BengSentiLex and BengSwearLex: creating lexicons for sentiment analysis and profanity detection in low-resource Bengali language

Salim Sazzed Corresp. 1

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

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

Bengali is a low-resource language that lacks tools and resources for various natural language processing (NLP) tasks, such as sentiment analysis and profanity identification. In Bengali, only the translated versions of English sentiment lexicons are available. Besides, no lexicon exists for detecting profanity in Bengali social media text. In this work, we introduce a Bengali sentiment lexicon, BengSentiLex, and a Bengali swear lexicon, BengSwearLex. To build BengSentiLex, we propose a cross-lingual methodology that utilizes a machine translation system, review corpus, English sentiment lexicons, pointwise mutual information (PMI), and supervised machine learning (ML) classifiers in different steps. For creating BengSwearLex, we introduce a semi-automatic methodology that leverages an obscene corpus, word embedding, and part-of-speech (POS) taggers. We compare the performance of BengSentiLex with the translated English lexicons in three evaluation datasets. BengSentiLex achieves 5%-50% improvement over the translated lexicons. For profanity detection, BengSwearLex achieves coverage of around 85% in document-level in the evaluation dataset. The experimental results imply that BengSentiLex and BengSwearLex are effective at classifying sentiment and identifying profanity in Bengali social media content, respectively.

BengSentiLex and BengSwearLex: Creating lexicons for sentiment analysis and profanity detection in low-resource Bengali language

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

Bengali is a low-resource language that lacks tools and resources for various natural language processing (NLP) tasks, such as sentiment analysis and profanity identification. In Bengali, only the translated versions of English sentiment lexicons are available. Besides, no lexicon exists for detecting profanity in Bengali social media text. In this work, we introduce a Bengali sentiment lexicon, BengSentiLex, and a Bengali swear lexicon, BengSwearLex. To build BengSentiLex, we propose a cross-lingual methodology that utilizes a machine translation system, review corpus, English sentiment lexicons, pointwise mutual information (PMI), and supervised machine learning (ML) classifiers in different steps. For creating BengSwearLex, we introduce a semi-automatic methodology that leverages an obscene corpus, word embedding, and part-of-speech (POS) taggers. We compare the performance of BengSentiLex with the translated English lexicons in three evaluation datasets. BengSentiLex achieves 5%-50% improvement over the translated lexicons. For profanity detection, BengSwearLex achieves coverage of around 85% in document-level in the evaluation dataset. The experimental results imply that BengSentiLex and BengSwearLex are effective at classifying sentiment and identifying profanity in Bengali social media content, respectively.

1 INTRODUCTION

The popularity of e-commerce and social media has surged the availability of user-generated content. Therefore, text analysis tasks such as sentiment classification and profanity or abusive content identification have received significant attention in recent years. Sentiment analysis identifies emotions, attitudes, and opinions expressed in a text (Liu, 2012). Extracting insights from user feedback data has practical implications such as market research, customer service, result predictions, etc. Profanity indicates the use of taboo or swear words to express emotional feelings and is prevalent in social media data (e.g., online post, message, comment, etc.) (Wang et al., 2014) across languages. The occurrences of swearing or vulgar words are often linked with abusive or hatred context, sexism, and racism. Hence, identifying swearing words has practical connections to understanding and monitoring online content. In this paper, we use the terms such as profanity, slang, vulgarity, and swearing interchangeably to indicate the usage of foul and filthy languages/words even though they have subtle differences in their meanings.

Lexicon plays an important role in both sentiment classification and profanity identification. For example, sentiment lexicons help to analyze key subjective properties of texts such as opinions and attitudes (Taboada et al., 2011). A sentiment lexicon contains opinion conveying terms (e.g., words, phrases, etc.), labeled with the sentiment polarity (i.e., *positive or negative*) and strength of the polarity. Some examples of positive sentiment words are *beautiful, wonderful, amazing* that express some desired states or qualities, while negative sentiment words, such as *bad, awful, poor, etc.* are used to represent undesired states. The profane list, on the other hand, contains words having foul, filthy, and profane meanings (e.g., *ass, fuck, bitch*). When labeled data are unavailable, sentiment classification methods

usually utilize opinion conveying words and a set of linguistic rules. As this approach relies on the polarity of the individual words, it is crucial to building a comprehensive sentiment lexicon. Similarly, creating a lexicon that consists of a list of swear and obscene words is instrumental for determining profanity in a text.

As sentiment analysis is a well-studied problem in English, many general-purpose and domain-specific sentiment lexicons are available. Some of the popular general-purpose sentiment lexicons are MPQI (Wilson et al., 2005), opinion lexicon (Hu and Liu, 2004), SentiwordNet (Esuli and Sebastiani, 2006), VADER (Hutto and Gilbert, 2014), etc. Besides English, other widely used languages such as Chinese, Arabic, Spanish, etc have their sentiment lexicons (Xu et al., 2010; Mohammad et al., 2016; Perez-Rosas et al., 2012). The presence of swearing in English social media has been investigated by various researchers (Wang et al., 2014; Pamungkas et al., 2020). Wang et al. (2014) found that the rate of swear word use in English Twitter is 1.15%, almost double compared to its use in daily conversation (0.5% - 0.7%) as observed in previous work (Jay, 1992). The work of Wang et al. (2014) also reported that a portion of 7.73% tweets in their random sampling collection contains swear words.

Although Bengali is the seventh most spoken language in the world, sentiment analysis or profanity identification in Bengali is still in its beginning. Limited research has been conducted on sentiment analysis in Bengali in the last two decades utilizing supervised machine learning (ML) techniques (Sazzed and Jayarathna, 2019; Das and Bandyopadhyay, 2010a; Chowdhury and Chowdhury, 2014; Sarkar and Bhowmick, 2017; Rahman and Kumar Dey, 2018), as they do not require language-specific resources such as sentiment lexicon, part-of-speech (POS) tagger, dependency parser, etc. Regarding profanity identification, although few works addressed the abusive content analysis, none of them focused on determining profanity or generating resources for identifying profanity.

There have been a few attempts to develop sentiment lexicons for Bengali by translating various English sentiment dictionaries. Das and Bandyopadhyay (2010b) utilized a word-level lexical-transfer technique and an English-Bengali dictionary to develop SentiWordNet for Bengali from English SentiWordNet. Amin et al. (2019) translated the VADER (Hutto and Gilbert, 2014) sentiment lexicon to Bengali for sentiment analysis. However, dictionary-based translation can not capture the informal language people use in everyday communication or social media. Regarding vulgar or swear words, there exist no resources in Bengali which can identify profanity in social media data. Therefore, in this work, we focus on generating resources for these two essential tasks.

To develop the Bengali sentiment lexicon BengSentiLex, we present a corpus-based cross-lingual methodology. We collected around 50000 Bengali drama reviews from Youtube; among them, we manually annotated around 12000 reviews (Sazzed, 2020a). Our proposed methodology consists of three phases, where each phase identifies sentiment words from the corpus and includes them in the lexicon. In phase 1, we identify sentiment words from the Bengali review corpus (both labeled and unlabeled) with the help of two English sentiment lexicons, Bing Liu's opinion lexicon and VADER. In phase 2, utilizing around 12000 annotated reviews and PMI, we identify top class-correlated (*positive* or *negative*) words. Using the POS tagger, we determine adjectives and verbs, which mainly convey opinions. In the final phase, we make use of unlabeled reviews to recognize the polar words. Utilizing the labeled reviews as training data, we determine the class of the unlabeled reviews. We then follow the similar steps of phase 2 to identify sentiment words from these pseudo-labeled reviews. All three phases are followed by a manual validation and synonym generation step. Finally, we provide the comparative performance analysis of the developed BengSentiLex with the translated English lexicons. As BengSentiLex is built from a social media corpus, it contains words that people use on the web, social media, and informal communication; Therefore, it is more effective in recognizing sentiments in text data compared to word-level translation of English lexicons. This sentiment lexicon creation framework is the extended version of the previous work of Sazzed (2020b).

