2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)

A Sentiment Classification in Bengali and Machine Translated English Corpus

Salim Sazzed and Sampath Jayarathna
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
(ssazz001@odu.edu, sampath@cs.odu.edu)

***Abstract*—The resource constraints in many languages have made the multi-lingual sentiment analysis approach a viable alternative for sentiment classification. Although a good amount of research has been conducted using a multi-lingual approach in languages like Chinese, Italian, Romanian, etc. very limited research has been done in Bengali. This paper presents a bilingual approach to sentiment analysis by comparing machine translated Bengali corpus to its original form. We apply multiple machine learning algorithms: Logistic Regression (LR), Ridge Regression (RR), Support Vector Machine (SVM), Random Forest (RF), Extra Randomized Trees (ET) and Long Short-Term Memory (LSTM) to a collection of Bengali corpus and corresponding machine translated English version. The results suggest that using machine translation improves classifiers performance in both datasets. Moreover, the results show that the unigram model performs better than higher-order n-gram model in both datasets due to linguistic variations and presence of misspelled words results from complex typing system of Bengali language; sparseness and noise in the machine translated data, and because of small datasets.**

***Keywords*-sentiment classification; machine translation; bilingual corpus;**

I. Introduction

Sentiment analysis [1] which is also known as opinion mining, is a contextual text mining process that extracts opinions, sentiments, attitudes, emotions, etc. from the textual data and classifies them based on their polarities. The text data can be retrieved from any sources such as product review websites, social media, blogs, and customer satisfaction survey. Through sentiment analysis, the underlying sentiment of the data can be classified into various categories such as binary (e.g., positive-negative) or multi-modal (e.g., positive-negative-neutral) or fine-grained sentiment (e.g., "very positive" and "very negative").

To classify the sentiment or polarity in a text, two broad categories of methods are available: (1) machine learning or statistical-based approach, and (2) unsupervised lexicon-based approach. The classical machine learning algorithms such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), etc. have been utilized extensively by the researchers for sentiment classification [2]. These machine learning algorithms use some form of supervised learning to train the classifiers which require labeling of training observations. The lexicon based approaches such as [3], [4], [5], determine the sentiment or polarity using some functions of opinion words in the document or sentence. The performance of a lexicon based approach depends on available language specific resources such as sentiment lexicons, Parts-of-speech (POS) tagger, modifiers, dependency parser, and context of the sentences. Both approaches typically use extracted unigrams (i.e., single word) or bigrams (i.e., consecutive word pairs) from the corpus as an input. There also exist hybrid methods, [6], [7] which are a combination of both approaches- based on labeled data and lexicons, optionally with unlabelled data.

Cross-lingual sentiment classification has gained popularity in the last decade. Especially, in resource-poor languages, due to the lack of available data and text analysis tools, cross-lingual sentimental analysis can play an important role. The ongoing improvement of machine translation makes the cross-lingual approach a viable option for sentiment analysis in resource-poor languages. Cross-lingual sentiment classification aims to leverage resources like labeled data, polarity lexicons, contextual valence shifters, modifiers, etc. from resource-rich language (such as English) to classify the sentiment polarity of texts in a low-resource language (such as Bengali). One of the biggest challenges for cross-lingual sentiment classification is the lexicon mapping between the source language and the target language.

This paper presents a comparative study of bilingual sentiment classification in Bengali and machine translated English corpora. We compare various machine learning algorithms performances in Bengali language and its machine translated English version. We utilize two Bengali datasets from distinct domains, a publicly available user comments dataset [8] of a popular sub-continent game and a drama review dataset mined from YouTube. We perform sentiment analysis at the document level on both datasets.

Our main contributions are to determine the applicability and performance of bilingual sentiment classification in Bengali language and a new annotated Bengali drama review dataset that we plan to make publicly available for other researchers. Using, (a) distinct feature sets (i.e., unigram, bigram), (b) multiple machine learning algorithms, and (c) two datasets from independent domains, we show that Bilingual approach can play a significant role in sentiment analysis.

978-1-7281-1337-1/19/$31.00 ©2019 IEEE
DOI 10.1109/IRI.2019.00029

II. Related Works

Sentiment analysis can be considered as a sub-field of information extraction, the research area within information and computer science that aims to summarize and draw inferences from collections of textual documents [9], [10]. Sentiment analysis started drawing the attention of computational linguistics communities only in the early 2000s [11]. Researchers performed sentiment analysis in various tasks such as election prediction [12], stock market prediction [13], opinion polling [14], customer feedback tracking [15] and at different levels of granularity such as document level [16], sentence level [4], phrase level [17] and aspect level [3].

