Hate Speech and Offensive Language Detection in Bengali

Mithun Das, Somnath Banerjee, Punyajoy Saha, Animesh Mukherjee
Indian Institute of Technology Kharagpur, West Bengal, India
mithundas@iitkgp.ac.in, som.iitkgpcse@kgpian.iitkgp.ac.in,
punyajoys@iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Social media often serves as a breeding ground for various hateful and offensive content. Identifying such content on social media is crucial due to its impact on the race, gender, or religion in an unprejudiced society. However, while there is extensive research in hate speech detection in English, there is a gap in hateful content detection in low-resource languages like Bengali. Besides, a current trend on social media is the use of Romanized Bengali for regular interactions. To overcome the existing research's limitations, in this study, we develop an annotated dataset of 10K Bengali posts consisting of 5K actual and 5K Romanized Bengali tweets. We implement several baseline models for the classification of such hateful posts. We further explore the interlingual transfer mechanism to boost classification performance. Finally, we perform an in-depth error analysis by looking into the misclassified posts by the models. While training actual and Romanized datasets separately, we observe that XLM-Roberta performs the best. Further, we witness that on joint training and few-shot training, MuRIL outperforms other models by interpreting the semantic expressions better. We make our code and dataset public for others[1].

1 Introduction

Social media websites like Twitter and Facebook have brought billions of people together and given them the opportunity to share their thoughts and opinions rapidly. On the one hand, it has facilitated communication and the growth of social networks; on the other, it has been exploited to propagate misinformation, violence, and hate speech (Mathew et al., 2019; Das et al., 2020) against users based on their gender, race, religion, or other characteristics. If such content is left unaddressed, it may result in widespread conflict and violence, raising concerns about the safety of human rights, the rule of law, and freedom of speech, all of which are crucial for the growth of an unprejudiced democratic society (Rizwan et al., 2020). Organizations such as Facebook have been blamed for being a forum for instigating anti-Muslim violence in Sri Lanka that resulted in the deaths of three individuals[2], and a UN report accused them of disseminating hate speech in a way that contributed significantly to the plausible genocide of the Rohingya population in Myanmar[3].

In order to reduce the dissemination of such harmful content, these platforms have developed certain guidelines[4] that the users of these platforms ought to comply with. If these rules aren't followed, the post can get deleted, or the user's account might get suspended. Even to diminish the harmful content from their forum, these platforms engage moderators (Newton, 2019) to manually review the posts and preserve the platform as wholesome and people-friendly. However, this moderation strategy is confined by the moderators' speed, jargon, capability to understand the development of slang, and familiarity with multilingual content. Moreover, due to the sheer magnitude of data streaming, it is also an ambitious endeavor to examine each post manually and filter out such harmful content. Hence, an automated technique for detecting hate speech and offensive language is extremely necessary and inevitable.

It has already been witnessed that Facebook vigorously eliminated a considerable amount of malicious content from its platforms even before users reported it (Robertson, 2020). However, the hindrance is that these platforms can detect harmful content in certain popular languages such as En-

[1] https://github.com/hate-alert/Bengali_Hate

[2] https://tinyurl.com/sriLankaRiots

[3] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate

[4] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 286–296
November 20–23, 2022.

glish, Spanish, etc. (Perrigo, 2019) So far, several investigations have been conducted to identify hate speech automatically, focusing mainly on the English language; therefore, an effort is required to determine and diminish such hateful content in low-resource languages.

With more than 210 million speakers, Bengali is the seventh most widely spoken language[5], with around 100 million Bengali speakers in Bangladesh and 85 million in India. Apart from Bangladesh and India, Bengali is spoken in many countries, including the United Kingdom, the United States, and the Middle East[6]. Also, a current trend on social media platforms is that apart from actual Bengali, people tend to write Bengali using Latin scripts(English characters) and often use English phrases in the same conversation. This unique and informal communication dialect is called code-mixed Bengali or Roman Bengali. Code-mixing makes it easier for speakers to communicate with one another by providing a more comprehensive range of idioms and phrases. However, as emphasized by Chittaranjan et al. (Chittaranjan et al., 2014), this has made the task of creating NLP tools more challenging. Along with these challenges, the challenges specific to identifying hate speech in Roman Bengali contain the following: *Absence of a hate speech dataset*, *Lack of benchmark models*. Thus, there is a need to develop open efficient datasets and models to detect hate speech in Bengali. Although few studies have been conducted in developing Bengali hate speech datasets, most of these have been crawled with comments from Facebook pages, and all of them are in actual Bengali. Hence, there is a need for developing more benchmarking datasets considering other popular platforms. To address these limitations, in this study, we make the following contributions.

