Cross-lingual Sentiment Analysis in Bengali Utilizing A New Benchmark Corpus

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

Abstract

Sentiment analysis research in low-resource languages such as Bengali is still unexplored due to the scarcity of annotated data and the lack of text processing tools. Therefore, in this work, we focus on generating resources and showing the applicability of the cross-lingual sentiment analysis approach in Bengali. For benchmarking, we created and annotated a comprehensive corpus of around 12000 Bengali reviews. To address the lack of standard text-processing tools in Bengali, we leverage resources from English utilizing machine translation. We determine the performance of supervised machine learning (ML) classifiers in machine-translated English corpus and compare it with the original Bengali corpus. Besides, we examine sentiment preservation in the machine-translated corpus utilizing Cohen's Kappa and Gwet's AC1. To circumvent the laborious data labeling process, we explore lexicon-based methods and study the applicability of utilizing cross-domain labeled data from the resource-rich language. We find that supervised ML classifiers show comparable performances in Bengali and machine-translated English corpus. By utilizing labeled data, they achieve 15%-20% higher F1 scores compared to both lexicon-based and transfer learning-based methods. Besides, we observe that machine translation does not alter the sentiment polarity of the review for most of the cases. Our experimental results demonstrate that the machine translation based cross-lingual approach can be an effective way for sentiment classification in Bengali.

1 Introduction

Sentiment analysis classifies the semantic orientation of a text. With the rapid growth of user-generated content, nowadays, it is essential to determine user opinions, attitudes, and feelings from the textual data. In literature, researchers identified sentiment orientations of the text in various levels, such as document, sentence, or aspect. Researchers employed both the machine learning-based and lexicon-based approaches for sentiment analysis. Utilizing labeled data, supervised ML classifiers such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machines (SVM), etc. (Pang et al., 2002; Gamon, 2004) and deep learning-based classifiers (Abdi et al., 2019; Araque et al., 2017) have been employed by the researchers for sentiment classification. Though the lexicon-based methods (Turney, 2002) do not require labeled data, they suffer from the lexicon coverage problem and are not robust to deal with the ambiguity and linguistic variations of natural languages.

Though English and few other languages enjoy ample resources for sentiment analysis, such resources are not available in many other languages. Cross-lingual sentiment classification aims to leverage resources like labeled data, polarity lexicons, contextual valence shifters, modifiers, etc. from resource-rich languages (typically English) to classify the sentiment polarity of the text written in a low-resource language (such as Bengali). For language mapping, several approaches such as machine translation (Banea et al., 2008a; Wan, 2009; Demirtas and Pechenizkiy, 2013; Zhou et al., 2016a,b; Abdalla and Hirst, 2017; Balahur and Turchi, 2014), cross-lingual word embedding (Barnes et al., 2018; Xu and Yang, 2017; Tang et al., 2014; AP et al., 2014), etc. have been used by the researchers.

1.1 Motivation

A limited amount of research in sentiment analysis has been conducted in Bengali in the last few decades; however, still, there is no benchmark dataset. Researchers used their curated datasets in various literatures that are not publicly available. The absence of publicly available datasets made

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 50–60

the research findings non-reproducible. Moreover, without a benchmark dataset, it is challenging to compare the performance of various approaches.

Though cross-lingual approaches have been successfully applied to several low-resource languages (Meng et al., 2012; Banea et al., 2008b), in Bengali only a few works utilized it for tasks like sentiment lexicon creation (Das and Bandyopadhyay, 2010a; Sazzed, 2020) and sentiment classification (Sazzed and Jayarathna, 2019). However, until now, no comprehensive study has been performed to explore the applicability of the cross-lingual sentiment classification approach in Bengali.

Therefore, in this work, we created and annotated a large Bengali review dataset for binary-level sentiment analysis. This corpus consists of around 12000 Bengali reviews collected from Youtube. We present a comprehensive study of the machine-translation based cross-lingual approach of sentiment analysis in Bengali.