To construct the Bengali swear lexicon, BengSwearLex, we propose a corpus-based semi-automatic approach. However, unlike the methodology used for sentiment lexicon creation, this approach does not use any cross-lingual resources as machine translation is not capable of translating language-specific swear terms. From an existing Bengali obscene corpus, utilizing word embedding and POS tagging, we create BengSwearLex. To show the efficacy of BengSwearLex for identifying profanity, we annotate a negative drama review corpus into profane and non-profane categories based on the presence of swear terms. We find that BengSwearLex successfully identifies 85.5% of the profane reviews from the corpus.

1.1 Motivation and Challenges

Since the existing Bengali sentiment dictionaries lack words people use in informal and social communication, it is necessary to build such a sentiment lexicon in Bengali. With the rapid growth of user-generated Bengali content on social media and the web, the presence of inappropriate content has become an issue. The content which is not in line with the social norms and expectations of a community needs to be censored. However, in Bengali, no such resources exist for identifying the presence of profanity; thus, we focus on building a swear lexicon for Bengali.

Some of the challenges to develop a sentiment lexicon in Bengali are-

1. One of the popular techniques to create a lexicon is to utilize corpora to extract opinion conveying words. However, the Bengali language lacks such corpus. Thus, we have to collect and annotate a large corpus.

2. One of the important tools for identifying opinion word is sophisticated part-of-speech (POS) tagger; However, in Bengali, there exists no sophisticated POS tagger; thus, we leverage POS tagger from English utilizing machine translation.

1.2 Contributions

Our main contributions in this paper can be summarized as follows-

- We introduce two lexical resources, BengSentiLex, a Bengali sentiment lexicon that consists of over 1200 opinion words created from a Bengali review corpus, and BenSwearLex, a Bengali swear lexicon, comprised of about 200 swear words. We have made both lexicons publicly available for the researchers.

- We show how the machine translation-based cross-lingual approach, the labeled and unlabeled reviews, English sentiment lexicons can be utilized to build a sentiment lexicon in Bengali.

- We present a semi-automatic methodology for developing a swear lexicon utilizing an obscene corpus and various natural language processing tools.

- We demonstrate that BengSentiLex and BengSwearLex are effective at sentiment classification and profane terms detection compared to existing tools.

The rest of the paper is structured as follows: In section 2, we review related literature. We explain the corpus creation and annotation process in section 3. Section IV describes various cross-lingual resources used for the Bengali lexicon generation. In section 4, we present the lexicon construction methodology. Section 5 and 6 provide experimental results and discussion. Finally, section 7 concludes and provides future directions.

2 RELATED WORK OF SENTIMENT LEXICON CREATION

Liu (2012) categorized the sentiment lexicon generation methods into three approaches: manual approach, dictionary-based approach, and corpus-based approach. Considerable time and resources are needed for the manual approach as the annotation is performed by humans; The dictionary-based methods usually start with a set of seed words, which are created manually and then expanded using a dictionary. The corpus-based techniques utilize both manually labeled seed words and corpus data. Since the proposed sentiment lexicon creation framework is corpus-based, we mention only the works which utilizes corpus for lexicon creation.

2.1 Corpus-based lexicon generation in English

Huang et al. (2014) proposed a label propagation-based method for generating domain-specific sentiment lexicon. In their work, the candidate sentiment terms are extracted by leveraging the chunk dependency information and prior generic sentiment dictionary. They defined the pairwise contextual and morphological constraints and incorporated the label propagation. Their experimental results suggested that constrained label propagation can improve the performance of the automatic construction of domain-specific sentiment lexicon.

Han et al. (2018) proposed a domain-specific lexicon generation method from the unlabeled corpus utilizing mutual information and part-of-speech (POS) tags. Their lexicon shows satisfactory performance on publicly available datasets.

Tai and Kao (2013) proposed a graph-based label propagation algorithm to generate a domain-specific sentiment lexicon. Their proposed approach considers words as nodes and similarities as weighted edges of the word graphs. Using a graph-based label propagation method, they assigned the polarity to unlabeled words. They conducted experiments on the Twitter dataset and achieved better performance than the general-purpose sentiment dictionaries.

Wang and Xia (2017) developed a neural architecture to train a sentiment-aware word embedding. To enhance the quality of word embedding as well as the sentiment lexicon, they integrated the sentiment supervision at both document and word levels. They performed experiments on the SemEval 2013-2016 datasets using their sentiment lexicon and obtained the best performance in both supervised and unsupervised sentiment classification tasks.

Hamilton et al. (2016) constructed a domain-sensitive sentiment lexicon using label propagation algorithms and small seed sets. They showed that their corpus-based approach outperformed methods that rely on hand-curated resources such as WordNet.

Wu et al. (2019) presented an automatic method for building a target-specific sentiment lexicon. Their lexicon consists of opinion pairs made from an opinion target and an opinion word. Their unsupervised algorithms first extract high-quality opinion pairs; Then utilizing general-purpose sentiment lexicon and contextual knowledge, calculates sentiment scores of opinion pairs. They applied their method on several product review datasets and found their lexicon outperformed several general-purpose sentiment lexicons.

Beigi and Moattar (2021) presented an automatic domain-specific sentiment lexicon construction method for unsupervised domain adaptation and sentiment classification. The authors first constructed a sentiment lexicon from the source domain using the labeled data. In the next phase, the weights of the first hidden layer of Multilayer Perceptron (MLP) are set to the corresponding polarity score of each word from the developed sentiment lexicon, and then the network is trained. Finally, a domain-independent Lexicon (DIL) is introduced that contains words with static positive or negative scores. The experiments on Amazon multi-domain sentiment datasets showed the superiority of their approach over the existing unsupervised domain adaptation methods.

2.2 Lexicon generation in Bengali and other languages

Al-Moslmi et al. (2018) developed an Arabic sentiment lexicon consists of 3880 positive and negative synsets annotated with the part-of-speech (POS), polarity scores, dialects synsets, and inflected forms. They performed the word-level translation of the English MPQA lexicon using google translation, which was followed by manual inspection for removing the inappropriate word. Besides, from two Arabic review corpora, they manually examined a list of opinion words or sentiment words and phrases.

Perez-Rosas et al. (2012), the authors presented a framework to derive sentiment lexicon in Spanish using manually and automatically annotated data from English. To bridge the language gap, they used the multilingual sense-level aligned WordNet structure.

Mohammad et al. (2016), the authors introduced several sentiment lexicons in Arabic that were automatically generated using two different methods: (1) by using distant supervision techniques on Arabic tweets, and (2) by translating English sentiment lexicons into Arabic using a freely available statistical machine translation system. They compared the performance of existing and their proposed sentiment lexicons in sentence-level sentiment analysis.

Asghar et al. (2019) presented a word-level translation scheme for creating an Urdu polarity lexicon using a list of English opinion words, SentiWordNet, English–Urdu bilingual dictionary, and a collection of Urdu modifiers.

Das and Bandyopadhyay (2010b) proposed a computational method for generating an equivalent lexicon of English SentiWordNet using an English-Bengali bilingual dictionary. Their approach used a word-level translation process, which is followed by the error reduction technique. From the SentiWordNet, they selected a subset of opinion words whose orientation strength is above the heuristically identified threshold of 0.4. They used two Bengali corpora, News, and Blog to show the coverage of their developed lexicon.

Amin et al. (2019), the authors compiled a Bengali polarity lexicon from the English VADER lexicon using a translation technique. They modified the functionalities of the VADER lexicon so that it can be

199 directly applied to Bengali sentiment analysis.

2.3 Comparison with Existing Sentiment Lexicons

201 We provide comparisons in both the methodological and evaluation phases with the Bengali lexicon-based
202 methods. Due to differences in language, it is not possible to compare the proposed framework with the
203 English lexicon-based methods in the evaluation step. Thus to show the novelty and originality of the
204 proposed framework, we discuss how the proposed framework is different from the existing sentiment
205 lexicon generation methods in English.

