Linguistic Taboos and Euphemisms in Nepali

NOBAL B. NIRAULA, Nowa Lab, USA
SAURAB DULAL, The University of Memphis, USA
DIWA KOIRALA, Nowa Lab, USA

Languages across the world have words, phrases, and behaviors - the taboos - that are avoided in public communication considering them as obscene or disturbing to the social, religious, and ethical values of society. However, people deliberately use these linguistic taboos and other language constructs to make hurtful, derogatory, and obscene comments. It is nearly impossible to construct a universal set of offensive or taboo terms because ofensiveness is determined entirely by different factors such as socio–physical setting, speaker-listener relationship, and word choices. In this paper, we present a detailed corpus-based study of offensive language in Nepali. We identify and describe more than 18 different categories of linguistic offenses including politics, religion, race, and sex. We discuss 12 common euphemisms such as synonym, metaphor and circumlocution. In addition, we introduce a manually constructed data set of over 1000 offensive and taboo terms popular among contemporary speakers. We describe the first experiments that provide baseline results in detecting offensive language in Nepali. This in-depth study of offensive language and resource will provide a foundation for several downstream tasks such as offensive language detection and language learning.

CCS Concepts: • Computing methodologies → Language resources; Discourse, dialogue and pragmatics.

Additional Key Words and Phrases: Offensive Language, Linguistic Taboo, Data Set, Offensive Language Detection, Low-resource Language, Nepali Language

1 INTRODUCTION

Disclaimer: Due to the nature of this topic, this paper contains highly offensive language, hate speech, and adult content. Most of these offensive languages are directly extracted from social media content. The offensive language used in this paper don’t reflect the views of the authors in any way. Our intention is not to stir a controversy from personal, religious, political, and geographical perspectives in any topics such as country, religion, language, and culture. Rather, the goal of this paper is to build a foundation to stop hate speech and discrimination in the use of language.

Note: Please refer Table 4 and Table 5 in the Appendix section to get the English translations of the Nepali words, phrases and sentences that do not have their inline translations in the manuscript.

Offensiveness is a socio-cultural and situational phenomenon. Constructing a universal set of offensive terms for all the languages is near impossible because a culture, language, or situation defines what terms, phrases, or actions are offensive. Furthermore, the semantics of terms keeps changing with time. These variations warrant the language and culture-specific study of taboos and offenses1. This paper presents the study of linguistic taboos

1Throughout this paper, the word “offense” is explicitly used to refer linguistic offense only

Authors’ addresses: Nobal B. Niraula, Nowa Lab, Madison, Alabama, USA, nobal@nowalab.com; Saurab Dulal, The University of Memphis, Memphis, Tennessee, USA, sdulal@memphis.edu; Diwa Koirala, Nowa Lab, Madison, Alabama, USA, diwa@nowalab.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.
2375-4699/2022/3-ART $15.00
https://doi.org/10.1145/3524111

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
and euphemisms in Nepali, a language belonging to the Indo-Aryan family. Nepali is the official language of Nepal and several eastern Indian regions including Sikkim, Darjeeling, and Kalimpong. It is spoken by more than 20 million people, mainly in Nepal, and other places in the world including Bhutan, India and Myanmar. To the best of our knowledge, this is the first study of linguistic taboos and euphemisms in Nepali. Thus, it provides a foundation for many Natural Language Processing (NLP) tasks such as offensive language detection, comment moderation and natural language generation.

Today social media, online forums and news portals are increasingly under pressure to stop offensive content. Manually reviewing streams of content is labor-intensive, costly and slow. NLP can greatly assist the moderation task. However, NLP for this task requires a language-specific understanding of offensive constituents. To this end, this study provides great insights into understanding how offensive languages are formed, what varieties of offensive content exist and how to generate non-offensive content in Nepali. Furthermore, it also provides benefits to Nepali language learners e.g. to avoid making offensive content and apply euphemisms to circumvent embarrassing situations.

Language is a medium through which one expresses thoughts and feelings to others. It is used in every aspect of human communication, expressing emotions such as love, hate and anger, revealing identity, disseminating information, and so on. Although language is used freely, society sometimes imposes restrictions and constraints on the use of certain terms, expressions, or behaviors, called “taboos”. Breaking a taboo is usually considered objectionable because it is harmful to the social, religious, or ethical values of the society. Taboo can be verbal or non-verbal. The non-verbal taboo (aka. behavioral taboo) restricts the use of behavioral patterns that are offensive to the social values or customs. For instance, entering a house with shoe on, and eating food during a classroom session are non-verbal taboos in Nepali culture. The verbal taboo (aka. linguistic taboo) restricts the use of certain words or parts of a language in public discourse due to social constraints e.g. kandoo (buttock). Therefore, instead of using linguistic taboos, euphemisms are used to alleviate or erase the harshness. They are roundabout ways of expressing unpleasant terms to neutralize or lessen the unpleasantness.

Nonetheless, people do deliberately use linguistic taboos to make hurtful, derogatory, and obscene comments. However, a language does not necessarily need linguistic taboos to be offensive. Some words or phrases that do not have taboo meaning lexically can also be offensive if used in certain contexts. For example, the sentence tan kukur hos (you are a dog) does not have any unmentionable linguistic taboo, yet it is an offensive sentence. In this study, we refer to both linguistic and contextual taboo terms as offensive terms.

The offensiveness is determined entirely by pragmatics variables such as socio-physical setting, speaker-listener relationship along with the words used and tone of voice [19]. Many actions or expressions that are considered normal in one society may not be considered normal in other societies [20]. For example, asking person’s age can be considered normal in Nepali culture due to its honorific system. People use appropriate pronouns and verbs in conversations based on the relative age. A younger person uses pronouns like tapai, hajur when talking to an elder, the latter uses pronouns like timi, tan to the younger. In contrast, asking person’s age isn’t considered normal in Western culture [14]. Similarly, touching a book with feet is fine in Western society while it is examined as sinful act in Nepali culture. Likewise, diverse animal names are used for personal insults in Nepali (more in Sections 3 & 4) compared to English which has fewer cases like bitch, pig, ass, snake, and monkey.

Different dialects of a language can have extremely different meanings of a term. The term coger, for example, is a normal word in Colombian Spanish which means to take but it is a taboo word in Argentine Spanish which refers to have sex. Similarly, cholnu is a normal word in Far-Western Nepal and means to scrub. This is a taboo

2https://en.wikipedia.org/wiki/Nepali_language
3Please refer glossary at Appendix in case translations of Nepali terms are not provided in the content itself.
4Branch of linguistics that deals with context of utterance [22]

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
term in Eastern Nepal and means *to masturbate*. Offensiveness level can also be different across the dialects e.g. *bloody* in Australian and American English is not as offensive as it is in British English [31, 39].

The offensiveness of a term evolves with society. The term *breast meat* was considered offensive in the 20th century and *white meat* was its euphemism [14]. It has lost the offensiveness in modern-day English. Similarly, in Nepal, the term *bhalu* had just a single meaning in the past referring to the animal *bear*. Nowadays it can also mean a *prostitute* in an offensive context.

Ambiguity is an inherent property of a language. Pragmatics and contextual evaluation are required to measure the offensiveness of a term. The word *gadha* (donkey) is offensive in the sentence *ta gadha hos* (you are a donkey) but is not offensive in *gadhale bhari bokchha* (donkey carries load). Likewise, the word *baby* is more offensive to young boys than to their parents [20].

In this paper, we present a detailed corpus-based study of offensive language in contemporary Nepali. We collected and analyzed thousands of offensive social media posts and comments from diverse sources. Specifically, we studied over 7000 social media posts and manually collected over 1000 tabooed and offensive terms. Furthermore, we identified 18 different varieties of linguistic offenses such as Bodily Excretion, Death, Disease, Relation, Sex, Gender, Plant, and Animal. We also provide baseline algorithms and results for offensive language detection in Nepali. Our major research contributions are as follows:

1. The detailed study of linguistic offenses in 18 different categories
2. A reference data set of over 1000 linguistic taboos and their categories
3. Common euphemism techniques
4. Baseline algorithms and results for offensive language detection in Nepali

2 BACKGROUND

In this section, we provide a general overview of the taboo and offensive language and related work.

2.1 Taboo Language

Wardhaugh defines taboo as the things that are prohibited or avoided in any society of behavior because they are believed to be harmful to its members and would cause them anxiety, embarrassment, or shame [41]. If necessary, they are expressed in very roundabout ways, i.e., using euphemisms. Popular English dictionaries also define taboo similarly: Cambridge Dictionary of English defines *taboo* as a subject, word, or action that is avoided for religious or social reasons. According to Longman Dictionary, a taboo subject, word, activity, etc. is one that people avoid because it is extremely offensive or embarrassing. It is not accepted as socially correct or it is too holy or evil to be touched or used.

The word *taboo* is derived from the Tongan word *tabu* towards the end of the eighteenth century [1]. Its first use was found in the log entry for 15 June 1777 by Captain James Cook who was sent to Tahiti. In Togan, the languages of Polynesia, the word *tabu* was used to reference things that are forbidden, holy, or untouchable. At present, there are wide varieties of tabooed subjects including sex, death, illness, excretion, bodily functions, and religious matters [19, 41].