Using a large and well-annotated dataset, we compare and provide detailed analysis regarding the performance of ML classifiers in the Bengali and machine-translated datasets. Besides, using Cohen's kappa and ML classifiers, we examine sentiment preservation in the machine-translated corpus.

As annotated data are not always obtainable, especially in low-resource languages, we investigate the performance of unsupervised lexicon-based methods in the machine-translated corpus. Popular lexicon-based sentiment analysis methods, VADER (Hutto and Gilbert, 2014), TextBlob[1], and SentiStrength (Thelwall et al., 2010) are applied and their relative performances are compared.

We investigate the applicability of the simple transfer learning-based approach to the machine-translated corpus. Resource-rich language such as English contains copious labeled data, which are not available in Bengali. Utilizing machine-translation and cross-domain labeled data, we show the performance of supervised ML classifiers in the translated corpus.

1.2 Contribution

Our major contributions can be summarized as follows:

- We introduce a large well-annotated benchmark dataset for sentiment analysis in Bengali.
- We perform a comparative evaluation of supervised ML classifiers in Bengali and machine-translated English corpus and provide a rigorous analysis of the results.
- We investigate cross-lingual lexicon-based methods, as well as a transfer learning-based approach to deal with the lack of labeled data in Bengali.

[1]https:textblob.readthedocs.io/

2 Literature Review

2.1 Sentiment Analysis in Bengali

English is the dominant language for sentiment analysis research due to commercial interest and a large research community. In recent years, with the popularity of e-commerce and social networking sites, review data is becoming available in other languages.

In Bengali, limited research has been performed using corpora collected from various sources such as Microblogs, Facebook, and other social media sources (Patra et al., 2015; Das and Bandyopadhyay, 2010b). Various supervised classifiers have been employed for Bengali sentiment analysis such as SVM with maximum entropy (Chowdhury and Chowdhury, 2014), Naive Bayes (NB) (Islam et al., 2016b), Deep Neural Network (Tripto and Eunus Ali, 2018), Convolutional Neural Network (CNN) (Sarkar, 2019). In (Al-Amin et al., 2017), the authors utilized word2vec and polarity score for the binary sentiment analysis problem. A word-embedding based approach was proposed by Islam et al. (2016a). Hassan et al. (2016) predicted sentiment orientation of Bengali and Romanized Bengali text using Long Short-Term Memory (LSTM).

2.2 Cross-lingual Sentiment Analysis

The cross-lingual sentiment analysis approaches have been studied in many languages. Mihalcea et al. (2007) leveraged the tools and resources available in English to generate subjectivity analysis resources in Romanian. They created a Romanian subjectivity lexicon translated from the English lexicon and utilized a corpus-based approach. Balamurali et al. (2012) presented an alternative approach to cross-lingual sentiment analysis (CLSA) using WordNet senses as features for supervised sentiment classification. They used the linked WordNets of two languages to bridge the language gap. They reported their results on two Indian languages, Hindi and Marathi. Balahur and Turchi (2014) in-