2.3.1 Bengali Sentiment lexicons

207 In contrast to the existing Bengali sentiment lexicons, which are the simple word-level translation of
208 English lexicons, BengSentiLex is created from a Bengali review corpus. Besides, BengSentiLex differs
209 in the way it has been developed and the aspects of the content. We use a cross-lingual corpus-based
210 approach utilizing labeled and unlabeled data, while the existing lexicons simply translate the English
211 lexicons to Bengali at the word level. Besides, as we utilize social media review data, BengSentilex is
212 capable of capturing opinion words people use in informal communication.

2.3.2 English Sentiment Lexicons

214 In this section, we discuss how the proposed methodology is different from some of the existing lexicon
215 creation methods in English. A number of corpus-based lexicon-generation methods employed label
216 propagation algorithms utilizing seed words (Velikovich et al., 2010; Hamilton et al., 2016; Tai and
217 Kao, 2013), while BengSentiLex does not use any seed word list. Some of the existing works utilized
218 PMI or modified PMI in some phase of lexicon generation framework (Yang et al., 2013; Turney and
219 Littman, 2003; Xu et al., 2012); however, other than using PMI, their entire framework is different from
220 the proposed methodology. Besides, they calculated PMI among various features, while the proposed
221 framework utilizes it between feature and target. The work of Beigi and Moattar (2021) utilized the word's
222 frequency in positive and negative comments and vocabulary size of the corpus to determine the polarity
223 score of the corresponding word, while BengSentiLex uses PMI based sentiment intensity (SI) score to
224 determine the semantic orientation of a word. Some other works utilized Matrix Factorization (Peng and
225 Park, 2011) or distant supervision for creating lexicon (Severyn and Moschitti, 2015). A comprehensive
226 literature review of the corpus-based lexicon creation method in English has been performed by Darwich
227 et al. (2019).

3 RELATED WORK OF PROFANITY AND ABUSIVE CONTENT ANALYSIS

229 Researchers studied the existence and sociolinguistic characteristics of swearing or cursing in social
230 media. Wang et al. (2014) investigated the cursing activities on Twitter, a social media platform. They
231 studied the ubiquity, utility, and contextual dependency of swearing on Twitter. Gauthier et al. (2015)
232 analyzed several sociolinguistic aspects of swearing on Twitter text data. Several studies investigated the
233 relationship between social factors such as gender the profanity and discovered males employ profanity
234 much more often than females (Wang et al., 2014; Selnow, 1985). Other social factors such as age,
235 religiosity, or social status were found to be related to the rate of using vulgar words (McEnery, 2004). Jay
236 and Janschewitz (2008) noticed that the offensiveness of taboo words depends on their context, and found
237 that usages of taboo words in conversational context is less offensive than the hostile context. Pinker
238 (2007) classified the use of swear words into five categories: dysphemistic; abusive, using taboo words to
239 abuse or insult someone; idiomatic, using taboo words to arouse the interest of listeners without really
240 referring to the matter; emphatic, to emphasize another word; cathartic, the use of swear words as a
241 response to stress or pain.

242 Research related to the identification of swearing or offensive words has been conducted mainly
243 in English; Therefore, lexicons comprised of offensive words are available in the English language.
244 Pamungkas et al. (2020) created SWAD (Swear Words Abusiveness Dataset), a Twitter English corpus,
245 where abusive swearing is manually annotated at the word level. Their collection consists of 1,511 unique
246 swear words from 1,320 tweets. Razavi et al. (2010) manually collected approximately 2,700 dictionary
247 entries including phrases and multi-word expressions, which is one of the earliest work offensive lexicon
248 creations. The recent work on lexicon focusing on hate speech was reported by (Gitari et al., 2015).
249 Currently, the largest English lexicon of abusive words was provided by (Wiegand et al., 2018).

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). 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, 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 and GloVe embedding by integrating them into several ML classifiers and showed the effectiveness of BengFastText for hate speech detection.

Sazzed (2021a) introduced an annotated Bengali corpus of 3000 transliterated Bengali comments categorized into two classes, abusive and non-abusive, 1500 comments for each. For the baseline evaluations, the author employed several supervised machine learning (ML) and deep learning-based classifiers. They observed support vector machine (SVM) shows the highest efficacy for identifying abusive content.

However, none of the existing works focused on creating resources to detect vulgarity or profanity in Bengali social media content. To the best of our knowledge, it is the first attempt to create resources to detect vulgarity or profanity in the context of Bengali social media data.

4 CREATION OF SENTIMENT LEXICON

4.1 Basic Terminology

This section describes some of the concepts used in this paper for sentiment lexicon creation.

4.1.1 Supervisory Characteristics

Supervised Learning   Supervised learning is one of the most popular approaches of machine learning defined by its usage of annotated data. The labeled data are used to train or "supervise" algorithms for classifying data accurately. Using annotated inputs and outputs, the model can assess its accuracy and learn over time.

Semi-supervised learning   Semi-supervised learning uses both labeled and unlabeled data. It is very useful when a high volume of data is available, but the annotation process is very challenging and requires a huge amount of time and resources.

4.1.2 Cross-lingual approach

The cross-lingual approach leverages resources and tools from a resource-rich language (e.g., English) to a resource-scarce language. Most of the research in sentiment analysis has been performed in English. Hence, resources from English can be employed in other languages using various language mapping techniques. The construction of a language-specific sentiment lexicon requires vast resources, tools, and an active research community, which are not available in the resource-scarce language. A feasible approach could be utilizing resources from the languages where sentiment resources are abundant (Sazzed, 2021b). In this work, we employ machine translation to leverage several resources from English.

4.1.3 Machine translation

Machine translation (MT) refers to the use of software to translate text or speech from one language to another. Over the decades, the machine translation system has evolved to a more reliable system, from the simple word-level substitution to sophisticated Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Zhu et al., 2020).

Machine translation has been successfully applied to various sentiment analysis tasks by researchers. Balahur and Turchi (2014) studied the possibility of employing machine translation systems and supervised methods to build models that can detect and classify sentiment in low-resource languages. Their evaluation showed that machine translation systems were rapidly maturing. The authors claimed that with appropriate ML algorithms and carefully chosen features, machine translation could be used to build sentiment analysis systems in resource-poor languages.

Bengali Reviews	Machine Translation	Polarity
শামিম ভাইয়ের কাছে এমন নাটক আশা করি নি!!	I did not expect such a drama from Shamim Bhai !!	Negative
ফালতু নাটক কবির সিং মুভির ট্রেইলার কপি করে সাওয়ার নাটক বানায়।	False drama Kabir Singh copied movie trailer and made Sawar drama.	Negative
যখন মন খুব থারাপ থাকে,তখন আপনাদের নাটক দেখি।তখন মনটা আরো থারাপ হয়ে যায়,আর একটা সময়ে মন থারাপে, মন থারাপে কাটাকাটি।সত্যি অনেক ভাল লাগে আপনাদের অভিনয় গুলা। দোয়া রইলো চালিয়ে যান ভাই।।।।।।	When the mind is very bad, then I watch your dramas. The prayer continued, brother	Positive
আফরান নিশো ভাই আমাদের টাংগাইলের অহংকার	Afran Nisho Bhai is the pride of our Tangail	Positive
অসাধারণ একটা জুটির অসাধারণ একটা নাটক ছিল ।।।। শেষ ৫ মিনিট ভীষণ কষ্ট লাগলো।।।। বারবার দেখতে ইচ্ছে করছে ।।।।	An extraordinary pair had an extraordinary drama ... The last 5 minutes were very difficult ... Wanting to see again and again ...	Positive
পরিচালক একটা গাঁজা থোর ছাগলের বাচ্ছা এইসব কী	The director is a cannabis-eating goat kid	Negative
অসাধারন আমার কাছে সেই লাকছে~	Extraordinary That Looks To Me	Positive

Figure 1. Sample reviews from training dataset Drama-Train

4.1.4 Pointwise Mutual Information (PMI)

Pointwise Mutual Information (PMI) is a measure of association used in information theory and statistics. The PMI between two variables X and Y is computed as,

$$PMI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)} \qquad (1)$$

The term P(X, Y) is the number of observed co-occurrences of event X and Y. P(X) represents the number of times X occurs, and P(Y) means the number of times Y occurs. When two variables X and Y are independent, the PMI between them is 0. PMI maximizes when X and Y are perfectly correlated.