There exists numerous sociolinguistic studies of taboos for several languages but Nepali. Gao [14] investigated English taboos and euphemisms. She described seven different varieties of taboos in English: bodily excretions, death and disease, sex, four-letter words, swear words, privacy, and discriminatory language. Similar studies also exist in Igbo society in Nigeria [13], Algerian Society [15], Yemani Culture [30], Karonese Culture in Indonesia [4], Russian [16] and Finnish [17].

---

5https://dictionary.cambridge.org/us/dictionary/english/taboo
6https://www.ldoceonline.com/dictionary/taboo
Closest to our work is by Behzad et al. who studied linguistic taboos in Pahari culture, the culture of the hilly northern part of India, Pakistan, and Nepal [5]. Pahari language is the mother tongue of people of Jammu and Kashmir. Since Nepali is the most spoken language in Nepal and covers diverse cultures, the study of linguistic taboos in Pahari does not apply to Nepali.

2.2 Offensive Language

Offensive language is used to make hurtful, derogatory, and obscene comments to a target — usually an individual, organization, or a social group. As mentioned previously, offensive language can contain unmentionable as well as contextual linguistic taboos.

Nowadays people heavily use computer-mediated communication (CMC) such as social media, emails, and blogs to express opinions and exchange messages. CMC differs in many ways with traditional communication. Users can partially or fully anonymize their identities behind the Internet. The anonymity provides strength to express things that one cannot express in someone’s face [36]. With anonymity and physically distant from their audiences, offensive language is more likely to occur on the Internet and has a far-reaching and determinative impacts [36]. The proliferation of hate speech in CMC has created a huge research interest and momentum towards automated offense detection in social media such as identifying offensive posts, comments, and memes [32].

To the extent of our knowledge, there is not a single study of identifying offensive content in the Nepali language. In other languages, this has been studied under different topics including detection of hate speech [11, 12], abusive language [28, 29], cyberbullying [6, 37, 42], aggression [2, 9], insults [38], and identification of slang and vulgar words [33, 34]. Recently, shared task initiatives such as SemEval Task 6 [45] for English and GermEval [43] for German have united researchers to develop and evaluate their approaches in consistent ways. These efforts have produced resources such as training and testing data sets that are critical for identifying offensive content and their types and targets.

Although the goal of this paper is to study and characterize offensive language, taboos and euphemisms in Nepali, the data set of tabooed and contextually offensive terms that we have produced can serve as a baseline system for offense detection. Such system can flag a content in CMC (e.g. a social media post) as offensive if any of the terms in the data set is present in the content.

We follow a corpus-based study in which we investigate a data set of human conversation to identify tabooed and offensive terms. Social media such as Twitter [7], YouTube [8], Facebook [9] and Blogs are excellent sources of human conversation where users discuss a wide range of topics including politics, gender, religion, movies and television shows, and so on. They participate in the discussions by writing, sharing, reacting, and commenting. Some users in these conversations express their opinions which are insulting or offensive to others. These offensive content are of our interest as they contain tabooed and offensive terms or phrases.

3 METHODOLOGY

Over 90% of current Nepali social media users are young (age < 54 years) [10]. Thus, the offensive terms that we identify from the current social media posts are most likely the contemporary offensive terms popular among the younger generation.
Table 1. Sample Nepali social media posts and their English translations with offensive terms in Italics. Posts are written in Devanagari as well as in Transliterated Roman Script.

| Nepali (Devanagari): मासाला पागल भए जस्ती छ! | English: it seems he got pagal (mad) |
|------------------------------------------------|-----------------------------------|
| Nepali (Transliterated): muji talai ta ma thokchhruu goli saala desh drohi | English: muji (pubic hair) I will shoot you, traitor |
| Nepali (Transliterated): saale khaate aaphu matra educated thando raichha | English: saale khate (people living in slum) he thinks he is the only educated person |
| Nepali (Transliterated): bahun le desh bigare | English: bahun (an upper caste) spoiled the country |

3.1 Corpus to Study Offensive Language

Social media provides a platform for people to discuss in public, private, or secret forums. We are particularly interested in the conversations at public forums because any offensive language used in such places is not acceptable to the general audience. Since most of the posts in social media are non-offensive (e.g. around 0.2% swear word rate for Myspace [19]), it is highly likely that we obtain just a few offensive posts even while annotating a very large number of posts. Researchers apply different strategies to improve the distribution of offensive content [44]. Following these strategies, we made a pool of 7000 records from different social media sources: Facebook, Twitter, YouTube, Blogs and News sites.

Depending upon the platform, we use a slightly different mechanism. For Facebook, we first made a list of potentially controversial posts posted to a general audience in open groups and public pages between 2017 and 2019. We then extracted the comments corresponding to those posts. For Twitter, we followed a bootstrapping approach as done by prior arts [44]. For this, we first created a small list of Nepali words (in both Devanagari and Romanized forms) that have higher chances of being used in hate speech. The words themselves are not explicitly offensive but can appear in hate speech depending on the context of their use. For example, the words “बाहुन” (bahun - an upper caste in Nepali society) and “भालु” (bhalu - bear) are non-offensive by themselves but can appear in offensive contexts. Offensively, bahun can be used to insult someone racially based on their cast, and bhalu can be used to call someone a prostitute. Using the list of keywords, we performed a targeted search on Twitter and collected the tweets. These tweets enhanced the pool with diverse and context-sensitive posts.

For YouTube, similar to Facebook, we manually created a list of potentially controversial, non-controversial, and neutral videos, and extracted corresponding comments. The YouTube video comments also helped to maintain the diversity of data set in the writing form as they were typed in transliterated, mixed, and pure Devanagari font and fulfill our categorical requirements. Besides, they captured the inputs from the diversity of people commenting on the posts. Finally, the rest of the records were gathered from several Nepali blogs and news websites.

Sample social media posts that contained offensive language are shown in Table 1 along with the corresponding English translations. The offensive terms in the original Nepali are highlighted in italics. If the translation seems to undermine the offensiveness level, we kept the offensive term as is in the English translation. Note that the last example in the table “bahun le desh bigare” does not contain a single offensive word but the overall meaning of this post is offensive i.e. racist comment targeting a caste, bahun.

3.2 Data Set Creation

Three human annotators, all native Nepali speakers, studied the social media posts and identified the offensive terms. Confusing and borderline terms were discussed together and the final decision was taken in consensus. As
expected, the social media posts were very noisy and contained many variants of the same term. For example, the offensive term *khate* was written as *khatee*, *khate*, *khte*, *khatey* etc. Besides, the social media posts were written in Devanagari script and Romanized text (using phonetics) for the ease of typing. We extracted all of these raw variants, normalized them, and included in the data set.

We normalized the raw offensive terms and obtained their synonyms using Nepali Shabdakosh\(^{11}\) dictionary service. However, since the offensiveness of synonyms can be different from its term’s offensiveness, manual revision was required. For instance, *pisab* is a non-offensive synonym of the offensive term *mut* (urine). In total, we discovered 1077 offensive terms and made them publicly available\(^{12}\). We believe this first public data set of linguistic taboos for a low-resource Nepali language can facilitate the research towards this important topic.

The snippet of the data set is shown in Table 2. It consists of *RawRom* and *RawNep* designating raw terms in Romanized form and in Devanagari script respectively. Their normalized forms are stored at *NormNep* and *NormRom* columns. In addition, we labeled “1” for taboo terms and “0” for contextually taboo terms in *IsTaboo* column. We also provided a primary class type (more on Section 4) for each entry in *Class* column.

The statistics of the manually derived data set are presented in Table 3. Out of 1077 identified offensive terms, 158 (about 15%) were taboo terms. Body Part, Race, Relation, Sex, and Vulgar are the top categories that contained a higher proportion taboos. Categories such as Action, Cloth, Animal, Death, Geographic Location, Plant, Politics, Religion, and Weapon have very few or no taboo terms at all. Note that the terms in these categories always require contexts to be offensive. Interestingly, Politics, with 115 offensive terms, is the largest category. Race with 77 offensive terms is the second-largest category. Unlike in English, use of diverse animal names are observed in Nepali offensive language. With 72 terms, Animal category is the third-largest category in our data set. As expected, universal taboo topics such as Sex, Vulgur, and Body Parts are also in the top most popular categories. Lastly, relations being very important in Nepali culture, the offensive terms related to kinship are also very common – a unique characteristic of Nepali culture compared to the Western cultures.

### 4 VARIETIES OF LINGUISTIC OFFENSES IN NEPALI

We carefully studied the data set of offensive terms that we constructed in Section 3 and identified over eighteen interesting varieties (categories) such as Animal and Plant names, Race, and Sex. While some of these categories exist in other cultures and languages, others are unique to Nepali culture. Besides, the identified varieties are not mutually exclusive, i.e, a term can belong to multiple categories. Importantly, the main rationale for the selection

\(^{11}\)http://nepalishabdakosh.com

\(^{12}\)https://github.com/nowalab/offensive-nepali

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
of these categories is to explore several dimensions of linguistic offenses existing in the language, keeping in mind that these categories will largely assist to do a fine-grained classification in offense delection task. In this section, we describe these categories in detail.

