ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230535

NEHATE: Large-Scale Annotated Data Shedding Light on Hate Speech in Nepali Local Election Discourse

Surendrabikram Thapa[§,a;*], Kritesh Rauniyar[§,b], Shuvam Shiwakoti[b], Sweta Poudel[c], Usman Naseem[d] and Mehwish Nasim[e,f]

[a]Department of Computer Science, Virginia Tech, USA
[b]Department of Computer Science, Delhi Technological University, India
[c]Kathmandu Engineering College, Tribhuvan University, Nepal
[d]College of Science and Engineering, James Cook University, Australia
[e]School of Physics, Mathematics and Computing, The University of Western Australia, Australia
[f]College of Science and Engineering, Flinders University, Australia

Abstract. The use of social media during election campaigns has become increasingly popular. However, the unbridled nature of online discourse can lead to the propagation of hate speech, which has far-reaching implications for the democratic process. Natural Language Processing (NLP) techniques are being used to counteract the spread of hate speech and promote healthy online discourse. Despite the increasing need for NLP techniques to combat hate speech, research on low-resource languages such as Nepali is limited, posing a challenge to the realization of the United Nations' Leave No One Behind principle, which calls for inclusive development that benefits all individuals and communities, regardless of their backgrounds or circumstances. To bridge this gap, we introduce **NEHATE**, a large-scale manually annotated dataset of hate speech and its targets in Nepali local election discourse. The dataset comprises 13,505 tweets, annotated for hate speech with further sub-categorization of hate speech into targets such as community, individual, and organization. Benchmarking of the dataset with various algorithms has shown potential for performance improvement. We have made the dataset publicly available at https://github.com/shucoll/NEHate to promote further research and development, while also contributing to the UN SDGs aimed at fostering peaceful, inclusive societies, and justice and strong institutions.

1 Introduction

Social media has proven to be a great means for people to express themselves by sharing information and ideas. With the advent of social media platforms, it was hoped that these platforms could provide a marketplace for open information dissemination and increased participation, especially in the form of political engagement, and that it would promote more civic engagement and participation in elections [10]. However, the ease of access and anonymity provided by these platforms and the desire of users to dominate discussions and defend beliefs has failed the *social media promise* as social media sites provided a suitable environment for the use of aggressive and harmful/hateful content [11]. Although hate speech is considered a complex phenomenon that relies on relations between groups and subtle language differences [11], Sellars [25] defines hate speech as verbal or written abuse that is directed towards a certain group of people, often because of their race, beliefs, or sexual orientation. Hate in social media is a growing problem as it not only induces short-term annoyance and terror, but it can also have long-term effects on the mental health of the victims which can discourage them from participating in any form of public discussion. Therefore, the detection and control of hate on online platforms have become crucial.

In a country like Nepal where political instability is a major issue, online hate against individuals and groups with certain political affiliations becomes inevitable. With the end of the 240-year-old monarchy on May 28, 2008 [2] and the formation of a democratic multi-party system in Nepal, the expectation of citizens rose and the perception of having a fair opportunity in the democratic process increased. But soon the inter and intra-political party conflicts increased and as a result of this, there were frequent changes in government over a short period of time [28]. During the campaign period leading to the local elections in Nepal on May 13, 2022 [26], the citizens have raised their dissatisfaction with the present political scenario of Nepal and wanted to see an increase in the number of youth coming forward to bring changes in the political sector. This dissatisfaction in the public gave rise to hate against political parties and their leaders in social media.

While users in the developed world can choose a technology that suits their needs, *emergent users* (users from developing countries) cannot afford this luxury. Several studies show that emergent users may adapt themselves to the technology that is readily available and indicate that studies on such user bases are highly valuable to understand the peculiarities in communication and the use of technology [5]. Studies on Nepal's user base are scarce. Nepali is the most spoken language in Nepal [12], so it is also widely used in social media in Nepal. Spoken by approximately 40 million people worldwide [34], Nepali is based on the Devanagari script, which consists of 36 consonants, 13 vowels, and 10 numerals [27]. With the widespread use of social media worldwide, work in the field of NLP in the Nepali language has been growing rapidly. Additionally, the growth of social media calls for the employment of automated NLP techniques for efficient meaningful information extraction from large amounts of textual data [27]. Driven by the idea of mitigating the use of hate

* Email: surendrabikram@vt.edu
§ The authors contributed equally and are joint first authors.

online and promoting work in a low-resource language like Nepali, we created a dataset that consists of 13,505 tweets in the Nepali language posted during the local election 2022 and its campaign period. These tweets are labeled as "Hate" or 'No Hate' with 'Hate' further divided into 3 targets - 'Community', 'Individual', and 'Organization'.