- First, we create a gold-standard dataset of 10K tweets among which 5K tweets are actual Bengali and 5K tweets are Roman Bengali.
- Second, we implement several baseline models to identify such hateful and offensive content automatically for both actual & Roman Bengali tweets.
- Third, we explore several interlingual transfer mechanisms to boost the classification performance.
- Finally, we perform in-depth error analysis by looking into a sample of posts where the models mis-classify some of the test instances.

2 Related Work

Over the past few years, research around automated hate speech detection has been evolved tremendously. The earlier effort in developing resources for the hate speech detection was mainly focused around English language (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018). Recently, in an effort to create multilingual hate speech datasets, several shared task competitions have been organized (HASOC (Mandl et al., 2019), OffensEval (Zampieri et al., 2019),, TRAC (Kumar et al., 2020), etc.), and multiple datasets such as Hindi (Modha et al., 2021), Danish (Sigurbergsson and Derczynski, 2020), Greek (Pitenis et al., 2020), Turkish (Çöltekin, 2020), Mexican Spanish (Aragón et al., 2019), etc. have been made public. There is also some work to detect hate speech in actual Bengali. Ismam et al. (Ishmam and Sharmin, 2019) collected and annotated 5K comments from Facebook into six classes-*inciteful*, *hate speech*, *religious hatred*, *communal attack*, *religious comments*, and *political comments*. However,the dataset is not publicly available. Karim et al. (Karim et al., 2021) provided a dataset of 8K hateful posts collected from multiple sources such as Facebook, news articles, blogs, etc. One of the problems with this dataset is that all comments are part of any hate class(*personal*, *geopolitical*, *religious*, and *political*), so we cannot build hate speech detection models using this dataset to screen out hate speech. Romim et al. (2021) curated a dataset of 30K comments, making it one of the most extensive datasets for hateful statements. The author achieved 87.5% accuracy on their test dataset using the SVM model. However, these datasets do not consider Roman Bengali posts, a prevalent communication method on social media nowadays.

With regards to the detection systems, earlier methods examined simple linguistic features such as character and word n-grams, POS tags, tf-idf with a traditional classifier such as LR, SVM, Decision Tree, etc (Davidson et al., 2017). With the development of larger datasets, researchers have shifted to data-hungry complex models such as deep learning (Pitsilis et al., 2018; Zhang et al., 2018) and graph embedding techniques to enrich

[5] https://www.berlitz.com/en-uy/blog/most-spoken-languages-world

[6] https://www.britannica.com/topic/Bengali-language

	Actual	Roman	Total
Hateful	825	510	1,335
Offensive	1,341	2,063	3,404
Normal	2,905	2,534	5,439
Total	5,071	5,107	10,178

Table 1: Dataset Statistics of both Actual and Roman tweets.

the classifier performance.

Recently, transformer-based (Vaswani et al., 2017) language models such as BERT, XLM-RoBERTa (Devlin et al., 2019) are becoming quite popular in several downstream tasks. It has already been observed that these transformer-based models outperform several earlier deep learning models (Mathew et al., 2021). Having observed these transformer-based models' superior performance, we focus on building these models for our classification task.

Further, researchers have begun to explore few shot classifications. One of the most popular techniques for few-shot classification is transfer learning - where a model (pre-trained in a similar domain) is further fine-tuned on a few labeled samples in the target domain (Alyafeai et al., 2020). Keeping these experiments in mind, we also examine the ability of transfer learning capabilities between actual and Roman Bengali data.

3 Dataset Creation

In this section, we provide the data collection procedure, annotation strategies we have followed and the statistics of the collected dataset.

3.1 Dataset collection and sampling

In this paper, we collect our dataset from **Twitter**. Despite Hatebase.org maintaining the most extensive collection of multilingual hateful words, it still lacks such lexicon base for Bengali[7]. To sample Bengali (actual and romanized) tweets for annotation, we create a lexicon of 74 abusive terms[8]). These lexicons consist of derogatory keywords/slurs targeting individuals or different protected communities. We also include words based on the name of the targeted communities. The choice to add names of targeted communities is made in order to extract random hateful/offensive tweets that do not contain any abusive words. Using Twitter API, we searched for tweets containing phrases from the lexicons, which resulted in a sample of 500K tweets for actual Bengali and 150K tweets for Roman Bengali. To evade problems related to user distribution bias, as highlighted by Arango et al. (Arango et al., 2019), we limit a maximum of 75 tweets per user. We also do not use more than 500 tweets per month to avoid event-specific tweets in our dataset.

3.2 Annotation procedure

We employed four undergraduate students for our annotation task. All undergraduate students are Computer Science majors and native Bengali speakers. They have been recruited voluntarily through departmental emails and compensated via an Amazon gift card. Two Ph.D. students led the annotation process as expert annotators. Both expert annotators had previous experience working with malicious content on social media. Each tweet in our dataset contains two kinds of annotations: first whether the text is hate speech, offensive speech, or normal; second, the target communities in the text. This additional annotation of the target community can help us measure bias in the model. Table 3 lists the target groups we have considered.