Although most of the research in subjectivity and sentiment analysis has been done for English, in recent years, sentiment-labeled data is becoming available for other languages. In Bengali, limited research has been done using publicly available Bengali corpora collected from various sources such as Microblogs, Facebook status, movie review websites, and other social-media sources. A limited number of classification methods have been applied for Bengali sentiment analysis such as SVM with maximum entropy [18], Naive Bayes (NB) [19][20], Multinomial Naive Bayes (MNB) with mutual information [21], Deep Neural Network [12] . In [22], authors used word2vec and polarity score based approach which gave 76% accuracy in two-class prediction. In [23], authors presented lexicon based approach for binary prediction. A word embedding based approach with Hellinger PCA was proposed by [24]. In [25], authors compared the performance of five machine learning approaches in Horoscope dataset, [26] performed sentiment analysis on Bengali and Romanized Bengali text using Long Short-Term Memory (LSTM) and achieved 70% accuracy for two-class prediction.

A number of studies [27] have been performed considering cross-lingual approaches which can be broadly classified into two main categories: (1) those that utilize parallel corpora to train bilingual word embeddings (BWE) [28], [29], and (2) methods that use bilingual lexicons [30], and machine translation (MT) systems [31] in order to learn features which work on both languages.

In [32], authors explored cross-lingual projections to generate subjectivity analysis resources in Romanian by leveraging on the tools and resources available in English. They have investigated two approaches: a lexicon-based approach based on Romanian subjectivity lexicon translated from the English lexicon, and a corpus-based approach based on Romanian subjectivity-annotated corpora obtained via cross-lingual projections. In [33], authors applied a bilingual system to improve the performance of chines sentiment analysis leveraging resources from English. To examine the polarity, authors focused on unsupervised sentiment polarity identification and only investigated the lexicon-based approach in their experiments including positive and negative lexicons to reverse the semantic polarity of specific terms; intensifier lexicon to determine the degree of the terms polarity. The results indicated that by applying the ensemble approach, classifier performance was improved by around 5%.

In [34], authors studied the possibility to employ machine translation systems and supervised methods for multilingual sentiment analysis. They used four languages English, German, Spanish, and French; three machine translation systems Google, Bing, and Moses; different supervised learning algorithms and various types of features and employed meta-classifiers to mitigate the noise introduced by the translation. Their extensive evaluations showed that machine translation systems could be used for multilingual sentiment analysis. In [35], authors proposed a bilingual approach for conducting social media sentiment analysis. Instead of processing English and Chinese comments separately, they considered review comments as a stream of text containing both Chinese and English words segmented and trimmed with the text stream using segment model and by the stop word lists. The stem words are then processed into feature vectors and applied with two exchangeable natural language models, SVM and N-Gram. In [36], authors proposed a cross-lingual mixture model (CLMM) to leverage unlabeled bilingual parallel data. From the bilingual parallel data, their proposed model learned previously unseen sentiment words and improved vocabulary coverage significantly. In [37], authors leveraged the resources available in English by employing machine translation to generate resources for subjectivity analysis in other languages (i.e., Romanian and Spanish) and showed a comparative evaluation. In [27], authors performed sentiment analysis utilizing English sentiment knowledge in Spanish and Chinese language with a translation matrix from one language to another and utilized binary sentiment word list from English. For learning, they collected 10,000 English words by scraping the most commonly used words in Google's "Trillion Word Corpus".

III. Methodology

The framework of our approach is illustrated in Figure 1, for the first dataset of sports (Cricket) comments. In the first step, Bengali comments from Cricket dataset are translated into the corresponding English comments using Google machine translation service. We split the data into a training set (80%) and test set (20%) using sci-kit learn stratified sampling method.