Research in the field of low-resource language promotes one of the core principles of Sustainable Development Goals (SDGs)- LNOB (Leave No One Behind) which aims at prioritizing actions for the most marginalized people and bridging the gap between them and other better-off groups [32]. Making the dataset publicly available invites further research while contributing to several United Nations Sustainable Development Goals (SDGs). Hate speech can have a lasting effect on the mental health of individuals by inducing terror, depression, and anxiety; therefore, its prevention aligns with SDG3: Good Health and Well-Being which aims at the overall physical and mental well-being of all. The development of effective hate speech models aligns with SDG9: Industry, Innovation, and Infrastructure which aims at the development of efficient and sustainable infrastructure. Detecting hate and identifying hate speech targets in political events can help identify and prevent discrimination among people. This aligns with SDG10: Reduced Inequalities which aims to reduce discrimination and promote equal opportunities for all. Finally, the mitigation of hate among people and political events aligns with SDG16: Peace, Justice, and Strong Institutions, which aims in achieving peaceful and inclusive societies.

Our contributions are:

- We release a large-scale original dataset of 13,505 tweets that are related to Nepali local election discourse.
- We manually annotated the dataset for Hate speech and its targets viz. community, individual, and organization using a comprehensive annotation schema.
- We set benchmarks with popular algorithms and also human-based evaluations. Our benchmarks show the scope for improvement in automated hate speech and target identification.

2 Related Work

Recent years have seen a plethora of studies that focused on computational techniques for identifying hate speech in social media. In the following subsections, we discuss the related works in the detection of hate speech.

2.1 Works in Non-Nepali Languages

The identification of hate speech and offensive language has received significant attention in languages with abundant resources, such as English [18]. Mathew et al. [13] collected and annotated 20,148 posts from Twitter and Gap in the English language. The posts were categorized into hate, offensive and normal speech with the identification of targeted communities for hate/offensive posts. Similarly, Mollas et al. [14] curated two datasets with comments from YouTube and Reddit. The first dataset contains 998 comments which are labeled as either hate or non-hate. The second dataset contains 433 hate comments categorized into 8 labels. The authors also present an active sampling annotation procedure for balancing the dataset in relation to the multiple aspects defined. Qian et al. [20] introduced two labeled hate speech datasets with manually written intervention responses. The two datasets were collected from Reddit (22K comments) and Gab (34K comments). Apart from the English language where extensive work has been done in the identification of hate speech, work on other languages has also been surfacing. Corazza et al. [8] proposed a multilingual hate speech detection model with 16,000 English tweets, 4,000 Italian tweets, and 5,009 German tweets. Each language's tweets were categorized into a separate set of labels. Political matters are often subject to hate speech in many languages. Realizing this, Mulki et al. [16] created a dataset of 5,846 tweets in the Arabic language related to political events. They divided the tweets into 3 classes - normal, abusive, and hate.

2.2 Works in Nepali Language

Being a low-resource language, only a few studies have been performed on the detection of hate speech in Nepali. Similarly, in general itself, for other tasks, we also have fewer resources available due to the limited research and data availability for the language [1, 33]. Shrestha et al. [29] annotated 3,490 sentences into two classes Positive and Negative and performed sentiment analysis using machine learning algorithms. Despite having a decent number of annotated data, they used an equal number of positive (814) and negative (814) sentences in the training dataset to address the class imbalance. Singh et al. [31] presented aspect-based abusive sentiment detection in Nepali Social Media Text. They extracted 3,068 comments from 37 different Youtube videos and performed benchmarks with classic machine learning and deep learning methods. Niraula et al. [17] collected and annotated 7,462 comments and performed sentimental analysis. In the analysis of performing benchmark classification, Multilingual BERT (M- BERT) which is trained using Wikipedia dump for multiple languages, did not perform well compared to traditional ML models. The M-BERT model's performance was found to be inadequate due to the limited size of Wikipedia content available for low-resource languages such as Nepali, which was used for training.

NLP research in a morphologically rich and complex language like Nepali [17], poses several challenges. One of the major challenges is the sentence structure of the Nepali language [31]. It differs from that of the commonly studied English language. In terms of social media texts, many of the Nepali language tweets also include a combination of other languages like Hindi and English which makes the automated NLP tasks more challenging. Although NLP has advanced in the English language, due to a lack of pre-training data, resource uniformity, and computational resources in the Nepali language, it has made a smaller contribution to NLP [34]. The major limitation that remains is the lack of a large enough corpus for the Nepali language. We believe, our annotated dataset which has over 13K annotated tweets is a big step towards the progress of NLP in the Nepali language. To shed light on the current state of hate speech datasets, Table 1 provides a comparison of hate speech datasets in different languages.

3 Dataset

Nepalese local elections were held on May 13, 2022, covering 6 metropolitan cities, 11 sub-metropolitan cities, 276 municipalities, and 460 rural municipalities. Prior to the 2022 Nepalese local elections, social media activity related to the electoral process saw a significant increase in engagement. To capture this, we collected tweets in the Nepali language from April 19, 2022, onward, with a view to monitoring online discourse and conversations leading up to the elections. Our dataset comprises tweets from this pre-election period as well as the period spanning from May 13, 2022, which was the date of the local elections, to May 18, 2022. We made the decision to extend the collection period post-election to capture conversations and sentiment analysis pertaining to the electoral outcomes and