4.2 Datasets for Lexicon Creation and Evaluation

4.2.1 Training dataset for sentiment lexicon

We use a drama review dataset (Drama-Train) collected from Youtube (Sazzed, 2020a) to build BengSentiLex. This corpus consists of around 50000 Bengali reviews, where each review represents the viewer's opinions towards a Bengali drama. Among the 50000 Bengali reviews, around 12000 are annotated by Sazzed (2020a), while the remaining are unlabeled. Figure 1 shows examples of drama reviews belong to the Drama-Train.

4.2.2 Evaluation dataset for sentiment lexicon

We show the effectiveness of BengSentiLex in three datasets from distinct domains. Table 1 provides the details of the evaluation datasets.

The first evaluation dataset is a drama review dataset (Drama-Eval) consisting of around 1000 annotated reviews. This dataset belongs to the same domain as the training dataset, Drama-Train; However, it has not been used for lexicon creation. This is a class-balanced dataset, consists of 500 *positive* and 500 *negative* reviews.

The second dataset is a news dataset (News1) that was collected from (Soc, 2020). It consists of 4000 news comments; among them, 2000 are *positive* and 2000 are *negative* comments.

The third dataset is also a news comment dataset (News2), collected from two popular Bengali newspapers, Prothom Alo and BBC Bangla (Taher et al., 2018). It consists of 5205 positive comments and around 5600 negative comments. For the evaluation, we select a class-balanced subset where each class contains 5205 comments.

Table 1. Evaluation Datasets for BengSentiLex

Dataset	Domain	Positive	Negative	Total
Drama-Eval	Drama Review	1000	1000	2000
News1	News Comments	2000	2000	4000
News2	News Comments	5205	5205	10410

4.3 Cross-lingual resources

4.3.1 Sentiment lexicon

To identify whether a Bengali word conveys an opinion, we employ a cross-lingual approach. Leveraging machine translation and English sentiment lexicons, we decide whether an extracted Bengali word bears opinion. The study by Sazzed and Jayarathna (2019) showed that though the Bengali to English machine translation (i.e., Google Translate) system is not perfect, it preserves semantic orientation in most cases.

We translate all the extracted Bengali words into English and then determine their polarities based on the English lexicon. If the translated word exists in an English lexicon, we include the corresponding Bengali word in our Bengali sentiment lexicon. Although we perform word-level translation between English and Bengali, it differs from existing works that translate words from English to Bengali, therefore only contain translated Bengali dictionary words rather than the words used by people in informal communication.

Our proposed approach supports the inclusion of informal Bengali words, which is not achievable using the dictionary-based translation of English lexicons. Furthermore, this approach can yield and include multiple synonymous opinion words instead of one. For example, by translating an English sentiment word, we only get the corresponding Bengali term. However, when words are extracted from the corpus and translated to English, due to the low coverage of the machine translation system, synonymous Bengali words can be mapped into the same English polarity word. Thus, it helps to identify and include more opinion words in the Bengali lexicon.

To determine the polarity of the translated words, we utilize the following English sentiment lexicons.

Opinion lexicon was developed by (Hu and Liu, 2004) and contains around 6800 English sentiment words (*positive* or *negative*). Besides the dictionary words, it also includes acronyms, misspelled words, and abbreviations. Liu's opinion lexicon is a binary lexicon, where each word is associated with either *positive* (+1) or *negative*(-1) polarity value.

VADER is a sentiment lexicon especially attuned to social media. VADER contains over 7,500 lexical features with sentiment polarity of either *positive* or *negative* and sentiment intensity between –4 to +4. VADER includes emoticons such as ':-)', which denotes a smiley face (positive expression), and sentiment-related initialisms such as 'LOL', 'WTF'.

4.3.2 Part-of-speech (POS) tagging

A Part-of-speech (POS) tagger is a tool that assigns a POS tag (e.g., noun, verb, adjective, etc.) to each of the words present in a text. As adjectives, nouns, and verbs usually convey opinions, the POS tagger can help to identify opinion words. In English, several standard POS taggers are available such as NLTK POS tagger Loper and Bird (2002), spaCy POS tagger Honnibal and Montani (2017). However, in Bengali, since no sophisticated POS tagger is available, we use the machine translation system to convert the probable Bengali opinion words to English. We then use the spaCy POS tagger to determine the POS tag of those English words, which allows us to label the POS tag of the corresponding Bengali words.

4.4 Methodology

The creation of the BengSentiLex involves several phases. We utilize various tools and resources to determine opinion words from the corpus and add them to BengSentiLex, as shown in Fig 2.

- Phase 1: Labeled and unlabeled corpus, machine translation system, English lexicons.

- Phase 2: Labeled corpus, PMI, machine translation system, English POS tagger, English lexicons, Bengali lexicon (constructed in phase 1).

- Phase 3: Unlabeled corpus, ML classifiers, PMI, machine translation system, English POS tagger, English lexicons, Bengali lexicon (constructed in phase 1 and phase 2).
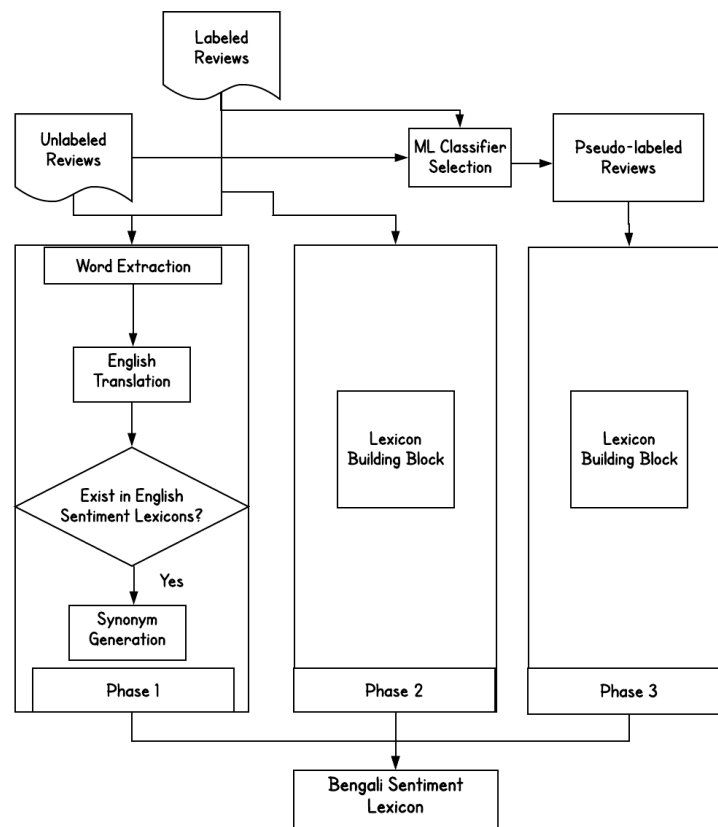
Figure 2. The various phases of sentiment lexicon generation in Bengali
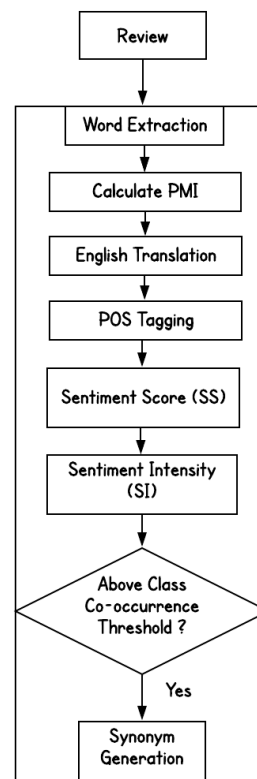
Figure 3. The lexicon building block

Each phase enlarges BengSentiLex with the newly recognized opinion-conveying words. A manual validation step is included to examine identified opinion words. Then, we generate synonyms for the validated words which are added to the lexicon.

For synonym generation, we utilize google translate[1], as no standard Bengali synonym dictionary is available on the web in digital format. We translate Bengali words into multiple languages such as English, Chinese, French, Hindi, Russian and Arabic and then perform back-translation. This approach assists to find synonyms as sentiments are expressed in different ways across the languages.

