ATLANTIS PRESS

Natural Language Processing Research
Vol. **1(3-4)**, 2021, *pp.* 34–45
DOI: https://doi.org/10.2991/nlpr.d.210316.001; eISSN: 2666-0512
https://www.atlantis-press.com/journals/nlpr/

Research Article

Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary

Nitish Ranjan Bhowmik[1], Mohammad Arifuzzaman[2,*], M. Rubaiyat Hossain Mondal[1], M. S. Islam[1]

[1]*Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka, Banglaadesh*
[2]*Department of Electronics and Communications Engineering, East West University, Dhaka, Banglaadesh*

ARTICLE INFO

Article History

Received 19 Sep 2020
Accepted 09 Mar 2021

Keywords

Sentiment analysis
Bangla NLP
Tf-Idf
SVM
BTSC
N-grams
Bi-grams

ABSTRACT

With the proliferation of the Internet's social digital content, sentiment analysis (SA) has gained a wide research interest in natural language processing (NLP). A few significant research has been done in Bangla language domain because of having intricate grammatical structure on text. This paper focuses on SA in the context of Bangla language. Firstly, a specific domain-based categorical weighted lexicon data dictionary (LDD) is developed for analyzing sentiments in Bangla. This LDD is developed by applying the concepts of normalization, tokenization, and stemming to two Bangla datasets available in GitHub repository. Secondly, a novel rule–based algorithm termed as Bangla Text Sentiment Score (BTSC) is developed for detecting sentence polarity. This algorithm considers parts of speech tagger words and special characters to generate a score of a word and thus that of a sentence and a blog. The BTSC algorithm along with the LDD is applied to extract sentiments by generating scores of the two Bangla datasets. Thirdly, two feature matrices are developed by applying term frequency-inverse document frequency (tf-idf) to the two datasets, and by using the corresponding BTSC scores. Next, supervised machine learning classifiers are applied to the feature matrices. Results show that for the case of BiGram feature, support vector machine (SVM) achieves the best classification accuracy of 82.21% indicating the effectiveness of BTSC algorithm in Bangla SA.

1.INTRODUCTION

Sentiment analysis (SA), also called opinion mining [1], is a field of study that predicts polarity in public opinion or textual data from microblogging sites [2] on a well-publicized topic by extracting people emotions, attitudes, emotions, etc. As, SA is becoming a relevant subject to natural language processing (NLP) in machine learning area, researchers are gradually finding interest in this topic because of having a large scale of opinionated data on the Internet. Nowadays people in social media sites, newspaper, blogs, etc., express their reviews on a specific product or items. There are also forum discussion, opinion on a specific posts, comments, and emotions. There may arise many obstructive in detecting binary or ternary class sentiment such as subjectivity or opinion based identification, if a phrase or text have not any core opinion word. So, lexicon-based [3] data dictionary approach is jointed with their semantic tendency with polarity and word strength. To determine these data with sentiment as a polarity, i.e., positive, negative, or neutral class, machine learning framework has acquired significant interest. This is because of the building model in many linguistic domains [4] with versatile feature extraction, alternating, predicting with probabilistic theory, and computing valuable feature matrix representations. Various types of features have been observed for this type of work such as bag of words (BoW) model, lexical analysis, and semantic feature [5]. This matrix feature is language-dependent. Bangla, an ancient Indo-European language, is the spoken language of over 250 million people [6]. So, extracting sentiment in Bangla language will surely be significant for NLP researchers to make substantive progress in machine learning. Among the three levels of SA, we worked on the sentence level polarity classification by using extended Bangla sentiment dictionary. This sentimental dictionary words are implying as opinion words which is an impetus for identifying polarity from text by implementing a set of rule-based automatic classifier algorithm [7]. In this paper, an effective and unique rule-based algorithm Bangla Text Sentiment Score (BTSC) is developed for detecting sentence polarity which provides the better sentiment extraction by giving a score from a chunk of Bangla text. We build an automated system which can extract opinion from Bangla dataset reviews with the help of extended Bangla sentimental dictionary with weighted value, and that automated system will be classified by supervised machine learning algorithm [8] with the help of N-grams model. This is because author [9] found this model performing well in text classification. The rest of the paper is organized as follows: Section 2 recapitulates the related research works. Section 3 demonstrates our system methodology. In Section 4, we discuss our proposed algorithm BTSC to find out score from the text. Section 5 illustrates the evaluation results with trained and tested datasets by a supervised classification approach like SVM,

*Corresponding author. Email: mazaman@ewubd.edu

logistic regression (LR), K-nearest neighbors (KNN), etc. Finally, the results of the research are summarized in Section 6.