1. **Action**: We identified many offensive comments that contained action verbs or expressions that were threatening, defaming, or described or called for harmful action to the target (e.g. a person or an organization). Examples:

   - Yeslai chhitai goli _thokna_ parcha (Need to shoot him soon)
   - Party karyalaya _aaago lagauna_ jaau (Let’s go to burn down the party office)

   Other actions include _katnu, pitnu, jalaunu, paataa kasnu, lakhetu, jel halnu, sutnu, chiknu, mutnu, chusnu, polnu, chhiraunu, latia hannu_ and so on. Most of the actions are context-sensitive i.e. offensiveness depends on their use in the language.

2. **Body Part**: Body parts are commonly tabooed across different cultures and languages. Female and male genital parts and pubic hair are tabooed outright while others need contexts to be tabooed. Our observation shows that body parts mostly appear in comments that are vulgar, racist, sexist, and personal insults. Commonly mentioned body parts include _puti, gula, chak, laando, geda, laantho, poka, thun, thun ko tuppo, junga, jhus, naak, kapal, taauko, muji_. Examples:

   - Sabaiko _taauko_ katera gift pathaidim (Let’s behead everyone and send the heads as a gift)
   - _Muji_ tero kaam chaina, turuntai rajinama de (_Muji_ you are useless, resign immediately)
Interestingly, the expressions following the pattern: (tera | tero | teri | mero | meri) + (baau | aama | hajurba | budi) + ko + body part (especially head—taauko or genital parts) are offensive, e.g., tero baau ko taauko (your father’s head), teri aama ko puti (your mother’s vagina), tero hajurba ko taauko (your grandfather’s head).

Because body parts are often required to be mentioned in normal conversations, euphemisms related to body parts are widely used in the Nepali. Wakṣhyastal and chhati are used for breast, guptaanga—private organs is used for male or female genitals, yoni is used for vagina, and so on. For instance, mahinawari bhako bela “yoni” safarakhnu jaruri cha (it is important to sanitize your vagina during menstruation). Interestingly, the term linga has multiple senses which can mean penis (a taboo term) or the word gender (non-taboo term). Shiva Linga (Lord Shiva’s penis), however, is a non-taboo term which is a revered icon in Shiva temple.

3. Cloth: Clothes that have direct connections with the internal body parts or sexual organs are contextually tabooed in Nepali language. Words such as thun choli, bra, petikot, taalo—sanitary pads, kanchuki, kachha, kattu etc. are commonly avoided in public and euphemisms are used in substitution. Aaimai ko bhitri basra or simply bhitri cholo is used instead of bra and thun choli. Similarly, bhitra lagaune is used for kachha, kattu, kanchuki etc. The use of alternatives are also contextual, for example bhitra lagaune can refer to underwear, but it can also refer to bra or petikot in a separate context. Example: mero bhitra lagaune khoi (where is my underwear).

Additionally, we identified clothes that are unrelated to internal body parts or sexual organs are contextually used in offensive languages. Words such as seto kapada or katro (shroud) are used to express personal attacks: talai katro odayera ghat ma purachau. Dhoti, topi, kandani, janai, etc are used to express racial offenses. Examples:

- **Dhoti** lai kholah khola lekhata parcha (Should chase this dhoti out towards the river belt)
- **Kandane (or toppe)** bahun

4. Animal: Animals are the commonly used metaphors in Nepali offense. Frequently used creatures in offensive dialect include baadar (monkey), gadha (donkey), kukur (dog), bheda (sheep), hanuman— a divine monkey in Hinduism—is used here to refer a fanatic follower, goru (ox), bheda (hyena), bwasha, bhaishi (buffalo), hatti (elephant), boka (he-goat), bakhri (she-goat) and pothi (hen). Most of these names are used in general offence, however, we also found some of them being used in sexist and racial comments. Examples:

- **Boka** haru bakhri khojna hide (bokas are moving in search of bakhri)
- Ta pani tah tai party ko hanuman hos ni (You are also a hanuman of the same party).
- Anuhar nai bhaishi ko jasto cha (Face looks like that of buffalo) (can be racist, depends on context)
- **Pothi** baseko suhaudaina (It does not suit a woman to raise her voice) (sexist idiom)

Interestingly, based on our observation, we found that the terms referring to female are more offensive than their male counterparts. For instance, kukur (dog) and baadar (monkey) are not offensive but their female versions kukurni (bitch) and baardash are taboo terms. This phenomenon is also observed in other cases e.g. (newar, newarni), (rai, raini) and (chor, chorni) – here, the masculine terms are generally non-offensive while the respective feminine terms are highly offensive.

5. Plant: Like animal names, plant names are also commonly used in offensive languages. Commonly used plant names include kera (banana), aalu (potato), mula (radish), pharsi (pumpkin), kandamul (wild plants), and paat (leaf). Examples:

- Oi gadha kera khanchas? (Hey donkey, do you want to eat my penis?)
- Tyo mula ko kura nagar (Don’t talk about that mula (radish))
- **Pharsi** ko poi (Husband of a fat women) (pharsi – pumpkin, here it means fat)

6. Death and Disease: Many words related to death and disease are contextually offensive and are frequently found in personal attacks. Examples include words such as pagal, cancer, haija related to diseases, and words such as hariyo bansh, katro, chihan, masan, maranchayshe, murda related to death. Examples:
Tero sarir le hariyo baas khojya jasto cha (Looks like your body is looking for a coffin) (context sensitive personal attack)

Yo pagal ho (He is a lunatic – direct offense)

Euphemisms related to death include bitnu, swargabash hunu, paralok hunu, and khasnu for marnu (die) and ghat for masan ghat (cemetery). Disease related euphemisms include manasik santulan thik nabhayeko for pagal (mentally ill person) and a generic term birami (sick) for all kind of diseases.

7. Disability: Disability is another category that accumulates words often tabooed across different cultures and languages. Disability terms such as lulo, langado, kaano, bahiyo, anhdo, psycho, kupro, lathebro, bhak-bhake, maranchayse are tabooed when associated with human. They are not tabooed when associated with non-human entities e.g. bahiyo sarkar (deaf government). Human rights and political movements have made many of these terms tabooed in recent years. Despite this, they are frequently used in offensive languages for personal insults and attacks. Examples:

- Bahiro chas ki kya ho? (Are you deaf or what?)
- Kulangar saape (You black-sheep lame) – extremely abusive phrase referring to a physically disable person
- Dekhdai khana na pako maranchyashe chas, k furti lagauchas. (You look emaciated, why are you showing off?)
- lati chhilli, lato jaat jani hal (dumb pervert, showing of her dumb caste)

Euphemisms are preferred when these terms are used for human. For example, euta aakha na bhako, kaan nasunne, manasik santulan thik na bhako are used instead of andho, bahiyo, and pagal respectively.

8. Excretions: Most of bodily excretions such as paad (fart), muut (urine), gu (shit), cher (watery diarrhea), and virya (sperm) are tabooed. However, some others such as aashu (tears), pasina (sweat), khakar (cough), dakar (burp) and ulti (vomit) are not. Verbs associated with the production of the tabooed bodily excretions are also tabooed e.g. padhnu (to fart) for paad (fart) and hagnu (excrete) for gu (shit). Examples:

- Kasto cher jasto aachar (This pickle looks like a shit)
- Ta gu ko poka hos (You are a pack of shit)
- Gobar bhariyeko tero dimag (Your mind is full of shit)

Euphemisms for tabooed excretions are commonly used in public and hospitals e.g. hawa instead of paad, aaka instead of gu / cher, susu instead of muut and seto pani instead of shukrakit. 

9. Marriage: Marriage is one of the most prominent traditions in Nepalese culture. There are several rituals related to marriage. These rituals along with the associated human participants are identified using specific terms such as dulaha (bridegroom), dulahi (bride), janti (participants from bridegroom side), daijo (dowry), kanya (unmarried young girl or damsel). These terms can appear in the offensive comments related to the same. Traditionally, divorce was a rare phenomenon in Nepali culture, and women seldom remarry after the death of the husband. Living together before marriage is still a taboo. These activities resulted in many linguistic taboos that are unique to Nepali culture compared to western. Examples include: bidhurwa (widow), bidhur (widower), batashe (a child whose father is unknown), rando (male client of prostitute). One interesting taboo expression is baau ko bihe, which means father’s marriage. The expression is used to threaten the target with the harm that is so unprecedented that it will be equivalent to seeing his (target’s) father’s marriage, a very rare event to perceive in Nepali culture. Examples:

- Bhaishi lai daijo ko lagi bihe garya hola (Married a bhaisi for dowry) (Bhaisi i.e. buffalo is used here as a metaphor to indicate a fat/dark girl)
- Kunai bidhawa ko aakha nalagos (May a widow’s curse not affect you)
- Batashe dherai nabol (Don’t you talk more, bastard)
10. Politics: Politics has been a long standing subject around the world for centuries. The advent of social media and the connectivity among the people have fostered the expression of frustrations and disagreements in several topics regarding policies, political parties and their activities. In Nepali language, we found the following words contextually used offensively in political conversations: deshdrohi (traitor), bhrastha (corrupt), bhastrachar (corruption), mandale (support of Nepali panchayat), maobadi (maoist), rastraghati (traitor), bikhandankari (separatist), hatyara (killer), puchhar (tail to refer a blind supporter), fascist, bheda (sheep – refer to fanatic follower), bha, and hanuman.

