Identifying vulgarity in Bengali social media textual content

Salim Sazzed Corresp. 1

1 Computer Science, Old Dominion University, Norfolk, VA, USA

Corresponding Author: Salim Sazzed
Email address: ssazz001@odu.edu

The presence of abusive and vulgar language in social media has become an issue of increasing concern in recent years. However, they remain largely unaddressed in low-resource languages such as Bengali. In this paper, we provide the first comprehensive analysis on the presence of vulgarity in Bengali social media content. We develop two benchmark corpora consisting of 7245 reviews collected from YouTube and manually annotate them into vulgar and non-vulgar categories. The manual annotation reveals the ubiquity of vulgar and swear words in Bengali social media content (i.e., in two corpora), ranging from 20% to 34%. To automatically identify vulgarity, we employ various approaches, such as classical machine learning (CML) algorithms, Stochastic Gradient Descent (SGD) optimizer, deep learning (DL) based architecture, and lexicon-based methods. We find although small in size, the swear/vulgar lexicon is effective at identifying the vulgar language due to the high presence of some swear terms in Bengali social media. We observe that the performances of machine leanings (ML) classifiers are affected by the class distribution of the dataset. The DL-based BiLSTM (Bidirectional Long Short Term Memory) model yields the highest recall scores for identifying vulgarity in both datasets (i.e., in both original and class-balanced settings). Besides, the analysis reveals that vulgarity is highly correlated with negative sentiment in social media comments.

Identifying vulgarity in Bengali social media content

Salim Sazzed[1]

[1]**Old Dominion University, 5115 Hampton Blvd, Norfolk, VA 23529, USA**

Corresponding author:
Salim Sazzed[1]

Email address: ssazz001@odu.edu

ABSTRACT

The presence of abusive and vulgar language in social media has become an issue of increasing concern in recent years. However, they remain largely unaddressed in low-resource languages such as Bengali. In this paper, we provide the first comprehensive analysis on the presence of vulgarity in Bengali social media content. We develop two benchmark corpora consisting of 7245 reviews collected from YouTube and manually annotate them into vulgar and non-vulgar categories. The manual annotation reveals the ubiquity of vulgar and swear words in Bengali social media content (i.e., in two corpora), ranging from 20% to 34%. To automatically identify vulgarity, we employ various approaches, such as classical machine learning (CML) algorithms, Stochastic Gradient Descent (SGD) optimizer, deep learning (DL) based architecture, and lexicon-based methods. We find although small in size, the swear/vulgar lexicon is effective at identifying the vulgar language due to the high presence of some swear terms in Bengali social media. We observe that the performances of machine leanings (ML) classifiers are affected by the class distribution of the dataset. The DL-based BiLSTM (Bidirectional Long Short Term Memory) model yields the highest recall scores for identifying vulgarity in both datasets (i.e., in both original and class-balanced settings). Besides, the analysis reveals that vulgarity is highly correlated with negative sentiment in social media comments.

1 INTRODUCTION

Vulgarity or obscenity indicates the use of curse, swear or taboo words in language (Wang, 2013; Cachola et al., 2018). Eder et al. (2019) conceived vulgar language as an overly lowered language with disgusting and obscene lexicalizations generally banned from any type of civilized discourse. Primarily, it involves the lexical fields of sexuality, such as sexual organs and activities, body orifices, or other specific body parts. Cachola et al. (2018) defined vulgarity as the use of swear/curse words. Jay and Janschewitz (2008) mentioned vulgar speech includes explicit and crude sexual references. Although the terms obscenity, swearing, and vulgarity have subtle differences in their meaning and scope, they are closely linked with some overlapping definitions. Thus, in this paper, we use them interchangeably to refer to the text that falls into the above-mentioned definition of (Cachola et al., 2018; Eder et al., 2019; Jay and Janschewitz, 2008).

With the rapid growth of user-generated content in social media, vulgar words can be found in online posts, messages, and comments across languages. The occurrences of swearing or vulgar words are often linked with abusive or hatred context, sexism, and racism (Cachola et al., 2018) ; thus, leads to abusive and offensive actions. Hence, identifying vulgar or obscene words has practical connections to understanding and monitoring online content. Furthermore, vulgar word identification can help to improve sentiment classification, as shown by various studies (Cachola et al., 2018; Volkova et al., 2013).