2. RELATED WORK

In the era of expansion of social media and microblogging sites, SA has become an interesting topic among researchers. Apparently, SA is done in many linguistic domains like English, French, Chinese, Arabic, etc. However, the depth of its progress in Bengali language is insignificant due to some technical and empirical constraint [10]. Our work is highly inspired by this paper [11], as far as we know SA in Bengali using extended dictionary has not been done in any research. Experiment results using lexicon- based data dictionary in Arabic language have been obtained so far better [12]. In Alshari *et al.* [13] authors described SentiWordNet (SW) as a curse of dimensionality, they used sentimental lexicon dictionary based on word2vec to perform SA. Besides, in Bangla text, author [14] preprocessed data to carry through a SA by taking TF-IDF vectorizer and classified the data with support vector machine (SVM) algorithm, however they did not measure the polarity by calculating the score of a text; hence it is required to detect the polarity of each sentence by a specific rule-based [15] algorithm. In Chowdhury and Chowdhury [16], the author proposed a semi-supervised bootstrapping approach in SVM and maximum entropy (MaxEnt) classifier to perform a SA using SW by translating Bangla word to English. In their bootstrapping rule-based approach, they have only counted positive, negative word polarity by SW which is only work for a low limited length text. In Azharul Hasan *et al.* [17], authors proposed a method of using XML based POS tagger and SW to identify the sentiment from Bangla text adopting valency analysis. They have used SW and WordNet (WN) which were designed for only English language. So, a lexicon weighted word dictionary for Bangla is needful for identifying the word score or polarity from the text. Besides, In Islam *et al.* [18], authors extracted positive, negative (bi-polar) polarity from facebook text by tokenizing adjective word using POS tagger, doing valence shifting negative words at the right side of a sentence and replace it with antonym word using SW. As SW has a weakness in giving proper polarity in Bangle text, the authors in [19] discussed an automated system for emotion detection by mapping each text to an emotion class, their accuracy was 90% however it was more time consuming for labeling the data and their phrase patterns were formed for only three sub categories sentiment not used for in complex sentences. In Tabassum and Khan [20], authors designed a framework for SA by counting only positive and negative words form their feature word list dictionary. In Zhang *et al.* [21], authors constructed an extended sentiment dictionary and a rule-based classifier was employed to classify the field of the text polarity by attaining the score of a sentence. In Akter and Aziz [22] authors described a lexicon-based dictionary model by checking the occurrences of a sentimental feature word in tagging each sentence.

3. METHODOLOGY

The main goal of this research is to analyze the sentiment from Bangla text in machine learning approach by an unique rule-based algorithm along with building a lexicon data dictionary (LDD). For detection of Sentiment polarity from raw of a text, we have divided our whole work into three parts.

In Figure 1, our proposed approach is illustrated and these steps are described below. To meet the goal, the following objectives have been identified:

a. To construct a specific domain-based categorical weighted LDD for analyzing sentiment classification from Bangla dataset.

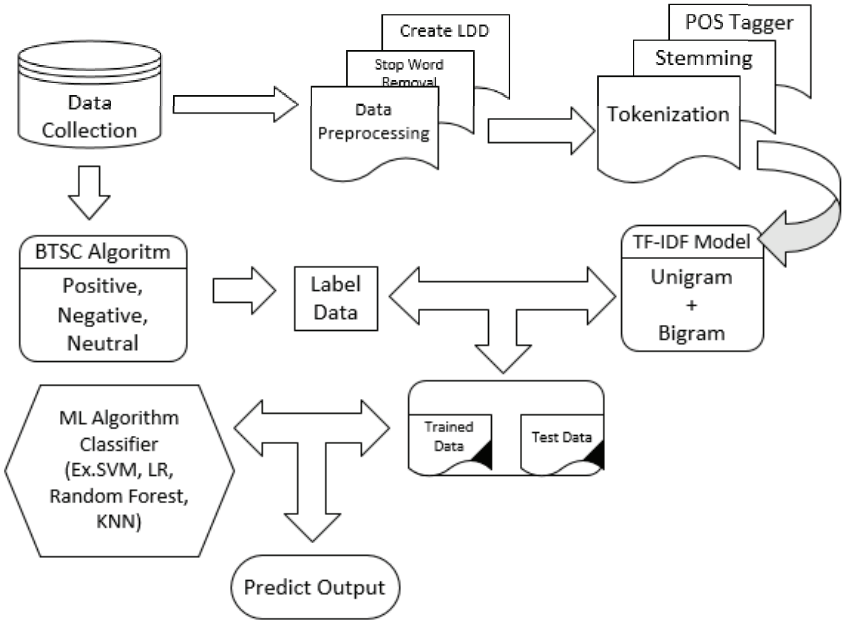


Figure 1 | Visualization of proposed system architecture.

b. To develop a novel and effective rule-based algorithm for detecting sentence polarity classification by extracting score from a chunk of Bangla text.

c. To investigate the feature matrix with target dataset and evaluate our theoretical claim and finally comparing the circumference of our work with some existing research paper in supervised machine learning algorithm.

3.1. Construction of Extended LDD

3.1.1. Creation of sentimental dictionary list

Extended lexicon dictionary means alphabetical list of words or phrases. As a Bengali sentence consists of many words, we have manually created a sentimental dictionary word list from the sentence. These words are applied for calculating the score from a sentence or phrase. We have collected data from [23], where there are two datasets based on two domains: Cricket and Restaurant. The construction of Bangla dataset are described in [24] extensively. Table 1 showing the statistical polarity for both datasets.

For performing SA, we have used those datasets to build up our extended sentimental dictionary, such as, যোগ্যতা [Competence], "অযোগ্যতা" [Inefficiency], "পরিষেবা" [Service,] "বাধা" [Hindrance], etc. In sentimental dictionary, a word can be intersecting in both datasets like "কল্পনাপ্রসূত" [Imaginary] word is an active word list in restaurant dataset but a contradict word list in cricket dataset. While SW works on the global domain data but to do SA in different domain data, sentimental weighted dictionary has to be created. The number of sentimental word list is composed of active (weight = +1) and contradict word (weight = −1), which is extracted by manually, represented in Table 2. Besides a negative word (weight = −1) list like, "না", "নেই", "নাই", "নয়" [not] vocabulary has been created so that negative words can be counted during score calculation from the text.

