Abusive content detection in transliterated Bengali-English social media corpus

Salim Sazzed
Old Dominion University
Norfolk, VA, USA
ssazz001@odu.edu

Abstract

Abusive text detection in low-resource languages such as Bengali is a challenging task due to the inadequacy of resources and tools. The ubiquity of transliterated Bengali comments in social media makes the task even more involved as monolingual approaches cannot capture them. Unfortunately, no transliterated Bengali corpus is publicly available yet for abusive content analysis. Therefore, in this paper, we introduce an annotated corpus of 3000 transliterated Bengali comments categorized into two classes, *abusive* and *non-abusive*, 1500 comments for each. For baseline evaluations, we employ several supervised machine learning (ML) and deep learning-based classifiers. We find support vector machine (SVM) classifier shows the highest efficacy for identifying abusive content. We make the annotated corpus publicly available for the researchers to aid abusive content detection in Bengali social media data.

1 Introduction

With the popularity of social media, nowadays, user-generated contents are available in many languages. In various social media platforms, such as review forums and social networking sites, users express their feelings, opinions, emotion, etc. Due to the open nature of social media, the presence of abusive, offensive, and hateful comments is common there (Schmidt and Wiegand, 2017; Wang et al., 2014).

Abusive language refers to the usage of demeaning, insulting, vulgar, or profane expression to attack individuals or groups (Nobata et al., 2016); However, there exist inconsistencies in the definitions of abusive language in various literature (Waseem et al., 2017). For example, Nobata et al. (2016) considered hate speech as a kind of abusive language, while Founta et al. (2018) distinguished it from abusive speech. As the presence of abusive and hatred content inflicts a negative impact on society and individuals (Nobata et al., 2016; Duggan, 2017; Park et al., 2018), it is important to identify them. While plenty of resources are available for abusive language detection in English (Poletto et al., 2020), limited research has been performed on abusive content analysis in low resource Bengali language.

Code-Mixing (CM) is a natural phenomenon of embedding linguistic units such as phrases, words, or morphemes of one language into an utterance of another (Myers-Scotton, 1993; Muysken et al., 2000). Transliteration can be considered as a special form of code-mixing where the phonetic transformations of the words from a source language to a target language is performed. The presence of code-mixing and transliterated Bengali (i.e., Bengali text using the Latin alphabet) is a common phenomenon in Bengali, as shown by the previous studies (Barman et al., 2014; Chanda et al., 2016).

The existing research on abusive content or hate speech detection in Bengali mainly investigated text written in Bengali (Kumar et al., 2021; Emon et al., 2019; Eshan and Hasan, 2017; Ishmam and Sharmin, 2019; Karim et al., 2020; Romim et al., 2020). Although a few works addressed the phenomenon of code-switching in word-level or sentence level (i.e., presence of both English and Bengali words written using the alphabet of the corresponding language), most of them did not consider transliterated Bengali. Only Jahan et al. (2019) utilized a small number of transliterated Bengali comments (around 200 abusive comments) in their study. English is the second language in Bangladesh, the country with the highest number of Bengali native speakers; therefore, transliterated Bengali is ubiquitous in Bengali social media content. Hence, to detect abusive content in Bengali social media, it is essential to consider the transliterated Bengali text. For example, ”Dor besorom mor tui” is an abusive comment which is written in transliterated Bengali;

Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pages 125–130
June 11, 2021. ©2021 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_016

the corresponding English translation is, " You are shameless, you die". The monolingual approaches can not identify it as an abusive comment as the transliterated words neither exist in the Bengali or English dictionary nor available in the monolingual training data.

Supervised ML classifiers are more effective for abusive content detection than the word-list based approaches, as shown in previous studies (Nobata et al., 2016; Park and Fung, 2017). However, ML classifiers require annotated training data, which are missing for transliterated Bengali text. Therefore, in this work, we develop an annotated corpus for transliterated Bengali for abusive content detection and make them publicly available [1].

We manually annotate around 3000 transliterated Bengali comments collected from YouTube into abusive and non-abusive categories, 1500 for each category. To the best of our knowledge, this is the largest annotated transliterated Bengali corpus for abusive content analysis. We then employ popular ML classifiers, logistic regression (LR), support vector machine (SVM), random forest (RF), and deep learning-based bidirectional long short-term memory (BiLSTM) architecture for baseline evaluations.

1.1 Contributions

The major contributions of this work can be summarized as follows:

- We introduce a large transliterated Bengali corpus consisting of 3000 comments collected from YouTube.
- We manually annotate the transliterated comments into abusive and non-abusive categories.
- We provide the comparative performances of various supervised ML and deep learning-based classifiers for recognizing abusive content in the transliterated Bengali corpus.