vestigated the performance and effectiveness of machine translation systems and supervised methods for multilingual sentiment analysis. In their experiment, they used four languages, English, German, Spanish, and French; three machine translation systems Google, Bing, and Moses; several supervised algorithms and various types of features. Yan et al. (2014) utilizing the SVM algorithm proposed a bilingual approach for sentiment analysis in the Chinese social media dataset. In (Meng et al., 2012), the authors proposed a cross-lingual mixture model (CLMM) to exploit unlabeled bilingual parallel corpus. In (Banea et al., 2008b), authors utilized a machine translation system for projecting resources from English to Romanian and Spanish and provided a comparative performance. Chen et al. (2015) proposed a semi-supervised learning model, CredBoost, to address cross-lingual sentiment analysis in English and Chinese. They introduced a knowledge validation step during transfer learning to reduce the noisy data caused by machine translation errors. Feng and Wan (2019) proposed a cross-lingual sentiment analysis (CLSA) model by leveraging unlabeled data in multiple languages and domains. Without using any supervised cross-lingual word embedding (CLWE), their model outperformed baseline methods on multilingual Amazon review datasets. Xu et al. (2018) proposed a learning approach that does not require any cross-lingual labeled data. Their algorithm optimizes the transformation functions of monolingual word-embedding space and uses a neural network. They evaluated their proposed approach on benchmark datasets for cross-lingual word similarity prediction and found competitive performance to other methods. Chen et al. (2018) introduced an Adversarial Deep Averaging Network (ADAN) to transfer the knowledge learned from source languages labeled data to the target language. Their experiments on Chinese and Arabic sentiment classification demonstrated the superior performance of ADAN. Rasooli et al. (2018) used multiple source languages to learn a robust sentiment transfer model. They explored the potential of using both the annotation projection approach and a direct transfer approach using cross-lingual word representations and neural networks.

The cross-lingual approach of sentiment analysis in Bengali is still largely unexplored, only a few works investigated it (Das and Bandyopadhyay, 2010a; Sazzed and Jayarathna, 2019; Sazzed, 2020). Das and Bandyopadhyay (2010a) translated English polarity lexicon to Bengali to create a Bengali sentiment dictionary. Sazzed and Jayarathna (2019) utilized two small datasets and n-gram (i.e.,unigram and bigram) feature vectors to compare the performance of supervised ML algorithms in Bengali and machine-translated English corpus. They found supervised ML algorithms showed better performance in the model trained on the translated corpus; however, they did not provide a thorough analysis of the results they reported.

Contrast to previous studies, we perform a comprehensive analysis of various cross-lingual sentiment analysis approaches in Bengali. We created a Benchmark dataset, explored several classification approaches utilizing labeled and unlabeled data, examine the applicability of transfer learning, investigate the sentiment preservation in the translated corpus, and finally provide the direction for future research. To best of our knowledge, this is the first extensive attempt to investigate the applicability of the cross-lingual approach in Bengali sentiment analysis.

3 Dataset

One of the barriers of sentiment analysis research in Bengali is the lack of publicly available review datasets. In literature, researchers reported results using their curated datasets that are not publicly available. The few publicly available datasets are either small in size or not well-annotated. Therefore, here, we have prepared a well-annotated Bengali review dataset that we made publicly available.[2]

3.1 Data Collection

We collected and manually labeled a large review dataset for sentiment analysis in Bengali. This dataset contains viewer opinions towards several Bengali dramas. Using a web scraping tool, we first downloaded the raw JSON data from Youtube that contains information such as user name, id, timestamp, comments, and like/dislike, etc. We use a parsing script to extract the viewer's comments from the JSON data. The comments are written in Bengali, English, Romanized Bengali, or use code-mixing. As we are only interested in reviews written in Bengali text, we excluded non-Bengali comments. We utilized a language detection li-

[2]https://github.com/sazzadcsedu/BN-Dataset.git

Bengali Reviews	Machine Translation	Polarity
এই ধরনের নাটক সমাজের কোন কাজে লাগবে।এই গুলো আর নীলছবি তো একি কথা।সাময়িক উত্তেজনা তৈরি করে,আর হিতাহিত জ্ঞান শূণ্য করে।	Such plays will be of no use to society. These blue and blue are the same thing. Creates temporary tension, and empties knowledge	Negative
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 খারাপ থাকে,তখন আপনাদের নাটক দেখি।তখন মনটা আরো খারাপ হয়ে যায়,আর একটা সময়ে মন খারাপে, মন খারাপে 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 খোর ছাগলের বাচ্চা এইসব কী	The director is a cannabis-eating goat kid	Negative
অসাধারন আমার কাছে সেই লাকছে~	Extraordinary That Looks To Me	Positive

Figure 1: Example of Bengali and machine translated reviews

brary[3] to identify Bengali comments. After removing the non-Bengali comments, the corpus contains around 15000 reviews, which are labeled using the procedure described in the next section.