As Cricket dataset is highly imbalanced, we make a class-balanced version of training data using Synthetic Minority Over-sampling Technique (SMOTE) [38] algorithm. We tokenize, vectorize data, and perform sentiment classification using machine learning classifiers. Bengali/English sentiment classification performance is compared in both original, and class-balanced datasets. For the Drama dataset,

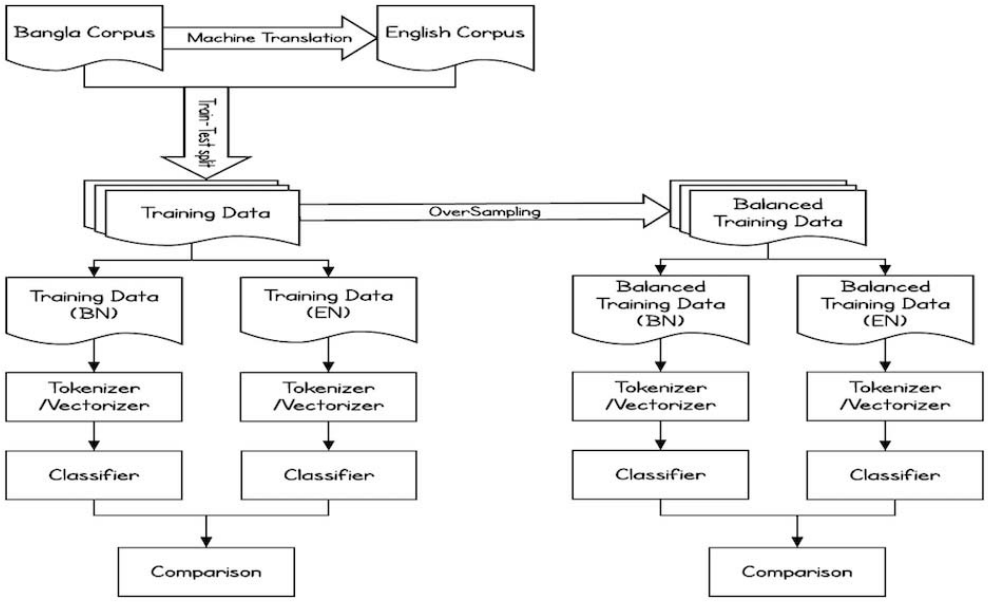

Figure 1. Overall Architecture of the Proposed Processing Pipeline

which is already class balanced, similar steps are followed except only the original dataset is used, and due to small dataset size, 10-fold cross-validation is used instead of 80%/20% split.

A. Datasets

We use two Bengali datasets from different domains-sports comments (Cricket) dataset and Drama review dataset. The Cricket (a popular game in Asian Sub-continent) dataset contains supporters comments that convey their thoughts, attitudes, and opinions towards Bangladesh National Cricket team. The comments were collected and manually annotated by [8], and publicly available. The dataset consists of 2489 annotated comments where each comment contains approximately 3-100 Bengali words. This highly imbalanced dataset consists of 1772 Negative comments, 494 Positive comments, and 223 Neutral comments (see Figure 2. for a sample content of the dataset and annotated polarity values).

The second dataset (Drama) consists of viewers opinions towards eight Bengali dramas. We use the website http://ytcomments.klostermann.ca to scrap opinion data from the YouTube links. The data is in JSON format, contains information like user name, id, timestamp, comments, and likes. The JSON data is parsed to extract user comments only.

We use langdetect library to distinguish Bengali and English comments. Since we are interested only in Bengali comments, we keep the comments written in Bengali. After removing the English comments, we obtain 1016 Bengali reviews from eight dramas. These reviews are labeled by a human annotator as Positive, Negative, or Neutral. This dataset contains 338 Negative reviews, 206 Neutral reviews, 472 Positive reviews.

Table I
CLASS DISTRIBUTIONS IN CRICKET AND DRAMA DATASET

	Negative	Neutral	Positive	Total
Cricket	1772	223	494	2489
Drama	338	206	472	1016

Due to the class imbalance problem (comprised of mostly negative comments) in the Cricket dataset, two variants of the dataset are used; original class imbalanced dataset and adjusted class balanced dataset. Class balancing is performed for both the Bengali and translated English corpora using imbalanced-learn package of [39] SMOTE implementation. As the Drama review dataset is already class balanced, we do not apply re-sampling for this dataset.

B. Pre-processing Pipeline

To convert the Bengali corpora to English, we utilize machine translation provided by Google Translate. We do not apply any correction filters mainly because the purpose of our study is to compare sentiment classification accuracy of Bengali corpus with the machine translated corpus.