2 Related Work

Researchers explored code-mixed and transliterated content for tasks like linguistic analysis, Part-of-Speech (POS) tagging, and sentiment analysis in various South Asian languages, such as Hindi and Bengali (Choudhury et al., 2010; Jamatia et al., 2015; Patwa et al., 2020). Mathur et al. (2018) introduced a Twitter dataset for the classification of offensive tweets written in the Hindi-English code-switched language. However, such a dataset for abusive content analysis in transliterated Bengali is not available yet.

2.1 Abusive Content Analysis in Bengali Text

Emon et al. (2019) applied linear support vector classifier (LinearSVC), logistic regression (LR), multinomial naïve bayes (MNB), random forest (RF), artificial neural network (ANN), and recurrent neural network (RNN) with a long short term memory (LSTM) to detect multi-type abusive Bengali text. They also introduced a stemming rule to improve the classifier performance. Eshan and Hasan (2017) investigated the performance of RF, MNB, SVM classifiers for abusive language detection using unigram, bigrams, and trigram based feature vectors. They found that the SVM classifier with linear kernel and tri-gram features showed the highest accuracy.

Ishmam and Sharmin (2019) employed traditional and deep learning-based ML algorithms for classifying different types of offensive comments collected from Facebook pages. They collected and annotated around 5000 Bengali comments and categorized them into six classes. They obtained the highest accuracy utilizing GRU based model, which is around 70.10%. Hussain et al. (2018) collected 300 comments from Facebook and an online newspaper for abusive content detection. They proposed a weighted-rule based method that utilized labeled data. Awal et al. (2018) employed Naïve Bayes (NB) classifier to detect the abusive content in Bengali; They collected text from YouTube and provided the performance of NB using 10-fold cross-validation. Chakraborty and Seddiqui (2019) employed MNB, SVM, Convolutional Neural Network (CNN) with LSTM classifiers. They leveraged both emoticons and Bengali characters as input. They found SVM with linear kernel performed best with 78% accuracy.

2.2 Abusive Content in Transliterated Bengali

In Jahan et al. (2019), the authors utilized Bengali-English code-mixed text and transliterated Bengali text in addition to the Bengali only text. They collected comments from several public Facebook pages. As input features, they used unigrams, bigrams, the number of likes, emojis along with their categories, sentiment

[1] https://github.com/sazzadcsedu/AbusiveCorpus.git

scores, offensive and threatening words used in the comments. They employed three Machine Learning classifiers, SVM, RF, and Adaboost for abusive speech detection.

As we mentioned earlier, most of the existing works considered only the Bengali text. Although few of them utilized code-switching text, transliterated Bengali is hardly explored. To the best of our knowledge, this is the first work that introduces a large annotated corpus of transliterated Bengali for abusive language detection and provides comparative performances of ML classifiers.

3 Corpus Creation

The developed corpus contains user-generated transliterated Bengali text regarding several Bengali dramas and celebrities (i.e., opinion data).

3.1 Data Collection

Using a web scraping tool, we first download the raw JSON data from YouTube that contains information such as user name, id, timestamp, comments, and like/dislike, etc. Utilizing a parsing script, we extract the viewer's comments from the JSON data.

3.2 Data Filtering

The comments are written in Bengali, English, transliterated Bengali, or using code-switching words. Since our goal is to create a corpus for transliterated Bengali (i.e., Bengali words in Latin alphabet), we exclude comments written using the Bengali alphabet (i.e., Bengali comments). We utilize a language detection tool[2] to distinguish comments written using the Latin alphabet and Bengali alphabet. However, the tool can not differentiate between English and transliterated Bengali words as both use the Latin alphabet. Since social media contains lots of non-dictionary and misspelled English words (especially when written by non-native speakers), checking the English dictionary is not a feasible option to distinguish English and Transliterated Bengali words. Therefore, we manually inspect all the comments to include them in the corpus. We discard comments which are written using only English words. Comments with both transliterated Bengali and English words are included in the corpus if they contain at least two transliterated Bengali words. Note that, unlike Bengali or English words, there is not fixed spelling for transliterated Bengali words; thus, the same transliterated word with different spellings can be present in the corpus.

3.3 Data Annotation

3.3.1 Annotation Guideline

For assigning the transliterated Bengali comments into abusive or non-abusive categories, we follow a similar guideline of Nobata et al. (2016). They labeled a piece of text as abusive if it contains either hate speech or derogatory language or profanity.

Based on that, we assign the class of the comments into two categories-

- *Abusive*: This class includes hate speech which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, etc. Besides, it consists of derogatory or demeaning remarks which attack an individual or a group and profanity towards individuals using sexually offensive and pornographic comments.

- *Non-abusive*: comments which do not fall into the abusive category. These comments could convey a positive, (non-abusive) negative or neutral opinion or could be objective in nature.