Works	Data Source	Language	Objective	Sub-classes/Targets	Context
Mossie et al. [15]	Facebook	Amharic	Hate Speech	✗	General discourse
Mulki et al [16]	Twitter	Arabic	Hate speech	✗	Politics
Qian et al. [20]	Reddit and Gab	English	Hate speech	✗	General discourse
Shrestha et al. [29]	Nepali News Portals	Nepali	Sentiment analysis	✗	News Media
Mathew et al. [13]	Twitter and Gab	English	Hate speech	Targeted Communities	General discourse
Armeu et al. [3]	Twitter	Arabic	Hate and Misinformation Identification	✗	COVID-19
Romim et al. [23]	YouTube and Facebook	Bengali	Hate Speech	✗	General Discourse
Niraula et al. [17]	Facebook, Twitter, YouTube	Nepali	Offensive Language	Sexist, Racist	General discourse
Toraman et al. [35]	Twitter	Turkish, English	Hate	✗	General Discourse
Arshad et al. [4]	Twitter	Urdu	Hate Speech	✗	Religious Hate
NEHATE (Our Dataset)	Twitter	Nepali	Hate speech	Individual, Organization, Community	Election in Nepal

Table 1: Summary of datasets used in the literature

Figure 1: Wordcloud for the words in NEHATE dataset

the ensuing discussions. By incorporating both the pre-election and post-election periods, our dataset provides a comprehensive and nuanced view of social media activity and discourse during the Nepalese local elections. The keywords selected for this study were chosen based on their relevance to the local elections in Nepal. These keywords included terms such as एमाले (*Translates to UML: the largest party in Nepal during the local election*), दल (*Translation: party*), सत्तारुढ (*Translation: ruling party*), जनप्रतिनिधि (*Translation: representative*), राजीनामा (*Translation: resignation*), मतदाता (*Translation: voters*), कांग्रेस (*Translates to Congress: The ruling party of Nepal during the election*), माओवादी (*Translates to Maoist: third largest party in Nepal during election*), निर्वाचन (*Translation: Election*). These keywords were selected to capture the important themes and topics related to the Nepalese local elections, including the major political parties and their representatives, election-related terminology, and other relevant keywords related to the election process. By selecting these keywords, we aim to capture a wide range of discussions and opinions related to the local elections in Nepal on social media. The tweets were collected using the Twitter API.

In order to ensure the relevance of the collected data, we conducted manual filtering by identifying and eliminating non-Nepali language tweets that were erroneously detected by Twitter's API. Furthermore, we removed tweets that were determined to be non-informative or highly irrelevant based on a set of predefined criteria, which are elaborated on below. The resulting data set consists of 13,505 manually annotated tweets, each with a unique tweet ID to ensure data integrity and to avoid duplication. Furthermore, in order to eliminate instances of repeated content, we also removed tweets that had different tweet IDs but identical text, as some users may have copied and pasted humorous or other non-original content.

3.1 Filtering Criteria

Filtering tweets is a fundamental task in annotation, as it removes irrelevant or misleading data that could distort the analysis results. Thus, to ensure that our dataset was both relevant and informative, we implemented a number of filtering criteria based on the following considerations:

- **Language filtering:** We manually excluded tweets written in languages other than Nepali. We retained tweets that contained only a few non-Nepali words or phrases that are commonly used, such as "link" or "share", as long as the majority of the tweet was in Nepali.
- **Non-informative tweets:** We eliminated tweets that contained little or no useful information on the local election, such as spam or advertisements. These tweets were deemed non-informative and were not conducive to our research goals.
- **Doubtful Tweets:** We also excluded doubtful tweets that lacked clear context or perception and might be influenced by local contexts related to the local election. This is because such tweets could potentially hinder the annotation process by introducing ambiguity and preventing accurate categorization of hate speech.
- **Unclear Targets:** We removed tweets that contained hate speech, but the target of the speech was unclear or ambiguous. This is because our work specifically focuses on annotating hate speech and its targets, and unclear targets would not contribute to the dataset's goal.

By implementing these filtering criteria, we were able to ensure that our dataset was both relevant and informative for our study on hate speech and its targets in the context of the Nepalese local election.

3.2 Annotation Process

The process of annotation involved labeling each tweet as either containing hate speech or not, as well as identifying the target of the hate speech. Our dataset was annotated manually by a team of four individuals with diverse educational qualifications, including undergraduate, MS, and Ph.D. degrees, as well as researchers with experience in NLP and data collection. All the annotators had a minimum of 10 years of formal Nepali education, ensuring that they were capable of providing high-quality annotations. The diversity of the annotators' backgrounds, originating from various regions of Nepal, served to minimize potential biases in the annotation process, which is an important aspect of data annotation.

Given the diverse backgrounds of our annotators, we recognized the possibility of some individuals experiencing a sense of personal offense or discomfort due to certain tweets potentially targeting their community or identity. To mitigate the risk of negative psychological impact on the human annotators, we provided them with a warning prior to the annotation process that the text may contain offensive or inappropriate language and content. This approach aimed to prepare the annotators for potentially sensitive material and to help them manage their emotional reactions during the annotation process.

3.2.1 3-Phase Annotation

Accurate and consistent annotations are critical to ensure the reliability and validity of any analysis or model development based on the labeled data. To ensure the accuracy of our tweet annotations, a three-phase annotation process was employed as described below.

