Development of Sentiment Lexicon in Bengali utilizing Corpus and Cross-lingual Resources

Salim Sazzed
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
ssazz001@odu.edu

***Abstract*—Bengali, one of the most spoken languages, lacks tools and resources for sentiment analysis. To date, the Bengali language does not have any sentiment lexicon of its own; only the translated versions of English lexica are available. Therefore, in this work, we focus on developing a Bengali sentiment lexicon from a large Bengali review corpus utilizing a cross-lingual approach. To build the sentiment dictionary, we first created a Bengali corpus of around 42000 drama reviews; among them, we manually annotated around 12000 reviews. Utilizing a machine translation system, labeled and unlabeled Bengali review corpus, English sentiment lexica, pointwise mutual information (PMI), and supervised machine learning (ML) classifiers in different phases, we develop a Bengali sentiment lexicon of around 1000 sentiment words. We compare the coverage of our lexicon with the translated English lexica in two evaluation datasets. The proposed lexicon achieves 70%-74% coverage in document-level and around 65% coverage in word-level, which is approximately 30%-100% improvement over the translated lexica in word-level and 30%-50% in document-level. The results demonstrate that our developed lexicon is highly effective in recognizing sentiments in the Bengali text.**

***Index Terms*—Bangla sentiment lexicon, Bangla sentiment analysis, cross-lingual sentiment analysis.**

I. Introduction

Sentiment analysis identifies emotions, attitudes, and opinions expressed in the text. Due to the popularity of e-commerce and the availability of a large amount of review data, sentiment analysis has received significant attention in recent years. A popular way of automatically identifying the semantic orientation of the text is to consider word-level polarity and apply 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. A sentiment lexicon is a linguistic resource that contains opinion conveying terms such as words or phrases. These terms are usually labeled with their sentiment polarity (i.e., *positive or negative*) and strength of the polarity.

Sentiment analysis has been studied extensively in English. Therefore, many general-purpose and domain-specific lexica are available in English. Some of the popular general-purpose sentiment lexica are MPQI [1], Bing Liu's opinion lexicon [2], SentiwordNet [3], VADER [4], etc. Besides English, other widely used languages such as Chinese, Arabic, Spanish, etc have their sentiment lexica [5]–[7].

Manual construction of sentiment lexicon requires resources that are not available in resource-constrained languages. Therefore, researchers utilized various translation techniques and WordNet synset mapping [7]–[9] for leveraging resources from English. They assumed that affective norms for sentiment words are stable across languages. However, the simple translation-based approach is not adequate when dealing with noisy content in various languages. Moreover, the linguistic differences between source and target language, and possible variations of polarity across different contexts could affect the simple translation-based approaches.

Although Bengali is the seventh most spoken language in the world, sentiment analysis research in Bengali is still in its beginning. Limited research has been conducted on sentiment analysis in Bengali in the last two decades. However, still, it lacks the two most essential resources for the lexicon-based sentiment analysis: sentiment lexicon and part-of-speech (POS) tagger. Hence, in most of the works, researchers utilized supervised machine learning (ML) techniques [10]–[14], as they do not require language-specific resources.

There have been a few attempts to develop sentiment lexicon for Bengali by translating various English sentiment dictionaries. In [8], the authors utilized a word-level lexical-transfer technique and an English-Bengali dictionary to develop SentiWordNet for Bengali from English SentiWordNet. In [15], the authors translated the VADER sentiment lexicon to Bengali for sentiment analysis. Unfortunately, dictionary-based translation can not capture the informal language people use in casual communication or social media.

In this paper, we present a corpus-based cross-lingual methodology for building a sentiment lexicon in Bengali. To construct the corpus, we collected around 42000 Bengali drama reviews from Youtube; among them, we manually annotated 12000 reviews. Our proposed methodology consists of three phases, where each phase identifies sentiment words from the corpus and includes them to the Bengali sentiment lexicon. In phase 1, we identify sentiment words from the Bengali review corpus (both labeled and unlabeled) with the help of two English sentiment lexica, Bing Liu's opinion lexicon and VADER. In phase 2, utilizing 12000 annotated reviews and PMI, we identify top-class relevant (*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 show the effectiveness of our developed lexicon in two evaluation datasets.

A. Objective and Contribution

The purpose of this work is to develop a sentiment lexicon for Bengali. 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.

