprehensive. After the auto-labeling of comments was done using the above-collected keywords, the comments were manually reviewed and we applied a two-level annotation scheme while classifying profane and offensive keywords for further processing. In the first level of manual annotation, we assumed that "*profane*" comments are inherently "*offensive*", when in fact not all "*offensive*" comments are "*profane*". In the second level of annotation, we further perform a finer-grained distinction between profane and offensive comments with taboo words like ("रण्डी","साला कुकुर", "मुजि"), now being categorized explicitly as "profane", while, the comments with non-taboo but hurtful words like("थिस", "बाउको बिहे") being categorized as "offensive" and the rest of the comments were categorized as "none".

[1] https://apify.com/
[2] https://www.kaggle.com/datasets/nischallal/nepali-name-dataset-in-devanagari-with-gender
[3] https://election.gov.np/np/page/voter-list-db
[4] https://pypi.org/project/ai4bharat-transliteration/
[5] https://github.com/nowalab/offensive-nepali
[6] https://github.com/oya163/nepali-sentiment-analysis

Romanized Nepali Text	Transliteration to Devanagari Nepali
Mukh samalara bola vai hami nepali ho tasto bolnu hudayn lo	मुख समलारा बोला वाई हामी नेपाली हो तस्तो बोल्नु हुदेन लो

Table 2: Conversion of Romanized Nepali Text to Devanagari script Nepali Text

We excluded emojis in the comments as our primary focus was on analyzing textual content. Further pre-processing of the text was done by applying stemming, and removal of punctuation symbols, special characters, and Nepali stop-words using the "Nepali_nlp"[7] toolkit. The final version of the annotated dataset (NepSense[8]) was proofread and checked for sanity by a domain expert and native speaker of the language.

3.3 Data Analysis and Description

After the annotation, 11681 comments were finalized to be used during model training. Among those data, 2229 comments were labeled "profane", 1313 comments were labeled "offensive" and 8139 comments were labeled as "none". On further analysis of the comments collected, 60.11%, 12.96%, and 26.91% of the comments were posted by "male", "female", and "unknown" respectively. Since classifying gender in the "unknown" category proved to be challenging, further analysis on this category was difficult to conduct. On the broader analysis, it was found that out of the 7022 instances of the comments posted by males, 21% of them were profane, and 10.45% of them were offensive. Similarly, out of the 1515 instances of comments posted by females, 14.91% of them were profane and 7.45% were offensive. From the above analysis, both the ratio and the number of profane and offensive comments posted by males were found to be higher than those posted by females, consistent with prior findings (Wang et al., 2014).

4 Experimentation and Result

The studies (Velankar et al., 2021) and (Yi et al., 2021) for profanity detection in low-resourced languages, Marathi and Korean, respectively, indicate that the LSTM model performs better than other sequential models namely: CNN and GRU. Additionally, the study by (Ezen-Can, 2020) emphasized that the LSTM model performs better than the BERT model for a smaller dataset. Based on above mentioned studies we selected the baseline Bi-LSTM model for our experimental setup. The model experimentation and results of different classification tasks are detailed as follows:

1. **Binary Classification of Profanity (Task A)**:

   The profanity classification task classifies the comments into two classes viz. "profane" and "non-profane". For the experimental purpose, the data subset was balanced by selecting equal instances (2229) from both classes. The model was trained to achieve an accuracy of 87.8%. The data was embedded using vocab embedding with the embedding dimension of 256 for contextual understanding. The classification results are shown in Table 3 indicating that the model is performing well for the profanity classification in the test set with high precision and recall for both classes.

2. **Binary Classification for Offensive (Task B)** :

   The offensiveness classification task classifies the comments into two classes viz. "offensive" and "non-offensive". The dataset for the model training of the offensive class classification was balanced by taking the same instances (1313) from both classes. The model was trained to achieve an accuracy of 78.2%. For embedding purposes, the data was embedded using the vocab embedding for understanding the context of the sentences. The classification results presented in Table 3 demon-

[7] https://pypi.org/project/Nepali-nlp/
[8] https://github.com/Eemayas/NepSense-Dataset

Task A			
Class	**Precision**	**Recall**	**F1-Score**
Profane	0.83	0.86	0.84
Non-Profane	0.85	0.82	0.84

(a) Binary Classification of Profanity

Task B			
Class	**Precision**	**Recall**	**F1-Score**
Offensive	0.87	0.56	0.68
Non-Offensive	0.68	0.92	0.78

(b) Binary Classification for Offensive

Task C(a)			
Class	**Precision**	**Recall**	**F1-Score**
Profane	0.851	0.681	0.756
Offensive	0.506	0.532	0.518
None	0.61	0.698	0.651

(c) Multi-classification using Multi-lingual BERT

Task C(b)			
Class	**Precision**	**Recall**	**F1-Score**
Profane	0.81	0.75	0.78
Offensive	0.57	0.51	0.54
None	0.54	0.64	0.58

(d) Multi-classification using Vocab Embedding

Table 3: Classification Report for Different Tasks

strate that the model exhibits lower recall for the "offensive" class, resulting in reduced overall performance compared to its effectiveness in **Task A**.