Annotation guidelines: The annotation scheme stated below constitute the main guidelines for the annotators, while a codebook ensured common understanding of the label descriptions. We construct our codebook (which consists the annotation guidelines[8] for identifying hateful and offensive tweets based on the definitions summarized as follows.

- **Hate speech:** *Hate speech is a language used to express hatred toward a targeted individual or group or is intended to be derogatory, humiliating, or insulting to the group members based on attributes such as race, religion, ethnic origin, sexual orientation, disability, caste, geographic location or gender.*
- **Offensive:** *Offensive speech uses profanity, strongly impolite, rude, or vulgar language expressed with fighting or hurtful words to insult a targeted individual or group.*
- **Normal:** *This contains tweets that do not fall into the above categories.*

3.3 Dataset creation steps

As a first step for creating the dataset, we required a pilot gold-label dataset to instruct the annotators. Initially, the expert annotators annotated 100

[7] https://hatebase.org/
[8] https://tinyurl.com/bengaliHate

Type	Tweet	Translation	Label	Target
Actual	এই জাতি যে বর্বর,মুর্খ,ইতর,ধর্মান্ধ সেটা আপনি আজকে বুঝলেন? ব্রিটিশরা কিন্তু আরও ৩০০ বছর আগেই বুঝেছিল। যারা সারাজীবন গোলামী করে আসছে অন্যের সেই বাঙালি জাতির সাথে সভ্য শব্দটা মিলানো উচিত না আসলে।	Do you understand today that this race is barbaric, stupid, mean, fanatical? But the British understood more than 300 years ago! The word 'civilized' should not be associated with the Bengali those who have been enslaved all their lives	Hate	Bengali
	@user স্বপ্নে তোকে চুদি, খানকির মেয়ে, এইজন্য স্বপ্নদোষ হয়	@user I fuck you in dream, daughter of a bitch, this is why I get nightmare	Offensive	Individual, Woman
	@user নাগরিকত্ব আইন নিয়ে প্রশ্ন তুলছে দলিত সংগঠনই https://url	@user Dalits are questioning the citizenship law https://url	Normal	Others
Romanized	@user @user 42 e 42 er ki holo re ganduchoda choti chata niche kata ??? Tor baper gnare dhukiye dilo 42 ta ?? Khankir pola... Kanglu mal... Suorer jaat... 🤣🤣	@user @user What happned to him ass fucker, shoe licker, circumcise man? Out of 42, 42 in your father's ass .. Son of a bitch .. Kanglu (derogatory term for Bangladeshi) .. Pig breed … 🤣🤣	Hate	Bangladeshi
	khankir chele dwijen barik. kal tui sesh. kal tui soshane. kal ami tor bou ke chudbo. kochi maal. LENOVO THE LAORA.	Son of a bitch dwijen barik. Tomorrow you are finish. Tomorrow you will be in the crematorium. I will fuck your wife tomorrow. Young wife. LENOVO THE LAORA.	Offensive	Individual
	@user He got best debutante wid #SBG!? 😐 Then wht abt his film #PaanchAdhyay? Sala amra audience ra ki bokachoda? r koto lobby cholbe!!	@user He got best debutante wid #SBG!? 😐 Then what about his film #PaanchAdhyay? Damm, are we fucking dumb audiences? How much longer will the lobby last?!!	Normal	Others

Table 2: Samples of Actual and Roman Bengali tweets for each label from the dataset

Target Groups	Categories
Gender	Men, Women, Trans.
Linguistic Community	Bengali, Bihari.
National Origin	Indian, Bangladeshi, Pakistani.
Religion	Hindu, Islam.
Miscellaneous	Individual, Political, Disabled, Dalit, Others.

Table 3: Target groups considered for the annotation.

tweets, out of which 30 were hateful, 35 were offensive, and the rest 35 tweets were normal.

Pilot annotation: Each annotator was given 30 tweets from the gold-label dataset in the pilot task. They were asked to classify hate/offensive speech and identify the target community (if any). They were provided the annotation codebook with multiple examples for the labeling process to understand the task clearly. They were asked to keep the annotation guidelines open while doing the annotation to have better clarity about the labeling scheme. After the annotators finished this set, we consulted the incorrect annotations in their set with them. This activity further trained the annotators and helped to fine-tune the annotation scheme. In addition, we collected feedback from annotators to enrich the main annotation task.