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

- We introduce a Bengali sentiment lexicon of around 1000 opinion words created from a Bengali corpus. To best of our knowledge, it is the first initiative to develop a corpus-based sentiment dictionary in Bengali.
- We collected and processed a large Bengali review corpus of around 42000 drama reviews from Youtube. We manually annotated around 12000 reviews to create a sentiment lexicon.
- We show how the labeled and unlabeled reviews, machine translation system, English sentiment lexica can be utilized to build a sentiment lexicon in Bengali.
- We make the review corpus and the developed lexicon publicly available for the researchers to perform sentiment analysis in Bengali.

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

II. Relevant Work

According to Liu [16], the sentiment lexicon generation techniques can be categorized into three approaches, manual approach, dictionary-based approach, and corpus-based approach. The manual approach requires annotations by humans; therefore, considerable time and resources are needed. In the dictionary-based methods, a set of seed words are created manually and then expanded using a dictionary. The corpus-based techniques use both manually labeled seed words and available corpus data.

A. Corpus-based lexicon generation in English

[17] proposed an automatic strategy for generating domain-specific sentiment lexicon based on constrained label propagation. 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 demonstrated that constrained label propagation improved the performance of the automatic construction of domain-specific sentiment lexicon.

[18] presented an automatic approach to building a target-specific sentiment lexicon. Their lexicon consists of opinion pairs made from an opinion target and an opinion word. They proposed an unsupervised algorithm to extract high-quality opinion pairs and utilized general-purpose sentiment lexicon and contextual knowledge to calculate sentiment scores of opinion pairs. Experimental results on product review datasets showed their lexicon performed better than several general-purpose sentiment lexica.

[19] proposed a domain-specific lexicon generation method from the unlabeled corpus based on mutual information and part-of-speech (POS) tags. They achieved good performance on publicly available datasets using their lexicon.

[20] proposed a graph-based label propagation algorithm to generate a domain-specific sentiment lexicon. They considered the 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 performed experiments on the Twitter dataset and found better performance than baseline approaches and general-purpose sentiment dictionaries.

[21] 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 state-of-the-art performance in both supervised and unsupervised sentiment classification tasks.

[22] constructed a domain-sensitive historical 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 (e.g., WordNet).

B. Lexicon generation in Bengali and other languages

[23] 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 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.

In [7], 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.

[6] authors introduced several large sentiment lexica 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 usefulness of existing and their proposed sentiment lexica in sentence-level sentiment analysis.

[24] 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.

[8] 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.

In [15], 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 directly applied to Bengali sentiment analysis.

Compared to the existing Bengali sentiment lexica, our constructed sentiment lexicon differs in the way it is generated and the nature of the content. As we utilize a review corpus collected from social media, it is more informal in content. Therefore, it is capable of capturing sentiments expressed in social media.

III. Corpus creation and annotation

A. Dataset for Developing Sentiment Lexicon

We constructed a large review corpus consisting of around 42000 Bengali reviews collected from Youtube. Each review in the corpus represents viewer opinions towards a Bengali drama. Utilizing web scraping and data parsing tools, we downloaded and extracted the viewer's comments from Youtube. In the corpus, we only included comments written in Bengali. We excluded comments that were written in English, Romanized Bengali, or using code-mixing language. We utilized a language detection tool [25] to distinguish the Bengali comments.

Among the 42000 collected reviews, around 12000 reviews were labeled by two native Bengali speakers. Each review was assigned to either *positive* and *negative* class by the annotators. The disagreements in the annotations were resolved by the third rater.

B. Evaluation Dataset

To show the efficacy of our developed sentiment lexicon, we utilize two datasets from varied domains. Table I provides the details of the evaluation datasets.

The first dataset is a drama review dataset consists of 2000 annotated reviews. We collected the reviews from Youtube and labeled them manually. This is a class-balanced dataset, consists of 1000 *positive* and 1000 *negative* reviews. Besides the document-level class assignment, we annotated opinion words present in the reviews, as it is required to identify word-level coverage.

TABLE I
Description of Evaluation Datasets

Dataset	Domain	Positive	Negative	Total
Drama	Drama Review	1000	1000	2000
News	News Comments	2000	2000	4000

The other dataset is a News dataset that was collected from [26]. This dataset consists of 4000 news comments; among them, 2000 are *positive* and 2000 are *negative* comments. We manually annotated the polar words of each comment in this dataset.