We examine the quality of machine translations in Drama review dataset to determine whether the quality of machine translations influences classifier performances in machine translated English corpora. We asked an expert Bengali reviewer to categorize the quality of every translated English

Bengali	English Translation	Polarity	Trans. Quality
মুমিনুল এর মতো সুন্দর ও একুরেট খেলার এবিলিটি তামিম এর নেই ।	Tamim's absence of beautiful and erect sports like Mominul does not exist.	Negative	2
জরিমানা করা হউক। ৩ মাসের বেতন কর্তন।	Fine. 3 months salary cut	Negative	4
টেষ্ট ক্রিকেটে রান আউট খুবই দুঃখজনক।	Test run out in the game is very sad.	Neutral	4
অসাধারণ এক ধরনের ছবি	Great Kind of Pictures	Positive	3
এই পিচে মুশফিকুর ছাড়া কেউ খেলতে খুব একটা স্বাছন্দবোধ করবে না	No one will feel very comfortable playing Mushfiqur on this pitch	Positive	4
অসাধারণ নাটক	Great Drama	Positive	5
তামিম ত টেস্টে খেলা জানে না	Tamim does not know how to play in Tests	Negative	5
সেলুট হানিফ সংকেত তোমাকে	Salt Hanif Signal To You	Positive	1
জামাই আদর খুব ভালো হয়ছে	Adult is very good	Positive	1
এই নাটকে একটা গান একটু দাড়াও এক মুঠো রৌদ এই গানটা কোন শিল্পীর?	In this play, a song will be a little Roud, this song is an artist's song ????	Neutral	2

Figure 2. Sample Bengali Dataset and Corresponding English Translation with Annotated Polarity

comment into one of the five categories in Likert scale; 1 (Poor, does not represent the Bengali comment at all), 2 (Not Very Good, few words are translated correctly, semantically translation does not make sense), 3 (Fair, partially correct based on semantic and lexicon), 4 (Good, represents corresponding Bengali comment well enough, semantically similar), 5 (Excellent, conveys same meaning as of the Bengali comments). Out of 1016 comments, quality labels after the review include 170 Poor, 279 Not Very Good, 229 Fair, 140 Good and 198 Excellent, with average translation score of 2.92 for the Drama dataset. From the expert ratings, it is evident that machine translation is not always accurate, and consists of translations that contain lexical, synthetic or semantic errors. The presences of misspelled words and differences in regional words make many word-to-word translations inaccurate. Linguistic complexity of Bengali language and machine translations inability to relate the word to the context make it difficult to align the semantic meaning in many cases.

Figure 2, shows some examples of Bengali comments from Drama dataset with the corresponding machine translated English comments, sentiment orientations, and translation quality scores. For example, one of the Bengali comments is translated to English as 'Test run out in a game is very sad', while the accurate translation would be- 'In Test Cricket run out is very unfortunate'. Although this translation is not entirely correct, it conveys similar meaning; therefore, scores rating 4 from the expert reviewer. Another machine translation is 'No one will feel very comfortable playing Mushfiqur on this pitch' where the accurate translation should be- 'No one except Mushfiqur will feel very comfortable playing on this pitch'. Even though this translation is nearly equivalent at word level, machine translation has changed the semantic meaning; therefore, get translation score of 3. An example of poor machine translation with translation score 1 is given to the comment- 'Salt Hanif Signal To You' where the correct translation should be- 'Hanif Shongket, salute to you'. As the English word 'Salute' is written in Bengali (also misspelled), machine translation cannot translate it. Besides, the person last name (shongket) in the comment is a Bengali noun which machine translation erroneously converted to English instead of recognizing it as a name.

In the pre-processing step, both the Bengali and English corpora are stemmed and tokenized using scikit-learn [40] machine learning library. The tokenized words are then converted to sparse term frequency-inverse document frequency (tf-idf) vector representation.

For deep learning model LSTM, we use Keras [41] built-in tokenizer and Embedding layer that represents words and comments using a dense vector representation.

C. Classification

To avoid bias in comparison due to the scarcity of lexical text analysis tools in Bengali which are easily available in English, we apply machine learning based classifiers to evaluate the performance of Bengali and machine translated English corpus. Moreover, machine learning based approaches can learn hidden patterns from the training data and generally more robust against noisy data compared to rigid rules; therefore, more suitable for our noisy machine translated English corpus. We employ six machine learning techniques to compare sentiment classification between Bengali and translated English corpus; Logistic Regression (LR), Ridge Regression (RR), Support Vector Machine (SVM), Random Forest (RF), Extremely Randomized Trees (ET) and recurrent neural network (RNN) based deep learning architecture Long Short-Term Memory (LSTM).