Social media platforms such as Twitter, Facebook, Instagram, YouTube have made virtual social interaction popular by connecting billions of users. In social media, swearing is ubiquitous according to various studies. Wang et al. (2014) found that the rate of swear word usage in English Twitter is 1.15%, almost double compared to its use in daily conversation (0.5% 0.7%) as reported by (Jay and Janschewitz, 2008; Mehl et al., 2007).Wang et al. (2014) also reported that 7.73% of tweets in their random sampling collection contain swear words. Based on (Jay and Janschewitz, 2008), offensive speech can be classified

into three categories: *vulgar*, which includes explicit and crude sexual references, *pornographic*, and *hateful*, which refers to offensive remarks targeting people's race, religion, country, etc. The categorization suggests that there exists a link between offensiveness and vulgarity.

Unlike English, research related to vulgarity is still unexplored in Bengali. As the vulgar word usage is dependent on the socio-cultural context and demography (Cachola et al., 2018), it is important to explore their usage in languages other than English. For example, the usage of f*ck, a*s, sh*t, etc. are common in many English speaking countries in an expression to emphasize feelings, to convey neutral/idiomatic or even positive sentiment as shown by (Cachola et al., 2018); However, the corresponding Bengali words are highly unlikely to be used in a similar context in Bengali, due to the difference in the socio-culture of the Bengali native speakers (i.e., people living in Bangladesh or India).

There is a lack of annotated vulgar or obscene datasets in Bengali, which are crucial for developing effective machine learning models. Therefore, in this work, we create resources for vulgarity analysis in Bengali. Besides, we investigate the presence of vulgarity, which is often associated with abusiveness and inappropriateness in social media. Furthermore, we focus on automatically distinguishing vulgar comments (e.g., usage of filthy language or curses towards a person), which should be monitored and regulated in online communications, and non-vulgar non-abusive negative comments, which should be allowed as part of freedom of speech.

We construct two Bengali review corpora consisting of 7245 comments and annotate them based on the presence of vulgarity. We find a high presence of vulgar words in Bengali social media comments based on the manual annotations. We provide the comparative performance of both lexicon-based and machine learning (ML)(i.e., CML and DL) based methods for automatically identifying the vulgarity in Bengali social media data. As a lexicon, we utilize a Bengali vulgar lexicon, BengVulLex, which consists of 184 swear and obscene terms. We leverage two classical machine learning (CML) classifiers, Support Vector Machine (SVM) (Cortes and Vapnik, 1995) and Logistic Regression (LR), and an optimizer, Stochastic Gradient Descendent (SGD) (Ruder, 2016), to automatically identify vulgar content. In addition, we employ a deep learning architecture, Bidirectional Long Short Term Memory (BiLSTM). We observe that BengVulLex provides a high recall score in one corpus and very high precision scores in both corpora. BiLSTM shows higher recall scores than BengVulLex in both corpora in class-balanced settings; however, they generate high false positives, thus yield a much lower precision score. The performances of the CML classifiers vary by the class distribution of the dataset.We observe that when undersampling is performed, CML classifiers provide much better performance. Class-balancing using over-sampling techniques like SMOTE (Chawla et al., 2002) or weighting class based on sample distributions does not improve the performance of CML classifiers significantly in two datasets.

1.1 Motivation

As vulgarity is often related to abusive comments on social media, it is required to identify its presence in the textual content. In Bengali, until now, no work has addressed this issue. Although a few papers tried to determine the offensive or hate speech in Bengali utilizing labeled data, none focused on recognizing vulgarity or obscenity. Since social media such as Facebook, Twitter, YouTube, Instagram are popular in Bangladesh, the country with the highest number of Bengali native speakers, it is necessary to distinguish vulgarity in the comments or reviews for various downstream tasks such as abusiveness or hate speech detection and understanding social behaviors. Besides, it is imperative to analyze how vulgarity is related to sentiment.

1.2 Contributions

The main contributions of this paper can be summarized as follows-

- We manually annotate two Bengali corpora consisting of 7245 reviews/comments into vulgar and non-vulgar categories and make them publicly available (the first of its kind in Bengali).[1]
- We provide a quantitative analysis on the presence of vulgarity in Bengali social media content based on the manual annotation.
- We present a comparative analysis of lexicon-based, CML-based, SGD optimizer, and deep learning-based approaches for automatically recognizing vulgarity in Bengali social media content.

[1] https://github.com/sazzadcsedu/Bangla-vulgar-corpus

- Finally, we investigate how vulgarity is related to sentiment in Bengali social media content.