Main annotation: After the training process, we proceeded with the main annotation task. For this task, we use the open-source platform Doccano[9], deployed on a Heroku instance. We provided a secure account to each annotator where they could annotate and track their progress. Two independent annotators annotated each tweet. Based on the guidelines, they were instructed to read the entire tweet and select the appropriate category (hate speech, offensive, or standard). Initially, we started with a small batch of 100 tweets and later expanded it to 500 tweets as the annotations became more efficient. We tried to preserve the annotators' agreement by pointing out some errors they made in the previous batch. Since hate/offensive speech is highly polarizing and adverse, the annotators were given plenty of time to complete the annotations. On completion of each set of annotations, if there was a mismatch between two annotators, one of the expert annotators annotated the same tweet to break the tie. For the cases where all the three annotators chose a separate class, we did not consider these tweets for further analysis. To determine the target community of a tweet, we combine the annotated targets.

Exposure to online abuse could lead to unhealthy mental health issues[10](Ybarra et al., 2006). Therefore, the annotators were recommended to take periodic breaks and not do the annotations in one sitting. Besides, we also had weekly meetings with them to ensure the annotations did not have any effect on their mental health.

Final dataset: Table 1 notes our final dataset statistics. It consists of 5,071 actual Bengali tweets (out of which 825 have been labelled as hateful, 1,341 are offensive, and 2,905 tweets are normal) and 5,107 Roman Bengali tweets (out of which 510 tweets are hateful, 2,063 tweets are offensive, and 2,534 tweets are normal). We achieved an inter-annotator agreement of 0.696 using Krippendorff's α which is better than the agreement score on other related hate speech tasks (Ousidhoum et al., 2019; Guest et al., 2021). In Table 2 we have shown some

[9] https://github.com/doccano/doccano

[10] https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd

examples of Bengali hate speech and offensive language that we have annotated.

4 Methodology

4.1 Baseline models

In this section, we discuss the models we implement for automatic detection of hate speech. We experimented with a wide range of models for our use case.

m-BERT (Devlin et al., 2019) is a stack of transformer encoder layers consisting of 12 "attention heads" with self-attention mechanisms. It is pre-trained on 104 languages using a masked language modeling (MLM) objective with the crawled Wikipedia data. To fine-tune m-BERT, we include a fully connected layer with the output corresponding to the CLS token in the input. Typically, the expression of the sentence provided to the model is retained in this CLS token output. In hate speech, the m-BERT model has been well studied, outperformed several baselines, and is considered state-of-the-art.

XLM-Roberta (Conneau et al., 2020) is another form of Transformer model, pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. XLM-R was trained using a lot more data compared to m-BERT. Similar to BERT, it is a stack of transformer encoder layers with 12 "attention heads" and can handle at max 512 tokens.

IndicBERT (Kakwani et al., 2020) is a multilingual ALBERT model (Lan et al., 2019) (a recent derivative of BERT) trained on large-scale corpora, covering 12 major Indian languages. It is pre-trained on 9 billion tokens and evaluated on a set of diverse tasks. Unlike m-BERT, XLM-Roberta, IndicBERT has around 10x fewer parameters and still manages to deliver state-of-the-art performance on several tasks.

MuRIL (Khanuja et al., 2021) stands for Multilingual Representations for Indian Languages and aims to enrich reciprocity from one language to another. This model uses a BERT base architecture pre-trained from scratch using the Common Crawl, Wikipedia, PMINDIA, and Dakshina corpora for 17 Indian languages and their transliterated counterparts.

4.2 Interlingual transfer mechanisms

One of the main attractions of transformer-based models is their potential to strengthen model transfer via several mechanisms. This can be especially beneficial for enhancing learning performance in low-resource languages like Bengali. In order to evaluate the extent to which language similarity improves transfer learning performance, we perform the following tests.[11]

ELFI (Each language for itself): In this situation, we use the same language's data for training, validation, and testing. This scenario typically appears in the real world, where monolingual datasets are frequently utilized to build classifiers for a particular language. Despite the anticipated high labeling costs, this gives an idea of the most achievable classification performance.

Joint training: In this setting, we integrate both actual & Roman Bengali posts to train all the transformer-based models. The notion is that even though the characters used to represent both languages are different, their semantic content is mostly the same. Hence, it gives an idea of whether jointly training the models can benefit learning the better semantic representation of a particular post for determining the corresponding label of the post.

Model transfer: In this scenario, the models are trained with one language (source language) and evaluated in another language (target language). In the zero-shot setting, no instances from the target language have been used while training (**MTx0**). In a related few-shot setting, we allow n = 32, 64, and 128 posts per label from the available gold target instances to fine-tune the existing models (trained in another language). These are named **MTx32, MTx64** and **MTx128**.

Language transfer: In this setting, we translate the Bengali posts to English using Google Translate tool[12] and do the entire training, testing on the translated instances. We do this to check if language space has been transformed for a task, how model's performance varies.