3.2 Data Annotation

Two native Bengali speakers classified these 15000 reviews into three categories, *positive, negative, and non-subjective*. From the annotator ratings, we observe an inter-rater agreement of around 0.83 using Cohen's kappa. We exclude all the reviews, which are marked as non-subjective by either of the annotators.

For each subjective reviews, we only include it to the corpus if both annotators assign it to the same category (i.e., *positive or negative*). Therefore, our dataset contains only highly polarized reviews. Reviews that are ambiguous or contain mixed sentiment are not included in the dataset.

The final labeled corpus consists of 11807 annotated reviews, where each review contains around 2-300 Bengali words. This corpus is class-imbalanced, comprised of 3307 *negative* and 8500 *positive* reviews. Figure 1 shows some examples of *negative* and *positive* reviews. We made this corpus publicly available for the researchers.

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

4 Cross-lingual Sentiment Analysis in Bengali

As Bengali is a resource-poor language, we leverage sentiment lexicon and labeled data from English for sentiment analysis in Bengali. We investigate the performances of various approaches (i.e., supervised, unsupervised, and transfer-learning based approaches) of sentiment analysis utilizing resources from English. Figure 2 shows the overview of various approaches we studied.

4.1 Language Mapping

The machine translation (MT) service is one of the most common ways to build the language connection (Wan, 2008a, 2009; Wei and Pal, 2010). Bautin et al. (2008) discussed the use of various Spanish translation systems, Wan (2008b) compared various Chinese machine translators and found Google Translate provided the best performance. Here, we use Google Translate[4] to translate our Bengali corpus into English.

4.2 Supervised Classification Approach

Supervised ML-based approaches have been successfully applied in English and other languages for sentiment classification. Since supervised ML classifiers do not rely on language resources such as sentiment lexicon, part-of-speech (POS) tagger,

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

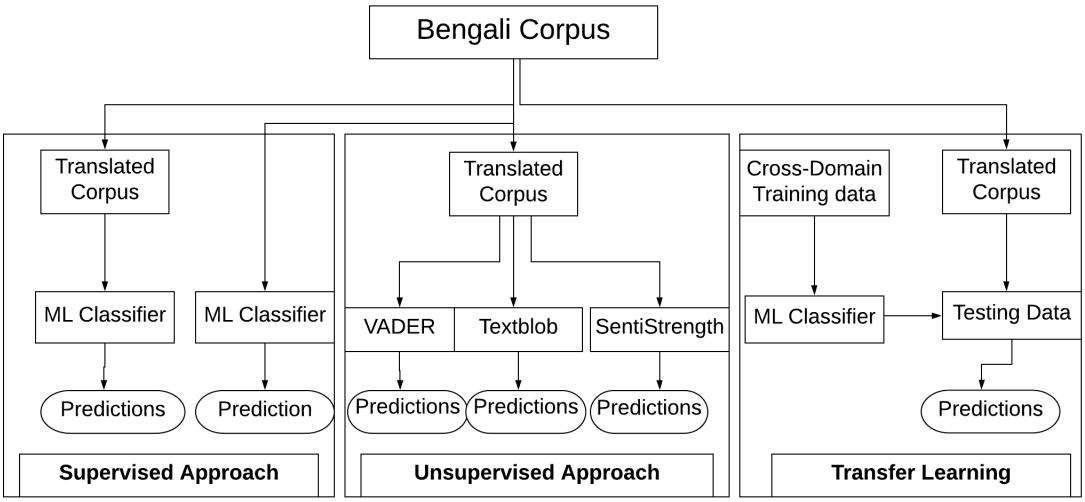

Figure 2: Various approaches of cross-lingual sentiment analysis in Bengali

etc., they can be applied to any language. In contrast to the rigid rule-based method, supervised ML algorithms learn hidden patterns from the training data; therefore, they can be more robust against noisy machine-translated English corpus.