The words in this category are temporal and can change with the dynamics of politics. A word that was neutral in the past may now be offensive and may become obsolete in the future. Hanuman, for example, had a single positive meaning in the past i.e. exemplifying the devotion towards his master, lord Ram. Currently, this word is also used to refer a fanatic follower of a party/ideology in an offensive context. Examples:

- Deshdrohi lai goli thok saala (Shoot the traitor saala)
- Ram ko bachhan jharna pa chaina, hanuman haru le chhati pitna thali sake (Fanatic followers started obeying their master in no time)

11. Vulgar: Words that are explicitly and offensively used to refer sex, sexual acts, and human biological process are categorised as vulgar. People usually tend to avoid vulgar words, specifically sex talks and sexuality (menstruation, masturbation, and intercourse) in public as they are likely to cause anxiety. In Nepali too, almost all of the words that fall under this category are tabooed. Examples include: chiknu (have sex), gula, lado (penis), puti (vagina), jhatha (pubic hair), phusi (semen), thankanu (to have an erection). In addition, imported terms from English such as fuck, condom, cock, suck and blue film are tabooed too. These terms are used very often in offensive remarks. Examples:

- Tero aama lai chickne mai hu bujhis? (I am the one who fucked your mother, do you understand?)
- Lado kha (Eat my penis)

As discussed above, the direct use of these words are eschewed and euphemisms are used instead. For example, garba nirdhak sadhan for condom, swapnadosh for wet dream, purus or mahila ko guptaanga for lado or puti respectively, seto pani for briya and so on.

Please note that our classification being mutually inclusive, the words in this category can appear in other categories as well.

12. Race: Nepal is a land of racial, ethnic and linguistic diversity. There are more than 125 ethnic groups and languages [26] in the country. Bahun, Chhetri, Tamang, Magar, Newar, Gurung, Kami, Muslim, Yadav, and Rai are some of the most populous races. Designating people by their ethnic groups has a long history, starting from religion and followed by rulers, dictators, aristocratic cultures, patricians, and so on. Several Hindu holy scripts categorize people as untouchable [40] (baishya, sudra, caste-less etc.) and educated or upper-caste (Brahmin, Chhetri etc.), to avoid eating food touched by lower-class people, and even touching them (e.g. Damai, Kami le choko khana hudaina). Traditionally, the upper-caste people are downgraded should they eat the food touched by the lower-caste people or marry them. Due to political awareness and human rights activities, the narratives are changing, and the caste system is slowly fading. But still, casteist comments using these terms are highly prevalent in contemporary society. Examples:

- Oi bhote, noon bechna chodis? (Hey bhote, have you stopped selling salt?)
- Bahun tero tuppi ukheldim? (Bahun, can I pluck your tuppi?). Tuppi – sikha is a tuft of hair left on the top of shaved head of traditional Hindu; stereotypically associated with Brahmin.
- Yesko buddhi kami kaha lagera sojhauna parcha (His head is not right, need to take him to kami – blacksmith to straighten it)
- Rajako kaam chhodi kamiko dewali (Going to kami’s festival over king’s assignment) is a traditionally non-tabooed idiom that is considered racist now.
We also found abundant words and phrases that are used to make racist comments. It includes dhoti, madhise, tapare, pate, chimre, bhote, toppa, paahade, ganaune, kote, sarki, jyami, sinu khane jaat (carcass eating people), sikarmi, dakarmi, and pode. The use of these words to a totally unrelated people are offensive outright. Substitution for caste names such as bahun baje for bahun, kaji for chhetri, sahu for newar, mijar for sarki, and bishwokarma for kami can serve as euphemisms if they are used properly.

13. Relation: Relational offensive terms in Nepali include both human-human relations such as ‘maa’ chickne (‘mother’ fucker) and randi ko ‘chhora’ (son of a prostitute) and human-nonhuman relation such as raandi ko ‘baan’ (widow’s ‘rider’). Other linguistic taboos in this category include harami (bastard), chutiya, rakhel, raandi, batashe (one without a known father) and rando which are outright offensive. Contextually offensive terms in this category include bidhuwa, dadha, logne, baau, kuputra, saali, and bhena. Examples:

- Talai kina chayieo? tero baule banako ho ra? (Why do you need it? did your father made it?)
- Raandi ko chhora, talai pakh na maile janya chu (Son of widow, you’ll see what I am going to do)

The popular euphemism for a widow is ekal mahila (single women) which also refers an unmarried/divorced woman.

14. Gender: Gender-related taboos and offensive terms exist in most of the cultures. Pothi baseko ramro hudaina, thangne aaimai, gireki aaimai are historically prevalent sexist phrases in Nepali household. Discrimination based on a person’s gender is extremely common in Nepal. Mostly, females and LGBT are found to be affected by this segregation. Several words for making gender bias such as ghopte, kukhuri, pothi, bakhri, budi kanya, boka, bokie, utauli, nakharmauli, chothale, maal (slang referring to a female) are routinely used in offensive remarks. The discriminatory and derogatory use of certain words such as aaimai, taruni, taruno, bidhuwa have changed their actual meanings over the course of time. Today, direct use of these words brings discomfort and are mostly avoided or substituted by euphemisms in public use. Additionally, observing thousands of social media comments, we found that a set of words such as ge, hijada, napubshah, lesbian – supposed to refer LGBTQ are misleadingly used to offend others. Examples:

- Bidhuwa le bisait pari (Widow brought a bad luck)
- Hijada jastai chhe (She looks like a transgender)
- Thangne aaimai (Dirty women)
- Yo aaimai ko jaat lai jati bhane pani lagdaina, jaat janai halchan (A sexist phrase which semantically means “All women are foolish”)

15. Weapon: Weapon names are rarely considered offensive when used in general conversation. Our observation shows that they are occasionally used to threaten or attack others. Commonly used terms includes khukuri (Nepali knife), chura (small knife), tarwar (long knife), goli (bullet), bhala (spear), and bikh (poison). Interestingly, we also observed these words being frequently used on posts related to political disagreements, social criminals, and racist bigotry. Examples:

- Bhala le hanna parcha yeslai (Need to strike him with a spear)
- Bhukdai gar, chadai khukuri khanchas taile (Keep barking, soon you will be struck with a khukuri – knife)

16. Religion: There are many religious taboo activities in Nepal such as wearing shoes in temples, touching god’s sculpture with the left hand and so on. The religious linguistic taboos in Nepali language are mostly related to death, devil, mystical words (associated with religious texts), and God names. And these taboos are mostly context-dependent e.g. raakshyas, bhoot, masan, daanab, raawan. Below are some examples of offensive comments in this category:

- Manav rupi daanab hun ini (He is an evil in the face of man).
12 • Niraula, Dulal and Koirala

- **Kukarmi daitya**, tero narka ma bas hos (Immoral evil, I wish you go to hell).

17. **Geographic Location**: Offending people based on their geographic location or country is a widely discussed topic in Nepal. There are routinely used racial slurs to discriminate people based on their origins. Slurs such as *bihari*, *indian* and *madhishe* are used for people from Terai (Southern plain) region and the slurs *bhothe* and *pahade* are used for people from Pahad (Northern Himalayan and hilly) region. In fact, we encountered hundreds of comments containing these abusive terms in our corpus. In addition, *jhapali*, *sangjali*, *jangali*, *khaire* were also abundant in the corpus. We also identified a few specific locations referring to red light areas such as *ratna park* and *balaju bus park*. Some examples of offensive comments in this category are given below:

- Herdai *ratna-park* ma beluki grahak khojne jasti che (She looks like the one who seeks for the costumers at ratna-park in the evening. In local context, the whole sentence is citing a woman as a whore)
- Thangne *bhoteni* (Dirty bhoteni). Here *bhoteni* refers to a women originated from Tibet.

18. **Other**: There cannot be a closed set of offensive varieties. Below we list some other interesting varieties that are not previously covered.

A. **Composite Offensive Terms**: Complex offensive terms are formed using the individual offensive terms such as *dhoti ko chhauro* and *muji bhatuwa*. More common case is to use a taboo term with a contextually offensive term such as *randi ko baan*, *saala kukur*, *gobar gwatch* and *tapare bahun*. Although these terms mostly collocate as bigrams, they might appear at different locations in a sentence. In the sentence “bahun jati sabai tapare nai bhaye”, the composite term *tapare bahun* is formed using *bahun* and *tapare* appearing at different locations.

In general, the degree of offensiveness increases with the increase in offensive terms count. For instance, the expression *muji bhatuwa* is more offensive than the individual offensive terms *muji* and *bhatuwa*.

B. **Imported Terms**: There is a large list of imported offensive terms from other local and foreign languages such as Hindi (*dekha jayega, murga, ulluka patha, behen chod, bhosadi*), Arabic (*aaulad, saitan*) and English (*over, virgin, blue, bra, saiko – psycho, fuck*).