3.1.2. Creation of adjective, adverb quantifier, and conjunction word dictionary weighted list

Since a major difference in English Bangla grammar, we need to create our own weighted dictionary word list of adjective "নাম বিশেষণ" [25] and adverb quantifier which is showed at Table 3. In Bangla grammar "বিশেষণের অতিশায়ন" [Exaggeration of adjectives] and degree of adverb "ক্রিয়া বিশেষণ" [26] is a segment of Adjective POS tagger. We partitioned the whole word set into 3 types: high, medium, low.

Although these words have no effect on determining the polarity of a sentence, however, sitting before a few words in a sentence can impact on the score of the whole sentence. These words can quantify the sentence score. For example, "ব্যাটসম্যানদের মধ্যে সাকিব সবচাইতে ভালো" [Shakib is the best among the batsmen] in that sentence "সবচাইতে" [Most of all] word quantify the word "ভালো" [good] which produce the word [best] in the translated English sentence. A Bangla sentence can have conjunction POS which is used to joint words, phrases, clauses. These types of word sit in the middle, beginning or at the end of a sentence, and connect one or more sentences together. This further increases the score of two sentences without effecting on polarity. As There are four main types of conjunction and many sub parts conjunctions in bangali grammar, for simplification of our work we generalize them into two categories named as "সমুচ্চয়ী", "অনুগামী" coordinating and subordinating or progressive conjunctions. However, in our research work, we simplify those words and assign appropriate weight values. It can be noted that the weight values of adjective and adverb in Table 3, and that of conjunctions in Table 4 are assigned carefully to make it particularly suitable for the Bangla language context. For this assignment, the GitHub Bangla dataset available in [23] is taken into consideration. Because of the difference in language structure, the weight values of Tables 3 and 4 for Bangla are different from the values mentioned for Chinese language in [11].

Table 1 Showing statistical polarity for both datasets with individual and total comments.

Dataset	Polarity			Total
	Positive	Negative	Neutral	
Restaurant	1216	478	365	2059
Cricket	620	2108	251	2979

Table 2 Showing sentimental word list with individual and total.

Dataset	Sentimental Dictionary		Total Words
	Active	Contradict	
Restaurant	1056	970	2026
Cricket	1115	2190	3035

3.2. Dataset Preprocessing

Our SA is a document sentiment classification based on supervised machine learning. After collecting corpus data, we need to preprocess these data for creating training and testing data. Because data preprocessing is an important part in the NLP domain. We use BLTK [27] version 1.2 in open source python PyPI package OSI Approved, MIT License to preprocess our data. Dataset preannotation or preprocessing step is described below. This step will be applied for removing ambiguity and redundancy from whole dataset.

3.2.1. Tokenization and normalization

Splitting the sentence into a word list is called a tokenization process. Each token is called a word. For example: "আচারের সংযোজন খুব ভালো ছিল।" [The addition of the pickle was very good], after tokenize this sentence it will create a list, as like ["আচারের" [pickle], "সংযোজন" [addition], "খুব" [very], "ভালো" [good], "ছিল" [was]]. While doing tokenization process we have also finished normalizing the data. Normalizing means removing characters [" , ", " . " , "!", "@", "#", "%"], etc. these and stop words [28] from the sentence. The characters nad stop word will no impact on creating training, test data, and machine learning model construction.

3.2.2. Stemming

Stemming means originating the root word from the given word list after doing tokenization process. In Bangla language removing "র", "এর", "গুলি", "গুলো", "টার", "টি" etc., words from sentence during

Table 3 Showing weighted list of Adjective, Adverb word dictionary.

Types	Example	Weight	Total Word
High	"সবচাইতে" [Most of all], "সর্বাধিক" [greatest], "যথেষ্ট" [enough], "অতিশয়" [too much] … }	3	18
Medium	"অধিক" [more than], "বেশী" [more], "অনেক" [lots of] … }	2	15
Low	"অতিশয়" [at least], "সামান্য" [a little] "প্রায়" [nearly] … }	0.5	20

Table 4 Showing weighted list of conjunction word dictionary.

Categories	Example	Weight	Total Word
Coordinating Conjunction	{"কিন্তু" [but], "আদপে" [in fact], "এবং" [and], "অথবা" [or], "বরং" [or] … }	2	25
Subordinating conjunctions	{"অধিকন্তু" [Furthermore], "বিশেষত" [in particularly], "এমনকি" [even], "এসত্তেও" [despite of] …	1.5	12

stemming process. For example: "স্বাধীনতার" [Independent], "বাংলাদেশের" [Bangladesh], "দুর্বলতাগুলির" [Weaknesses] words convert the root word into respectively "স্বাধীনতা", "বাংলাদেশ", "দুর্বলতা" by stemming process.