Utilizing the annotated data, we employ four supervised ML classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Extremely Randomized Trees (ET) on Bengali and its machine-translated English corpus. We use the scikit-learn (Pedregosa et al., 2011) implementation of the aforementioned ML classifiers. For all the ML classifiers, we utilize the default parameter settings. To deal with the class imbalance problem, we set the weight of a class inversely proportional to the number of instances it contains. Both the unigram and bi-grams features are used as input for the ML classifiers. We perform 10-fold cross-validation in both Bengali and translated English corpus.

4.3 Lexicon-based Approach

To deal with the scenario when annotated data are not available, we study the performances of lexicon-based methods in machine-translated English corpus. In Bengali, no standard lexicon-based tool is publicly available for sentiment analysis; therefore, we could not compare the performance with the English counterpart.

Three popular lexicon-based methods from English: VADER, TextBlob, and SentiStrength are employed to find the effectiveness of the cross-lingual unsupervised approach.

4.4 Transfer Learning-based Approach

Annotated data are hard to achieve in low-resource languages such as Bengali. But resource-rich languages like English owns a vast amount of labeled data. Hence, we explore the applicability of a transfer learning-based approach to the machine-translated corpus. However, in this work, we did not introduce any new transfer learning method. We examine whether utilizing existing cross-domain labeled data assist in achieving an acceptable performance of sentiment classification in Bengali when labeled data are not available.

In the transfer setting, a classifier is trained on one distribution while applied to a different distribution. The idea is to leverage labeled data from distinct domains but use in a similar task, as annotated in-domain data are not always available.

We employ multiple cross-domain datasets from the English language, IMDB (Maas et al., 2011), Yelp[5], TripAdvisor (Thelwall, 2018), Clothing[6], UCI Drug[7], WebMD[8] as shown in Table 1. We train the Logistic Regression (LR) classifier using cross-domain datasets and use the trained model to predict the semantic orientations of reviews in our machine-translated corpus. The default parameter settings of the LR classifier of scikit-learn (Pedregosa et al., 2011) library is used with a class-balanced weight.

[5]https://kaggle.com/omkarsabnis/yelp-reviews-dataset

[6]https://kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

[7]https://kaggle.com/jessicali9530/kuc-hackathon-winter-2018

[8]https://kaggle.com/nataliele/webmd-contraceptives-reviews-file

Dataset	Domain	Positive	Negative	Total
IMDB	Movie	12500	12500	25000
YELP	Restaurant	6860	1676	8536
TripAdvisor	Hotel	9520	9520	19040
Clothing	Clothing	18540	4101	22641
UCI Drug	Drug	35437	11838	47275
WebMD	Drug	7461	1808	9269

Table 1: Cross-domain review datasets from English

Classifier	Precision	Recall	Macro F1	Accuracy
	BN/EN	BN/EN	BN/EN	BN/EN
SVM	0.908/0.912	0.924/0.934	0.916/0.923	93.0/93.5%
LR	0.889/0.893	0.922/0.927	0.905/0.910	91.8/92.2%
ET	0.893/0.882	0.882/0.865	0.888/0.874	91.0/90.0%
RF	0.878/0.889	0.870/0.881	0.874/0.885	89.9/90.8%

BN= Bengali, EN= English

Table 2: Performances of supervised ML classifiers in Bengali and machine-translated English corpus

5 Experimental Results

5.1 Evaluation Criteria

To compare the performances of various classifiers, we compute precision, recall, macro F1 score, and accuracy. As our dataset is class-imbalanced, the macro F1 score the better metric than the accuracy for the evaluation.