IV. Cross-Lingual Approach

The construction of language-specific sentiment lexicon requires vast resources 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.

The cross-lingual approach leverages resources and tools from a resource-rich language such as 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. In this work, we utilize machine translation to leverage several resources from English.

A. 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).

Machine translation has been successfully applied to various sentiment analysis tasks by researchers. [27] 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. They claimed that with appropriate ML algorithms and carefully chosen features, machine translation could be used to build sentiment analysis systems in resource-poor languages. [10] utilized google translate to convert Bengali reviews to English and compared the performance of supervised ML classifiers in Bengali and translated English corpus. They found similar accuracy on both corpora. Their study implied that though Bengali to English machine translation system is not perfect, it is capable of utilizing resources from English.

B. Sentiment Lexicon

To determine the subjectivity of the words extracted from the Bengali corpus, we employ a cross-lingual approach. With the help of machine translation and English sentiment lexica, we decide whether an extracted word conveys opinion. However, we do not perform the translation of English sentiment lexica to Bengali. Instead, we translate all the extracted Bengali words into English and then determine their polarities

based on the English lexica. If we find the translated word in an English lexicon, we include the corresponding Bengali word to our sentiment lexicon.

Our proposed approach supports the inclusion of informal Bengali words, which dictionary-based translation of English lexica can not perform. Moreover, this approach can generate multiple opinion words. 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 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 to the Bengali lexicon.

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

1) Bing Liu's Opinion Lexicon: Bing Liu's opinion lexicon 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.

2) VADER: VADER is a sentiment lexicon especially attuned for 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'.

C. Part-of-speech (POS) Tagging

Part-Of-Speech (POS) tagger is an important tool for sentiment analysis. A POS tagger reads the text written in a language and then assigns a POS tag to each word, such as noun, verb, adjective, etc. As adjectives, nouns, and verbs usually convey opinions, the POS tagger can help to identify opinion words. In English, some of the popular POS taggers are NLTK POS tagger [28], spaCy POS tagger [29]. In Bengali, no standard POS tagger is publicly available, thus, 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 allow us to label the POS tag of the corresponding Bengali words.

V. Methodology

The construction of the proposed sentiment lexicon involves several phases. We utilize various resources in different stages to identify opinion words from the corpus and include them into the lexicon, as shown in Fig 1.

- Phase 1: Labeled and unlabeled corpus, machine translation system, English lexica.
- Phase 2: Labeled corpus, PMI, machine translation system, English POS tagger, English lexica, Bengali lexicon (constructed in phase 1).
- Phase 3: Unlabeled corpus, ML classifiers, PMI, machine translation system, English POS tagger, English lexica, Bengali lexicon (constructed in phase 1 and phase 2).

Each phase expands the Bengali sentiment lexicon with the newly identified opinion conveying words. We manually validate the sentiment words identified in each stage. Then, we generate synonyms for the validated words which are added to the lexicon.

To generate synonyms, we employ google translate, as no standard Bengali synonym dictionary is available on the web. We translate Bengali words into multiple languages and then retranslate them to Bengali. This approach helps to produce synonyms as sentiments are expressed in different ways across the languages.

A. Phase-1: Utilizing English Sentiment Lexica

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 corpus using 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 the 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 into English. Two English sentiment lexica, Bing Liu's 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.

B. Phase 2: Lexicon Generation from Labeled Data

Phase 2 focuses on extracting sentiment words from the annotated corpus by leveraging the PMI formula. Pointwise mutual information (PMI) is a measure of association used in information theory and statistics. The PMI between two variables X and Y is calculated as,

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

When two variables X and Y are independent, the PMI between them is 0. PMI maximizes when X and Y are perfectly correlated.

From the labeled reviews, we derive the terms which are highly correlated with the class label. We exclude the words