3. **Multi-classification (Task C)**:

   The multi-class classification task classifies the comments into three classes viz. "profane", "offensive", and "none". The embedding for multi-class classification was done in two ways:

   (a) BERT Embedding (Task C(a))

   (b) Vocab Embedding (Task C(b))

   The dataset was balanced by taking 1313 instances from each of the classes which was the count of the least number of instances among the three classes i.e. "offensive" with 1313 instances. An accuracy of 65% was achieved using BERT embedding and 66.3% while using vocab embedding, showing relatively similar performance. The classification results of Task C(a), and Task C(b), are shown in Table 3 for the test set, showing that both models have a lower recall for "offensive" class compared to "profane" and "none" class.

Model	**Validation Accuracy**	**Training Accuracy**
Task A	0.84	0.878
Task B	0.74	0.782
Task C(a)	0.64	0.65
Task C(b)	0.63	0.663

Table 4: Accuracy of Models on Different Tasks

All models were trained in Google Colab using T4 GPU. It should be noted that the parameter values from the Table 5 represent the optimal values achieved during the experiments.

5 Discussion

The promising result of the model in **Task A** can be attributed to higher instances (2229) of profane comments on the dataset. The profanity of the comment does not depend on the context as much compared to the offensive words. The common profane words like **"मुजी", "फक", "माचिक्ने"** etc. are used more frequently. The inferior performance of the model for **Task B** when compared to **Task A** can be attributed to lower instances (1313) of offensive comments for training. Also, the offensiveness of the comment depends upon the context and intent. E.g. कुकुर - represents the animal "dog" in the layman's understanding but when the context is of offensive sentiment it represents the person is being insulted as "a dog". Furthermore, significant low recall values for the offensive class across models **Task B**, **Task C(a)**, and **Task C(b)** indicate that the models make more errors by inaccurately predicting offensive comments as non-offensive, which is due to contextual complexities regarding sentiment associated with

Parameters	Task A	Task B	Task C(a)	Task C(b)
Epoch	30	30	100	30
Batch Size	128	128	128	128
Regularization	1e-04	1e-04	1e-04	1e-04
Learning Rate	6e-05	6e-05	1e-03	6e-05
Early Stopping Patience	5	5	3	5
Optimizer	Adam	Adam	Adam	Adam
Dropout	0.7	0.7	0.4	0.7

Table 5: Tuning Parameters for models used in experimentation.

offensive comments and their lower instances in the dataset.

6 Conclusion

In this study, we developed an annotated dataset for profanity and offensiveness detection in the Nepali language to address the scarcity of labeled datasets. The dataset was developed by labeling 11681 Facebook comments collected from various controversial Facebook posts. Also, we proposed a method to pre-process the social media text via transliteration from Romanized English Latin script) to Devanagari script instead of transliteration from Devanagari to Romanized English (Latin script) as done in the previous study (Niraula et al., 2021).

The observation from the experiments indicated that the models performed better in the classification of profane comments than offensive comments. It was also noted that the binary classification models outperformed the multi-class classification models. The study also points out that male users are more likely to use offensive and profane language compared to females, supporting the study done by (Wang et al., 2014).

The dataset and models developed during this study open doors for future research and eradication of antisocial behavior like bullying, harassment, etc. in social media platforms. The first-level annotated dataset can further be potentially used for hate speech detection. In the future, we plan to enrich and enlarge our dataset by incorporating Nepali texts from other social media platforms covering regional slang as well, while implementing the inter-annotator agreement for better data reliability. The enlarged dataset is to be further revalidated by experts. We also plan to develop models using advanced architecture, such as Transformers, once sufficient data is gathered. We also plan to improvise the pre-processing by introducing the lemmatization technique.

Limitations

1. The models developed in this study remain prone to inaccurately flagging genuinely offensive comments as non-offensive, which needs explicit human verification and consequently further model tuning.

2. Due to the ever-evolving nature of language based on environment, age, generation, ethnicity, etc., our models will likely fail due to this e.g. Millennial slang v/s GenZ slang.

Acknowledgments

We extend our sincere gratitude to the Information and Language Processing Research Lab (ILPRL), Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Bagmati, Nepal, for providing lab facilities and support essential to this research.