LR is a classification method that assumes there are one or more independent variables that determine an outcome. Although LR is mainly a binary classifier, it can be modified to handle multi-class problems using one-vs-rest logistic

```
function ComputeResults (benCorp: inputData) :
    /* Machine Translation                    */
    engCorp = machine-translation(benCorp);
    /* Class Balancing                        */
    benCorpBal = class-balancing(bengaliCorp);
    engCorBal = class-balancing(engCorp);
    /* Test-train split                       */
    bengCorpTrain = train-test-split(benCorp);
    engCorpTrain = train-test-split(engCorp);
    bengCorpBalTrain =
      train-test-split(benCorpBal);
    engCorpBalTrain = train-test-split(engCorBal);
    /* Vectorization                          */
    bengCorpTrain = vectorize(bengCorpTrain);
    engCorpTrain = vectorize(engCorpTrain);
    bengCorpBalTrain =
      vectorize(bengCorpBalTrain);
    engCorpBalTrain =
      vectorize(engCorpBalTrain);
    /* All algorithms                         */
    algorithms =
      ['LR','RR','SVM','RF','ET','LSTM'];
    /* All training dataset                   */
    allTrainData = (bengCorpTrain,
      engCorpTrain, bengCorpBalTrain,
      engCorpBalTrain);
    results = [][]
    /* Apply machine learning algorithm
                                              */
    forall trainData in allTrainData do
        forall algo in algorithms do
            results [trainData][algo] =
              get-result-using-algorithm(trainData,
              algo)
        end
    end
    return results;
end
```

Algorithm 1: Data Wrangling Pipeline

regression (OVR) or multinomial logistic regression.

RR is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, even though least squares estimates are unbiased, their variances can be far from the true value. To address this issue, a bias is added to the regression estimates.

SVM is a discriminative classifier defined by a separating hyperplane. Given the labeled training data, the algorithm outputs an optimal hyperplane which categorizes unseen observations. For SVM, we use 'linear' kernels as it performs better in our datasets compared to non-linear kernels such as Radial Basis Function (RBF) or Polynomial.

RF is a decision tree based classifier, usually trained by recursively splitting the data. It is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class having the highest voting.

ET is a variant of a random forest that uses the entire sample at each step, and decision boundaries are picked at random, rather than the best one. ExtraTrees classifier is typically faster and performs better compared to random forest in presence of noisy data.

LSTM is a type of recurrent neural network for text analysis which can model sequence dependent behavior. As they are designed for persistent memory, they can detect long-term dependencies. We use Keras deep learning framework to employ LSTM in our sentiment analysis problem. For the optimizer, we apply Stochastic gradient descent (SGD), for activation (Rectified Linear Units) ReLU and Softmax, and categorical-cross-entropy is used as a loss function. We use batch size of 64, learning rate of 0.01 with momentum and decay. For regularization, a drop-out value of 0.5 is applied. We train the classifier for 100 epochs for Cricket dataset and 150 epochs for Drama review dataset.

Algorithm 1 provides the complete pseudo-code for our data wrangling pipeline.

Table II
COMPARISON OF EVALUATION METRICS IN CRICKET DATASET USING BENGALI AND TRANSLATED ENGLISH CORPORA

Method	P(B/E)	R(B/E)	F1 (B/E)	Accuracy (B/E)
LR	0.37/0.42	0.34/0.34	0.35/0.37	70.9/72.2
RR	0.41/0.43	0.36/0.38	0.37/0.40	69.6/72.7
SVM	0.37/0.42	0.34/0.37	0.35/0.39	70.1/73.2
RF	0.39/0.42	0.36/0.37	0.37/0.39	69.1/72.1
ET	0.38/0.46	0.36/0.38	0.37/0.41	68.3/72.6
LSTM	0.37/0.41	0.44/0.37	0.40/0.39	59.0/72.2

B= Bengali E = English

IV. EXPERIMENTAL RESULTS

To compare the performance of various classifiers, we utilize several standard performance assessment measures. We compare classifiers relative performance based on accuracy, precision, recall, and F1 score. We consider accuracy to compare the predicted label of every instance with the ground-truth label. Accuracy does not always provide the complete picture of the imbalanced dataset, as a large number of samples forms a bias towards majority classes. Hence, precision, recall, and F1 measure are utilized.