3.3.2 Inter-annotator Agreement

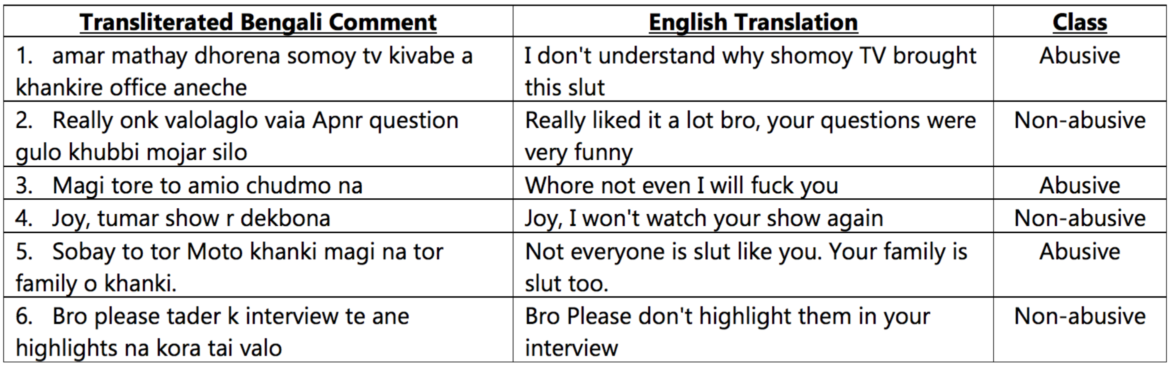
Two native Bengali speakers assign the class of the transliterated comments into two categories, *abusive* and *non-abusive*. Among the 3764 transliterated comments, both annotators assign 3000 comments into the same class; 323 comments are identified as abusive by the first annotator only, while 141 comments are rated abusive by the second annotator only. We observe a Cohen's kappa score of 0.733 between two annotators, which refers to substantial agreement.

3.4 Corpus Statistics

The final corpus includes the 3000 comments which are assigned to the same class by both annotators, 1500 from each category. To avoid ambiguity, we exclude the comments in which annotators disagree on class assignment (Awal et al., 2020; Waseem, 2016).

Each comment contains one or multiple sentences and 1-300 words. We purposely make the

[2] https://github.com/Mimino666/langdetect



Transliterated Bengali Comment	English Translation	Class
1. amar mathay dhorena somoy tv kivabe a khankire office aneche	I don't understand why shomoy TV brought this slut	Abusive
2. Really onk valolaglo vaia Apnr question gulo khubbi mojar silo	Really liked it a lot bro, your questions were very funny	Non-abusive
3. Magi tore to amio chudmo na	Whore not even I will fuck you	Abusive
4. Joy, tumar show r dekbona	Joy, I won't watch your show again	Non-abusive
5. Sobay to tor Moto khanki magi na tor family o khanki.	Not everyone is slut like you. Your family is slut too.	Abusive
6. Bro please tader k interview te ane highlights na kora tai valo	Bro Please don't highlight them in your interview	Non-abusive

Figure 1: Examples of annotated abusive and non-abusive reviews

corpus class-balanced to avoid introducing any bias in the classifier. Figure 1 shows some examples of original transliterated Bengali comments, corresponding English machine translations, and annotations.

3.4.1 Word Frequency Distribution

We manually investigate the presence of English and transliterated Bengali words in the comments by randomly selecting 100 abusive and 100 non-abusive comments. After tokenizing the 100 abusive comments, we find 1720 words. A manual inspection on them identifies 412 English words and 1308 Transliterated Bengali words, which indicates that 76% of the words are transliterated Bengali words. Among the 1088 words in the 100 non-abusive comments, we notice 858 transliterated Bengali and 230 English words, which reveals that nearly 80% of words are transliterated Bengali. As we discard the words written using the Bengali alphabet in the data filtering step, no Bengali words are present in the final corpus.

4 Baseline Classifiers

4.1 Traditional ML Classifiers

Three popular supervised ML classifiers, LR, RF, and SVM are employed to identify abusive comments. We extract unigrams and bigrams from the text and calculate the term frequency-inverse document frequency (TF-IDF) scores for them, which are used as an input for the ML classifiers. TF-IDF is a numerical statistic that aims to reflect the importance of a word to a document in a corpus. We utilize the LR, RF, and SVM implementation of scikit-learn (Pedregosa et al., 2011) library. The default parameter settings of ML classifiers are used.