4.4.1 Phase-1: Utilizing English Sentiment Lexicons

The development of a sentiment lexicon typically starts with a list of well-defined sentiment words. A well-known approach for identifying the initial list of words (often called seed words) is to use a dictionary. However, dictionary words denote mostly formal expressions and usually do not represent the words people use in social media or informal communication. On the contrary, words extracted from a corpus represent terms people use in regular communication, hence, more useful for sentiment analysis.

We tokenize words from the review corpus (both labeled and unlabeled) using the NLTK tokenizer and calculate their frequency in the corpus. Only the words with a frequency above 5 are added to the candidate pool. However, not all high-frequency words convey sentiments. For example-'drama' is a high-frequency word in our drama review dataset, but it is not a sentiment word.

As Bengali does not have any sentiment dictionary of its own, we utilize resources from English. Using a machine translation system, we convert all the words from the candidate pool to English. Two English sentiment lexicons, Opinion Lexicon, and VADER are employed to determine the polarity of the translated words.

The assumption is that if a translated English word exists in the English sentiment lexicon, then it is an opinion conveying word; therefore, the corresponding Bengali word can be added to the Bengali sentiment dictionary.

---

[1]https://translate.google.com

4.4.2 Phase 2: Lexicon generation from labeled data

Phase 2 retrieves opinion words from the labeled corpus by leveraging the pointwise mutual information (PMI) and a POS tagger, as shown in Figure 2 and 3.

From the labeled reviews, we derive the terms which are highly correlated with the class label. The words or terms that already exist in the lexicon (from earlier phases) are not considered. The remaining words are translated into English using the machine translation system. We utilize the spaCy POS tagger to identify their POS tags. Since usually adjectives and verbs convey opinions, we only keep them and exclude the other POS.

The sentiment score of a word, w, is calculated using the formula shown below,

$$SentimentScore(w) = PMI(w, pos) - PMI(w, neg) \qquad (2)$$

where, $PMI(w, pos)$ represents the PMI score of word w corresponding to *positive* class and $PMI(w, neg)$ represents the PMI score of word w corresponding to *negative* class.

We then calculate the sentiment intensity (SI) of w, using the following equation,

$$SI(w) = \frac{SentimentScore(w)}{PMI(w, pos) + PMI(w, neg)} \qquad (3)$$

We use the sentiment strength along with the threshold value to identity opinion conveying words from the labeled reviews.

If the sentiment intensity of a word, w, is above the threshold of 0.5, we consider it as a *positive word*. if sentiment strength is below -0.5, we consider it as a *negative* word.

$$Class(w) = \begin{cases} Positive, & \text{if } SI(w) > 0.50 \\ Negative, & \text{if } SI(w) < -0.50 \\ Unassigned, & Otherwise \end{cases} \qquad (4)$$

4.4.3 Phase 3: Lexicon generation from unlabeled (pseudo-labeled) data

In addition to annotated reviews, our review corpus consists of a large number of unlabeled reviews. For the labeled reviews, we use PMI to identify the top class-correlated words. However, for the unannotated reviews, the true class labels are not available; thus, automatic labeling is required. To automatically annotate the unlabeled reviews, we employ various several ML classifiers and select the classifier with the highest accuracy. The unigram and bigram-based tf–idf (term frequency–inverse document frequency) scores are used as input features for the ML classifiers. The following ML classifiers are employed:

SVM (Support Vector Machine) is a popular supervised ML algorithm used for classification and regression problems. Originally, SVM is a binary classifier that decides the best hyperplane to separate the space into binary classes by maximizing the distance between data points belong to different classes. However, SVM can be used as multi-class classifier following same principle and employing *one-versus-one* or *one-vs-the-rest* strategy.

SGD (Stochastic gradient descent) is a method that optimizes an objective function iteratively. It is a stochastic approximation of actual gradient descent optimization since it calculates gradient from a randomly selected subset of the data. For SGD, hinge loss and l2 penalty with a maximum iteration of 1500 are employed.

LR (Logistic regression) is a statistical classification method that finds the best fitting model to describe the relationship between the dependent variable and a set of independent variables.

Random Forest (RF) is a decision tree-based ensemble learning classifier. It makes predictions by combining the results from multiple individual decision trees.

K-nearest neighbors (k-NN) algorithm is a non-parametric method used for classification and regression. In k-NN classification, the class membership of a sample is determined by the plurality vote of its neighbors. Here, we set k=3, the class of a review depends on three of its closest neighbors.

We use scikit-learn (Pedregosa et al., 2011) implementation of the aforementioned ML classifiers. For all of the classifiers, we use the default parameter settings. Using 10-fold cross-validation, we assess their

Table 2. Performances of supervised ML classifiers in annotated corpus

Classifier	Precision	Recall	F1 Score	Accuracy
SGD	0.939	0.901	0.920	93.61%
SVM	0.908	0.924	0.916	93.00%
LR	0.889	0.922	0.905	91.80%
k-NN	0.901	0.849	0.875	90.18%
RF	0.878	0.870	0.874	89.91%

Steps for generating pseudo-labeled data
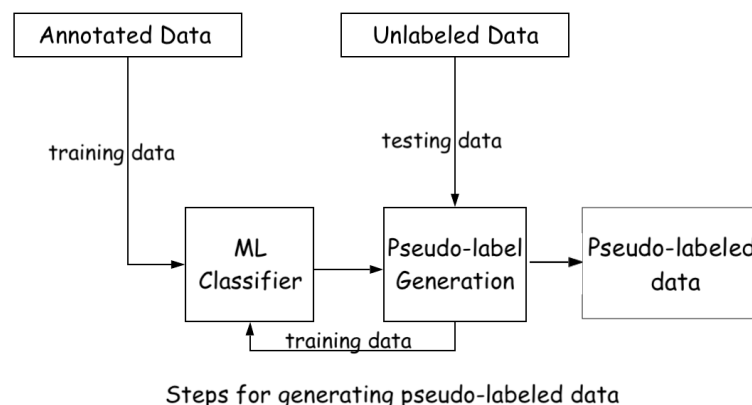
Figure 4. Class-label assignment of unlabeled reviews using supervised ML classifier and labeled data

performances. The purpose of this step is to find reliable classifiers that can be used for automatic class labeling.

Table 2 shows the classification accuracy of various ML classifiers using 10-fold cross-validation. Among the five classifiers we employ, SGD and SVM show higher accuracy. Both of them correctly identify around 93% of the reviews, which is close to the accuracy of manual annotations. LR shows similar accuracy of around 92%. We use these three classifiers to determine the class of the unlabeled reviews. The following procedures are considered for automatically generating class-label of the unannotated reviews utilizing the ML classifiers,

1) Use all the labeled reviews as training data and all the unlabeled reviews as testing data.

2) Iteratively utilize a small unlabeled set as testing data. After assigning their labels, we add these pseudo-labeled reviews to the training set and select a new set of unlabeled reviews as testing data. This procedure continues until all the data are labeled.

To determine the performance of the approach (1), we conduct 4-fold cross-validation on the labeled reviews. We use 1-fold as training data and the remaining 3-folds as testing data. The training-testing data ratio is selected based on the ratio of labeled (around 12000) and unlabeled data (around 38000) reviews. For approach (2), similar way to approach 1, we split the 12000 labeled data into four subsets. Initially, we use one subset (around 3000 reviews) as a training set. In each iteration, a group of reviews is selected from the other three subsets (from around 9000 reviews), and used as a testing set. The size of the chosen group is equal to 10% of the current training set. After assigning the class of the reviews that belong to the testing set, they are added to the training set. This process continues until all the reviews (9000 reviews) are annotated.

We find that gradually expanding the training set by adding the predicted reviews from the testing set provides better performance. After applying approach 2 (shown in Figure 4), our dataset contains around 38000 pseudo-labeled reviews. We then employ PMI and POS tagger in a similar way to phase 2. However, since this phase utilizes pseudo-labeled data instead of the true-label data, we set a higher threshold of 0.7 for the class label assignment.