3.2.3. Parts of speech (POS) tagger

Detecting the word pos tagger in a sentence have a great significance calculating the score. Our Bangla text sentiment algorithm require pos tagger to find out word weighted value from LDD. For example, "এটি খুব বেশি চিত্তাকর্ষক এবং খুব সুস্বাদু নয়।" [It's not too impressive and not too tasty.]. After generating in python pos tagger, we will get a list of word with POS [এটি_PR, খুব_RB, বেশি_JJ, চিত্তাকর্ষক_NN, এবং_CC, খুব_RB, সুস্বাদু_VB, নয়_NA. Here, বেশি_JJ] word quantify the word [চিত্তাকর্ষক_NN], therefore it will amplify the score of the text and [এবং_CC] word connects two sentence which will be tracked by our BTSC algorithm.

4. BTSC ALGORITHM

4.1. Discussion of Algorithmic Pseudocode

This section discusses the proposed novel BTSC algorithm which has a total of 30 steps. This BTSC algorithm is unique for Bangla and for any other language. Out of the 30 steps, the unique steps are from step 11 to 26 that manages the POS conjunction, adjective, adverb, punctuation, and question marks (QMs). This Algorithm 1 termed as *BTSC* which is used for generating score from the text. The inputs, notations, output, and pseudocode are below:

Inputs & Notations:

DD: Dataset Dictionary

LDD: Lexicon Dataset Dictionary [Active (Score = +1) & Contradict word (Score = −1)].

JJ / RB: Adjective or Adverb Word Quantifier Dictionary (3 types in dataset) [HIGH = 3, MID = 2, LOW = 0.5]

CC: Conjunction Type Pos Tagger, ***CD***: Co-ordinating & ***CS***: Subordinating Conjunction Word

POS: Parts of Speech, ***PR***: Pronoun, ***VB***: Verb, ***NN***: Noun, ***RB***: Adverb Type POS Tagger Word

Algorithm 1: Bangla Text Sentiment Score Calculation (BTSC)

1: **for** each *Sentence*[*i*] in Dataset **do**
2: **for** each *Tokenize*(*word*[*j*]) in Sentence **do**
3: Remove(TP, PR)
4: Scanning List of Word[*j*] from LDD
5: **if** *word*[*j*] is Active word in LDD **then**
6: *SC*[*Word*[*j*]] = +1
7: **else if** *word*[j] is Contradict word in LDD **then**
8: *SC*[*Word*[*j*]] = −1
9: **else if** *word*[j] is a negative word in LDD **then**
10: *k* = *k* + 1
11: **else if** *word*[*j*] is a *CC* type of POS tagger **then**
12: **if** *CD* type *word*[*j*] occurs in a sentence **then**
13: *SC*[*Word*[*j*]] = +2
14: **if** *CS* type *word*[*j*] at the beginning of a sentence **then**
15: *SC*[*Word*[*j*]] = +1.5
16: **else if** *word*[*j*] is a *JJ/RB* POS tagger **then**
17: *SC[Word*[*j*]] = explore in *JJ/RB* type of POS tagger in DD to get *word*[*j*] score value
18: **else if** *PU* occurs at the *Sentence*[i] **then**
19: **if** *word*[*j* − 1] *of PU is a Contradict word in LDD* **then**
20: *SC*[*Word*[*j*]] = −2
21: **else if** *word*[*j* − 1] *of PU is a VB type of POS tagger* **then**
22: *SC*[*Word*[*j*]] = −1
23: **else if** *QM* towards the end *Sentence*[*i*] and *word*[*j* − 1] of *QM* is a *VM* type POS **then**
24: *SCS*[*Sentence*[*i*]] = −1
25: **break**
26: **else**
27: $SCS[Sentence[i]] = (-1)^{k*}$ [$SC[Word[j]] * SC[Word[j+1]] \ldots * SC[Word[j_n]]]$
28: **end for**
29: $SCS = SCS[Sentence[i]] + SCS[Sentence[i+1]] + \ldots + SCS[Sentence[i_n]]$
30: **end for**

PU: Punctuation (!) Character, ***QM***: Question Mark (?) Character

TP: Transitional Preposition Word, *k*: count of negative word (initial, k = 0), *SC*[*Word*]: Score of a Word

Output:

SCS: Score(*SC*) calculation form a Sentence (per sentence by sentence)

1. If SCS is > 0, Sentence polarity is Positive.
2. If SCS is = 0, Sentence polarity is Neutral.
3. If SCS is < 0, Sentence polarity is Negative.

4.2. Score Calculation

To demonstrate our Algorithm 1: *BTSC*, we have considered here five examples form the both dataset. We consider five tables for simulating our examples score, besides showing each word scores, English translations, POS tagger, and total final score. These tables are formatted below according to algorithm 1.

Ex 01: "শুধুমাত্র রান্নাই যে সেরা তা নয়, সেবা সবসময় মনোযোগী এবং ভাল হয়েছে।" [Not only is cooking great, the service has always been attentive and good.]

Ex 02: "বাংলাদেশের ব্যাটিং বিপর্যয়।। ভালো লক্ষণ নয়।।" [Bangladesh's batting disaster. Not a good sign.]

Ex 03: "সময় বাংলাদেশের ভাগ্যে ড্র রেখেছে, নিশ্চয়ই হার ছাড়া উপায় ছিলোনা!" [Time has left a draw for the fate of Bangladesh, of course there wasn't a way without a defeat!!]

Ex 04: "খুব সীমিত আসন আছে এবং ঠিক সময়ে খাদ্য পাওয়ার জন্য যথেষ্ট অপেক্ষা করতে হবে।" [There are very limited seats and you have to wait long enough to get food on right time].