Joint training with language transfer: In this scenario, we combine the translated Bengali and Roman Bengali posts, to train all the transformer based models. The motivation behind this experiment is that, in case of romanized Bengali data, people use English words/sentences in their posts for ease of writing. Thus, we perform this experiment to determine whether adding translated Bengali data points will further improve the performance of the classification or not.

[11] Although the discussed models have been pre-trained using multiple languages, fine-tuning has been done using the Bengali language dataset.

[12] https://cloud.google.com/translate

4.3 Experimental setup

All the models are evaluated using the same 70:10:20 train, validation, and test split, stratified by class across the splits. For the model transfer evaluation, we use 32, 64, and 128 training data points from each class to train the model in another language. We create three such different random sets for the target dataset to have a more robust assessment and report the average performance. The models were run for 10 epoch with Adam optimizer, batch_size = 16, learning_rate = $2e-5$ and adam_epsilon = $1e-8$. In addition, we set the number of tokens n = 400 for all the models.

4.4 Evaluation metric

To remain consistent with existing literature, we evaluate our models in terms of **accuracy**, **F1-score** and **AUROC** score. These metrics together should be able to thoroughly evaluate the classification performance in distinguishing among the three classes, e.g., hate, offensive and normal. For zero-shot and few-shot settings, we report only **macro F1-score** due to paucity of space. We also highlight the best performance using **bold** and second best using underline.

5 Results

In this section, we discuss the findings of our experiments.

5.1 Performnace of ELFI

In Table 4, we report the performance of all the models for actual & Roman Bengali. We observe for both of these, XLM-Roberta performs the best in terms of accuracy and macro-F1 score. Followed by XLM-Roberta, MuRIL performed the second best for the actual Bengali. For the hate class m-BERT does slightly better than XLM-Roberta in terms of F1-score. For Roman Bengali, IndicBERT performs next to XLM-Roberta.

5.2 Performance of joint training

Here we investigate the importance of joint training. Even though both the actual & Roman Bengali is written using different characters, semantic expression of both the languages are same. Table 5 summarizes the performance of different models when trained jointly. We observe some improvements in the joint training models. In particular, MuRIL, which is pretrained on both Indian languages and their transliterated counterparts, is able to interrelate the semantics of the actual & Roman Bengali sentences. We notice that for actual Bengali, MuRIL performs the best with Macro-F1 score of 0.808 (and accuracy of 0.833), followed by m-BERT with Macro-F1 score of 0.800 (and accuracy of 0.829). For the Roman Bengali though, XLM-Roberta still performs the best (with Macro F1-score of 0.810), MuRIL performs very close to it and in fact better for the hate class F1-score.

5.3 Performance of model transfer

In this scenario, we investigate the power of existing fine-tuned models. The idea is to understand how these models are generalized across the same language, having same semantic content, but are written using different characters/words. We report our results in Table 6.

In **zero-shot** setting we observe, when the model is trained on actual Bengali and tested on Roman Bengali, m-BERT performs the best (with macro F1 score of 0.390) among all the models followed by IndicBERT (Macro F1 score 0.319). On the other hand, when trained on Roman Bengali and tested on actual Bengali, MuRIL performs the best (macro F1 score 0.414) among all the models followed by IndicBERT (macro F1 score 0.397). An interesting thing to note is that although the XLM-Roberta performs best in monolingual settings, it is not performing well in the model transfer setup.

To further investigate, how the performance of these models would vary, we conduct a second stage of fine-tuning. In this setting we use the existing trained model in actual Bengali and further fine-tune it with n samples of Roman Bengali data points per label (and vice-versa). we repeat the subset sampled data selection with 3 different random sets and report the average performance. This will help to reduce performance variations across different sets. In general We observed with the increasing data points the performance of all models has improved.

- **Actual** $\rightarrow$ **Roman**: We observe further fine-tuning the model with 32 instances, m-BERT performs the best followed by MuRIL. While increasing these instances, MuRIL outperforms all other models. Only with 128 instances per label, MuRIL achieves macro F1-Score of 0.751.
- **Roman** $\rightarrow$ **Actual**: We see MuRIL outperforms all other models. Followed by MuRIL, XLM-Robera performed the second best for

	Actual Bengali					Roman Bengali				
Model	**Acc**	**M-F1**	**F1(H)**	**F1(O)**	**AUROC**	**Acc**	**M-F1**	**F1(H)**	**F1(O)**	**AUROC**
m-BERT	0.813	0.795	**0.824**	0.693	0.917	0.840	0.789	0.658	0.840	0.910
XLM	**0.830**	**0.803**	0.812	**0.717**	**0.919**	**0.858**	**0.805**	**0.666**	**0.857**	**0.924**
MuRIL	0.817	0.797	0.816	0.704	0.887	0.843	0.788	0.646	0.841	0.897
IndicBERT	0.790	0.767	0.788	0.656	0.896	0.846	0.793	0.651	0.846	0.908

Table 4: Performance on Both Actual & Roman Bengali Datasets. XLM:XLM-Roberta, M: Macro, Acc: Accuracy, H: Hate, O: Offensive.