2 RELATED WORK

Researchers studied the existence and socio-linguistic characteristics of swearing, cursing, incivility or cyber-bullying in social media (Wang et al., 2014; Sadeque et al., 2019; Kurrek et al., 2020; Gauthier et al., 2015; Agrawal and Awekar, 2018). Wang et al. (2014) investigated the cursing activities on Twitter, a social media platform. They studied the ubiquity, utility, and contextual dependency of swearing on Twitter. Gauthier et al. (2015) analyzed several sociolinguistic aspects of swearing on Twitter text data. Wang et al. (2014) investigated the relationship between social factors such as gender with the profanity and discovered males employ profanity much more often than females. Other social factors such as age, religiosity, or social status were also found to be related to the rate of using vulgar words (McEnery, 2004). McEnery (2004) suggested that social rank, which is related to both education and income, is anti-correlated to the use of swear words. The level of education and income are inversely correlated with the usage of vulgarity on social media with education being slightly more strongly associated with a lack of vulgarity than income (Cachola et al., 2018). Furthermore, liberal users tend to use vulgarity more on social media, an association on Twitter revealed by (Cachola et al., 2018; Sylwester and Purver, 2015; Preoţiuc-Pietro et al., 2017).

Eder et al. (2019) described a workflow for acquisition and semantic scaling of a lexicon that contains lexical items in the German language, which are typically considered as vulgar or obscene. The developed lexicon starts with a small seed set of rough and vulgar lexical items, and then automatically expanded using distributional semantics.

Jay and Janschewitz (2008) noticed that the offensiveness of taboo words depends on their context, and found that usages of taboo words in conversational context is less offensive than the hostile context. Pinker (2007) classified the use of swear words into five categories. Since many studies related to the identification of swearing or offensive words have been conducted in English, several lexicons comprised of offensive words are available in the English language. Razavi et al. (2010) manually collected around 2,700 dictionary entries including phrases and multi-word expressions, which is one of the earliest work offensive lexicon creations. The recent work on lexicon focusing on hate speech was reported by (Gitari et al., 2015).

Davidson et al. (2017) studied how hate speech is different from other instances of offensive language. They used a crowd-sourced lexicon of hate language to collect tweets containing hate speech keywords. Using crowd-sourcing, they labeled tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. They analyzed when hate speech can be reliably separate from other offensive language and when this differentiation is very challenging.

In Bengali, several works investigated the presence of abusive language in social media data by leveraging supervised ML classifiers and labeled data (Ishmam and Sharmin, 2019; Banik and Rahman, 2019). Sazzed (2021) annotated 3000 transliterated Bengali comments into two classes, abusive and non-abusive, 1500 comments for each. For baseline evaluations, the author employed several traditional machine learning (ML) and deep learning-based classifiers.

Emon et al. (2019) utilized linear support vector classifier (LinearSVC), logistic regression (LR), multinomial naïve Bayes (MNB), random forest (RF), artificial neural network (ANN), recurrent neural network (RNN) with long short term memory (LSTM) to detect multi-type abusive Bengali text. They found RNN outperformed other classifiers by obtaining the highest accuracy of 82.20%. Chakraborty and Seddiqui (2019) employed machine learning and natural language processing techniques to build an automatic system for detecting abusive comments in Bengali. As input, they used Unicode emoticons and Unicode Bengali characters. They applied MNB, SVM, and Convolutional Neural Network (CNN) with LSTM and found SVM performed best with 78% accuracy. Karim et al. (2020) proposed BengFastText, a word embedding model for Bengali, and incorporated it into a Multichannel Convolutional-LSTM (MConv-LSTM) network for predicting different types of hate speech. They compared BengFastText against the Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embedding by integrating them into several ML classifiers.

However, none of the existing works focused on recognizing vulgarity or profanity in Bengali social media data. To the best of our knowledge, it is the first attempt to identify and provide a comprehensive analysis of the presence of vulgarity in the context of Bengali social media data.

3 SOCIAL MEDIA CORPORA

We create two datasets consisting of 7245 comments written in Bengali. Both datasets are collected from social media, YouTube [2].

3.1 Drama Review Dataset

The first corpus we utilize is a drama review corpus. This corpus was created and deposited by (Sazzed, 2020a) for sentiment analysis; It consists of 8500 positive and 3307 negative reviews. However, there is no distinction between different types of negative reviews. Therefore, we manually annotate these 3307 negative reviews into two categories; one category contains reviews that convey vulgarity, while the other category consists of negative but non-vulgar reviews.

3.2 Subject-Person Dataset

The second corpus is also collected from YouTube. However, unlike the drama review corpus which represents the viewer's feedback regarding dramas, this corpus consists of comments towards a few controversial female celebrities.

We employ a web scraping tool to download the comment data from YouTube, which comes in JSON format. Then utilizing a parsing script, we retrieve the comments from the JSON data. Utilizing a language detection library [3], we recognize the comments written in Bengali. We exclude reviews written in English and Romanized Bengali (i.e., Bengali language in the Latin script).