Ex 05: "টেস্টে মাশরাফির কি দোষটা ছিল? সে তো টেস্ট খেলেই না । এখানেও তাকে টেনে আনতে হবে?" [What was Mashrafe's fault in the test? He doesn't even play Tests. Should he be dragged here too?]

4.3. Simulation of BTSC Algorithm

The input for our Algorithm 1 a list of sentences considered in the dataset. Line 1 in the algorithm considers each text or sentence whose score will be calculated. At lines 2 and 3, tokenizing sentences along with stemming, removing transitional preposition (TP) word, parts of speech (POS) tagging processes are performed. Here TP word, i.e., "শুধুমাত" [only], "না হয়" [or else], "না তো" [not at all], "তা নয়" [not that], "সেইজন্য" [that's why], "তবুও কেন" [yet why] have not any significance in Bangla sentence for calculating score. At line 4, we scan every preprocessed word in each sentence from the LDD. With the help of LDD, we have found the weight score values of each active and contradict words at lines 5 to 8. As "না", "নেই", "নাই" [not] are negative dictionary word in the LDD, k counter is automatically incremented at line 9 to 10. At lines 11 to 17, the POS conjunction, adjective and adverb are managed. The rules for punctuation and QM characters are set at lines 18 to 24. The sentiment of a single sentence is calculated by multiplying each word score at line 27, and the total polarity of a whole paragraph score is calculated at line no 29 by adding each sentence score.

A number of examples are shown to demonstrate the BTSC algorithm. In the first example shown in Table 5, "মনোযোগী" [attentive], "সেবা" [service] are active words and "বিপর্যয়" [disaster], "হার" [defeat] are negative words in LDD, those word score are calculated at lines 6 and 8, respectively. In this case there is one sentence and the score value is the multiplication of individual scores resulting in (+2). Now, example 2 is demonstrated in Table 6 from the cricket dataset. Here are two sentence, first (i = 1) sentence ["বাংলাদেশের ব্যাটিং বিপযর্য়" [Bangladesh's batting disaster]] score is (−1) and second (i = 2) sentence ["ভালো লক্ষণ নয়" [Not a good sign]] score is (−1) and final total score of this phrase is (−1) + (−1) = (−2) which is calculated at line 29. In the third example as shown in Table 7, one word "নিশ্চয়ই" [of course], is a CC (CS type word) POS tagger and this score value is obtained from lines 14 to 15. There is a contradict word, i.e., "ছিলোনা" [was not] before the punctuation(!) character, the score of this word is calculated at line 21 to 22. In this case there is only one sentence and the score is (−3).

In example 4 shown in Table 8 there is one sentence, "এবং" [and] is a CC (CD type word). This is calculated at lines 11 to 13. There

Table 5 Showing Ex: 01 score calculation.

List of Word	রান্না	সেরা	সেবা	মনোযোগী	এবং	ভাল	Final Score
English Translation	cook	great	service	attentive	and	good	(+2)
Word Score Value	(+1)	(+1)	(+1)	(+1)	(+2)	(+1)	

Table 6 Showing Ex: 02 score calculation.

List of Word	বাংলাদেশের	ব্যাটিং	বিপর্যয়	ভাল	লক্ষণ	নয়	Final Score
English Translation	Bangladesh	batting	disaster	good	sign	not	(−2)
Word Score Value	(+1)	(+1)	(−1)	(+1)	(+1)	(−1)	

Table 7 Showing Ex: 03 score calculation.

List of Word	বাংলাদেশ	ভাগ্য	ড্র	নিশ্চয়ই	হার	উপায়	ছিলোনা	Final Score
POS Tagger	NN	NN	NN	CC	NN	VB	VB	
English Translation	Bangladesh	fate	draw	of course	defeat	way	was not	(−3)
Word Score Value	(+1)	(+1)	(−1)	(+1.5)	(−1)	(+1)	(−2)	

Table 8 Showing Ex: 04 score calculation.

List of Word	খুব	সীমিত	আসন	এবং	ঠিক	খাদ্য	পাওয়া	যথেষ্ট	অপেক্ষা	Final Score
POS Tagger	RB	NN	NN	CC	VB	NN	VB	JJ	VB	
English Translation	very	limit	seat	and	right	food	get	enough	Wait	(−18)
Word Score Value	(+3)	(−1)	(+1)	(+2)	(+1)	(+1)	(+1)	(+3)	(+1)	

Table 9 Showing Ex: 05 score calculation.

List Of Word	টেস্ট	মাশরাফি	দোষ	ছিল	খেলা	না	টেনে	আনা	হবে	Final Score
POS Tagger	NN	NN	NN	VB	NN	NA	VB	VB	VB	
English Translation	test	Mashrafe's	fault	was	play	not	dragged	Should		(−3)
Word Score Value	(+1)	(+1)	(+1)	(+1)	(+1)	(−1)	(−1)	(+1)		

is "যথেষ্ট" [enough] as adjective (JJ) and "খুব" [very] as adverb (RB) quantifier POS tagger word. The score of the words is calculated from line 16 to 17. The final score is calculated as (−18) after multiplying individual scores.