Bengali Comments	English Translations
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.
ব্লাইঞ্চোদ থানকির ছেলে। এতো অ্যাড চুদাও কে,,,	Fucker whore's son. why so many advertisements?

Figure 5. Examples of Bengali obscene comments and corresponding English translation

5 CREATION OF SWEAR LEXICON

5.1 Corpus

We utilize two Bengali corpora, one for creating the swear lexicon, BengSwearLex, which we refer to as *training corpus (SW)*, and the other one for analysis and evaluating the performance of BengSwearLex, which we refer to as *evaluation corpus (SW)*.

5.1.1 Training corpus (SW)

We use a Bengali corpus deposited by Abu (2020) for constructing the BengSwearLex. Originally, this corpus consists of 10221 text reviews/comments belong to different categories, such as toxic, racism, obscene, and insult. However, this corpus is noisy, consisting of many empty and punctuation only comments or erroneous annotation. We manually validate and exclude comments having the above-mentioned issues.

From the modified corpus, we only select the reviews labeled as *obscene*. After excluding erroneous reviews and reviews that belong to other classes, the corpus consists of 3902 obscene comments. The length of each comment ranges from 1-100 words.

Figure 5 shows some examples of the obscene comments from the *training corpus (SW)*.

5.1.2 Evaluation corpus (SW)

The evaluation corpus we utilize is a drama review corpus collected from Youtube (the same corpus that is used for sentiment lexicon creation).

This corpus was created and deposited by (Sen, 2019) 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 label these 3307 negative reviews into two categories, profane and non-profane. The annotation of these 3307 negative reviews was conducted by three Bengali expert annotators (A1, A2, A3). The first two annotators (A1 and A2) initially annotated all the reviews. In case of disagreement in annotation, it was resolved by a third annotator (A3).

Table 3. Description of Drama Review Evaluation Corpus

Profane	Non-Profane	Total
664	2643	3307

After annotation, this corpus consists of 2643 non-profane negative reviews and 664 profane reviews, as shown in Table 3. The kappa statistic for the two raters (A1 and A2) is 0.81, which indicates a high agreement in the annotation.

5.2 Text Processing Tools

5.2.1 POS tagger

Similar to sentiment lexicon (described in previous section), here we utilize a POS tagger to identify opinion word [2] [3]. However, in Bengali, limited research has been conducted for developing a sophisticated POS tagger; therefore, the existing Bengali POS taggers are not as accurate as of its English counterpart. Hence, manual validation is needed to check the correctness of the POS tag assigned by the POS taggers.

[2] https://github.com/AbuKaisar24/Bengali-Pos-Tagger-Using-Indian-Corpus/
[3] https://www.isical.ac.in/ utpal/docs/POSreadme.txt
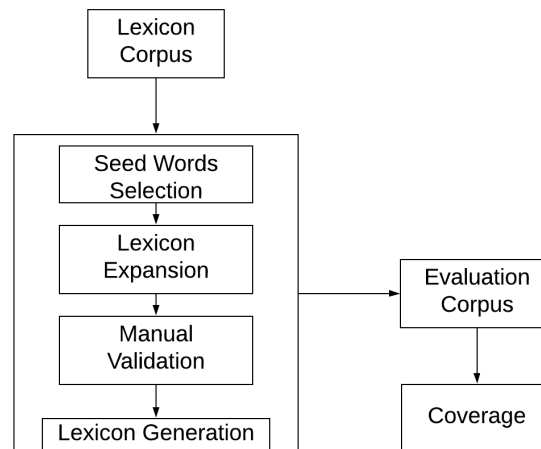
Figure 6. The Proposed Methodology

5.2.2 Word Embedding

A word embedding is a learned representation for text, where related words have similar representations. The word-embedding provides an efficient way to use the dense representation of words of varying lengths. The values for the embedding of a word are learned by the model during the training phase.

There exist two main approaches for learning word embedding, count-based and context-based. The count-based vector space models heavily rely on the word frequency and co-occurrence matrix with the assumption that words in the same contexts share similar or related semantic meanings. The other learning approach, context-based methods, build predictive models that predict the target word given its neighbors. The best vector representation of each word is learned during the model training process.

The Continuous Bag-of-Words (CBOW) model is a popular context-based method for learning word vectors. It predicts the center word from surrounding context words.

5.3 Swear Lexicon Creation Framework

Lexical resources can help to identify the presence of profanity in Bengali social media. Here, we present a semi-automatic approach for creating a swear lexicon utilizing an annotated corpus, word-embedding, and POS tagger. The complete lexicon development stage consists of three phases as shown in Figure 6,

1. Seed word selection

2. Lexicon expansion

3. Manual Validation

5.3.1 Seed word selection

The lexicon creation process usually begins with a list of seed words. Our proposed approach utilizes an annotated obscene corpus to generate a seed word list. We extract and count the occurrence of individual words in the corpus. Based on the word-occurrence count in the corpus, we select the top 100 words. However, we find that the list contains some non-vulgar words, which we exclude utilizing POS tagger and manual validation.

5.3.2 Lexicon expansion

The lexicon expansion step involves utilizing word embedding to identify similar words of the already recognized swear words. We use the training corpus (SW) as a training dataset and utilize the Gensim (Řehůřek and Sojka, 2010) Continuous Bag-of-Words (CBOW) implementation to find similar words. The entire procedure consists of the following steps.

- In the first step, we find the words which are the most similar to the seed words.

527 • The second step iteratively finds words similar to vulgar words recognized in step 1. We remove the
528 duplicate word automatically. In addition, we remove words that are not a noun, adjective, or verb.
529 Several iterations are performed until we notice no significant expansion of the swear word list.

5.3.3 Manual validation

531 In the final step, we manually exclude non-swear words that exist in the swear lexicon. As lexical resources
532 such as POS tagger in Bengali are not sophisticated enough, a manual validation step is necessary to
533 eliminate non-swear words. Moreover, we find that vulgar comments often do not follow the usual
534 sentence structure; Therefore, the POS tagger fails to label them correctly.

6 EXPERIMENTAL RESULTS

6.1 Sentiment Classification

6.1.1 Baselines and evaluation metrics

538 We compare the performances of our corpus-built lexicon BengSentiLex (716 negative words and 519
539 positive words) with the translated versions of three English sentiment lexicons: VADER (7518 words),
540 AFINN (2477 words), and Opinion Lexicon (6800 words) by integrating them into a lexicon-based
541 classifier. We compute the accuracy of all the four lexicons in three cross-domain evaluation datasets to
542 show the effectiveness of BengSentiLex in the varied domains and distributions.

Table 4. Performance of various lexicons for sentiment classification

Dataset	Lexicon	#Neg Class	#Pos Class	Total
Drama	AFINN	145/1000 (14.15%)	488/1000 (48.80%)	633/2000 (31.65%)
	Opinion Lexicon	241/1000 (24.10%)	598/1000 (59.80%)	839/2000 (41.95%)
	VADER	225/1000 (22.5%)	707/1000 (70.70%)	932/2000 (46.60%)
	BengSentiLex	533/1000 (53.30%)	775/1000 (77.50%)	1308/2000 (65.40%)
News1	AFINN	626/2000 (31.3%)	628/2000 (31.4%)	1254/4000 (31.35%)
	Opinion Lexicon	590/2000 (29.5%)	833/2000 (41.65%)	1423/4000 (35.57%)
	VADER	566/2000 (28.30%)	1070/2000 (53.50%)	1636/4000 (40.90%)
	BengSentiLex	932/2000 (46.6%)	960/2000 (48.80%)	1892 4000 (47.30%)
News2	AFINN	2004/5660 (35.4%)	1826/5205 (35.08%)	3830/10865 (35.2%)
	Opinion Lexicon	1763/5660 (31.14%)	2274/5205 (43.68%)	4037/10865 (37.15%)
	VADER	1662/5205 (31.93%)	2827/5205 (54.31%)	4489/10410 (43.12%)
	BengSentiLex	2086/5205 (40.08%)	2721/5205 (52.27%)	4807/10410 (46.17%)

543 As no Bengali lexicon-based sentiment analysis tool is publicly available, we develop a simple
544 lexicon-based sentiment analysis tool, BengSentiAn (Bengali Sentiment Analyzer). The polarity score of
545 a review is computed by adding up the polarity score of individual opinion words (based on the opinion
546 lexicon) present in a review. Besides, negation words are considered to shit the polarity of the opinion
547 words. If the total polarity score of a review is above 0, we consider it as a positive prediction; if the final
548 score is below 0, we consider it as negative; when the polarity score is 0, we consider the prediction as
549 wrong.