Besides, we assess the agreement of the predictions of various supervised ML classifiers in Bengali and machine-translated English corpus utilizing Cohen's kappa and Gwet's AC1 statistics. Cohen's kappa and Gwet's AC1 are statistical measures used to gauge inter-rater reliability, where a score of 1 refers to perfect agreement. The purpose of evaluating the agreement is to determine the sentiment preservation in the machine-translated English corpus. .

5.2 Supervised Approach

In this section, we provide the comparative performances of ML classifiers in Bengali and machine-translated English corpus and agreement of the predictions.

5.2.1 Performance Comparison

Supervised ML classifiers show similar performance in both Bengali and translated English corpus, as shown in Table 2. The best macro F1 score and accuracy are obtained using the SVM classifier, which is 0.923 and 93.5% for English and 0.916 and 93.0% for Bengali. A similar performance is observed when the LR classifier is applied to the English and Bengali corpus. The decision tree-based methods, RF and ET show lower F1 scores and accuracies compared to SVM and LR.

Classifier	Cohen-kappa	AC1
SVM	0.819	0.868
LR	0.820	0.860
RF	0.694	0.800
ET	0.703	0.809

Table 3: The Cohen's kappa and AC1 scores of various ML classifiers in Bengali and translated corpus

5.2.2 Agreement of Predictions

We compute the agreement of the predictions of ML classifiers in Bengali and machine-translated English corpus. The purpose is to examine whether the noise induced by machine translation changes the sentiment orientations of the translated reviews. When the sentiment orientation is maintained in the translated corpus, we can expect a high agreement between the predictions of an ML classifier in Bengali and its machine-translated version.

Table 3 provides the Cohen's kappa and AC1 scores applying various ML algorithms. SVM and LR show kappa scores above 0.80 and AC1 score above 0.85, while RF and ET provide around 0.70 kappa score and 0.80 AC1 scores.

Method	Precison	Recall	Macro F1	Accuracy
VADER	0.846	0.707	0.771	82.56%
TextBlob	0.863	0.705	0.776	82.79%
SentiStength	0.787	0.645	0.708	78.61%

Table 4: The performances of lexicon-based methods in the machine-translated corpus

5.3 Lexicon and Transfer learning-based Approaches

Table 4 shows the results of the lexicon-based methods in the translated corpus. VADER and TextBlob exhibit similar F1 scores and accuracies, while SentiStrength performs relatively worse. Using VADER, we achieve an F1 score of 0.771 and an accuracy of 82.56%, while TextBlob obtains 0.776 and 82.79%, respectively.

Table 5 provides the results of LR classifier utilizing cross-domain data. The best performance is obtained by combining all cross-domain datasets, which is 0.78 for the F1 score and 82% for the accuracy.

6 Discussion

6.1 Supervised Approach

Table 2 shows that supervised ML classifiers provide similar performance in the translated corpus and the original Bengali corpus. We found that several factors influence the comparable performance on the machine-translated corpus.

6.1.1 Error Correction

Misspelling is a common scenario in online Bengali content due to the complexity of the Bengali writing system and the education level of most of the internet users. Modern machine translation tools are trained on a huge amount of data and are capable of correcting misspelling. Although the Bengali-English machine translation system is not that sophisticated compared to some major language pairs, occasionally, it can identify misspelled words in Bengali text, and translate to correct English word. For those cases, machine translation improves the quality of data, so the classifier performance is improved.

6.1.2 Word Mapping

The current Bengali-English machine translation system still lacks enough coverage. We observed in some cases, Bengali synonym words are mapped to the same English word. This word-mapping assists supervised ML classifiers to perform well in the machine-translated corpus.

6.1.3 Regional Variety of Bengali

The Bengali language contains a large variety of dialects that are widely used on the web, especially in social media. The machine translation service that is trained on thousands of corpora can identify them as a variant of the same words and translate them to the same English word that positively impacts the performances of ML classifiers in the translated corpus.