To maintain consistency and reliability in the annotation process, clear guidelines and regular quality checks were employed. These measures helped to ensure that the annotations were of high quality and could be used for further analysis. To assess the inter-annotator agreement quantitatively, we used Fleiss' Kappa (κ) as our inter-rater agreement measure.

We initiated the annotation process by preparing clear and concise instructions that were iteratively revised until all annotators were entirely familiar with the instructions. To ensure that the instructions were unambiguous, we followed a three-phase annotation process.

- **Pilot Run:** The first phase involved a pilot run of 50 tweets to ensure that everyone understood the annotation instructions. Given that labeling tweets can be a challenging task, it was essential to have a shared understanding of what constitutes hate speech. During this phase, there was some confusion among the annotators, and the instructions were revised to address all the confusion.
- **Revised Instructions:** In the second phase, all four annotators annotated 200 tweets to verify that the instructions revised after the first stage were clear enough. During this phase, the annotators were given the revised instructions and asked to label the tweets, and this stage confirmed that the revised instructions were unambiguous and that the annotators could consistently identify hate speech.
- **Consolidation Phase:** In the third phase, the annotators engaged in a group discussion of conflicts identified in the second phase of annotation, during which they discussed any discrepancies in their annotations and reached a consensus. This phase was vital in resolving any disagreements and ensuring that all tweets were labeled consistently. The group discussion also helped to make the instructions more apparent and provided an opportunity to identify any further ambiguities or inconsistencies in the instructions.

3.2.2 Annotation Guidelines

During political events like elections, hate speech can manifest in several ways, including the use of targeted language, memes, and expressions of hostility and aggression towards specific political groups or individuals. The annotation guidelines are mentioned below.

Hate Speech: A text contains hateful content such as a personal attack, homophobic abuse, racial abuse, or attack on minorities.

- **Targeted language**: Hate speech during the election in Nepal often targets specific groups based on their political beliefs or affiliations. This can include language that demeans, degrades, or dehumanizes a particular political group or individual.

> एमाले,काङ्ग्रेस जस्तो भ्रष्टाचारि नारी बलात्कारि समाज बलात्कारि देश बेचारी पार्टी अरु कोहि छैन।
> Translation: *There is no party like UML or Congress that are corrupt, women rapists, society rapists, and who sell the country.*

- **Hostility and aggression**: Hate speech during political events like the election in Nepal often expresses hostility or aggression towards a particular political group or individual. This can include language that promotes or glorifies violence or hatred against a particular political group or individual.

> गठबन्धन लाइ भोट हालेर जिताउनु भनेको देश दुर्घटनामा पार्नु हो
> Translation: *Electing the alliance by voting for them is equivalent to pushing the country into an accident.*

- **Use of Hateful Satires**: Hate speech during political events like the election in Nepal often uses satires to disseminate harmful messages that are intended to demean, degrade or dehumanize a particular political group or individual.

> पाँच दल गठबन्धन गरेर चुनावी उमेदवार उठ्दा पनि जित्न सक्दिन भनेर पैसा बाड्ने जस्ता अनैतिक व्यवहार गरेर चुनाव जित्नु भन्दा त गेरुवस्त्र लगाएर नदीको किनारमा खरानी घसेर बस्न ठीक होला नि!
> Translation: *If you are winning an election through immoral activities like distributing money, etc. even after making a five-party alliance, it would rather be better if you wear saffron cloth and sit on the riverside.*

Further, it is important to note that sarcasm and political satire can be used to express hate speech and can be difficult to identify. Sarcasm and satire can be used to mask hate speech, making it more subtle and harder to detect. Sarcasm can be used to express hate speech in a way that is less obvious and less likely to be flagged as hate speech. Satire can also be used to express hate speech in a way that is intended to be humorous or satirical but can still be hateful. Annotation guidelines included clear examples of sarcasm and satire and how they can be used to express hate speech.

No Hate Speech: A text reports the events or others' opinions objectively and contains no offensive or hateful content. To make guidelines clear, the following points were discussed as the significant characteristics of non-hate speech.

- **Constructive criticism:** Non-hate speech during political events like the Nepal election often includes constructive criticism of political figures, policies or parties. It can also include criticism of political events and happenings.
- **Factual and informative:** Non-hate speech during political events like the Nepal election often includes factual and informative content, it can be news, updates, and analysis of the political events.
- **Lack of hostility:** Non-hate speech during political events like the Nepal election does not express hostility or aggression towards a particular political group or individual.
- **Lack of misinformation or fake news:** Non-hate speech during political events like the Nepal election does not spread misinformation or fake news, it is based on facts and credible sources.