	Actual Bengali					Roman Bengali				
Model	**Acc**	**M-F1**	**F1(H)**	**F1(O)**	**AUROC**	**Acc**	**M-F1**	**F1(H)**	**F1(O)**	**AUROC**
m-BERT	0.829	0.800	0.831	0.684	**0.914**	0.845	0.789	0.647	0.830	**0.928**
XLM	0.819	0.794	0.805	0.701	0.912	**0.865**	**0.810**	0.666	**0.867**	0.918
MuRIL	**0.833**	**0.808**	**0.835**	**0.704**	0.895	0.850	0.800	**0.670**	0.842	0.904
IndicBERT	0.785	0.769	0.807	0.658	0.900	0.817	0.767	0.637	0.808	0.890

Table 5: Performance of Both Actual & Roman Bengali Datasets on Joint Training. XLM:XLM-Roberta, M: Macro, Acc: Accuracy, H: Hate, O: Offensive.

Actual Bengali Model's Performance on Roman Bengali				
Model	**Zero-Shot (MTx0)**	**Few-Shot (MTx32)**	**Few-Shot (MTx64)**	**Few-Shot (MTx128)**
m-BERT	**0.390**	**0.530**	0.655	0.692
XLM-Roberta	0.230	0.456	0.570	0.668
MuRIL	0.269	0.507	**0.671**	**0.751**
IndicBERT	0.319	0.332	0.355	0.462
Roman Bengali Model's Performance on Actual Bengali				
Model	**Zero-Shot (MTx0)**	**Few-Shot (MTx32)**	**Few-Shot (MTx64)**	**Few-Shot (MTx128)**
m-BERT	0.268	0.449	0.608	0.691
XLM-Roberta	0.299	0.542	0.613	0.664
MuRIL	**0.414**	**0.575**	**0.645**	**0.709**
IndicBERT	0.397	0.463	0.508	0.557

Table 6: Performance of Zero-shot & Few-shot Learning.

32 and 64 instances and for 128 instances m-BERT is the second best.

5.3.1 Performance of language transfer

Here we investigate the importance of gold instances[13] in a low resource language. We do so by transforming the language space. We translate[12] the Bengali datasets to English and do training, testing on the translated dataset. In Table 7 we report the results of all the models. Although XLM-Roberta outperforms all other models, an important point to note is that its performance (Macro-F1 score 0.764) is much lower compared to the model trained on the gold (i.e., actual Bengali) instances (Macro-F1 score 0.803).

Model	**Acc**	**M-F1**	**F1 (H)**	**F1 (O)**	**AUROC**
m-BERT	0.777	0.754	**0.775**	0.647	**0.893**
XLM-Roberta	**0.796**	**0.764**	0.757	**0.649**	0.891
MuRIL	0.771	0.722	0.728	0.586	0.830
IndicBERT	0.723	0.671	0.650	0.540	0.826

Table 7: Performance on Translated Data. M: Macro, Acc: Accuracy, H: Hate, O: Offensive.

[13] raw labeled data

Model	**Acc**	**M-F1**	**F1(H)**	**F1(O)**	**AUROC**
m-BERT	**0.856**	**0.811**	0.694	**0.852**	**0.930**
XLM-Roberta	0.849	0.799	0.670	0.847	0.910
MuRIL	0.845	0.791	0.647	0.844	0.895
IndicBERT	0.839	0.787	0.649	0.830	0.911

Table 8: Performance of Roman Bengali on Joint Training with the Translated Data. M: Macro, Acc: Accuracy, H: Hate, O: Offensive.

5.4 Performance of joint training with language transfer

In this scenario we investigate, even though models trained on translated Bengali instances cannot outperform the monolingual models trained on gold labels, can it be useful to improve the performance of Roman Bengali data? This is motivated by the fact that in a romanized(code-mixed) scenario, people mix English words/phases while writing. Table 8 shows the results on the code-mixed test set. We monitor the performance of m-BERT (Macro-F1 score: earlier (0.790), now (0.811)) and MuRIL (Macro-F1 score: earlier (0.788), now: (0.791)) and observe that these have improved for the detection in the Roman Bengali dataset. However, for XLM-Roberta (Macro-F1 score: earlier (0.805), now (0.799)) and IndicBERT (Macro-F1 score: earlier (0.793), now (0.787)) the models perform slightly worse compared to those trained on only Roman Bengali gold data. Overall, it can be concluded that while some models are able to leverage the strength of the translated Bengali data while predicting the labels of the Roman Bengali posts, others are not. This might hint at the differences in the generalizability powers of these models. To understand this better, in section 7 we deep dive into the models further using error analysis techniques.