C. **Code Mixing**: Code switching is a process of forming a composite term by mixing more than one language. Examples of offensive terms due to code switching include *over nabol* (English + Nepali), *randiko aaulad* (Nepali + Arabic), and *ta murga* (Nepali + Hindi)

D. **Exclamation**: Also known as Bismayadibodhak in Nepali, exclamations in communication pertain to strong emotions, especially anger, surprise, dissatisfaction, or agony. Words such as *thukka, thuiya, chya, murdabad, jindabaad, thu-thu, bajo paros, mari ja* etc. are a few examples of exclamatory terms in Nepali that can appear in offensive comments.

5 **EUPHEMISMS IN NEPAL**

Sometimes the use of taboo terms is unavoidable, e.g., patients describing doctors about their genital-related issues, talking with kids about bodily excretions, mentioning disabled persons in a conversation, and so on. Euphemisms are the ways people use to lessen or completely remove the unpleasantness of taboo expressions. They are often the replacement words or phrases that substitute the offensive terms and expressions to neutralize or reduce the nuisance [14]. For example, in English the phrase *make love* is preferred over *fuck* in a conversation with mixed participants [20, 23]. Similarly, instead of using *underdeveloped nations* for Third World countries, the United Nations calls them *less-developed countries* [14]. In this section, we discuss common types of euphemisms used in Nepali. Many of these euphemisms are applicable to the offense varieties we listed in Section 4.

- (1) **Synonym**: Synonyms of a term can help to eliminate the offensiveness of a taboo term. The non-offensive terms *rajkhani, pisab* and *disa* are the synonyms for taboo terms *geda* (testis), *mut* and *gu* respectively.

---

13 Bhote, also known as Bhotiya are group of people related with Tibetan ethnically or linguistically
Word Replacement: The changes in sociopolitical structure and human right movements have helped to coin and encouraged to use non-offensive terms instead of tabooed terms: Ykal mahila for bidhuwa, Jestha Nagarik for buda-budi, Tesro Lingi for Hijada, Sadak bal-balika for Khaate and Yaunkarmi for Besya etc.

Circumlocution: Circumlocution is an evasive way of expression using multiple words rather than being explicit with fewer and apter words. It is commonly practiced to avoid the use of taboo terms. One popular way of circumlocution is to use the definitions of offensive terms. Examples include Pisab ferne anga / Pisab aaune thaau instead of laado or turi for man and puti for women.

Word Translation: Sometimes translation to another language can enfeeble the ofensiveness of a taboo term. In Nepali, the translation is typically done in English or Sanskrit for euphemism. It is easier to use mens or period instead of nachhune menstruation. Similarly, stool is commonly preferred over gu in hospitals. Occasionally, the translated forms are shorter than the offensive terms and thus are preferred. One such example is nipple instead of dukh ko tupper.

Child Language: People in any society have diferent terms that replace the corresponding tabooed terms when talking with a child. Here are few examples in Nepali: aaka for gu, susu for mut, turi-turi or simply turi for lado (male genital), hawa for paad etc.

Suffix: Addressing a person with his first name is very common in western societies. In Nepali culture, which has an honorific system, addressing people having more honor (e.g. parents) with their names is considered offensive. Use of person names, however, is common among the peers or to the persons with lower honors, e.g., children. There are some suffixes (e.g. jee and jyu) that can be attached to person names to increase the politeness. For example, using Ramjee for Ram and Sitajee for Sita in a conversation can provide more honor than using the direct names. Adding jee after a person’s name can also mean the speaker has a good relationship with the person and therefore is common across friends. Also, jee can be used after the surnames (e.g. Dahaljee) or nicknames (e.g. Prachanda jee).

Use of jyu provides respects to superiors, e.g., ministers, presidents and managers. Unlike the suffix jee, this suffix is mostly used after the complete name of a person e.g. Madhav Nepal jyu. It is also commonly used in formal writing, official letters, legalese and public speech.

Family Alias: Using alias for names inside a family is a common practice in Nepali culture. Instead of using husband’s name directly (which is considered abnormal), a wife uses aliases such as “X ko Baba” (where “X” refers to a child), baba, and daddy. Similarly, aliases for wife include sani, kanchhi, and “X ki ama”. However, parents can directly use the first names of their children. Also, it is common to use rank alias for children e.g. jetho/jethi (the first male/female child), mailo/maili (the second male/female child), kancho/kanchhi (the last male/female child), etc.

Relation Name: This is probably the most widely used approach in Nepali culture to avoid using person name in a conversation. Examples include relations to superiors in the family tree (hajurba/hajurama, buwa/ama), elder siblings (daju/didi), younger siblings (bhai/bahini), rank of siblings by order (jetha/jethi, maila/maili, saila/saili, kaila/kaili, antare/antari, thaila/thaili, jantare/jantarni, kanchha/kanchhi), peers (sathi, mitra), inferiors (chhora/chhori, bhanja/bhanjee, keta/keti) and other aliases (sane/sani, maya, mayalu, maiya, mitjiyu, mitini) etc.

Title: People of high political rank or that are of a prominent post of an organization are addressed using their titles or titles followed by their names. Examples include Pradhan Mantri (Prime Minister), Upapradhan mantri (Vice-president), Rastrapati (President of a country), Sabhamukh (House Speaker), Adyakchya (President), Swastha Mantrin (Health Minister), Purba Mantri (Ex-minister), etc. Adding “Sir” after the name is also a popular practice e.g. Hari sir, Madhav sir, etc.

Metaphor: Metaphoric expressions are also used as euphemisms. If someone is very talkative (negative connotation), he is called Narad, a Vedic sage (in Hindu religion) who is a traveling musician, storyteller,
news carries, and enlightening wisdom. Other examples include: *dhunga jasto mutu* (heart is like a stone) for *nisthuri* (cruel), *Gobar Ganesh* for *pat murkha* (stupid).

11) **Idiom**: Idioms are very common in both written and spoken languages. Some examples in this category include: *sangai uth bas garnu* (up down together) for sex, *para sarnu* or *nachhune hunu* for menstruation, *swarga ma baas hunu* (move to heaven) for death, and *chura phutnu* (to get bangle broken) for bidhuwa hunu (to be a widow).

12) **Proverb**: Proverbs can be used as euphemisms. One such example is *bhukne kukurle tokdaina* (barking dog seldom bites) which can contextually mean *bataure* (braggart or blowhard).

Interestingly, euphemisms more than needed can reverse the purpose of euphemism [14]. The expression “*Yaauna karmi ko pisab pherne anga*” has a negative meaning although euphemistic terms *Yaauna karmi* and *pisab pherne anga* are used in the expression.

6 OFFENSIVE LANGUAGE DETECTION

Detection of hate speech and offensive languages is accelerating in recent years [35]. The major works have been conducted in resource-rich languages such as English and German. Such research is further facilitated by recent competitions and shared tasks that make availability of gold training examples. Toxic Comment Classification Challenge by Kaggle[14], for example, provides thousands of human-labeled examples for detecting toxic behaviors in Wikipedia comments. Similarly, Shared Tasks on Aggression Identification for Hindi and English [24], and Germeval [43] for German provide gold data sets for detecting offensive languages. Schmidt and Wiegand [35] provide a general survey of hate speech detection including different approaches, data sets and annotation strategies. Offensive language detection in resource-poor language Nepali, however, is still in its infancy. Although research in related languages such as Hindi [7, 21, 25] and Bangla [3, 18] exists, due to different cultural and ethnic diversities, and vocabulary differences, such methods do not work for Nepali. In this section, we describe the first experiments on detecting offensive language in Nepali social media content using Natural Language Processing.

Offensive language detection problem is typically modeled as a text classification problem. For instance, a record can be categorized into offensive or non-offensive using a binary classifier. In this study, we experiment with a rule-based system and a machine learning approach as binary classifiers to discover Nepali offensive comments.

Our rule-based approach utilizes the list of 1077 Nepali offensive terms that we collected and described in Section 3. Each term, as shown in Table 2, has it’s original, romanized, and transliterated form. We augmented this data set with an offensiveness score for each term. The score ranges from 1 (non-offensive) to 5 (very likely offensive, e.g. taboo terms). To get the offensiveness scores for terms, three native speakers initially assigned their scores independently for each of these terms. The speakers then discussed together to finalize the final offensiveness scores on consensus. It is important to emphasize to readers that the offensiveness (score) of a term depends on its use in the sentence and we should not completely rely on it to automatically reject comments without looking at the context. However, the score helps to develop baseline algorithms that can be used to flag records for human review as well as to compare the performances of sophisticated algorithms. For instance, in the sentiment analysis task, Nielsen’s AFINN-111 word list is hugely popular which has 2477 sentiment-bearing words with corresponding sentiment scores [46]. The AFINN-111 data set has been used to make several baseline methods for sentiment analysis.