4 CORPORA ANNOTATION

It is common practice to compare annotations of a single source by multiple people which helps validating and improving annotation schemes and guidelines, identifying ambiguities or difficulties in the source, or assessing the range of valid interpretations (Artstein, 2017). The comparison can be performed using a qualitative examination of the annotations, calculating agreement measures, or statistical modeling of annotator differences.

4.1 Annotation Guideline

For annotating a corpus for various NLP tasks (e.g., hate speech detection, sentiment classification, profanity detection), it is required to utilize a set of guidelines (Khan et al., 2021; Mehmood et al., 2019; Pradhan et al., 2020; Fortuna and Nunes, 2018; Sazzed, 2020a).

Here, to distinguish the comments into vulgar and non-vulgar class, annotators are asked to consider the followings guideline-

- Vulgar comments: The presence of swearing, obscene language, vulgar slang, slurs, sexual and pornographic terms in a comment (Eder et al., 2019; Cachola et al., 2018; Jay and Janschewitz, 2008).
- Non-vulgar comments: The comments which do not have above mentioned characteristics.

4.2 Annotation Procedure

The annotation is performed by three annotators (A1, A2, A3); Among them, two are male and one female (A1: male, A2: female, A3: male). All of them are Bengali native speakers. The first two annotators (A1 and A2) initially annotate all the reviews. In case of disagreement in annotation, it is resolved by a third annotator (A3) by majority voting.

4.3 Annotation Results

The annotation of the reviews by two reviewers (A1, A2) results two cases.

1. Agreement: The two annotators (A1, A2) assign the same label to a review.
2. Conflict: Each annotator (A1, A2) assigns a different label to a review.

Drama Review Dataset	
বাংলা নাটকের গোয়া মোশাররফ করিম গং রাই মারতাছে,	Mosharraf Karims gang's are fucking Bengali drama,
চুদনাগিরি স্ক্রিপ্ট ছাড়া আর কোন স্ক্রিপ্ট ছিলো না।মাদারচোদ মার্কা নাটক এইটা	Wasn't there any other script except this fucking one. This is a motherfucker drama.
ব্লাইক্কোদ খানকির ছেলে। এতো অ্যাড চুদাও কে,,,	Fucker whore's son. why so many advertisements?
Subject-Person Dataset	
কুত্তার বাচ্চা তরে পাইলে দুইটা হাত কাটতাম নটি	Son of a Bitch, If I find you, I will chop your two hands, slut
শাহরিয়ার নাজিম ভাই খানকি নিয়া শো বন্ধ করুন	Sharir Nazim vai, please stop making tv show with whore
কিন্তু দুধের সাইজ বড়ো করে মনে হচ্ছে দুধ না ফুটবল	The enlarged tits look like a football, not tit

Figure 1. Sample vulgar reviews from annotated datasets

Table 1. Annotation of drama review corpus by two annotators (A1, A2)

	Vulgar	**Non-vulgar**
Vulgar	592	160
Non-vulgar	53	2502

Table 2. Annotation of subject-person dataset by two annotators (A1, A2)

	Vulgar	**Non-vulgar**
Vulgar	1282	120
Non-vulgar	163	2373

From the table 1, we see Cohen's kappa (κ) statistic of two raters (A1, A2) is 0.8070 in the Drama review dataset, which indicate almost perfect agreement. Regarding the percentages, we find both reviewers agreed on 93.55% reviews.

As shown by Table 2, in the subject-person dataset, an agreement of 92.81% is observed. Cohen's κ (Cohen, 1960) provides a score of 0.8443, which refers to almost perfect agreement.

Table 3. Description of two corpora after final annotations

Dataset	**Vulgar**	**Non-vulgar**	**Total**
Drama	664	2643	3307
Subject-person	1331	2607	3938

4.4 Corpora Statistics

After annotation the drama review corpus consists of 2643 non-vulgar negative reviews and 664 vulgar reviews (Table 3). The presence of 664 vulgar reviews out of 3307 negative reviews reveals a high presence of vulgarity in the dataset, around 20%. The annotated subject-person dataset consists of 1331 vulgar reviews and 2607 non-vulgar reviews, a total of 3938 reviews. This dataset contains even higher percentages of reviews labeled as vulgar, around 34%.