Using imbalanced Cricket dataset, we evaluate and compare the performance of sentiment classifiers in Bengali and corresponding machine-translated English corpus which are shown in Table II. Considering accuracy, SVM is the top performing classifier for the English corpus with an accuracy of 73.2%. In Bengali corpus, we notice the highest accuracy from the LR classifier, which is 70.9%. Based on F1 score, ET performs best in English corpus, which is 0.417. For Bengali corpus, LSTM classifier shows highest F1 score

Table III
COMPARISON OF EVALUATION METRICES IN BENGALI AND TRANSLATED ENGLISH CORPUS CRICKET DATASET(CLASS BALANCED)

Method	P(B/E)	R(B/E)	F1(B/E)	Accuracy(B/E)
LR	0.38/0.42	0.40/0.44	0.39/0.43	54.2/65.8
RR	0.38/0.43	0.39/0.44	0.38/0.43	54.4/66.6
SVM	0.38/0.42	0.39/0.43	0.39/0.43	53.3/67.2
RF	0.38/0.45	0.38/0.40	0.38/0.43	59.8/71.6
ET	0.39/0.43	0.41/0.40	0.40/0.41	59.9/71.3
LSTM	0.42/0.41	0.40/0.44	0.41/0.43	45.5/47.9

B= Bengali E = English

Table IV
COMPARISON OF EVALUATION METRICS IN BENGALI AND TRANSLATED ENGLISH CORPUS IN DRAMA REVIEW DATASET

Method	P(B/E)	R(B/E)	F1(B/E)	Accuracy(B/E)
LR	0.56/0.65	0.54/0.57	0.55/0.61	64.2/68.2
RR	0.57/0.63	0.56/0.62	0.56/0.62	65.0/70.0
SVM	0.58/0.65	0.57/0.62	0.57/0.64	65.4/70.1
RF	0.57/0.58	0.55/0.54	0.56/0.56	63.0/61.7
ET	0.55/0.58	0.54/0.55	0.54/0.56	63.0/64.6
LSTM	0.60/0.61	0.57/0.60	0.59/0.60	58.2/64.2

B= Bengali E = English

0.407 while other classifiers LR, SVM, RF, ET provide similar performance.

We also present the class-balanced version of Cricket dataset to asses the performances of the classifiers. We make the Cricket dataset class-balanced for two reasons: (a) to examine how it improves the prediction of minority classes, and (b) to verify whether it shows similar performance improvement in English corpus as of the original imbalanced dataset and support our findings. Table III shows that class-balancing using SMOTE improves F1 scores across all classifiers.

In Table IV, the comparison between Bengali and machine translated English corpora of Drama review dataset is shown. Since the Drama dataset is nearly balanced, re-sampling is not employed before comparison. In both Bengali and translated English corpora the highest accuracy is obtained by applying SVM classifier, which is 65.4% and 70.1% respectively. SVM also provides highest F1 score for translated English corpus, which is 0.641. For Bengali corpus, LSTM classifier provides the best F1 score, which is 0.593. If we compare the F1 scores of different classifiers in Bengali and English corpus, we can see improvement of classifiers performance in English corpus in all the cases. The highest performance gain from 0.577 to 0.641 is achieved in best performing SVM classifier, which is 11% improvement from the Bengali dataset.

We also compare the classifiers relative performance using different n-grams in a class-balanced subset of Cricket dataset and Drama datasets. The class-balanced subset of Cricket dataset contains all the Neutral (223) and Positive (494) comments from the original Cricket dataset along with 500 randomly selected Negative comments; in total, a set of 1017 comments. For both datasets, we use the unigram model (i.e., single word) that assumes independence among the words. To see whether the adjacent word-pairs can capture better semantic relationship in both Bengali and machine translated English corpora, thus improves the performance of the classifiers, we apply the bigram model. Table V presents the accuracy comparison between unigram and bigram models in both datasets. The results show that for the Drama dataset, utilizing the bigram model decreases performances for all the classifiers. When bigram is used, the accuracy of the LR classifier drops by 20% and 15% in English corpus and Bengali corpus respectively. For the class-balanced Cricket dataset when using bigram, classifiers accuracy fall by 10%-20% in both Bengali and English corpora.