With the word list containing offensiveness scores, several strategies can be implemented for making rule-based systems as done for sentiment analysis by Nielsen [46] (e.g. sum of scores, an average of scores, and normalized scores by the length of tokens). Our rule-based baseline algorithm uses the “sum of scores” approach [14]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
and works as follows. Given a post, it looks for the tokens that are present in our dictionary of offensive terms and sums the corresponding offensiveness scores. If the sum is 5 or more, it declares the post as offensive.

We experimented with two machine learning approaches: (a) Logistic Regression (LR) with L2 regularization and limited-memory BFGS optimization, and (b) Support Vector Machine (SVM) with a linear kernel. The models use unigram and bigram word features as well as trigram character features. We performed feature reduction to limit the features to 10000 using Chi-squared stats. To train the models, we utilized a human-annotated data set for offensive language detection that we have prepared in a separate study [27]. The data set consists of 7248 Nepali social media posts with human assigned labels for each as offensive (positive class) and non-offensiveness (negative class). The data set has been already split into train and test sets. The train data set consists of 3562 negative and 2236 positive examples and the test data set consists of 896 negative and 554 positive examples.

We reported precision, recall, and F1 scores, commonly used metrics for text classification problems, of the baseline, LR, and SVM models on the test data set in Figure 1. The rule-based system obtained F1 scores of 0.73 and 0.58 for non-offensive and offensive categories, respectively. For non-offensive category, both the LR and
SVM outperformed the baseline model with 0.85 and 0.83 $F_1$ scores, respectively. For the offensive category, the LR model performed very well compared to the other approaches with $F_1$ the score of 0.68. Although the SVM model had the best precision for this category (0.92), it did not get higher $F_1$ score due to its lowest recall. Interestingly, the baseline system had competitive recall (0.58) for offensive category but it suffered heavily in the precision score (0.57) compared to the SVM and LR models. This is expected as the machine learning approaches leverage the context through the features while the baseline approach relies solely on the assigned offensiveness scores without looking at the contextual information. In addition, we have used character features for machine learning models to address spelling variations of terms to some extent. The low precision of the baseline model strongly suggests that we should not simply flag and reject social media posts solely using the collected terms because legitimate posts can still contain these terms in non-offensive context.

Although the performance of a hate speech detection system depends on many factors such as the language, available resources, annotated examples and approaches, our results are on par with similar studies in other closely related languages such as Hindi [8] and Bangla [10]. Bohra et al. [8] obtained an overall accuracy of 71.7% in detecting offensive Hindi tweets. Similarly, Das et al. [10] reported $F_1$ score of 78% in detecting offensive Bangla Facebook comments.

Our experiment here is one of the first experiments in detecting offensive language in Nepali and provides strong baselines for future studies. The $F_1$ score for the offensive category is not high (e.g. the highest score is 0.68 for LR) but it is still very encouraging given the difficulty of the task itself. Classifiers are making mistakes due to spelling variations of terms (out-of-vocabulary issues), lack of representation for word and offensive context. Additional steps can be utilized to further enhance the performance e.g. advanced preprocessing, additional training examples, and leveraging deep learning approaches. We explore these options in our future studies.

7 CONCLUSION AND FUTURE WORK

Social media are excellent platforms for sharing information. They are also misused to spread hate speech through offensive posts and comments. With enormous amount of content generated every day, automated methods are required to prevent them from spreading. Because offensiveness is a socio-cultural and situational phenomenon, language agnostic automated methods do not work for hate speech detection. Language specific study is necessary to understand what can be considered offensive. To this end, we presented our corpus-based study of offensive language in Nepali. We primarily focused on linguistic taboos and euphemisms. We believe this is the first study of its kind.

We collected and studied over 7000 social media posts containing high degree of linguistic taboos and offensive terms. Furthermore, we carefully extracted and characterized over 1000 tabooed and contextually offensive terms into 18 different varieties. The annotated data set is made publicly available at Github\textsuperscript{15}. Besides the universal taboo topics like Sex, Body Part, and Excretion, we identified many varieties that are unique to Nepali language and culture e.g. Relation and Plant Names. Since no related prior work or resource exists for Nepali, the study and the data set lay a foundation for future research on this important topic. For example, in hate speech detection task, a baseline system can be built from the aforementioned data set that can flag content (e.g. a Tweet or Reddit post) as offensive if the terms in the data set appear in the content. It can further help to do a fine grained classification of linguistic offenses.

We also studied euphemisms – ways for people to lessen or completely remove the offensiveness. We discussed 12 common types of euphemisms in Nepali. While euphemisms such as Circumlocution and Child Language are common across the cultures in the world, many other discovered euphemisms are specific and unique to Nepali culture.

\textsuperscript{15}https://github.com/nowalab/offensive-nepali

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
We described experiments on detecting offensive languages in Nepali using rule-based and machine learning based approaches. Our experiment provided baseline results for offensive language detection problem in Nepali to fight against racism, discrimination and cyberbullying.

Our future work includes constructing an ontology of offensive categories. For example, many of the identified varieties such as Disability, Race, Religion, Gender, and Location can be put together under a higher-order Discriminatory category. Likewise, Plant and Animal can be grouped, and so on. The study of behavioral taboos is another future topic in this direction. Future work in offensive language detection includes applying advanced preprocessing techniques, generating additional training examples, leveraging state-of-the-art deep learning approaches and exploring fine-grained classification of offensive comments.

ACKNOWLEDGMENTS

We are thankful for the anonymous reviewers who spent their time reviewing our paper and providing critical feedback. We are grateful to Professor Dr. Kumar Prasad Koirala and Mr. Daniel Whyatt for their valuable inputs and inspiration. We would also like to acknowledge Nepali Shabdakosh (http://nepalishabdakosh.com) for providing synonyms of the offensive terms.

REFERENCES

[1] Keith Allan and Kate Burridge. 2006. Taboos and their origins. Forbidden words: Taboo and the censoring of language (2006), 1–28.
[2] Segun Taeove Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). 90–97.
[3] Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Aggression identification in english, hindi and bangla text using bert, roberta and svm. In Proceedings of the second workshop on trolling, aggression and cyberbullying. 76–82.
[4] Jumati Barus, Robert Sibarani, Amirn Saragh, et al. 2018. Linguistic Taboos in Karenese Culture. KnE Social Sciences (2018), 411–421.
[5] Anwar Behzad, Waseem Malik, and Sara Azam. 2017. Linguistic Taboos in the Pahari Culture: A Sociolinguistic Analysis. ARIEL-An International Research Journal of English Language and Literature 27 (2017).
[6] Amy Bellmore, Angela J Calvin, Jun-Ming Xu, and Xiaojin Zhu. 2015. The five W’s of “bullying” on Twitter: Who, what, why, where, and when. Computers in human behavior 44 (2015), 305–314.
[7] Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tannoy Chakraborty. 2020. Hostility detection dataset in hindi. arXiv preprint arXiv:2011.03588 (2020).
[8] 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. 36–41.
[9] Despoina Chatzakou, Nicolas Kourkellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In Proceedings of the 2017 ACM on web science conference. 13–22.
[10] Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems 30, 1 (2021), 578–591.
[11] Thomas Davidson, Dana Warmshley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Eleventh international aaai conference on web and social media.
[12] Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). 86–95.
[13] Gbenga Fakuade, Ngozi Kendirim, Ikechukwu Nnaji, and Florence Nwosu. 2013. Linguistic taboos in the Igbo society: A sociolinguistic investigation. Language Discourse & Society 2, 2 (2013), 117–152.
[14] Chunming Gao. 2013. A sociolinguistic study of English taboo language. Theory and Practice in Language Studies 3, 12 (2013), 2310.
[15] Nadia Ghounane. 2014. A sociolinguistic view of linguistic taboos and euphemistic strategies in the Algerian Society: Attitudes and beliefs in Tlemcen speech community. International Journal of Research in Applied, Natural and Social Sciences 2, 3 (2014), 73–88.
[16] Carole Teresa Greene. 2001. The use of euphemisms and taboo terms by young speakers of Russian and English. (2001).
[17] Minna Hjort. 2017. Swearing in Finnish. Advances in Swearing Research: New Languages and New Contexts. Amsterdam/Philadelphia: John Benjamins (2017), 231–56.
[18] Md Gulzar Hussain, Tamim Al Mahmud, and Waheda Akhtar. 2018. An approach to detect abusive bangla text. In 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE, 1–5.
[19] Timothy Jay. 2009. The utility and ubiquity of taboo words. Perspectives on psychological science 4, 2 (2009), 153–161.

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
[20] Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture* 4, 2 (2008), 267–288.

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

[22] Kepa Korta and John Perry. 2020. Pragmatics. https://plato.stanford.edu/archives/spr2020/entries/pragmatics/. (2020).

[23] Johan Tobias Kristiano and Priyatno Ardi. 2020. Swear words in Bad Boys II: A semantic analysis. *LLT Journal: A Journal on Language and Language Teaching* 21, 2 (2018), 191–198.

[24] Ritik Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 1–11.

[25] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. 18–26.

[26] Nepal Government NG. 2014. Population Monograph of Nepal, Volume II, Social Demographics. *Central Bureau of Statistics* II (2014).

[27] Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive Language Detection in Nepali. In *Press. The 5th Workshop on Online Abuse and Harms, ACL* (2021).