In example 5 shown in Table 9, there are three sentences. A QM occurs at the end of the first (i = 1) ["টেস্ট মাশরাফির কি দোষটা ছিল?" [What was Mashrafe's fault in the test?]], and there is a "ছিল" [was] VB type POS tagger before a QM. So, this sentence has negative meaning due to the presence of a QM after VB POS tagger. The score is of the first sentence is (−1). The score of the second sentence is (−1) executed at lines 5 to 10. In the third (i = 3) sentence ["এখানেও তাকে টেনে আনতে হবে?" [Should he be dragged here too?]] sentences there is a "হবে" [Should] VB type POS tagger before a QM, So, this sentence has negative meaning due to the presence of a QM after VB POS tagger. The score of this sentence is (−1). The score of the first sentence and the third sentences are (−1) executed as lines 23 to 24. Finally, the total calculated score of the three sentences are summed as (−1) + (−1) + (−1) = (−3).

5. EVALUATION OF EXPERIMENTAL RESULT

After applying the BTSC algorithm on both dataset, we construct a confusion matrix (CM) based on positive, negative, and neutral polarity label showed on Tables 10 and 11. From Table 10, total 1067 and 398 comments is identified out of 1216 positive and negative comments at restaurant dataset. Similarly form Table 11, 547 and 1905 comments is identified out of 620 positive and negative comments at cricket dataset. From these both CM, it can be inferred that the BTSC rule-based algorithm has been able to detect sentiment fairly accurately except the neutral sentiments. Because of, the total dataset comments polarity are voted by categorical based. The maximum neutral data is manually generated above the aspect-based category. It means, a comments has a positive on x category and negative on y category or neutral on z category.

Consider that example from Table 12: row 1 is taken from restaurant dataset that has three categories on three polarities and row 2 is taken from cricket dataset that has two categories on two polarities. But BTSC algorithm only extracts sentiments according to a global extended lexicon dictionary not to use in categorical sentimental dictionary. For this reason, neutral sentiments checking will be difficult to check.

Table 10 Polarity detection by Bangla Text Sentiment Score (BTSC) on restaurant data.

True\Predicted		Predicted Label			Total
		+1	−1	0	
True Label	+1	1067	140	9	1216
	−1	74	398	6	478
	0	242	63	60	365
Total		1383	601	75	2059

Table 11 Polarity detection by Bangla Text Sentiment Score (BTSC) on cricket data.

True\Predicted		Predicted Label			Total
		+1	−1	0	
True Label	+1	547	67	6	620
	−1	186	1905	17	2108
	0	61	39	151	251
Total		794	2011	174	2979

From Tables 10 and 11, we calculate some parameter measurements named as true positive rate (***TPR***), true negative rate (***TNR***), and false positive rate (***FPR***). In order to keep consistency with the relevant literature of NLP [29,30], we have used ***TPR*** which is also known as recall or sensitivity indicated as Equation (1) below. It is measured by the ratio of the true positive (***TP***) of a particular label to the sum of its ***TP*** and false negative (***FN***). ***TNR*** is also known as specificity which is measured as the ratio of true negative (***TN***) of a particular label to the sum of the ***TN*** and false positive (***FP***), shown at Equation (2). ***FPR*** is known as type II error which is calculated by the ratio of the ***FP*** of a particular label to the sum of the ***FP*** and ***TN***, shown at Equation (3).

$$TPR(\ label\ ) = \frac{TP}{TP + FN} \tag{1}$$

$$TNR(\ label\ ) = \frac{TN}{TN + FP} \tag{2}$$

$$FPR(\ label\ ) = \frac{FP}{FP + TN} \tag{3}$$

These formula are extracted by having a sharp concept on ***TP***, ***TN***, ***FP***, ***FN***. Basically, ***TP*** is correctly predicted class, ***TN*** is correctly predicted non-class, ***FP*** is incorrectly predicted class, and ***FN*** is

Table 12 Problems of not being able to detect neutral data.

No.	Comments	Category	Polarity	Bangla Text Sentiment Score (BTSC) Algorithm Polarity
1	"পরিমিত থাই খাদ্য - যদিও একটি নরম টুকরা - সামান্য~ ঘুরা ফিরা, কিন্তু সেবা ভাল ।" Moderate Thai food—although a soft piece—turns slightly, but the service is good.	Food Service Ambience	Positive Negative Neutral	Positive
2	"বোলিং পিচ তবে আমাদের ব্যাটসম্যানদের আউটগুলো আত্মহত্যা ছাড়া আর কিছুই নয়" Bowling pitch, but the batsmen outs are nothing but suicide.	Batting Neutral	Negative Bowling	Negative

Table 13 $C(x, y)$ Notation for indicating the parameters of confusion matrix (CM).

True\Predicted		Predicted Label		
		+1	−1	0
True Label	+1	C(1, 1)	C(−1, 1)	C(0, 1)
	−1	C(1, −1)	C(−1, −1)	C(0, −1)
	0	C(1, 0)	C(−1, 0)	C(0, 0)

incorrectly predicted nonclass. Before calculating Equations (1–3), we need to consider a Table 13 for stepping out these formulas. Here, ***C(x,y)*** notation for each box is introduced to measure the parameters for CM. In ***C(x,y)***, ***x*** is a predicted label or class and y is a true label or class. Calculation of ***TPR***, ***TNR***, and ***FPR*** for negative label (−1) is shown below at Equations (4–6).