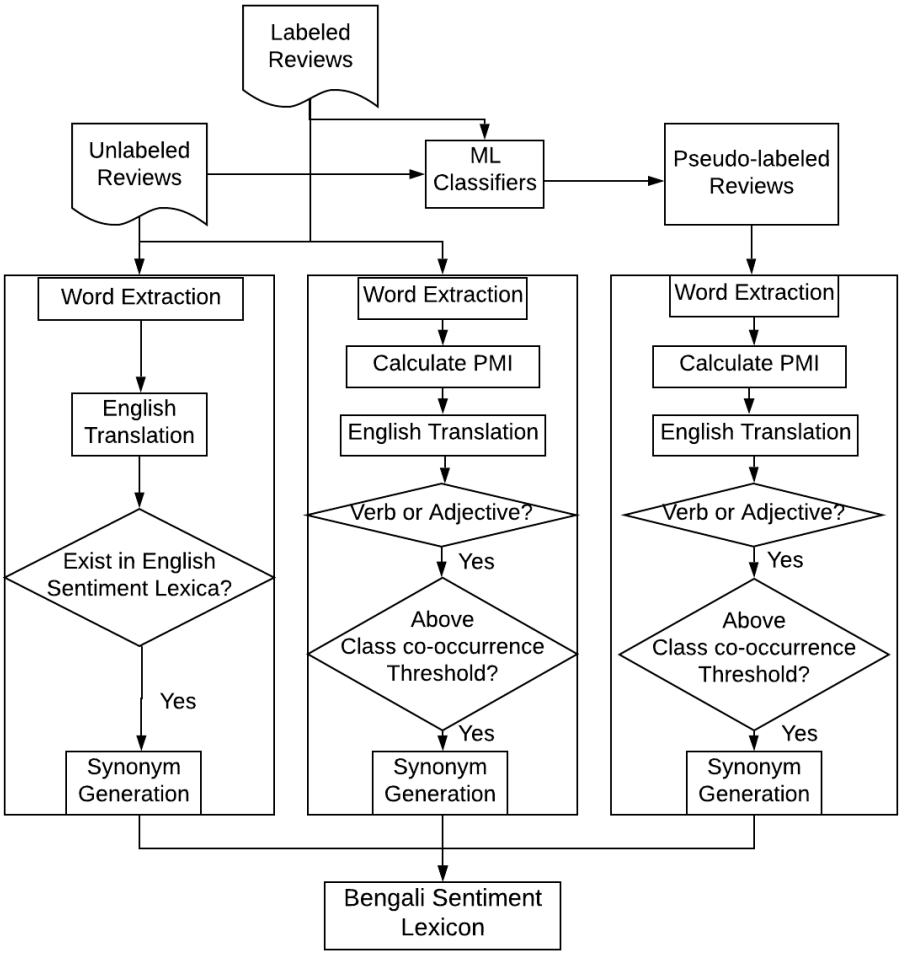

Fig. 1. The various phases of sentiment lexicon generation in Bengali

that already added to the lexicon in the earlier phase. The remaining words are translated into English using a 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. For all adjectives and verbs, we compute their PMI scores correspond to the *positive* and *negative* class.

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

$$SentimentScore(w) = PMI(w, pos) \text{ - } PMI(w, neg)$$

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

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

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.

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

C. Phase 3: Lexicon Generation from Unlabeled Data

In addition to the annotated reviews, our review corpus consists of a large number of unlabeled reviews. For the labeled reviews, we use PMI to obtain highly class-correlated words. However, for the unlabeled reviews, class information is not available; thus, we need to assign their class labels. To automatically determine the class labels of the unannotated reviews, we first identify the approach that provides high accuracy in automatic labeling.

We apply several ML classifiers to the annotated data to determine the most accurate classifiers. The following ML classifiers are employed:

1) Support Vector Machine (SVM): SVM is a supervised ML algorithm used for classification and regression problems. SVM finds the best hyperplane to separate the space into multiple classes. The hyperplane is determined by maximizing the distance between data points of different classes.

2) Stochastic Gradient Descendent (SGD): Stochastic gradient descent (SGD) 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.

3) Logistic Regression (LR): Logistic regression (LR) is a statistical method for classification. The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable and a set of independent variables.

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

5) K-Nearest Neighbors (k-NN): 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 [30] 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 performances. The purpose of this step is to find reliable classifiers that can be used for automatic class-labeling.

TABLE II
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%

Table II shows the classification accuracy of various ML classifiers using 10-fold cross-validation. Among the five classifiers we employ, SGD and SVM show very high 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.

We consider the following two ways to utilize the ML classifiers for automatic class-label assignment of the unlabeled reviews,

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 30000) reviews. For approach (2), in each iteration, we randomly select 10% reviews from the unlabeled dataset and use them as a testing dataset. Then we add these predicted reviews to the training set. This process continues until all the data are annotated.

We find that gradually expanding the training set by adding the predicted results of the testing set provides better performance. After applying approach 2, our dataset contains around 30000 pseudo-labeled reviews. We then employ PMI and POS tagger in a similar way of 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.