4.2 Deep Learning Classifier

Furthermore, we apply the deep learning-based BiLSTM architecture for identifying abusive content. For BiLSTM, we use word embedding of 100-dimensional vectors trained on the transliterated corpus. A dropout rate of 0.25 is applied in the dropout layers; Rectified Linear Unit (ReLU) activation is used in the intermediate layers. In the final layer, softmax activation is employed. As an optimization function, Adam optimizer (Kingma and Ba, 2014), and as a loss function, binary-cross entropy is utilized. We set the batch size to 64, use a learning rate of 0.001, and train the model for 6 epochs. We use the Keras (Chollet et al., 2015) library for implementing BiLSTM model.

5 Results and Discussion

We report the precision (P_{abus}), recall (R_{abus}) and F1 scores ($F1_{abus}$) of various classifiers for identifying abusive comments. The TP, FP, and FN values of the abusive class are defined as follows-

TP = abusive review classified as abusive

FP = non-abusive review classified as abusive

FN = abusive review classified as non-abusive

$$R_{abus} = \frac{TP}{TP+FN}\ ,\ P_{abus} = \frac{TP}{TP+FP}$$

$$F1_{abus} = \frac{2*P_{abus}*R_{abus}}{P_{abus}+R_{abus}}$$

We perform 10-fold cross-validation on the transliterated corpus. We run each classifier 10 times and provide the range of R_{abus}, P_{abus} and $F1_{abus}$ scores.

Table 1 shows the performance of various ML classifiers for abusive language detection in the translated Bengali corpus. We observe that among the three traditional ML classifiers, LR and SVM

Table 1: Performance of various classifiers for abusive language detection in the transliterated corpus

Classifier	R_{abus}	P_{abus}	$F1_{abus}$
SVM	0.790 ±0.008	0.865 ±0.015	0.827 ±0.010
LR	0.779 ±0.006	0.876 ±0.004	0.823 ±0.006
BiLSTM	0.781 ±0.031	0.800 ±0.036	0.790 ±0.031
RF	0.781 ±0.013	0.762 ±0.028	0.770 ±0.020

show similar R_{abus} and P_{abus} scores. RF classifier provides a similar R_{abus} score of LR and SVM; however, it attains a lower P_{abus} score. We find both LR and SVM yield lower R_{abus} scores than the P_{abus} scores. BiLSTM, the deep learning-based architecture, obtains a relatively lower $F1_{abus}$ score compared to LR and SVM, which could be attributed to the small size (i.e., 3000 comments) of the corpus.

6 Conclusion and Future Work

Identifying abusive content in social media is of paramount importance due to its detrimental impact. Not addressing this problem can lead to the increasing growth of harassment and cyberbullying in social media. While there have been few works for abusive speech detection in Bengali social media content, they mostly ignored the presence of transliterated Bengali. In this paper, we present the most comprehensive study of abusive content detection in transliterated Bengali text by providing an annotated corpus and baseline evaluations. Our future works will focus on expanding the size of the transliterated corpus, providing more rigorous analysis, and introducing a customized deep learning model to improve the performance of abusive content detection in transliterated Bengali text.

References

Md Abdul Awal, Md Shamimur Rahman, and Jakaria Rabbi. 2018. Detecting abusive comments in discussion threads using naïve bayes. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 163–167. IEEE.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2020. On analyzing annotation consistency in online abusive behavior datasets. *arXiv preprint arXiv:2006.13507*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Puja Chakraborty and Md Hanif Seddiqui. 2019. Threat and abusive language detection on social media in bengali language. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6. IEEE.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016. Unraveling the english-bengali code-mixing phenomenon. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 80–89.

François Chollet et al. 2015. Keras. https://keras.io.

Monojit Choudhury, Kalika Bali, Tirthankar Dasgupta, and Anupam Basu. 2010. Resource creation for training and testing of transliteration systems for indian languages. LREC.

Maeve Duggan. 2017. Online harassment 2017.

Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, and Tanni Mittra. 2019. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

Shahnoor C Eshan and Mohammad S Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Antigoni 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 *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Md Gulzar Hussain, Tamim Al Mahmud, and Waheda Akthar. 2018. An approach to detect abusive bangla text. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5. IEEE.

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.

Maliha Jahan, Istiak Ahamed, Md Rayanuzzaman Bishwas, and Swakkhar Shatabda. 2019. Abusive comments detection in bangla-english code-mixed and transliterated text. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–6. IEEE.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. Association for Computational Linguistics.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ritesh Kumar, Bornini Lahiri, and Atul Kr Ojha. 2021. Aggressive and offensive language identification in hindi, bangla, and english: A comparative study. *SN Computer Science*, 2(1):1–20.

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

Pieter Muysken, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Carol Myers-Scotton. 1993. Common and uncommon ground: Social and structural factors in codeswitching. *Language in society*, pages 475–503.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

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

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. *arXiv e-prints*, pages arXiv–2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *arXiv preprint arXiv:2012.09686*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.