Figure 2 presents the top 10 vulgar words from each dataset. We find a high presence of some vulgar words in the reviews, as shown in the top few rows. Besides, we observe a high number of misspelled vulgar words, which makes identifying them a challenging task. Among the top 10 vulgar words in the subject-person dataset, we notice all of them except the last word (last row) are female-specific sexually

[2] https://www.youtube.com/
[3] https://github.com/Mimino666/langdetect

Bengali	English	Count
মাগি	Slut	245
দুধ	Tit	181
মাগির	Slut's	180
খানকি	Whore	125
মাগী	Slut	93
পতিতা	Prostitute	69
দুধের	Tit's	47
খানকি মাগি	Whore slut	34
খানকির	Whore's	32
বেশ্যা	Hooker	30
কুত্তা	Bitch	29

Bengali	English	Count
বালের	-	235
বাল	-	66
আবাল	Stupid	27
শালা	-	26
কুত্তার	Bitch	23
খানকির	Whore's	21
মাগির	Slut's	14
শালার	-	14
আচোদা	Fucking dumb	13
চোদা	Fuck	12
শাউয়ার	-	11

Figure 2. (a) Top 10 vulgar words in drama review dataset. (b) Top 10 vulgar words in subject-person dataset

vulgar terms. As the subjects of this dataset are female celebrities, this is expected. In the drama review dataset, among the top 10 vulgar words, we find five terms as generic (not gender-specific) vulgar words, three are male-specific vulgar, and two are female-specific vulgar. The two female-specific vulgar terms also exist in the Subject-person dataset.

5 BASELINE METHODS

5.1 Lexicon-based Methods

We utilize two publicly available Bengali lexicons for identifying vulgarity in a text. The first lexicon we use is a vulgar lexicon, BenVulLex [4]. The other lexicon is a sentiment lexicon, which contains a list of positive and negative sentiment words (Sazzed, 2020b). The BenVulLex consists of 184 Bengali swear and vulgar words, semi-automatically created from a social media corpus. The sentiment lexicon consists of 690 opinion words. The goal of utilizing a sentiment lexicon for vulgarity detection is to investigate how well the negative opinion word present in sentiment lexicon can detect vulgarity. The few other Bengali sentiment lexicons are a dictionary-based word-level translation of popular English sentiment lexicons; thus, not capable of identifying swearing or vulgarity in Bengali text.

5.2 Classical Machine Learning (CML) algorithms and SGD optimizer

Two popular CML classifiers, Logistics Regression (LR) and Support Vector Machine (SVM), and an optimizer, Stochastic Gradient Descendent (SGD), are employed to identify vulgar comments.

LR is a predictive analysis model that assigns observations into a discrete set of classes. LR assumes there are one or more independent variables that determine the outcome of the target.

SVM is a discriminative classifier defined by a separating hyperplane. Given the labeled training data, SVM generates an optimal hyperplane that categorizes unseen observations. For example, in two-dimensional space, this hyperplane is a line dividing a plane into two parts where each class lays on either side (for linear kernel).

SGD is an optimization technique and does not correspond to a specific family of machine learning models. SGD can be used to fit linear classifiers and regressors such as linear SVM and LR under convex loss functions.

5.2.1 Input

We extract unigrams and bigrams from the text and calculate the tf-idf scores, which are used as an input for the CML classifiers. tf-idf refers to the term frequency-inverse document frequency, which is a numerical statistic that is aimed to reflect the importance of a word to a document in a corpus.

[4] https://github.com/sazzadcsedu/Bangla-Vulgar-Lexicon

5.2.2 Parameter settings and library used

For LR [5] and SVM [6], the default parameter settings of scikit-learn library (Pedregosa et al., 2011) are used. For SGD, hinge loss and l2 penalty with a maximum iteration of 1500 are employed. We use the scikit-learn library (Pedregosa et al., 2011) to implement the SVM, LR and SGD.

5.3 Deep Learning Classifier

BiLSTM (Bidirectional Long Short Term Memory) is a deep learning-based sequence processing model that consists of two LSTMs (Hochreiter and Schmidhuber, 1997). BiLSTM takes input in both forward and backward directions, thus, provides more contextual information to the network.

5.3.1 Network architecture, hyperparameter settings and library used

The BiLSTM model starts with the Keras embedding layer (Chollet et al., 2015). The three important parameters of the embedding layer are *input dimension*, which represents the size of the vocabulary, *output dimensions*, which is the length of the vector for each word, *input length*, the maximum length of a sequence. The *input dimension* is determined by the number of words present in a corpus, which vary in two corpora. We set the *output dimensions* to 64. The maximum length of a sequence is used as 200.