[28] Chikashi Nobata, Joel Tetreault, Achithomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online social media. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[29] Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206* (2017).

[30] Nada Qanbar. 2011. A sociolinguistic study of the linguistic taboos in the Yemeni society. *Modern Journal of Applied Linguistics* 3, 2 (2011), 86–104.

[31] Allen Walker Read. 2004. The geolinguistics of verbal taboo. *ETC: A Review of General Semantics* 61, 4 (2004).

[32] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. *arXiv preprint arXiv:1910.02334* (2019).

[33] JI Saini and AA Desai. [n.d.]. Identification of Slang Words Used in Pornographic Unsolicited Bulk Emails. *Journal of SCI-TECH Research (JSTR)* [n.d.], 4–9.

[34] Jatinderkumar R Saini and Apurva A Desai. 2011. Identification of Hindi Words Used in Pornographic Unsolicited Bulk E-Mails. *IUP Journal of Systems Management* 9, 2 (2011).

[35] 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*. 1–10.

[36] LaShel Shaw. 2011. Hate speech in cyberspace: Bitterness without boundaries. *Notre Dame JL Ethics & Pub. Pol’y* 25 (2011), 279.

[37] Peter K Smith. 2015. The nature of cyberbullying and what we can do about it. *Journal of Research in Special Educational Needs* 15, 3 (2015), 176–184.

[38] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.

[39] Teresa Stone and Lyn Francis. 2010. What’s the bloody law on this? Nurses, swearing, and the law in New South Wales, Australia. *Contemporary nurse* 34, 2 (2010), 246–257.

[40] Trincenter. [n.d.]. Racism In Hinduism. http://www.trincenter.com/more/India/invasionsandracism.htm. Accessed: 2020-07-06.

[41] Ronald Wardhaugh. 2011. An Introduction to sociolinguistics. Vol. 28. John Wiley & Sons.

[42] Elizabeth Whitaker and Robin M Kowalski. 2015. Cyberbullying via social media. *Journal of school violence* 14, 1 (2015), 11–29.

[43] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. (2018).

[44] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. (2019). 1415–1420.

[45] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983* (2019).

[46] Finn Aruprup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR abs/1103.2903* (2011). arXiv:1103.2903 http://arxiv.org/abs/1103.2903

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
### A GLOSSARY

#### A.1 Phrase and Sentence-level

| Phrase/Sentence                  | English Meaning                                      |
|----------------------------------|------------------------------------------------------|
| maa chickne                      | mother fucker                                        |
| aaimai ko bhitri bastra          | bra; an undergarment worn by women to support the breasts |
| talai katro odayera ghat ma purauchu | will cover you using with shroud and will take you to cemetery |
| anuhar nai bhaishi ko jasto cha   | face looks like that of buffalo, can be racist, depends on context |
| baau ko bihe                     | father’s marriage                                    |
| baho r chas ki kya ho?           | are you deaf or what?                                |
| baho sarkar                      | deaf government                                      |
| bahun jati sabap paid na bhaaye   | all bahur are greedy                                 |
| bahun le desh bigare             | bahun (an upper caste) spoiled the country           |
| bahun tero tuppi ukheldim?       | bahun, can I pluck your tuppi?                       |
| bajo paros                       | curse to have no child’s                             |
| batahes dherai nabol             | don’t you talk more, bastard                         |
| bhaishi lai dajo ko lagi bihe garyo hola | married a bhaishi for dowry                         |
| bhala le hanna parcha yeslai     | need to strike him with a spear                      |
| bhitra lagaune                   | underwear, bra, petikot                              |
| bhukdai gar, chadai khukuri khanchas taile | keep barking, soon you will be struck with a khukuri – knife |
| bhukne kuchurle tokdaina         | barking dog seldom bites                             |
| bidhuwa le bisait pari           | widow brought a bad luck                             |
| boka haru bakhri khojna hide      | bokas are moving in search of bakhri                 |
| damai, kami le choko khana hudaina | we should not eat foods touched by lower cast         |
| dekhdaik khana na pako maranchyashe chas, k furti laagachas. | you look emaciated, why are you showing off          |
| deshdhoani lai goli thok saala   | shoot the traitor saala                              |
| dhoti lai kholai khola lekhatna parcha | should chase this dhoti out towards the river belt |
| dhunga jasto mutu                | heart is like a stone, cruel                         |
| elder siblings (daju/didi)       | younger siblings(bhai/bahini)                        |
| gadhale bhi bokchha               | donkey carries load                                  |
| garba nirodhak sadhan            | condom                                               |
| ghat for masan ghat              | cemetery                                             |
| gobar bhariyeko tero dimag       | your mind is full of shit                            |
| gobar ganesh/pat murkha          | stupid                                               |
| goli hanno                       | shot him                                             |
| herdai ratna-park ma beluki graham khojne jasti che | she looks like the one who seeks for the cos- tumers at ratna-park in the evening, woman as a whore |
| English                                                                 | Nepali                                                                 |
|-----------------------------------------------------------------------|------------------------------------------------------------------------|
| she looks like a transgender                                         | hijada jastai chhe                                                     |
| Khaire                                                                | jangali                                                               |
| this pickle looks like a shit                                         | kasto cher jasto aachar                                                |
| immoral evil, I wish you go to hell                                   | kukarmi daiyta, tero narka ma bas hos                                 |
| you black-sheep lame                                                  | kulangar saape                                                         |
| may a widow’s curse not affect you                                    | kunai bidhuwa ko aakha nalagos                                         |
| eat my penis                                                          | lado kha                                                                |
| dumb pervert, showing off her dumb caste                              | lati chhilli, lato jaat janai hali                                     |
| it is important to sanitize your vagina during menstruation          | mahinawari bhako bela “yoni” safra rakhnu jaruri cha                   |
| lunatic; mad                                                          | manasik santulan thik nabhayeko                                        |
| he is an evil in the face of man                                      | manav rupi daanab hun ini                                              |
| where is my underwear                                                 | mero bhita lagaune khoi                                                 |
| muji (pubic hair) I will shoot you, traitor                           | muji talai ta ma thokchhu goli saala desh drohi                       |
| hey bhote, have you stopped selling salt?                             | oi bhote, noon bechna chodis?                                           |
| hey donkey, do you want to eat my penis?                              | oi gadha kera khanchas?                                                |
| and chura phutnu (to get bangle broken) for hidhuwa hunu (to be a widow) | para sarnu or nachhune hunu for menstruation                           |
| let’s go to burn down the party office                                | party karyalaya aaago lagauna jaau                                      |
| husband of a fat women, pumpkin, here it means fat                    | pharsi ko poi                                                           |
| genital                                                               | pisab ferne anga / pisab aaune thau                                  |
| dirty women                                                           | thangne aaimai, gireki aaimai                                          |
| it does not suit a woman to raise her voice                           | pothi baseko suhaudaina                                                |
| widow’s ‘rider’                                                       | raandi ko ‘baan’                                                       |
| son of widow, you’ll see what I am going to do                        | raandi ko chhora, talai pakh na male janka chu                        |
| going to kami’s festival over king’s assignment                       | rajako kaam chhodi kamiko dewali                                      |
| fanatic followers started obeying their master in no time             | ram ko bachhan jharna pa chaina, hanuman haru le chhuti pitma thali sake |
| son of a prostitute                                                   | randi ko ‘chhora’                                                      |
| saale khate (people living in slum) he thinks he is the only educated person | saale khaate aaphu matra educated thando raichua                      |
| shroud                                                                | seto kapada or katro                                                   |
| lord shiva’s penis                                                   | shiva linga                                                            |
| died                                                                  | swargabash hunu ,paralok hunu                                           |
| you are a donkey                                                      | ta gadha hos                                                           |
| you are a pack of shit                                                | ta gu ko poka hos                                                       |
| you chicken                                                           | ta murga                                                               |
| you are also a hanuman of the same party                              | ta pani tah tai party ko hanuman hos ni                                |
| why do you need it? did your father made it?                          | talai kina chayieo? tero baule banako ho ra?                           |
|  المركبية للغة الإنجليزية واللغة النيبالية، بما في ذلك الجمل التي تم ترجمتها إلى اللغة النيبالية من خلال المكتبة، تم تحويلها إلى النص اليدوي واستخدامه في نص الشبكة في الأجهزة البلدية. |
| Original        | Meaning                  | Original        | Meaning                  |
|-----------------|--------------------------|-----------------|--------------------------|
| tan kukur hos   | you are a dog            | teri aama ko puti | your mother’s vagina     |
| tero aama lai chickne mai hu bujhis? | I am the one who fucked your mother, do you understand? | tero baau ko taauko | your father’s head       |
| tero hajurba ko taauko | your grandfather’s head | tero sarir le hariyo baas khojya jasto cha | looks like your body is looking for a coffin |
| thangne aaimai  | dirty women              | thangne bhoteni  | dirty bhoteni (women originated from tibet) |
| tuo mula ko kura nagar | don’t talk about that mula (radish) | yesko buddhi kami kaha lagera sojhauna parcha | his head is not right, need to take him to kami – blacksmith to straighten it |
| yeslai chhitai goli thokna parcha | need to shoot him soon | yo aaimai ko jaat lai jati bhane pani lagdaina, jaat janai halchan | a sexist phrase which semantically means “all women are foolish” |
| yo pagal ho     | he is a lunatic          | yeslai chhitai goli thokna parcha | need to shoot him soon |

Table 4. Transliterated Nepali sentences and corresponding English translations of their contextual meaning as used in this paper.