Some of the examples are as follows:

➢रुपन्देहीको देवदह नगरपालिकामा एमाले विजयी।
Translation: *UML emerged victorious in Devdah Municipality of Rupandehi.* ***(Fact)***
➢पार्टीको झोला बोक्न छोडेर तपाईं जस्तो देशप्रेमी मान्छे स्वतन्त्रहरुको साथमा आउन पर्ने देखिन्छ।
Translation: *Nation-loving people like you should come together to support independent candidates instead of running after parties.* ***(Constructive Criticism)***
➢सम्पूर्ण स्थानीय तहमा गठबन्धन गर्ने सत्तारुढ दलबीच सहमति
Translation: *Agreement between the ruling parties to form an alliance at all local levels* ***(No Misinformation)***

Hate speech was further divided into three sub-categories viz. "Community", "Organization", and "Individual". The annotation guidelines for targets are following:

- **Community:** In the context of our Nepal election dataset, a community refers to a group of individuals who share common beliefs, or characteristics. They can have the same caste, religion, place of origin, etc.
- **Organization:** An organization in the Nepal election context refers to a structured group of individuals created to achieve a specific political goal or set of goals. Examples of organizations in this context could include political parties such as UML or the Nepali Congress, or interest groups advocating for specific policies or social issues.
- **Individual:** In the context of our dataset, an individual refers to a person as an autonomous entity who is involved in politics in some way. This can include politicians, political candidates, activists, journalists, and other individuals who are involved in political discourse or have a stake in the outcome of the election. Some of the most frequently mentioned individuals in our dataset include Sher Bahadur Deuba, KP Sharma Oli, and Pushpa Kamal Dahal (Prachanda).

The annotation guidelines for the dataset were comprehensive, and the team of annotators regularly communicated to address any issues that arose during the labeling process. Collaborative meetings and annotation sessions facilitated the resolution of any labeling discrepancies. The annotators demonstrated their ability to differentiate between tweets targeting an organization versus those targeting a community by analyzing the linguistic cues and contextual information present in the tweets. Overall, the annotation process was well-organized and efficiently executed by annotators.

3.2.3 Inter-Annotator Agreement and Statistics

In order to assess the consistency of our annotations, we used a statistical measure called *Fleiss' Kappa*. Fleiss' Kappa is particularly useful when dealing with multiple raters and multiple categories because it corrects for the possibility of agreement occurring by chance alone. It provides a more robust measure of inter-rater reliability, allowing researchers to better assess the consistency of annotations in situations where there are more than two raters or annotators involved.

In our annotations, we obtained a high level of inter-annotator agreement, with a Fleiss' Kappa of 0.82 for the 2-class annotation of "Hate" vs "Non-Hate" and 0.76 for the 3-class annotation.

3.3 Dataset Statistics

Our dataset, NEHATE contains 13,505 labeled tweets labeled as 'Hate' and 'No Hate' with 1,888 (13.98%) labeled as 'Hate' and 11,617(86.02%) as 'No Hate'. The tweets labeled as hate speech are further divided into 3 targets - 'Individual', 'Organization', and 'Community'. Of all hate speech, 'Individual' contains 931 (49.3%) tweets, 'Organization' - 780 (41.31%), and Community - 177 (9.37%). These data statistics along with the average character count and average word count are mentioned in Table 2. It is also worth noting that the average world count for hate speech is significantly higher than that of non-hate speech.

Problem	Labels	Tweets	Avg. Char	Avg. words
Hate Speech	Hate	1,888	166.39 (146.37)	25.85 (24.22)
	Non-Hate	11,617	130.23 (105.25)	18.45 (16.57)
Targets	Individual	931	176.62 (156.04)	27.55 (25.88)
	Organization	780	156.70 (136.93)	23.99 (22.41)
	Community	177	155.28 (137.07)	25.03 (23.41)

Table 2: Statistics for NEHATE data. Once the text has been preprocessed, average value of characters per tweet (Avg. Char) and words per tweet (Avg. Words) are determined.

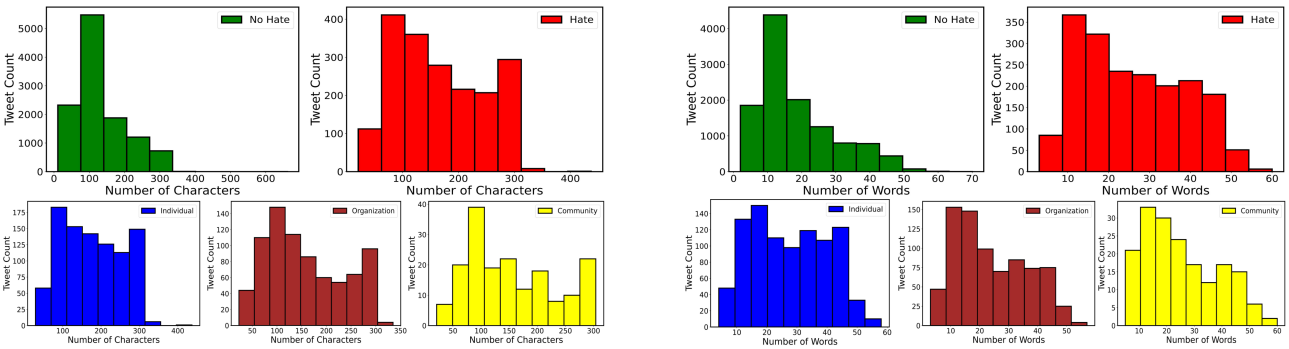
3.4 Exploratory Data Analysis

Table 3 displays the top 5 words that occur most frequently in our overall dataset as well as for the hate speech and non-hate speech classes. Words, translation, or transliteration along with their corresponding TF-IDF scores are given. Similarly, Table 4 shows the top 5 words for each of the target classes. TF-IDF is a statistical method used to measure the significance of a word to a document in a collection of documents. The TF-IDF score comprises 2 parts - TF (Term Frequency) tells us how often the word occurs in a document. IDF (Inverse Document Frequency) gives us how common or rare the word is in the entire set of documents. The resulting TF-IDF score for a word is the product of its TF and IDF scores. Words with high TF-IDF scores are considered more important and relevant to the document, as they are frequent within the document and rare across the entire collection of documents. Table 3 and Table 4 reveal that the words Vote (भोट), UML (एमाले), and Election (निर्वाचन) have a high significance in most of the classes. Figure 1 gives a general overview of the words present in our dataset.