Table V
COMPARISON OF ACCURACY USING UNIGRAM AND BIGRAM MODELS IN CRICKET (BALANCED SUBSET) AND DRAMA DATASET

	Cricket		Drama	
Method	$(U/B)^{BN}$	$(U/B)^{EN}$	$(U/B)^{BN}$	$(U/B)^{EN}$
LR	48.2/35.6	49.4/39.7	64.2/56.1	68.2/55.4
RR	48.5/39.7	50.1/45.3	65.0/57.7	70.0/58.5
SVM	48.6/38.4	51.6/44.9	65.4/58.4	70.1/58.2
RF	45.8/42.2	42.1/36.1	63.0/57.8	61.7/54.8
ET	47.3/41.6	45.7/36.3	63.0/57.6	64.6/55.1
	U= Unigram	B = Bigram	BN=Bengali	EN =English

Table VI
THE CONFUSION MATRIX OF LOGISTIC REGRESSION (LR) BEFORE AND AFTER CLASS-BALANCING IN CRICKET DATASET (BENGALI)

		Predicted		
		Negative (I/B)	Neutral (I/B)	Positive (I/B)
Actual	Negative	365/227	0/61	4/81
	Neutral	32/16	0/12	4/8
	Positive	86/28	0/21	7/44
		I= Imbalanced	B = Balanced	

V. DISCUSSION

In this paper, we apply multiple machine learning approaches in Bengali and translated English corpora to compare the performance. Based on the F1 score, it is evident that in most cases, machine learning algorithms show better performance in translated English corpus compared to the original Bengali corpus.

The machine learning algorithms provide higher F1 scores in Drama review dataset compared to Cricket dataset even with less amount of data. The cricket dataset contains a large number of descriptive comments, consists of many domain-specific words. Moreover, users opinions are diverse and directed towards various aspects of the game. On the other hand, the Drama review dataset contains many short

comments, and opinions follow similar patterns. Besides, This dataset is more subjective compared to Cricket dataset.

The results show class balancing using SMOTE improves the F1 scores in both the originally imbalanced Bengali and translated English Cricket dataset. Imbalanced data is a common problem in many natural language processing tasks including sentiment analysis. The low number of observations from minority classes can make feature learning difficult for a machine learning algorithm that affects the overall performance. For example, The confusion matrix in Table VI shows that LR classifier does not predict any Neutral class labels due to low numbers of Neutral samples in training data. Utilizing the balanced dataset improves the average recall scores for all the classifiers, though the average precision scores remain almost the same. Overall, the impact of class-balancing is reflected in higher F1 scores for all of the classifiers. The results also suggest that even after class-balancing, classifiers perform better in machine-translated English corpus.

The results also indicate that using higher n-gram (i.e., bigram) degrades performances of the classifiers in both datasets. The variation of linguistics dialects in Bengali, misspelling, usage of special regional words, inaccuracy in machine translation, and small datasets negatively affect the performance of bigram based model as it produces more sparse feature vectors compared to unigram based method.

The wide variety of the linguistic expressions in Bengali often diverge from the standard language that makes sentiment analysis in Bengali a challenging task. The usage of informal language and the presence of Romanized word in textual data compound the problem. Moreover, the complex writing system of Bengali makes typing difficult, which results in misspelling. Therefore, to learn language features requires lots of training data. The experimental results demonstrate that machine translation is capable of retaining sentiments in translated English comments even though the translation itself is not accurate. Therefore, even with inaccurate machine translation, classifiers perform better in English corpora. The results also suggest that to capture semantic information/linguistic patterns in Bengal language requires a large amount of labeled training data, which is obvious in the superior performance of unigram based model compared to bigram model in both datasets.

VI. Conclusions and Future Outlook

In this study, we analyze the performance of sentiment classification in Bengali and corresponding machine-translated English corpus using multiple machine learning algorithms. We evaluate models performance on two datasets from different domains. From the experimental results, it is apparent that class-balancing shows performance improvement in both the Bengali and translated English version of imbalanced Cricket dataset. The results also suggest that machine translation improves classifiers performance in both datasets.