Here Considering for negative labels,

$$TP = C(1,1)$$

$$TN = C(1,1) + C(0,1) + C(1,0) + C(0,0)$$

$$FP = C(1,1) + C(1,0)$$

$$FN = C(1,1) + C(0,1)$$

Finally we get from Equations (1–3) respectively

$$TPR(1) = \frac{C(1,1)}{C(1,1) + C(1,1) + C(0,1)} \tag{4}$$

$$TPR(1) = \frac{C(1,1) + C(0,1) + C(1,0) + C(0,0)}{C(1,1) + C(0,1) + C(1,0) + C(0,0) + C(1,1) + C(1,0)} \tag{5}$$

$$TPR(1) = \frac{C(1,1) + C(1,0)}{C(1,1) + C(1,0) + C(1,1) + C(0,1) + C(1,0) + C(0,0)} \tag{6}$$

Similarly, other labels will be calculated in this way. Full Summary of the calculation is showed at Figure 2. In these measurements, ***TPR*** is above average 85% which signifies our dictionary and BTSC algorithm efficacy. At most 90% ***TPR*** at negative label is obtained in the restaurant dataset and 87% ***TPR*** at positive label is obtained in cricket dataset. As high rate of *TPR* have the low rate of ***TNR*** however in better performance both will be high is preferable. ***TNR*** and ***TPR*** is better on positive and negative labels but not in neutral dataset because of having categorical identification polarity. In neutral comments 60% and 16% ***TPR*** is carried in cricket and restaurant dataset respectively.

Having a higher value of precision and recall vindicates a good model. Precision is measurement of the accuracy with respect to the prediction of a specific label or class. It is measured by the ratio of ***TP*** of a particular label or class in the sum of ***TP*** and ***FP***, indicates on Equation (7). F1-score is a combined formula of precision and recall showed at Equation (8).

$$Precision(label) = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - score\,(\,label\,) = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

To evaluate our experiment, we used supervised machine learning classification algorithm to classify our data. The evaluation of our result is measured through a CM including classifier metrices called accuracy, precision, recall, and F1-Score with the help of using Spyder, python IDE environment. Among the classifier, SVM with linear kernel trick (c = 1) is the best for giving proper result in new observations because SVM has found better accuracy in finding text classification.

At least 20% dataset have been randomly chosen for testing dataset and rest of the data is trained for classifying the polarity. A standard feature matrix called term frequency-inverse document frequency (Tf-Idf) vectorizer is used to calculate the feature matrix. It maps text or word into a significant representation number.

Since we have used BTSC algorithm to calculate our sentence score, Table 14 shows algorithm results in classifying polarity with much expected accuracy. Around 78% accuracy in both cricket and restaurant dataset is achieved on UniGram model. Having a multi class CM, we use weighted average to define our metrices, because marco and micro average cannot give proper result on same number of instances. As weighted average precision in restaurant is 78% and 80% in cricket, our extended Bangla sentiment dictionary construction is quietly proved in both score and polarity determination.

Here Figure 3a and 3b show the percentage of classifying polarity during SVM classification. At Figure 3a, at most 62.86% positive, 14.08% negative, and 0.97% neutral comments are identified as ***TP*** during SVM classification. Here ***FP*** is much lower than the ***TP***. Total 4.37% positive and 3.7% negative comments are incorrectly identified those class, which has lower ***FP*** than ***TP***. At Figure 3b, at most 61.58% negative, 15.60% positive, and 1.51% neutral comments are identified as *TP*. Total 4.37% positive and 3.7% negative

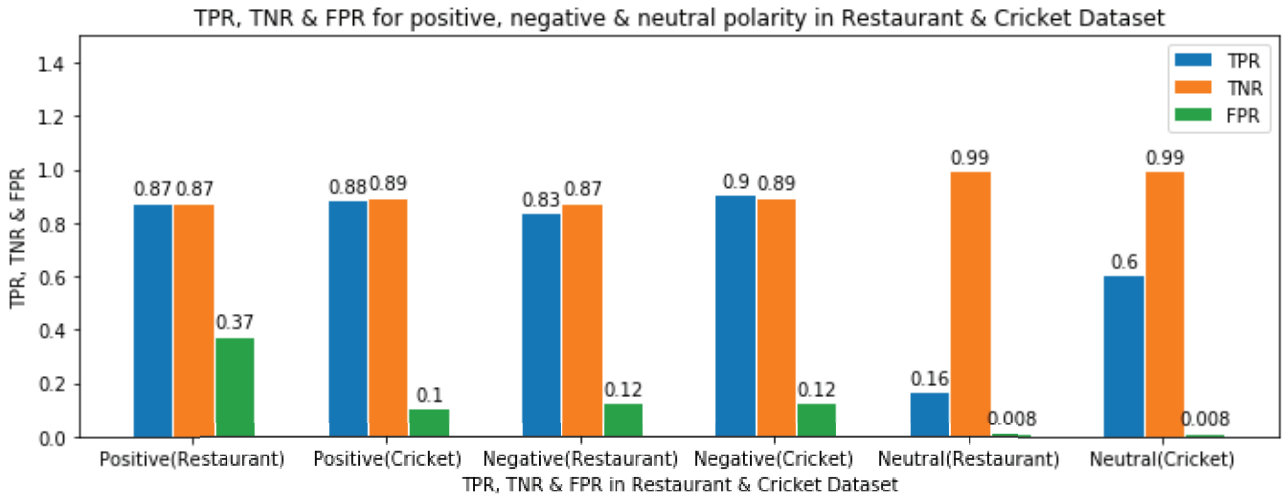
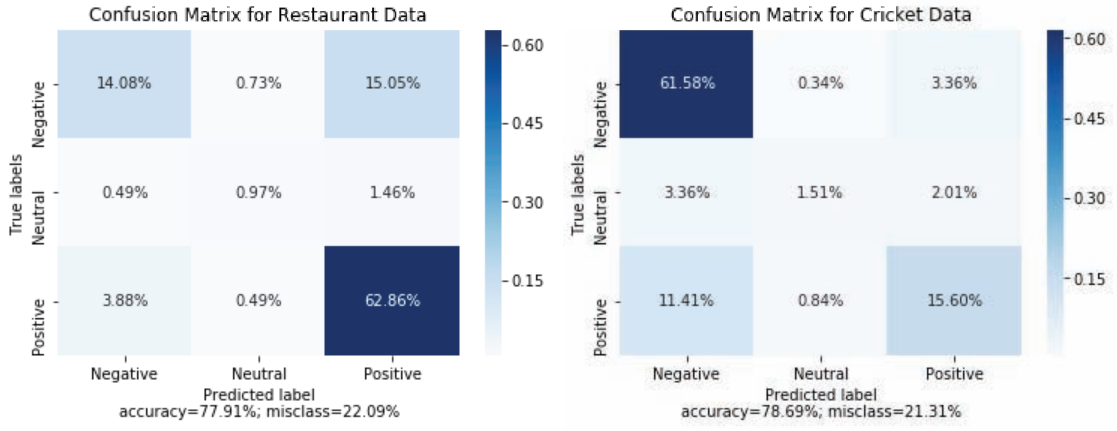