References

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz V. Seiler, Dennis M. Riehle, and Heike Trautmann. 2021. RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2021. Antisocial behavior in online discussion communities. *Proceedings* *of the International AAAI Conference on Web and Social Media*, 9(1):61–70.

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):386–393.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes-y Gómez, Juan Aguilera, and Luis Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136, Marseille, France. European Language Resources Association (ELRA).

Aysu Ezen-Can. 2020. A comparison of LSTM and BERT for small corpus. *CoRR*, abs/2009.05451.

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, Online. Association for Computational Linguistics.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *Conference on Natural Language Processing*.

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.

Phoey Lee Teh and Chi-Bin Cheng. 2020. Profanity and hate speech detection. *International Journal of Information and Management Sciences*, 31(3):227–246.

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

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 415–425, New York, NY, USA. Association for Computing Machinery.

MoungHo Yi, MyungJin Lim, Hoon Ko, and JuHyun Shin. 2021. Method of profanity detection using word embedding and lstm. *Mobile Information Systems*, 2021(1):6654029.

A Dynamic Knowledge Graph of Interaction Between Pollution and Cardiovascular Diseases

Sudeshna Jana[1], **Anunak Roy** [2], **Manjira Sinha** [1], **Tirthankar Dasgupta** [1],
[1]TCS Research, [2] IIT Kharagpur,
Correspondence: { sudeshna.jana, sinha.manjira, dasgupta.tirthankar} @tcs.com, anunakroy@kgpian.iitkgp.ac.in

Abstract

In recent decades, environmental pollution has become a pressing global health concern. According to the World Health Organization (WHO), a significant portion of the population is exposed to air pollutant levels exceeding safety guidelines. Cardiovascular diseases (CVDs) — including coronary artery disease, heart attacks, and strokes — are particularly significant health effects of this exposure. In this paper, we investigate the effects of air pollution on cardiovascular health by constructing a dynamic knowledge graph based on extensive biomedical literature. This paper provides a comprehensive exploration of entity identification and relation extraction, leveraging advanced language models. Additionally, we demonstrate how in-context learning with large language models can enhance the accuracy and efficiency of the extraction process. The constructed knowledge graph enables us to analyze the relationships between pollutants and cardiovascular diseases over the years, providing deeper insights into the long-term impact of cumulative exposure, underlying causal mechanisms, vulnerable populations, and the role of emerging contaminants in worsening various cardiac outcomes.

1 Introduction

Over the past few decades, the increasing levels of environmental pollution have emerged as a formidable global health crisis. The World Health Organization (WHO) has reported that nearly 99% of the world's population breathes air that contains pollutant levels exceeding established safety guidelines (Organization et al., 2016). This exposure is particularly severe in low- and middle-income countries, where industrial activities, urbanization, and insufficient regulatory measures exacerbate air quality issues. Among the various health effects attributed to pollution, its impact on cardiovascular diseases (CVDs) is particularly concerning (Rajagopalan et al., 2018). CVDs encompass a wide range of conditions, including coronary artery disease, heart attacks, strokes, and heart failure. These diseases are not only prevalent but also represent a leading cause of morbidity and mortality worldwide. The mechanisms by which air pollutants influence cardiovascular health are multifaceted, involving both direct effects on the cardiovascular system and indirect effects through exacerbating existing risk factors such as hypertension and diabetes.

Recent studies have demonstrated that pollutants such as fine particulate matter (PM), heavy metals, and toxic gases have a significant impact on cardiac health (Basith et al., 2022; Zhang et al., 2022, 2016). Research conducted by the United States Environmental Protection Agency (EPA) indicates that exposure to elevated concentrations of $PM_{2.5}$, even over a short duration of a few hours to weeks, can trigger heart attacks and fatalities associated with cardiovascular diseases. Prolonged exposure to these pollutants is linked to an increased risk of cardiovascular mortality and a reduction in life expectancy (Beelen et al., 2014). Furthermore, a substantial body of epidemiological evidence reveals a strong correlation between air pollutants and rising rates of cardiovascular diseases, including heart failure (Jia et al., 2023). Animal studies also corroborate these findings, illustrating that exposure to pollutants can elevate the likelihood of conditions such as thrombosis and atherosclerosis (Sun et al., 2005). These insights underscore the urgent need for comprehensive public health interventions aimed at reducing air pollution and mitigating its adverse effects on cardiovascular health.

However, current findings do not demonstrate how cumulative exposure impacts cardiovascular health over decades, the precise causal pathways by which these pollutants lead to cardiovascular damage. Additionally, it is crucial to identify which populations are most vulnerable to specific pollutants and the reasons behind their susceptibil-