A drop-out rate of 0.5 is applied to the dropout layer; ReLU activation is used in the intermediate layers. In the final layer, softmax activation is applied. As an optimization function, Adam optimizer, and as a loss function, binary-cross entropy are utilized. We set the batch size to 64, use a learning rate of 0.001, and train the model for 10 epochs. We use the Keras library (Chollet et al., 2015) with the TensorFlow backend for BiLSTM implementation.

6 EXPERIMENTAL SETTINGS AND RESULTS

6.1 Settings

6.1.1 Lexicon-based method

If a review contains at least one term from BengVulLex, it is considered vulgar. As BengVulLex is comprised of only manually validated slang or swear terms, referring a non-vulgar comment to vulgar (i.e., false positive) is highly unlikely; thus, a very high precision score close to 1 is expected.

6.1.2 ML-based classifiers/optimizer

The results of ML classifiers are reported based on 10-fold cross-validation. We provide the performance of various ML classifiers in four different settings based on class distribution,

1. Original setting: The original setting is class-imbalanced, where most of the comments are non-vulgar.

2. Class-balancing using class weighting: This setting considers the distribution of the samples from different classes in training data. The weight of a class is set inversely proportional to the number of samples it contains.

3. Class-balancing using undersampling: In this class-balanced setting, we use all the samples of vulgar class; however, for the non-vulgar class, we randomly select the equal number of non-vulgar comments from a pool of all non-vulgar comments.

4. Class-balancing using SMOTE: SMOTE (synthetic minority over-sampling technique) (Chawla et al., 2002) is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

6.2 Evaluation Metrics

We report the comparative performances of various methods utilizing precision, recall and F1 score.

The TP, FP, FN for is defined as follows-

TP = vulgar review classified as vulgar

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

FP = non-vulgar review classified as vulgar
FN = vulgar review classified as non-vulgar

The recall (R_V), precision (P_V) and F1 score ($F1_V$) of vulgar class are calculated as-

$$R_V = \frac{TP}{TP+FN} \tag{1}$$

$$P_V = \frac{TP}{TP+FP} \tag{2}$$

$$F1_V = \frac{2*R_V*P_V}{R_V+P_V} \tag{3}$$

Table 4. Performance of various methods for vulgarity detection in drama review dataset

Type	Method	# Correctly Identified Vulgar Review	R_V	P_V	$F1_V$
Lexicon	**Sentiment Lexicon**	204 (664)	0.307	-	-
	BengVulLex	564 (664)	0.849	0.998	0.917
ML Classifier (Original Setting)	**LR**	161 (664)	0.245	1.0	0.394
	SVM	345 (664)	0.534	0.994	0.686
	SGD	386(664)	0.588	0.985	0.736
	BiLSTM	462(664)	**0.704**	0.783	**0.741**
ML Classifier (Undersampling)	**LR**	609(664)	0.917	0.801	0.855
	SVM	593(664)	0.893	0.859	0.876
	SGD	592(664)	0.891	0.876	0.883
	BiLSTM	624(664)	**0.940**	0.851	**0.893**
ML Classifier (SMOTE)	**LR**	367(664)	0.552	0.970	0.704
	SVM	386 (664)	0.581	0.982	0.730
	SGD	385 (664)	0.579	0.987	0.730
	BiLSTM	563(664)	**0.850**	0.707	**0.772**
(Class weighting) ML Classifier	**LR**	385(664)	0.579	0.96	0.723
	SVM	388(664)	0.584	0.934	0.719
	SGD	438(664)	0.659	0.964	**0.783**
	BiLSTM	564(664)	**0.854**	0.667	0.749

6.3 Comparative results for Identifying Vulgarity

Table 4 shows that among the 664 vulgar reviews present in the drama review corpus, the sentiment lexicon identifies only 204 vulgar reviews (based on the negative score). The vulgar lexicon BengVulLex registers 564 reviews as vulgar, with a high recall score of 0.85. In the original class-imbalanced dataset, all the CML classifiers achieve very low recall scores. However, when a class-balanced dataset is selected by performing undersampling to the dominant class, the recall scores of CML classifiers increase significantly to 0.90. However, we notice precision scores decrease in the class-balanced setting due to a higher number of false-positive (FP). BiLSTM provides the highest recall scores in both original and class-balanced setting, which is 0.70 and 0.94, respectively.