6.1.4 Feature Importance and Sentiment Preservation

Supervised ML algorithms utilize the bag-of-words model to train the classifiers. The term frequency-inverse document frequency (tf-idf) score is calculated and used as an input feature vector. tf-idf is a numerical statistic that reflects the importance of a word considering a collection of documents.

tf-idf score refers that not all the words in a document are equally important for classification. (Abdalla and Hirst, 2017) showed that sentiment is highly preserved even in the face of poor translation accuracy. Therefore, low-quality translation does not always affect classifier accuracy.

The Cohen's kappa and AC1 scores reveal the sentiment consistency between original Bengali reviews and its machine translated version as shown in Table 3. The Cohen-kappa and AC1 scores from SVM and LR show nearly perfect agreements on the results from Bengali and translated English corpus. For RF and DT, Cohen's kappa and AC1 statistics are a bit lower compared to SVM and LR, which could be affected by the inferior performance of those classifiers, however, still, agreements are substantial.

6.2 Lexicon-based Approach

TextBlob and VADER exhibit similar accuracy, precision, recall, and F1-scores, while SentiStrength performs worse. The results demonstrate that lexical-rule based methods are not as robust as supervised ML approaches, exhibited by the lower

Training Dataset	#Reviews	Precision	Recall	Macro F1	Accuracy
IMDB	25000	0.73	0.71	0.72	78.0%
Clothing	22641	0.62	0.64	0.63	67.0%
TripAdvisor	19040	0.68	0.72	0.70	66.0%
UCI Drug	47275	0.71	0.70	0.71	76.7%
WebMD	9269	0.61	0.64	0.62	63.7%
Yelp	8536	0.67	0.65	0.66	73.6%
Aggregated Dataset	135121	0.78	0.77	0.78	82.0%

Table 5: The performance of LR classifier in the translated corpus utilizing the multi-domain training datasets

scores in all categories. Particularly, the recall scores, due to the non-comprehensive coverage of lexicon, are quite low. The poor performance of the rule-based approach mainly comes from the intrinsic nature (e.g., lexicon/rule coverage) of lexicon-based methods.

6.3 Transfer Learning with Cross-domain Datasets

The results obtained using the LR classifier and the cross-domain datasets indicate that the classifier's performance depends on both the-

- Data distribution and
- Size of the training dataset

The IMDB movie review dataset is the most similar to our translated drama review dataset considering the essence of the reviews. However, still, they differ in the aspects of data, languages used in the reviews, and the presence of noise due to machine translation. The translated drama reviews are much shorter in length and contain simple English words compared to IMDB reviews, which are written mostly by native English speakers. Utilizing 25000 reviews from IMDB, we achieve the best performance among all the cross-domain datasets used. Leveraging data from different domains, such as clothing or drug, yields worse performance despite using similar or larger size training dataset, which demonstrates the domain specificity in the sentiment analysis dataset.

We consolidate all the six cross-domain datasets to create a large corpus of over 130k reviews. The supervised LR classifier exhibits performance improvement utilizing this aggregated dataset. The results indicate that though datasets from the different domains show poor performance in isolation when aggregated, they can enhance the classifier performance.

With over 130k consolidated cross-domain reviews, the transfer learning-based approach shows noticeably worse performance compared to in-domain data, an F1 score of 0.773 compared to 0.910 using the LR classifier. It provides similar performance to the best lexicon-based method, VADER, which yields an F1 score of 0.771. Word level polarity is heavily influenced by context and domain, which was reflected in the classifier's performance when cross-domain data are used.

6.4 Findings and Implications

- We find that online content in Bengali consists of lots of misspelled and regional words, which affects the performance of sentiment classifiers. Therefore, it is necessary to build sophisticated tools that can fix misspellings and recognize regional variants of Bengali words.
- Although the existing Bengali-to-English machine translation system is still far from perfect, it is capable of preserving sentiment information; hence can be utilized for cross-lingual sentiment analysis.
- We find that the lexicon-based method performs poorly compared to the supervised ML methods in the machine-translated corpus. Therefore, it is imperative to develop an automatic or semi-automatic data annotation method.
- We find that a large number of cross-domain labeled data provides similar performance of the lexicon-based approach. Therefore, transfer learning can help when in-domain labeled data are unavailable.
- Our study reveals that the cross-lingual approach can be effective in Bengali sentiment analysis. Therefore, future research should

focus on exploring and developing new methods for the cross-lingual sentiment analysis in Bengali.