A.2 Word-level

| Original        | Meaning                  | Original        | Meaning                  |
|-----------------|--------------------------|-----------------|--------------------------|
| aachar          | pickle                   | kaji            | minister or similar level |
| aaimai          | women                    | kanchha/kanchhi | the last male/female child |
| aaka            | stole or poop            | kanchuki        | corselette               |
| aalu            | potato                   | kandamul        | tuber                    |
| aama            | mother                   | kandane/toppe   | slang use for bahun      |
| aashu           | tears                    | kandani         | waistband; waist-band    |
| aashu           | tears                    | kandani         | suspenders (belt)        |
| aatankabadi     | terrorist                | kandoo          | buttock                  |
| aadad           | son                      | kanya           | unmarried young girl     |
| adyakchyaya     | president                | kapal           | hair                     |
| andho           | blind                    | katnu           | cut                      |
| angha           | body parts               | katro           | shroud                   |
| antare, antari  | fifth born child         | kattu           | underwear                |
| baadar          | monkey                   | kera            | banana                   |
| baau            | father                   | keta/keti, sane/sani | children |
| bairro          | inattentive; deaf        | khaate          | slum dwellers            |
| bairro          | deaf                     | khaire          | white (westerners)       |
| bahun       | brahmin, refer to an upper caste | khakar     | cough          |
|------------|----------------------------------|------------|---------------|
| bahun      | brahmin                          | khasnu, marnu | die            |
| bakhri     | she-goat                          | khat       | slum dwellers  |
| bal-balika | children                          | khoj       | search         |
| balatkari  | rapist                            | khukuri    | nepali knife   |
| batashe    | child of unknown father           | kukarmi    | Immoral evil   |
| behen chod | sister fucker                     | kukur      | dog            |
| bhai/bahini| younger siblings                  | kukurni    | bitch          |
| bhaishi    | buffalo                           | kulangar   | black-sheep    |
| bhak-bhake | Stutterer                         | kupro      | bent; humpbacked |
| bhalal     | spear                             | kuputra    | bastard        |
| bhalu      | bear, prostitute (slang)          | laando     | penis          |
| bhanja/bhanjee | sister’s son/sister’s daughter | laantho    | penis          |
| bhari      | load                              | lado       | penis          |
| bhastrachar| corrupted                         | lakhetnu   | chase away     |
| bhastrachar| corruption                        | langado    | crippled/limping person |
| bhaate     | one who eats a lots of rice       | lathebro   | ignorant; half-dumb |
| bheda      | sheep                             | latta hannu | to kick        |
| bhena      | elder sister’s husband            | linga      | penis          |
| bhitra     | inside                            | logne      | husband        |
| bhoot      | ghost                             | logne      | husband        |
| bhosadi    | motherfucker (slang)              | lulo       | weak           |
| bhoite / bhoteni | of Tibetan ethnically    | maal       | female (slang) |
| bhrastha   | corrupt                           | madhise    | people living in Madhesh |
| bidhur     | widower                           | mailo/maili | the second child |
| bidhuwa    | widow                             | maila       | human          |
| bikh       | poison                            | mandale    | support of nepali panchayat |
| bikhantakari | separatist                       | maobadi    | maoist         |
| birami     | sick                              | maranchayse| emaciated, skinny |
| birya      | sperm                             | maranchayshe| emaciated    |
| bisait     | bad luck                          | maranchyashe| emaciated    |
| bismayadibodhak | interjection                      | mari ja    | go die         |
| bitnu      | died, passed away                 | masan      | cemetery; crematorium |
| boka       | he-goat; pervert (slang)          | maya       | love           |
| bokie      | pervert (slang, for women)        | mayalu     | valentine; affectate; |
| bokshi     | witch; evil woman                 | meri       | mine           |
| bokshi     | witch                             | mero       | mine           |
| English | Nepali Meanings |
|---------|----------------|
| budi | wife; old woman; |
| budi kanya | spinster |
| buwa/ama | father/mother |
| bwaso, bwasha | hyena |
| chak | buttock |
| cher | shit |
| chhati | breast |
| chhiraunu | to insert |
| chhora | son |
| chihan | grave |
| chiknu | have sex |
| cholnu | masturbate |
| chor | thief |
| chor, chorni | chief |
| chothale | harp-tongued; acerbic |
| chura | knife |
| chura (small knife) | small knife |
| chusnu | absorb; suck; exploit others |
| chutiya | idiot, fucker |
| chyaa | yuck |
| daanab | demon |
| daanab | devil |
| dadha | brother; (slang) gangster |
| daijo | dowry |
| daijo | dawry |
| daitya | evil |
| dakar | burp |
| damini | refer to a lower cast women |
| damwit | refer to a lower cast women |
| dekha jayega | will see |
| deshdrohi | traitor |
| dhoti | long |
| dhunga | stone |
| dimag | mind |
| disha | stool; excretion |
| dulaha | groom |
| dulahi | bride |
| fat | swapnadosh |
| furti | show off |
| gadha | donkey |
| chhati | breast |
| buwa/ama | father/mother |
| budi | wife; old woman; |
| budi kanya | spinster |
| chak | buttock |
| cher | shit |
| chhati | breast |
| chhiraunu | to insert |
| chhora | son |
| chihan | grave |
| chiknu | have sex |
| cholnu | masturbate |
| chor | thief |
| chor, chorni | chief |
| chothale | harp-tongued; acerbic |
| chura | knife |
| chura (small knife) | small knife |
| chusnu | absorb; suck; exploit others |
| chutiya | idiot, fucker |
| chyaa | yuck |
| daanab | demon |
| daanab | devil |
| dadha | brother; (slang) gangster |
| daijo | dowry |
| daijo | dawry |
| daitya | evil |
| dakar | burp |
| damini | refer to a lower cast women |
| damwit | refer to a lower cast women |
| dekha jayega | will see |
| deshdrohi | traitor |
| dhoti | long |
| dhunga | stone |
| dimag | mind |
| disha | stool; excretion |
| dulaha | groom |
| dulahi | bride |
| fat | swapnadosh |
| furti | show off |
| gadha | donkey |

**Note:** Some terms are slangs or colloquial forms in Nepali.
| Garb | Curved and ugly faced | Monument; ex-minister | Prostitute, derog., call girl |
|---|---|---|---|
| Geda | Testis | Purba Mantri | Woman |
| Ghopte | | | |
| Gobar | Manure | Puti, Yoni | Vagina |
| Gobar Gwatch | Blockhead; dull | Rakshyas | Wicked, devil, monster |
| Goli | Bullet | Rakshyas | Monster |
| Goli (bullet) | Bullet | Raandi | Prostitute |
| Goru | Bull; ox | Raandi | Whore |
| Gu | Feces | Rajina | Resin |
| Gu/Cher/Susu | Stool; excretion | Rajkhan | Testicles |
| Gula | Testicles | Rakhe | Mistress; kept woman |
| Guptanga | Male or female genital | Rakhani | To keep |
| Guptanga | Genital | Randi ko baan | Son of a prostitute |
| Gwatch | Idiot | Rando | Prostitutes client (male) |
| Haija | Pestilence; cholera | Rastraghati | Traitor |
| Hajar | Jol | Rastrapati | President of a country |
| Jangali | Uncouth, uncivilized | Swastha Mantri | Health minister |
| Jangali, Jantare/Jantari | Wild, sixth child | Talo, Talo | Clout, sanitary pads |
| Janti | Procession | Taako | Head |
| Janti | Bridegroom participants | Tan | Physique, life, body |
| Jee/Jyu | Greeting; salutation | Jel Halnu | Put to prison |
| Jestha | Senior | Taruni, Taruno | Young, youthful |
| Jestha Nagarik | Senior citizen | Tarwar | Long knife |
| Nepali Term | English Translation | Context (used in this paper) |
|-------------|---------------------|-----------------------------|
| jetho/jethi | the first child     | tera, teri, tero            |
| jharna      | drop                | thangne, dirty person       |
| jhatha      | pubic hair          | thankanu, erection         |
| jhus        | awn                 | thun choli, bra             |
| jindabaad   | immortality; long live! | tira, towards             |
| jindabaad   | hail                | topi, cap                  |
| junga       | mustache; whiskers  | tun-turi, turi             |
| junga       | mustache            | tuppi, sikha               |
| kaan        | ear                 | turuntai, immediately      |
| kaano       | one-eyed            | ullu ka pattha, foolish man|
| kaano       | blind               | ulti, vomit                |
| kachha      | underwear           | upapradhan mantri, vice-president |
| kaila/kaili | fourth child        | uthbas, sticking together |
| uttauli     | act shamelessly;    | wakṣhyastal, breast        |
| virya       | sperm               | yaunkarmi, prostitute      |

Table 5. Transliterated Nepali words and their contextual (as used in this paper) meaning.