Table 5. Performance of various methods for vulgarity detection in subject-person dataset

Type	Method	# Correctly Identified Vulgar Review	R_V	P_V	$F1_V$
Lexicon	**Sazzed (2020b)**	239 (1331)	0.180	-	-
	BengVulLex	917(1331)	0.689	0.998	0.815
ML Classifiers (Original Setting)	**LR**	551(1331)	0.394	0.992	0.563
	SVM	788(1331)	0.594	0.962	0.746
	SGD	860(1331)	0.660	0.940	**0.775**
	BiLSTM	1050(1331)	**0.793**	0.724	0.757
ML Classifiers (Undersampling)	**LR**	954(1331)	0.717	0.870	0.786
	SVM	969(1331)	0.728	0.893	0.802
	SGD	1027(1331)	0.772	0.884	**0.824**
	BiLSTM	1064(1331)	**0.786**	0.866	**0.824**
ML Classifier (SMOTE)	**LR**	826(1331)	0.620	0.892	0.731
	SVM	847(1331)	0.636	0.941	0.759
	SGD	866(1331)	0.650	0.938	0.768
	BiLSTM	1075(1331)	**0.809**	0.737	**0.771**
ML Classifier (Class Weighting)	**LR**	911(1331)	0.684	0.814	0.743
	SVM	824(1331)	0.619	0.912	0.737
	SGD	935(1331)	0.702	0.904	**0.790**
	BiLSTM	1070(1331)	**0.807**	0.742	0.773

Table 5 shows the performances of various methods in subject-person dataset. We find that the sentiment lexicon shows a very low recall score, only 0.18. The BengVulLex yields a recall score of 0.69. SVM, LR, and SGD exhibit low recall scores below 0.60 in the original class-imbalanced setting. However, in the class-balanced setting with undersampling (i.e., 1331 comments from both vulgar and non-vulgar categories), a higher recall score is observed. SGD yields a recall score of 0.77. BiLSTM shows the highest recall scores in both original and all the class-balanced settings, which is around 0.8. BiLSTM provides lower precision scores compared to CML classifiers in both settings (i.e., original class-imbalanced and class-balanced).

6.4 Vulgarity and Sentiment

We further analyze how vulgarity is related to user sentiment in social media. As a social media corpus, we leverage the entire drama review dataset, which contains 8500 positive reviews in addition to 3307 negative reviews stated earlier. Using the BenVulLex vulgar lexicon, we identify the presence of vulgar words in the reviews. We perform a comparative analysis of the presence of vulgar words in both positive and negative reviews. We find only 37 positive reviews out of 8500 positive reviews contain any vulgar words, which is only 0.4% of the total positive reviews. Out of 3307 negative reviews, we observe the presence of vulgar words in 553 reviews, which is 16.67% of total negative reviews. Figure 3 shows examples of several positive reviews that contain vulgar terms.

7 DISCUSSION

The results show that the sentiment lexicon yield poor performance in identifying vulgarity in Bengali textual content, as shown by its poor performance in both datasets. The poor coverage of the sentiment lexicon is expected as it contains different types of negative words, thus may lack words that are particularly

1. বালের প্রেম ভালোবাসা সস্তা আবেগ ছাড়াও যে এত সুন্দর নাটক করা যায় তা আবারও দেখিয়ে দিলো বৃন্দাবন দা।অসাধারণ।
2. নাটক টা যে বানাইছে শালা একটা মাল । ভালো লাগছে ।
3. জাস্ট অসাধারণ! ৯৩০ জন মাদারচোদ ডিসলাইক কি কারণে দিলো ওরা জানে!
4. কুত্তার বাচ্চা বলে যে সেটা মজার,,,আর শেষ খুব দারুণ

Figure 3. Examples of positive reviews with vulgar words in drama review corpus

associated with vulgarity. Besides, vulgarity is often linked with the usage of internet slang words that may not exist in small-sized sentiment lexicon. For example, the sentiment lexicon we use is contains around 700 opinion words.

The vulgar lexicon, BengVulLex, on the other hand, provides a significantly higher recall scores than sentiment lexicon as it was specially curated to identify vulgarity, obscenity or swearing. The high presence of some of the vulgar words, as shown in figure 2 also helps BengVulLex to achieve a good coverage (i.e., recall score) for vulgarity detection. We observe that the recall score of BengVulLex varies in two corpora. In the smaller drama review data (664 vulgar review), it shows a recall score of 0.85, while in the other dataset which contains a much higher number of vulgar review, BengVulLex achieve much lower recall score of 0.69. Since BengVulLex contains less than 200 words, its performance can be affected by the characteristics and size of the dataset. BengVulLex achieves almost a perfect precision score, close to 1, in both corpora. Since BengVulLex was manually validated to assure that it contains only vulgar or swear words, the almost perfect precision score is expected.