Train	Test	Acc	MF1
Romin	Romin	0.905	0.894
Romin	Ours	0.646	0.646
Ours	Ours	0.846	0.843
Ours	Romin	0.774	0.754
Joint	Romin	0.910	0.899
	Ours	0.837	0.835

Table 9: Comparison with existing dataset (Romim et al., 2021). Acc: Accuracy, MF1: Macro-F1

6 Additional experiment

In addition, we perform another experiment to further compare the quality of our dataset with the existing dataset of Romim et al. (Romim et al., 2021). Using their dataset, we train the XLM-Roberta model[14] and test its performance on our dataset. Likewise, we test the performance on their dataset when the model is trained on our dataset. We only conduct this experiment with the actual Bengali tweets to have valid comparison with their dataset. We combine hate and offensive into a single class for this experiment, as the authors in (Romim et al., 2021) have considered these two labels as same. In Table 9 we summarize the results. We observe our model achieves macro-F1 score of 0.754 on their dataset, while the model trained on their dataset achieves 0.646 macro-F1 score on our dataset. Further, we train the XLM-Roberta model jointly with both datasets. We observe jointly training the model further improved the performance on the Romim et al. (Romim et al., 2021) test data; however, we do not see any improvement in our test data.

7 Error analysis

In order to deep dive into the models further, we conduct a manual error analysis on our models by using a sample of 50 posts where the model incorrectly categorizes some test instances. We analyze common errors and classify them into the following five categories.

- **Sarcastic content consisting emojis**: Communication via emojis is becoming extremely popular these days. Sometimes these emojis completely change the interpretation of the post by making it sarcastic/ambiguous. This naturally results in mis-classification.
- **Sequence of obscene words**: Some instances of a series of swear words not targeting individuals or communities are mis-classified. This indicates that the presence of hateful, obscene keywords should not be the only decisive factor for a model to make its predictions.
- **Viewpoints**: Some instances mostly relating to a political or religious sense cannot be fully binary or ternary. The annotators' viewpoint plays a key role in such instances and makes the models to mis-classify these instances. All the models suffer similarly here.
- **Code-mixed linguistic structure**: Instances following the grammatical structure of Bengali but written using English words sometimes get mis-classified due to the code-mixed nature of data at hand because there is a heavy between the tokens from Bengali and English.
- **Tentatively wrong ground truths**: Some instances containing slur words many not be targeting any group as such. However annotators tentatively marked it hateful leading to the model mis-classifying the post.

In Table 10 we present example instances for the above categories and the predictions thereof. Though we show the predictions for XLM-Roberta, all the other models also produce similar results.

8 Discussion

In this section we discuss the key insights from our results. We observe that depending on the availability of training data points, the performance of the model varies. When we have sufficient number of training instances XLM-Roberta model performs the best. Further we argue that when actual & Roman Bengali instances are merged together for joint training, models like MuRIL performs the best by leveraging the semantic connection between actual and Romanized instances. This is, to some extent, expected from MuRIL due to the nature of its pretraining mechanism, where both actual language and its transliterated counterpart have been used.

Further exploring the performance of these models in zero-shot setting shows, although XLM-Robera performs best while trained with standalone data, it performs very poorly for unseen data with similar semantic content but a different orthography. In such scenarios, models like m-BERT, IndicBERT exhibit better performance. To improve

[14]We consider XLM-Roberta, as this performs the best while training standalone.

Posts	Translation	Ground Truth	Predicted Label	Category
@user আপনি কি বিপথগামী রাস্তার কুকুরের মেয়ে? 🐕🤦‍♂️😀😜	@user Are you a misguided street dog girl? 🐕🤦‍♂️😀😜	Offensive	Non-Hate	**Sarcastic content consist emoji**
শালা গুদমারানি বোকাচোদা ট্রেন	Damn pussy fucking stupid train	Non-Hate	Offensive	**Sequence of obscene words**
@user Sir Sudhu mullader noy. TMC CPIM er hinduder o hoy🤣🤣🤣	@user Sir not only mullahs. TMC CPIM's hindus too 🤣🤣🤣	Hate	Non-Hate	**Angle of viewpoints**
@user Idiots BJP, & There Blind Vokto not be understood Your language! দিনরাত তো জয়শ্রার, গোমূত্র, হিন্দু মুসলিম আর পাকিস্তান করেই ক্লান্ত, অসভ্য বর্বর বিজেপির সুশিক্ষা থাকলে না হয় সভ্যদের মর্যাদা দেবে!!	@user Idiots BJP, & There blind supporter will not understood Your language! Tired of Jai Shri Ram, cow urine, Hindu Muslims and Pakistan day and night, If the uncivilized barbarian BJP had better education, it would have given dignity to the civilized!!	Offensive	Non-Hate	**Code-mixed linguistic nature**
@user তারা এই দেশে মদিকে আনবে এবং মদির সাপট করবে তারা সবাই হিন্দু গরু মদির চোদা	Those who will bring Modi in this country and support Modi are all Hindu cows, Modi's fucker	Hate	Offensive	**Tentatively wrong ground truths**