550 A polarity score of 0 can result when the word-level polarity score of a lexicon can not distinguish a
551 review as positive or negative, or the lexicon lacks coverage of opinion words present in the review. It
552 is more appropriate to consider this scenario as a misprediction rather than a positive or negative class
553 prediction.

6.1.2 Comparative results

555 Table 4 shows the comparative performances of various translated lexicons and BengSentiLex when
556 integrated with BengSentiAn. In the drama review dataset, BengSentiLex classifies 1308 reviews out of
557 2000 reviews with an accuracy of around 65%. Among the three translated lexicons, VADER classifies
558 46.60% reviews correctly, while AFINN and Opinion Lexicon provide 31.65% and 41.95% accuracy,
559 respectively. In the News1 dataset, BengSentiLex exhibits an accuracy of 47.30%, while the VADER,

Opinion Lexicon, and AFINN provide an accuracy of 40.90%, 35.57%, 31.65%, respectively. In the News2 dataset, BengSentiLex shows an accuracy of 46.17%, while the VADER provides the second-best performance with an accuracy of 43.12%.

6.2 Profanity Identification

6.2.1 Evaluation metric

To show the effectiveness of BengSwearLex, we utilize document-level coverage. The document-level coverage (or recall) of a lexicon corresponding to a review corpus is calculated as follows-

From the corpus, we first count the number of reviews that contain at least one word from the lexicon, which is then divided by the total number of reviews present in the corpus. Finally, it is multiplied by 100. The following equation is used to calculate document-level coverage (*DCov*) of a lexicon-

$$DCov = \frac{Number\ of\ reviews\ with\ (>0)\ swear\ word\ identified}{total\ number\ of\ reviews\ in\ corpus} * 100$$

The purpose of creating BengSwearLex is to identify comments and reviews that contain swear or slang words, not to identity non-profane comments; thus, we show document-level coverage for only the profane reviews. Regarding the false positive, as BengSwearLex is manually validated at the final step, it contains only swear words; hence, there is no possibility that it identifies a non-profane comment as profane (false positive).

As no swear lexicon exists in Bengali, we compare the performance of BengSwearLex with several supervised classifiers (that use in-domain labeled data) for profanity detection in the evaluation corpus.

Two popular supervised ML classifiers: Logistics Regression (LR) and Support Vector Machine (SVM), and an optimization method, Stochastic Gradient Descendent (SGD) is employed in the evaluation corpus to identify profane reviews. As a feature vector, we use the unigram and bigram-based tf-idf score. 10-fold cross-validation is performed to assess the performance of various ML classifiers. For all the classifiers, default parameter settings are used. For SGD, hinge loss and l2 penalty with a maximum iteration of 1500 are employed.

Furthermore, we employ Deep Neural Network (DNN) based architecture, Convolutional Neural Network (CNN), Long short-term memory (LSTM), and Bidirectional Long short-term memory (BiLSTM) to identify profanity. The DNN based model starts with the Keras (Chollet et al., 2015) embedding layer. 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 300. 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 implementing DNN based model.

6.2.2 Comparison results

Table 5 shows that among the 664 profane reviews present in the evaluation corpus (SW), BengSwearLex registers 564 reviews as profane by identifying the presence of at least one swear term in the review, document-level coverage of 84.93%.

Table 5 shows the coverage of these ML classifiers in the evaluation corpus. We provide their performances in two different settings: class-balanced setting and class-imbalanced setting. In the class-imbalanced setting, we utilize all the 664 profane comments and 2643 non-profane negative comments. In the class-balanced setting, we use all the 664 profane comments; however, for the non-profane class, we randomly select 664 non-profane comments from the pool of 2643 non-profane comments.

From Table 5, we observe that when the original class-imbalanced data is used, all the three ML classifiers achieve coverage of around 60%. However, when a class-balanced dataset is utilized, the performances of classifiers dramatically increase, achieve coverage of around 90%.

Table 5. Document-level coverage of various methods for profanity detection

Type	Method	# Identified	DCov
Unsupervised	**BengSwearLex**	564/664	84.93%
Supervised (Unbalanced)	**LR**	161/664	24.5%
	SVM	345/664	53.4%
	SGD	366/664	58.8%
	LSTM	433/664	65.21%
	BiLSTM	462/664	70.4%
	CNN	444/664	66.86%
Supervised (Balanced)	**LR**	609/664	91.71%
	SVM	594/664	89.45%
	SGD	589/664	88.70%
	LSTM	610/664	91.67%
	BiLSTM	624/664	94.0%
	CNN	609/664	91.71%

7 DISCUSSION

7.1 Sentiment Lexicon

The results suggest that translated lexicons are not good enough to capture the semantic orientation of the reviews as they lack coverage of opinion words presents in Bengali text. We find that BengSentiLex performs considerably better than the translated lexicons in the drama review dataset, with over 40% improvements. Since BengSentiLex is developed from the corpus that belongs to the same domain, it is very effective at classifying sentiments in this drama review evaluation corpus.

Also, for the two other cross-domain evaluation corpus, News1 and News2, BengSentiLex yields better performance compared to translated lexicons; especially, for classifying negative reviews, which can be attributed to the presence of a higher number of *negative* opinion words (716) in BengSentiLex compared to 519 *positive* sentiment words.

The results indicate that utilizing corpora in the target language for automated sentiment lexicon generation is more effective as opposed to translating words directly from another language such as English. As BengSentiLex is built from a social media corpus, it is comprised of words that people use in the web, social media, and informal communication; Thus, it is more effective in recognizing sentiments in text data compared to word-level translated lexicons.

Although supervised ML classifiers usually perform better in sentiment classification, they require annotated data, which are largely missing in low-resource languages such as Bengali. Thus, the developed lexicon can help sentiment classification in Bengali.

7.2 Swear Lexicon

The results of Table 5 reveal that BengSwearLex is capable of identifying profanity in Bengali social media content. It shows higher document-level coverage than in-domain labeled data when a class-imbalanced training set is used. However, a class-balanced training set performs better than BengSwearLex. Labeled data is scarcely available in low-resource languages such as Bengali; therefore, although small in size, BengSwearLex can be an effective tool for profanity identification in the inadequacy of labeled data. Besides, since BengSwearLex consists of only swear or obscene terms, there is a very low chance that it would refer to non-obscene comments as obscene (False Positive), thus capable of achieving a very high precision score.

8 SUMMARY AND CONCLUSION

In this paper, we present two methodologies for creating lexical resources for Bengali (i.e., sentiment lexicon and swear/obscene lexicon). The first methodology leverages the Bengali review corpus, machine-

translation system, English lexicons, and ML classifiers to develop the Bengali sentiment lexicon, BengSentiLex. We demonstrate the effectiveness of BengSentiLex in both in-domain and cross-domain datasets. When integrated into a lexicon-based tool, BengSentiLex yields better performance compared to translated English lexicons. The other methodology creates a swear lexicon, named BengSwearLex, utilizing corpus and other text processing tools resources. We show that BengSwearLex is capable of identifying vulgar language that exists in social media content. Besides, we provide an annotated dataset for profanity analysis in Bengali social media data. We have made both BengSentiLex and BengSwearLex publicly available for the researchers [4].

The superior performance of the BengSentiLex suggests that a corpus-based lexicon can capture the language-specific features and connotations related to the language, which translated sentiment lexicon can not accomplish. Similarly, BengSwearLex can be utilized to distinguish profanity in Bengali social media content when annotated data are unavailable. The proposed methodologies can be adapted to other resource-limited languages to create lexical resources. The future work will involve expanding the size of both lexicons utilizing larger and multi-domain training corpora.