Table 4 and 5 reveal that the performances of ML classifiers can be affected by the class distribution of the training data. Specially for the CML classifiers, when a class-imbalanced training data is used, the result is biased toward the dominating class (i.e., non-vulgar category) and achieves low recall and high precision score, as shown by Table 4 and 5. Due to the much higher number of non-vulgar comments in the original dataset, CML classifiers yield a high number of false-negatives (FN) and a low number of false-positives (FP) for the vulgar class, which is reflected in the low recall score and high precision score. Whenever a class-balanced training set is employed, all the CML classifiers yield a higher recall score.

We find that the deep learning-based method, BiLSTM is less affected by class imbalance. Only when the difference of class proportion is very high, such as 18% vs 82% in the drama review dataset, we observe BiLSTM shows a high difference in recall score.

Besides, we analyze the motivation behind using vulgar words in Bengali social media data. Although the usage of vulgar words can be non-offensive such as when used in informal communication between closely-related groups or expressing emotion such as Twitter or Facebook status (Holgate et al., 2018), we observe when it is used in review or targeted towards a person with no personal connection, it is inappropriate or offensive most of the time.

8 CONCLUSION

With the surge of user-generated content online, the detection of vulgar or abusive language has become a subject of utmost importance. While there have been few works in hate speech or abusive content analysis in Bengali, to the best of our knowledge, this is the first attempt to thoroughly analyze vulgarity in Bengali social media content.

This paper introduces two annotated datasets in Bengali with 7245 reviews to address the resource scarcity for Bengali vulgar language analysis. Besides, we investigate the prevalence of vulgarity in social media comments. Our analysis reveals a high presence of swearing or vulgar words in social media, ranges from 20% to 34% in two datasets. We explore the performance of different automatic approaches for vulgarity identification of Bengali and present a comparative analysis. The analysis reveals the strengths and weaknesses of different approaches and provides directions for future research.

REFERENCES

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Artstein, R. (2017). Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer.

Banik, N. and Rahman, M. H. H. (2019). Toxicity detection on bengali social media comments using supervised models. In *International Conference on Innovation in Engineering and Technology (ICIET)*, volume 23, page 24.

Cachola, I., Holgate, E., Preoţiuc-Pietro, D., and Li, J. J. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938.

Chakraborty, P. and Seddiqui, M. H. (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.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (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.

Eder, E., Krieg-Holz, U., and Hahn, U. (2019). At the lower end of language—exploring the vulgar and obscene side of german. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 119–128.

Emon, E. A., Rahman, S., Banarjee, J., Das, A. K., and Mittra, T. (2019). A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5. IEEE.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Gauthier, M., Guille, A., Rico, F., and Deseille, A. (2015). Text mining and twitter to analyze british swearing habits. *Handbook of Twitter for Research*.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., and Li, J. J. (2018). Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414.

Ishmam, A. M. and Sharmin, S. (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.

Jay, T. and Janschewitz, K. (2008). The pragmatics of swearing.

Karim, M., Chakravarthi, B. R., Arcan, M., McCrae, J. P., Cochez, M., et al. (2020). Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. *arXiv preprint arXiv:2004.07807*.

Khan, M. M., Shahzad, K., and Malik, M. K. (2021). Hate speech detection in roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–19.

Kurrek, J., Saleem, H. M., and Ruths, D. (2020). Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149.

McEnery, T. (2004). *Swearing in English: Bad language, purity and power from 1586 to the present*. Routledge.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., and Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834):82–82.

Mehmood, K., Essam, D., Shafi, K., and Malik, M. K. (2019). Sentiment analysis for a resource poor language—roman urdu. *ACM Transactions on Asian and Low-Resource Language Information*

Processing (TALLIP), 19(1):1–15.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Pradhan, R., Chaturvedi, A., Tripathi, A., and Sharma, D. K. (2020). A review on offensive language detection. In *Advances in Data and Information Sciences*, pages 433–439. Springer.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., and Ungar, L. (2017). Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., and Bethard, S. (2019). Incivility detection in online comments. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 283–291.

Sazzed, S. (2020a). Cross-lingual sentiment classification in low-resource bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60.

Sazzed, S. (2020b). Development of sentiment lexicon in bengali utilizing corpus and cross-lingual resources. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 237–244. IEEE Computer Society.

Sazzed, S. (2021). Abusive content detection in transliterated bengali-english social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130.

Sylwester, K. and Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PloS one*, 10(9):e0137422.

Volkova, S., Wilson, T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Wang, N. (2013). An analysis of the pragmatic functions of "swearing" in interpersonal talk. *En: Griffith Working Papers in Pragmatics and Intercultural Communication*, 6:71–79.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2014). Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425.