Figure 2 shows the histogram for the number of characters and the number of words across all classes. Similarly, Table 2 contains the average words per tweet and average characters per tweet for each class. It can be observed that the average word count for tweets labeled as 'Hate' is significantly higher than that of tweets labeled as 'No Hate'. This aligns with the fact that the majority of the 'No Hate' tweets that the annotators encountered were simply informative tweets. The 'Hate' tweets on the other hand were often subject to people talking in length about how the current political system is failing. Further investigation is warranted to determine the underlying causes of this disparity in tweet length.

4 Experimental Results and Analysis

We created benchmarks using a range of approaches, consisting of classical machine learning algorithms as well as advanced transformer-based models.



(a) Number of Characters (b) Number of Words

Figure 2: Histogram of number of characters and words per tweet in our dataset

All Posts			No Hate Speech Posts			Hate Speech Posts		
Words	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**
भोट	Vote	0.1678	भोट	Vote	0.1588	एमाले	UML	0.2027
एमाले	UML	0.1507	निर्वाचन	Election	0.1408	भोट	Vote	0.1979
निर्वाचन	Election	0.1333	एमाले	UML	0.1376	देउवा	Deuba	0.1279
नेपाली	Nepali	0.1105	नेपाली	Nepali	0.1051	नेपाली	Nepali	0.1248
उम्मेदवारी	Candidacy	0.0806	उम्मेदवारी	Candidacy	0.0911	निर्वाचन	Election	0.0698

Table 3: Top-5 most frequent words in the overall dataset and also for each class belonging to Hate Speech and Non-Hate Speech.

4.1 Baselines

For traditional ML algorithms, we used Naive Bayes [21], Decision Tree [22], XGBoost [7], and AdaBoost [24] with TF-IDF vectorizer. For transformer-based models, except for DistillBERT, we took FillMask models and modified them for classification tasks. For establishing baselines, we employed Nepali DistillBERT [30], Nepali RoBERTa [6], NepaliBERT [9], NepNewsBERT [19], and NepBERTa [34]. The split for training, test, and validation data is given in Table 6.

We further evaluated a random sample among our test data using human evaluators. We employed two evaluators who were fluent in Nepali. The evaluators were given access to the test data and were asked to label the tweets based on whether they contained hate speech or not. They were also made to further label target classes. Instructions were provided to our human evaluators.

4.2 Experimental Settings

We employed pre-existing models for every baseline and assessed their performance using the F1-score, Macro-averaged Mean Absolute Error (MMAE), and accuracy. All models were trained using Tensorflow on a GPU with Tesla T4 architecture, which has a dedicated memory of 25 GB. We used the hugging-face library to import the pre-trained transformers models. The FillMask language models were finetuned for downstream classification tasks. In order to change the fill-mask task in transformer models for classification, we modified the model head by adding a classification layer on top of the model and trained the model on our data for the classification task.

4.2.1 Text Preprocessing

Text preprocessing is a critical step in any NLP task, as pointed out by recent research efforts. In order to facilitate subsequent analysis, we performed a preprocessing step on the tweet text to remove non-alphanumeric elements, including special characters, hyperlinks, mentions, and emojis. Special characters and other symbols can contribute to the noise in the data, which could ultimately impact the accuracy of any subsequent analysis. Additionally, hyperlinks and mentions are irrelevant to the tweet's content and can also contribute to the noise. Emojis, though popular in social media, were removed as they are not typically used in standard NLP techniques and can result in errors or inaccuracies during subsequent analyses. Our preprocessing step ensures that the text data is cleaned and standardized and that only meaningful content is retained for further analysis.

4.3 Results and Analysis

The results with baseline models show that transformer-based models are quite promising. Among the models we experimented with, NepBERTa achieved the highest F1-score of 0.68. The relatively low F1-score achieved by the best-performing model, NepBERTa, compared to that of the human evaluators indicates that there is still a need for further improvement in the development of hate speech detection models for the Nepali language. This is consistent with previous studies that have shown the challenge of developing accurate models for hate speech detection in low-resource languages.

Looking at some of the misclassification cases by the algorithms, it was interesting to notice that sarcastic yet non-hate tweets were misclassified as shown below.