Table 10: Error analysis on XLM-Roberta (we found similar trend on other models).

the performance of these models, when some instances from the target language are used, MuRIL shows an increase in performance at a rate higher than the other models. Observing these results it may be safe to say that when there is a data scarcity for a particular language, it is better to reuse existing fine-tuned models in the same domain. Also careful selection of model is needed. In our case, actual Bengali and Roman Bengali use different characters for writing, but their semantic expressions are same, which is why MuRIL performed best overall.

While doing the in-depth error analysis, we also found that for some cases it can be even difficult for a model to find the actual label correctly. Not only models, as hate speech is complex in nature, sometimes annotators make mistake while labelling them due to differing viewpoints.

Limitation: There are a few limitations of our work. First is the lack of external context. We have not considered any external context such as profile bio, history of user's posting pattern, gender etc., which might be helpful for the hate speech detection task. Although the effectiveness of these transformer-based models are quite good, they have not been tested against adversarial examples.

9 Conclusion

This paper presents a new benchmark dataset for Bengali hate speech detection, consisting of 10K posts from Twitter, covering both actual & Roman scenarios. Each tweet was annotated with one of the hate/offensive/normal labels. We assessed different transformer-based architectures for hate speech detection. We also experimented with several interlingual transfer mechanisms. Our experiments show how few-shot techniques could be beneficial. Besides, we saw how joint training performs better than training on standalone data. We further notice that joint transliterated training performs best in the case of the Roman Bengali dataset. Our error analysis reveals some of the typical shortcomings of the transformer models.

As part of the future work, we plan to evaluate the robustness of these models' under adversarial attack as hateful users keep contriving newer ways to deceive the standard hate speech detection models. Another direction could be lessening the biases that can be present in the dataset/model.

Ethical considerations

We only analyzed publicly available data crawled via Twitter API. We followed standard ethical guidelines (Rivers and Lewis, 2014), not making any attempts to track users across platforms or deanonymize them. We have added a data statement (Bender and Friedman, 2018) in the appendix. Although we achieved good performance and the results look promising, these models cannot be deployed directly on a social media platform without rigorous testing. Further study might be needed to track the presence of unintended bias towards specific target communities.

References

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Mario Ezra Aragón, Miguel Angel Alvarez Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberLEF@ SEPLN*, pages 478–494.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *ACM SIGWEB Newsletter*, (Autumn):1–8.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 555–560. IEEE.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of WebSci*. ACM.

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

S Modha, T Mandl, GK Shahi, H Madhu, S Satapara, T Ranasinghe, and M Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indoaryan languages and conversational hate speech. In *FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021*.

Casey Newton. 2019. The terror queue.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684.

Billy Perrigo. 2019. Facebook's hate speech algorithms leave out some languages.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119.

Georgios K. Pitsilis, H. Ramampiaro, and H. Langseth. 2018. Detecting offensive language in tweets using deep learning. *ArXiv*, abs/1801.04433.

Caitlin Rivers and Bryan Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman urdu. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2512–2522.

Adi Robertson. 2020. Facebook says ai has fueled a hate speech crackdown.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468. Springer.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, 118(4):e1169–e1177.

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). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

A Data statement

A.1 Curation rationale

The dataset consists of a collection of Tweets in actual and roman Bengali. To crawl the dataset, Twitter API has been used.

A.2 Language variety

The languages of the dataset are in Bengali (bn), Roman Bengali (bn-En).

A.3 Speaker demographic

- Twitter users
- Age: Unknown – mixed.
- Gender: Unknown – mixed.
- Race/Religion: Unknown – mixed.
- Native language: Unknown; Bengali speakers.
- Socioeconomic status: Unknown – mixed.
- Geographical location: Unknown; mostly from Bangladesh & India.

A.4 Annotator demographic

- Age: 22-29.
- Gender: 2 male & 2 female.
- Race/Religion: prefer not to disclose.
- Native language: Bengali.
- Socioeconomic status: undergraduate students.

A.5 Speech situation

Discussions held in public on Twitter platform.

A.6 Text characteristics

All the sentences in this dataset come from Twitter.

A.7 Other

N/A