The efficacy of sentiment analysis in Bengali is often negatively affected by the linguistic complexity, complicated writing system, inadequate labeled data, and lack of lexical tools. The results indicate that machine translation can provide better accuracy compared to the original language (Bengali) and can be used as a way of sentiment analysis for the resource-poor language like Bengali. Moreover, Our comparative results demonstrate that even though the current machine translation system is not perfect in Bengali-English translation, it can be reliably used for bilingual sentiment analysis. In our future work, we will focus on extending the bilingual sentiment analysis into the multilingual analysis.

References

[1] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD*, 2004.

[4] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[5] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: ACM, 2008, pp. 231–240.

[6] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," 2011.

[7] V. Sindhwani and P. Melville, "Document-word co-regularization for semi-supervised sentiment analysis," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 1025–1030.

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

[9] J. Cowie and W. Lehnert, "Information extraction," *Commun. ACM*, vol. 39, no. 1, pp. 80–91, Jan. 1996.

[10] S. Sarawagi, "Information extraction," *Found. Trends databases*, vol. 1, no. 3, pp. 261–377, Mar. 2008.

[11] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[12] N. Tripto and M. Eunus Ali, "Detecting multilabel sentiment and emotions from bangla youtube comments," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 09 2018, pp. 1–6.

[13] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *J. Comput. Science*, vol. 2, pp. 1–8, 2011.

[14] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. C. M. Fong, "Quantitative study of individual emotional states in social networks," *IEEE Transactions on Affective Computing*, vol. 3, pp. 132–144, 2012.

[15] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th International Conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004.

[16] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424.

[17] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354.

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

[19] M. S. Islam, M. A. Islam, M. A. Hossain, and J. J. Dey, "Supervised approach of sentimentality extraction from bengali facebook status," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, Dec 2016, pp. 383–387.

[20] N. Banik and H. H. Rahman, "Evaluation of nave bayes and support vector machines on bangla textual movie reviews," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–6.

[21] A. K. Paul and P. C. Shill, "Sentiment mining from bangla data using mutual information," in *2016 2nd International Conference on Electrical, Computer Telecommunication Engineering (ICECTE)*, Dec 2016, pp. 1–4.

[22] M. Al-Amin, M. S. Islam, and S. D. Uzzal, "Sentiment analysis of bengali comments with word2vec and sentiment information of words," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Feb 2017, pp. 186–190.

[23] S. Akter and M. T. Aziz, "Sentiment analysis on facebook group using lexicon based approach," in *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Sep. 2016, pp. 1–4.

[24] M. S. Islam, M. A. Amin, and S. D. Uzzal, "Word embedding with hellinger pca to detect the sentiment of bengali text," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, Dec 2016, pp. 363–366.

[25] T. Ghosal, S. K. Das, and S. Bhattacharjee, "Sentiment analysis on (bengali horoscope) corpus," in *2015 Annual IEEE India Conference (INDICON)*, Dec 2015, pp. 1–6.

[26] A. Hassan, N. Mohammed, and A. K. al Azad, "Sentiment analysis on bangla and romanized bangla text (BRBT) using deep recurrent models," *CoRR*, vol. abs/1610.00369, 2016.

[27] M. Abdalla and G. Hirst, "Cross-lingual sentiment analysis without (good) translation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 506–515.

[28] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 1555–1565.

[29] S. Chandar A P, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha, "An autoencoder approach to learning bilingual word representations," in *Advances in Neural Information Processing Systems 27*.

[30] A. Balamurali, A. Joshi, and P. Bhattacharyya, "Cross-lingual sentiment analysis for indian languages using linked wordnets," in *COLING*, 2012.

[31] X. Zhou, X. Wan, and J. Xiao, "Cross-lingual sentiment classification with bilingual document representation learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1403–1412.

[32] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 976–983.

[33] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis," in *EMNLP*, 2008.

[34] 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.

[35] G. Yan, W. He, J. Shen, and C. Tang, "A bilingual approach for conducting chinese and english social media sentiment analysis," *Comput. Netw.*, vol. 75, no. PB, pp. 491–503, Dec. 2014.

[36] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang, "Cross-lingual mixture model for sentiment classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 572–581.

[37] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan, "Multilingual subjectivity analysis using machine translation," in *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2008, pp. 127–135.

[38] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011.

[39] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[40] 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.

[41] F. Chollet *et al.*, "Keras," https://keras.io, 2015.