VI. RESULTS AND DISCUSSION

To show the effectiveness of our corpus-built lexicon, we compare it with the translated versions of two English sentiment dictionaries, VADER, and Bing Liu's opinion lexicon.

One of the common ways of attesting the effectiveness of a lexicon is to incorporate it into an existing lexicon-based sentiment analysis tool and report the classification accuracy. However, due to the unavailability of the lexicon-based sentiment analysis tool in Bengali, we can not follow this approach. Hence, we contrast the coverage of our lexicon with the translated lexica in two cross-domain evaluation datasets.

The purpose of utilizing datasets from multiple domains is to show the performance of our corpus-based lexicon in cross-domain data. Using the manual annotation as a gold standard, we calculate the coverage of the various sentiment lexica in both document and word level.

A. Document-level Coverage

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

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

B. Word-level Coverage

The word-level coverage ($WCov$) of a lexicon refers to the ratio between opinion words present in both corpus and lexicon and the total number of opinion words present in the corpus, which is then multiplied by 100. It is calculated as follows-

$$WCov = \frac{number\ of\ opinion\ words\ identified}{total\ number\ of\ opinion\ words\ in\ corpus} * 100.$$

Table III shows the comparative performances of translated lexica and our corpus-built lexicon in two datasets. In the drama review dataset, our lexicon identifies the presence of at least one opinion word in 1469 reviews out of 2000 reviews,

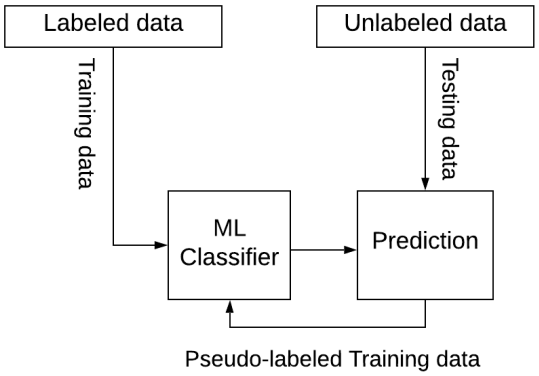

Fig. 2. Class-label assignment of unlabeled reviews using supervised ML classifier and labeled data

TABLE III
DOCUMENT-LEVEL COVERAGE OF VARIOUS LEXICA

Dataset	Lexicon	# Documents	DCov
Drama	**VADER**	871 (2000)	43.35%
	Bing-Liu	973 (2000)	48.65%
	Proposed	1469 (2000)	73.45%
News	**VADER**	2147 (4000)	53.67%
	Bing-Liu	2191 (4000)	54.77%
	Proposed	2819 (4000)	70.47%

which is around 74% coverage in the document-level. Among the two translated lexica, VADER shows 43.35% coverage, while Bing Liu's opinion lexicon provides 48.65% coverage. In the News dataset, our developed lexicon exhibits a coverage of 70.47%, while the VADER and Bing Liu's lexica provide coverage of around 54%.

TABLE IV
WORD-LEVEL COVERAGE OF VARIOUS LEXICA

Dataset	Lexicon	# Opinion Words	WCov
Drama	**VADER**	1507(4664)	32.23%
	Bing-Liu	1371(4664)	29.39%
	Proposed	3028(4664)	64.92%
News	**VADER**	4695 (8972)	52.32%
	Bing-Liu	3783 (8972)	42.16%
	Proposed	5882 (8972)	65.55%

Besides the document-level coverage, we compare the word-level coverage of various lexica in the evaluation datasets, as shown in Table IV. In the Drama review dataset, our developed lexicon identifies approximately 65% of the opinion words, while translated lexica capture only 30%-35% of them. Similarly, in the News dataset, our developed lexicon shows the coverage of 65%, while other lexica provide coverage between 42% and 52%.

The results suggest that our lexicon, although contains a much smaller number of opinion words compared to translated lexica, can detect more sentiment words present in the reviews. Since our developed lexicon comprised of words that people use in web and social media, they are more effective in recognizing sentiments compared to dictionary-based translation.

The results also reveal that even though our lexicon was built from the Drama review corpus, it performs well in the evaluation corpus from a different domain (News dataset). The two translated lexica display lower coverage in both the News and Drama review evaluation datasets compared to our developed lexicon, which infers that word-level translation from the English lexicon is not effective for identifying sentiment in Bengali text.