7 Conclusion

To facilitate sentiment analysis research in Bengali, in this work, we introduce a benchmark dataset and explore the adaptation of resources and tools from English. We notice that due to misspellings, usage of regional varieties of Bengali, and advancement of the machine translation system, supervised ML algorithms perform comparably in the Bengali and machine-translated corpus. The agreements of the predictions suggest that Bengali-English machine translation can preserve the sentiment information. The mediocre performances of the lexicon-based methods infer that annotated data are essential to achieve better classification accuracy.

We present the performance of simple transfer learning utilizing cross-domain data. We note that with enough cross-domain training data, supervised ML classifiers provide a comparable performance of the lexicon-based methods, though lag behind the performance achieved through in-domain data. We report our findings regarding cross-lingual sentiment classification approaches in Bengali, which provide directions for future research.

References

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 506–515, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Asad Abdi, Siti Mariyam Shamsuddin, Shafaatunnur Hasan, and Jalil Piran. 2019. Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, 56(4):1245–1259.

M. Al-Amin, M. S. Islam, and S. Das Uzzal. 2017. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 186–190.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in neural information processing systems*, pages 1853–1861.

Oscar Araque, Ignacio Corcuera-Platas, J Fernando Sánchez-Rada, and Carlos A Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

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

Aiswarya Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *COLING*.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008a. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008b. 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*, pages 127–135.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493, Melbourne, Australia. Association for Computational Linguistics.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *ICWSM*.

Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. 2015. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 419–429.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

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

Amitava Das and Sivaji Bandyopadhyay. 2010a. Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task*, 2:1–8.

Amitava Das and Sivaji Bandyopadhyay. 2010b. Topic-based bengali opinion summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 232–240. Association for Computational Linguistics.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8.

Yanlin Feng and Xiaojun Wan. 2019. Towards a unified end-to-end approach for fully unsupervised cross-lingual sentiment analysis. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1035–1044, Hong Kong, China. Association for Computational Linguistics.

Michael Gamon. 2004. 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*, page 841. Association for Computational Linguistics.

Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Mohammed. 2016. Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 51–56. IEEE.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

M. S. Islam, M. A. Amin, and S. Das Uzzal. 2016a. Word embedding with hellinger pca to detect the sentiment of bengali text. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 363–366.

M. S. Islam, M. A. Islam, M. A. Hossain, and J. J. Dey. 2016b. Supervised approach of sentimentality extraction from bengali facebook status. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*, pages 383–387.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 572–581, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer.

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

Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, Tao Yu, and Kathleen McKeown. 2018. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1-2):143–165.

Kamal Sarkar. 2019. Sentiment polarity detection in bengali tweets using deep convolutional neural networks. *Journal of Intelligent Systems*, 28(3):377–386.

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

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

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. 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)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Mike Thelwall. 2018. Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42:343–354.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558.

Nafis Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6.

Peter D Turney. 2002. 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*, pages 417–424. Association for Computational Linguistics.

Xiaojun Wan. 2008a. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561.

Xiaojun Wan. 2008b. Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics.

Bin Wei and Christopher Pal. 2010. Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the ACL 2010 conference short papers*, pages 258–262. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1415–1425. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Gongjun Yan, Wu He, Jiancheng Shen, and Chuanyi Tang. 2014. A bilingual approach for conducting chinese and english social media sentiment analysis. *Comput. Netw.*, 75(PB):491–503.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 247–256.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. 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)*, pages 1403–1412, Berlin, Germany. Association for Computational Linguistics.