> जसले जे भने पनि आफ्नो परीक्षा भोलि प्रचण्डले गएर दिने होईनन् क्यारे! के ट्विटरमा चुनावको बारेमा लेखेर बस्नु ?
> **Translation:** Whatever is being said, Prachanda is not gonna give my exams tomorrow. There is no point in writing on Twitter about the election.
> ➢ Label: **No Hate** Predicted: **Hate**

The misclassification of sarcastic yet non-hate tweets by the algorithms is an interesting finding that suggests the need for further research to develop models that can accurately distinguish between

Target: Individual			Target: Organization			Target: Community		
Words	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**	**Words**	**Translation**	**TF-IDF**
भोट	Vote	0.2153	एमाले	UML	0.3252	नेपाली	Nepali	0.3066
देउवा	Deuba	0.1865	भोट	Vote	0.1743	भोट	Vote	0.1714
नेपाली	Nepali	0.1085	कांग्रेस	Congress	0.1033	नेपाल	Nepal	0.1023
मेयर	Mayor	0.1042	निर्वाचन	Election	0.0847	जनता	People	0.0898
एमाले	UML	0.1011	नेपाली	Nepali	0.0832	देश	Nepal	0.0763

Table 4: Top-5 most frequent words in each target class. The TF-IDF scores are given for each word.

Model	Hate vs Non Hate			Targets		
	Acc↑	**MMAE↓**	**F1$_{macro}$↑**	**Acc↑**	**MMAE↓**	**F1$_{macro}$↑**
Naive Bayes	0.86	0.50	0.46	0.66	0.79	0.45
XGBoost	0.86	0.44	0.56	0.66	0.72	0.51
AdaBoost	0.85	0.40	0.61	0.64	0.74	0.48
Decision Trees	0.83	0.42	0.58	0.58	0.74	0.50
DistillBERT (Nepali)	0.85	0.33	0.66	0.65	0.32	0.55
RoBERTa (Nepali)	0.73	0.28	0.62	0.60	0.52	0.58
NepaliBERT	0.73	0.29	0.62	0.60	0.67	0.52
NepNewsBERT	0.79	0.27	0.67	0.67	0.51	0.60
NepBERTa	0.79	0.23	0.68	0.69	0.46	0.60
Human Evaluator-A	0.89	0.12	0.88	0.87	0.21	0.86
Human Evaluator-B	0.93	0.09	0.91	0.89	0.11	0.89

Table 5: Baseline Results with different algorithms

Tasks	Train	Validation	Test
Hate Identification	9,454	2,026	2,025
Targets Identification	1,321	283	284

Table 6: Train/Test/Val of NEHATE for different tasks

sarcastic language and hate speech. Overall, the results suggest that there is a need for further research and development in the area of hate speech detection in the Nepali language, and use of transformer-based models can be a promising approach.

5 Limitations and Ethical Concerns

Limitations: In this paper, we present a large-scale dataset for hate speech detection and target identification in the Nepali language. We also present baselines for detecting hate speech and identifying targets using this dataset. However, there are several limitations to our work that should be acknowledged. First, our dataset is limited to tweets from a specific time period surrounding the local election in Nepal, and may not be representative of hate speech in other contexts. Additionally, our dataset is based on tweets from a single microblogging platform. Second, our annotation scheme for targets is based on broad categories (Individuals, Organizations, and Communities), and may not capture more specific or nuanced targets. Furthermore, the annotation process is subjective, and different annotators may have different opinions on whether certain tweets should be considered hate speech or not. Third, the baselines we provide are based on a limited set of features, and it is possible that other features or architectures could lead to improved performance. Finally, it's important to note that hate speech detection and target identification technologies can raise ethical concerns, such as potential bias. These ethical concerns should be considered and addressed in the development and deployment of such technology.

Ethics Statement: The dataset does not contain direct identifiers. It contains tweet IDs. Tweet IDs can be used to retrieve the tweets. The tweet becomes unavailable if the user deletes the tweet. This gives the original author of the tweet full control over their content. All the tweets presented in the examples have been anonymized and obfuscated for user privacy and to avoid misuse. Thus, no ethical approval is required. The annotation is very subjective and hence we can expect some bias in the annotation. To address these issues, examples from various users and groups are collected, along with clear instructions for annotation. Due to excellent inter-annotator agreement (κ score), we are confident that annotation instructions are mostly valid.

Reproducibility: The dataset and resources for this work are available at https://github.com/shucoll/NEHate.

6 Conclusion

Our work presents the NEHATE dataset, a valuable resource for developing and evaluating hate speech detection models in Nepali local election discourse. Our dataset consists of tweets in Nepali language, which is a very understudied language in AI-based scholarly research. Despite the subjectivity of annotations, the high inter-annotator agreements show that the annotations are mostly uniform, which is an attribute of a good dataset. In the future, we plan to explore new avenues to improve hate speech detection, including developing novel NLP models customized to detect hate speech. Expanding the NEHATE dataset to include posts from other social media platforms would also be a promising area of exploration. Additionally, we suggest exploring hate speech detection for more specific or nuanced targets. Finally, it is worth noting that the NEHATE dataset can serve as a starting point for further annotation efforts in the Nepali language, with the potential to add other dimensions of annotations beyond the ones presented in this work. Overall, we hope that our dataset will contribute to the development of effective hate speech detection models, ultimately promoting a more inclusive and respectful online discourse.

References

[1] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad, 'Exploiting linguistic information from nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques', *International Journal of Human-Computer Studies*, **160**, 102761, (2022).