VII. CONCLUSIONS AND FUTURE DIRECTION

In this paper, we present a semi-automatic methodology for building a sentiment lexicon in Bengali. Leveraging various resources such as Bengali review corpus, machine-translation system, English lexica, and ML classifiers, we develop a Bengali sentiment lexicon. We demonstrate the efficacy of our sentiment lexicon in two cross-domain datasets. Our lexicon yields higher coverage compared to translated lexica in both document-level and word-level sentiment detection. We make the sentiment lexicon publicly available for the researchers in [31].

The superior performance of the developed lexicon 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. Our proposed methodology can be adapted to other resource-limited languages to reduce the time and cost required for manual annotation. In the future, we wish to improve the quality and coverage of the lexicon by utilizing larger review corpora from multiple domains.

REFERENCES

[1] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 347–354, 2005.

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, 2004.

[3] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining.," in *LREC*, vol. 6, pp. 417–422, Citeseer, 2006.

[4] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[5] G. Xu, X. Meng, and H. Wang, "Build chinese emotion lexicons using a graph-based algorithm and multiple resources," in *Proceedings of the 23rd international conference on computational linguistics*, pp. 1209–1217, Association for Computational Linguistics, 2010.

[6] S. Mohammad, M. Salameh, and S. Kiritchenko, "Sentiment lexicons for arabic social media," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 33–37, 2016.

[7] V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in spanish.," in *LREC*, vol. 12, p. 73, 2012.

[8] A. Das and S. Bandyopadhyay, "Sentiwordnet for bangla," *Knowledge Sharing Event-4: Task*, vol. 2, pp. 1–8, 2010.

[9] A. Bakliwal, P. Arora, and V. Varma, "Hindi subjective lexicon: A lexical resource for hindi polarity classification," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pp. 1189–1196, 2012.

[10] S. Sazzed and S. Jayarathna, "A sentiment classification in bengali and machine translated english corpus," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 107–114, IEEE, 2019.

[11] A. Das and S. Bandyopadhyay, "Phrase-level polarity identification for bangla," *Int. J. Comput. Linguist. Appl.(IJCLA)*, vol. 1, no. 1-2, pp. 169–182, 2010.

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

[13] K. Sarkar and M. Bhowmick, "Sentiment polarity detection in bengali tweets using multinomial naïve bayes and support vector machines," in *2017 IEEE Calcutta Conference (CALCON)*, pp. 31–36, IEEE, 2017.

[14] M. Rahman, E. Kumar Dey, *et al.*, "Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation," *Data*, vol. 3, no. 2, p. 15, 2018.

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

[16] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[17] S. Huang, Z. Niu, and C. Shi, "Automatic construction of domain-specific sentiment lexicon based on constrained label propagation," *Knowledge-Based Systems*, vol. 56, pp. 191–200, 2014.

[18] S. Wu, F. Wu, Y. Chang, C. Wu, and Y. Huang, "Automatic construction of target-specific sentiment lexicon," *Expert Systems with Applications*, vol. 116, pp. 285–298, 2019.

[19] H. Han, J. Zhang, J. Yang, Y. Shen, and Y. Zhang, "Generate domain-specific sentiment lexicon for review sentiment analysis," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 21265–21280, 2018.

[20] Y.-J. Tai and H.-Y. Kao, "Automatic domain-specific sentiment lexicon generation with label propagation," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, pp. 53–62, 2013.

[21] L. Wang and R. Xia, "Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 502–510, 2017.

[22] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016, p. 595, NIH Public Access, 2016.

[23] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, "Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis," *Journal of information science*, vol. 44, no. 3, pp. 345–362, 2018.

[24] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, "Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language," *Expert Systems*, vol. 36, no. 3, p. e12397, 2019. e12397 EXSY-Apr-18-123.R2.

[25] N. Shuyo, "Language detection library for java," 2010.

[26] "socian-bangla-sentiment-dataset-labeled." https://github.com/socian-ai/socian-bangla-sentiment-dataset-labeled. Accessed: 2020-04-30.

[27] A. Balahur and M. Turchi, "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis," *Computer Speech & Language*, vol. 28, no. 1, pp. 56–75, 2014.

[28] E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[29] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, 2017.

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

[31] "Bengali lexicon." https://github.com/sazzadcsedu/BNLexicon.git. Accessed: 2020-04-30.