REFERENCES

(2019). Sentiment analysis bengali dataset. https://github.com/sazzadcsedu/BN-Dataset . Accessed 31 Mar. 2020.

(2020). Sentiment analysis bengali dataset. https://github.com/aimansnigdha/Bangla-Abusive-Comment-Dataset.git . Accessed 10 Sep. 2020.

(2020). socian-bangla-sentiment-dataset-labeled. https://github.com/socian-ai/socian-bangla-sentiment-dataset-labeled. Accessed: 2020-04-30.

Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., and Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of information science*, 44(3):345–362.

Amin, A., Hossain, I., Akther, A., and Alam, K. M. (2019). Bengali vader: A sentiment analysis approach using modified vader. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE.

Asghar, M. Z., Sattar, A., Khan, A., Ali, A., Masud Kundi, F., and Ahmad, S. (2019). Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language. *Expert Systems*, 36(3):e12397. e12397 EXSY-Apr-18-123.R2.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

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.

Beigi, O. M. and Moattar, M. H. (2021). Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. *Knowledge-Based Systems*, 213:106423.

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.

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

Chowdhury, S. and Chowdhury, W. (2014). Performing sentiment analysis in bangla microblog posts. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 1–6. IEEE.

Darwich, M., Mohd, S. A., Omar, N., and Osman, N. A. (2019). Corpus-based techniques for sentiment lexicon generation: A review. *J. Digit. Inf. Manag.*, 17(5):296.

Das, A. and Bandyopadhyay, S. (2010a). Phrase-level polarity identification for bangla. *Int. J. Comput. Linguist. Appl.(IJCLA)*, 1(1-2):169–182.

Das, A. and Bandyopadhyay, S. (2010b). Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task*, 2:1–8.

Emon, E. A., Rahman, S., Banerjee, J., Das, A. K., and Mittra, T. (2019). A deep learning approach to de-

---

[4]https://github.com/sazzadcsedu/BNLexicon.git. Accessed: 2021-06-25.

693   tect abusive bengali text. In *2019 7th International Conference on Smart Computing & Communications*
694   *(ICSCC)*, pages 1–5. IEEE.

695   Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion
696   mining. In *LREC*, volume 6, pages 417–422. Citeseer.

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

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

701   Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment
702   lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural*
703   *Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume
704   2016, page 595. NIH Public Access.

705   Han, H., Zhang, J., Yang, J., Shen, Y., and Zhang, Y. (2018). Generate domain-specific sentiment lexicon
706   for review sentiment analysis. *Multimedia Tools and Applications*, 77(16):21265–21280.

707   Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings,
708   convolutional neural networks and incremental parsing. *To appear*, 7(1).

709   Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM*
710   *SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

711   Huang, S., Niu, Z., and Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon
712   based on constrained label propagation. *Knowledge-Based Systems*, 56:191–200.

713   Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of
714   social media text. In *Eighth international AAAI conference on weblogs and social media*.

715   Ishmam, A. M. and Sharmin, S. (2019). Hateful speech detection in public facebook pages for the
716   bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications*
717   *(ICMLA)*, pages 555–560. IEEE.

718   Jay, T. (1992). *Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the*
719   *Movies, in the Schoolyards, and on the Streets*. John Benjamins Publishing.

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

721   Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of*
722   *the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

723   Karim, M. R., Chakravarthi, B. R., McCrae, J. P., and Cochez, M. (2020). Classification benchmarks for
724   under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE*
725   *7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

726   Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technolo-*
727   *gies*, 5(1):1–167.

728   Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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

731   Mohammad, S., Salameh, M., and Kiritchenko, S. (2016). Sentiment lexicons for arabic social media. In
732   *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*,
733   pages 33–37.

734   Pamungkas, E. W., Basile, V., and Patti, V. (2020). Do you really want to hurt me? predicting abusive
735   swearing in social media. In *The 12th Language Resources and Evaluation Conference*, pages 6237–
736   6246. European Language Resources Association.

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

741   Peng, W. and Park, D. H. (2011). Generate adjective sentiment dictionary for social media sentiment
742   analysis using constrained nonnegative matrix factorization. In *Fifth International AAAI Conference on*
743   *Weblogs and Social Media*.

744   Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in spanish. In *LREC*,
745   volume 12, page 73.

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

747   Rahman, M. and Kumar Dey, E. (2018). Datasets for aspect-based sentiment analysis in bangla and its

748   baseline evaluation. *Data*, 3(2):15.

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

751   Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In
752   *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50,
753   Valletta, Malta. ELRA.

754   Sarkar, K. and Bhowmick, M. (2017). Sentiment polarity detection in bengali tweets using multinomial
755   naïve bayes and support vector machines. In *2017 IEEE Calcutta Conference (CALCON)*, pages 31–36.
756   IEEE.

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

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

762   Sazzed, S. (2021a). Abusive content detection in transliterated bengali-english social media corpus. In
763   *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages
764   125–130.

765   Sazzed, S. (2021b). Improving sentiment classification in low-resource bengali language utilizing cross-
766   lingual self-supervised learning. In *International Conference on Applications of Natural Language to*
767   *Information Systems*, pages 218–230. Springer.

768   Sazzed, S. and Jayarathna, S. (2019). A sentiment classification in bengali and machine translated english
769   corpus. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data*
770   *Science (IRI)*, pages 107–114. IEEE.

771   Selnow, G. W. (1985). Sex differences in uses and perceptions of profanity. *Sex Roles*, 12(3-4):303–312.

772   Severyn, A. and Moschitti, A. (2015). On the automatic learning of sentiment lexicons. In *Proceedings of*
773   *the 2015 Conference of the North American Chapter of the Association for Computational Linguistics:*
774   *Human Language Technologies*, pages 1397–1402.

775   Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for
776   sentiment analysis. *Computational linguistics*, 37(2):267–307.

777   Taher, S. A., Akhter, K. A., and Hasan, K. A. (2018). N-gram based sentiment mining for bangla text
778   using support vector machine. In *2018 International Conference on Bangla Speech and Language*
779   *Processing (ICBSLP)*, pages 1–5. IEEE.

780   Tai, Y.-J. and Kao, H.-Y. (2013). Automatic domain-specific sentiment lexicon generation with label
781   propagation. In *Proceedings of International Conference on Information Integration and Web-based*
782   *Applications & Services*, pages 53–62.

783   Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation
784   from association. *acm Transactions on Information Systems (tois)*, 21(4):315–346.

785   Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. (2010). The viability of web-derived
786   polarity lexicons.

787   Wang, L. and Xia, R. (2017). Sentiment lexicon construction with representation learning based on
788   hierarchical sentiment supervision. In *Proceedings of the 2017 Conference on Empirical Methods in*
789   *Natural Language Processing*, pages 502–510.

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

793   Wiegand, M., Ruppenhofer, J., Schmidt, A., and Greenberg, C. (2018). Inducing a lexicon of abusive
794   words–a feature-based approach.

795   Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment
796   analysis. In *Proceedings of human language technology conference and conference on empirical*
797   *methods in natural language processing*, pages 347–354.

798   Wu, S., Wu, F., Chang, Y., Wu, C., and Huang, Y. (2019). Automatic construction of target-specific
799   sentiment lexicon. *Expert Systems with Applications*, 116:285–298.

800   Xu, G., Meng, X., and Wang, H. (2010). Build chinese emotion lexicons using a graph-based algorithm and
801   multiple resources. In *Proceedings of the 23rd international conference on computational linguistics*,
802   pages 1209–1217. Association for Computational Linguistics.

803 Xu, T., Peng, Q., and Cheng, Y. (2012). Identifying the semantic orientation of terms using s-hal for
804   sentiment analysis. *Knowledge-Based Systems*, 35:279–289.
805 Yang, A. M., Lin, J. H., Zhou, Y. M., and Chen, J. (2013). Research on building a chinese sentiment
806   lexicon based on so-pmi. In *Applied Mechanics and Materials*, volume 263, pages 1688–1693. Trans
807   Tech Publ.
808 Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating bert into
809   neural machine translation. *arXiv preprint arXiv:2002.06823*.