[2] Ministry Of Foreign Affairs. History Of Nepal. https://mofa.gov.np/about-nepal/history-of-nepal/, 2022. [Online; accessed 23-Feb-2023].

[3] Mohamed Seghir Hadj Ameur and Hassina Aliane, 'Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset', *Procedia Computer Science*, **189**, 232–241, (2021).

[4] Muhammad Umair Arshad, Raza Ali, Mirza Omer Beg, and Waseem Shahzad, 'Uhated: hate speech detection in urdu language using transfer learning', *Language Resources and Evaluation*, 1–20, (2023).

[5] Anas Bilal, Aimal Rextin, Ahmad Kakakhel, and Mehwish Nasim, 'Analyzing emergent users' text messages data and exploring its benefits', *IEEE Access*, **7**, 2870–2879, (2018).

[6] Amit Chaudhary. RoBERTa(Nepali). https://huggingface.co/amitness/roberta-base-ne, 2021. Accessed: 2023-02-25.

[7] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al., 'Xgboost: extreme gradient boosting', *R package version 0.4-2*, **1**(4), 1–4, (2015).

[8] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata, 'A multilingual evaluation for online hate speech detection', *ACM Transactions on Internet Technology (TOIT)*, **20**(2), 1–22, (2020).

[9] Rajan Ghimire. NepaliBERT. https://huggingface.co/Rajan/NepaliBERT, 2022. Accessed: 2023-02-25.

[10] Tim Hwang, Ian Pearce, and Max Nanis, 'Socialbots: Voices from the fronts', *interactions*, **19**(2), 38–45, (March 2012).

[11] Md Saroar Jahan and Mourad Oussalah, 'A systematic review of hate speech automatic detection using natural language processing', *arXiv preprint arXiv:2106.00742*, (2021).

[12] Rajendra Khanal, 'Linguistic geography of nepalese languages', *The Third Pole: Journal of Geography Education*, 45–54, (2019).

[13] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee, 'Hatexplain: A benchmark dataset for explainable hate speech detection', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14867–14875, (2021).

[14] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas, 'Ethos: an online hate speech detection dataset', *arXiv preprint arXiv:2006.08328*, (2020).

[15] Zewdie Mossie and Jenq-Haur Wang, 'Social network hate speech detection for amharic language', *Computer Science & Information Technology*, 41–55, (2018).

[16] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani, 'L-hsab: A levantine twitter dataset for hate speech and abusive language', in *Proceedings of the third workshop on abusive language online*, pp. 111–118, (2019).

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

[18] Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra, 'Hate speech detection using natural language processing: Applications and challenges', in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1302–1308. IEEE, (2021).

[19] Shushant Pudasaini. NepNewsBERT. https://huggingface.co/Rajan/NepaliBERT, 2021. Accessed: 2023-02-25.

[20] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang, 'A benchmark dataset for learning to intervene in online hate speech', *arXiv preprint arXiv:1909.04251*, (2019).

[21] Irina Rish et al., 'An empirical study of the naive bayes classifier', in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pp. 41–46, (2001).

[22] Lior Rokach and Oded Maimon, 'Decision trees', *Data mining and knowledge discovery handbook*, 165–192, (2005).

[23] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam, 'Hate speech detection in the bengali language: A dataset and its baseline evaluation', in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pp. 457–468. Springer, (2021).

[24] Robert E Schapire, 'Explaining adaboost', *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37–52, (2013).

[25] Andrew Sellars, 'Defining hate speech', *Berkman Klein Center Research Publication*, (2016-20), 16–48, (2016).

[26] SetoPati. Local election on May 13. https://en.setopati.com/political/157891, 2022. [Online; accessed 2022-02-07].

[27] Tej Bahadur Shahi and Chiranjibi Sitaula, 'Natural language processing for nepali text: a review', *Artificial Intelligence Review*, 1–29, (2022).

[28] Kishor Sharma, 'Politics and governance in nepal', *Asia Pacific Journal of Public Administration*, **34**(1), 57–69, (2012).

[29] Birat Bade Shrestha and Bal Krishna Bal, 'Named-entity based sentiment analysis of nepali news media texts', in *Proceedings of the 6th workshop on natural language processing techniques for educational applications*, pp. 114–120, (2020).

[30] Dipesh Shrestha. DistillBERT(Nepali). https://huggingface.co/dexhrestha/Nepali-DistilBERT, 2021. Accessed: 2023-02-25.

[31] Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi, '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)*, pp. 301–308. IEEE, (2020).

[32] Elizabeth Stuart and Emma Samman, 'Defining 'leave no one behind'', *ODI Briefing Note. London: Overseas Development Institute*, (2017).

[33] Surendrabikram Thapa, Surabhi Adhikari, Usman Naseem, Priyanka Singh, Gnana Bharathy, and Mukesh Prasad, 'Detecting alzheimer's disease by exploiting linguistic information from nepali transcript', in *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pp. 176–184. Springer, (2020).

[34] Sulav Timilsina, Milan Gautam, and Binod Bhattarai, 'Nepberta: Nepali language model trained in a large corpus', in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 273–284, (2022).

[35] Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yılmaz, 'Large-scale hate speech detection with cross-domain transfer', *arXiv preprint arXiv:2203.01111*, (2022).