Figure 2 | Visualization performance of Bangla Text Sentiment Score (BTSC) algorithm in restaurant & cricket dataset.

Table 14 | Showing weighted average of precision, recall, F1-score & accuracy in UniGram model for both dataset.

Dataset	Polarity	precision	recall	f1-score	Accuracy	Support	Feature Matrix	No. of Feature Word
Restaurant	−1	0.76	0.47	0.58	77.91%	123	UniGram	3454
	0	0.44	0.33	0.37		12		
	+1	0.72	0.90	0.80		259	BiGram	6673
	Weighted Avg.	0.78	0.77	0.76	Total	412		
Cricket	−1	0.80	0.94	0.86	78.69%	389	UniGram	3751
	0	0.56	0.21	0.30		41		
	+1	0.74	0.56	0.63		166	BiGram	12854
	Weighted Avg.	0.80	0.78	0.74	Total	596		



(a) Polarity Prediction on Restaurant Data (b) Polarity Prediction on Cricket Data

Figure 3 | Bangla Text Sentiment Score (BTSC) algorithm polarity prediction on both data.

comments are incorrectly identified those class, which has much lower ***FP*** than ***TP***.

Besides, other classifier like LR, K-nearest neighbors (KNN), random forest (RF) algorithm is applied on our UniGram model. Among these classifiers, SVM shows better accuracy. Figure 4a and 4b depicts the performance of different classifier. At Figure 4a, we have achieved best accuracy 77.91% and precision 78.61% at restaurant dataset and at Figure 4b, 78.69% accuracy and 80% precision is achieved in cricket dataset in SVM classification. Both dataset have much better accuracy and precision rather than other classification.

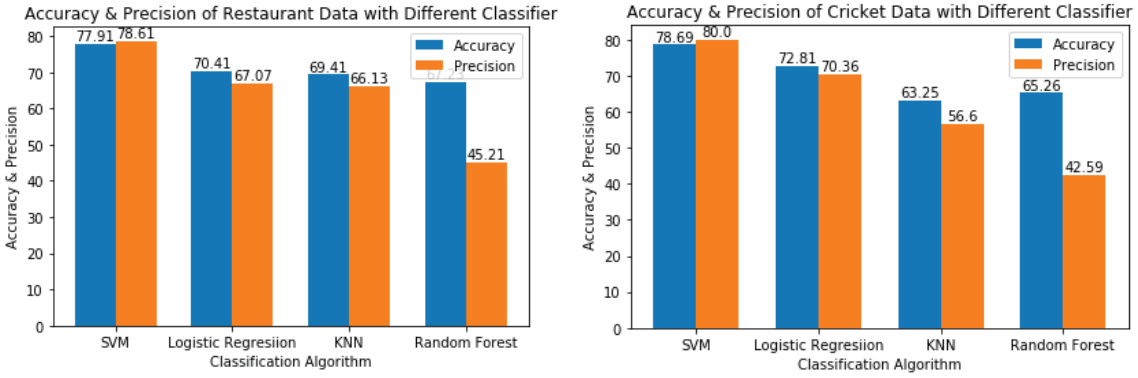
After finding quite improvement in UniGram approach in Tf-Idf model, we created another model BiGram in Tf-Idf word vectorization. In this model we performed Linear SVM classification

algorithm, finally accuracy is attained in both dataset 80.58% and 82.21% respectively which are greater than UniGram approach and also having precision 80.92 and 81.64 in both dataset. Figure 5 shows the performance and summary of the SA of UniGram and Bigram model. This analysis shows that cricket data has higher accuracy than restaurant dataset, because cricket dataset has trained more data than the restaurant data.

Figure 6 shows a comparative summary of our results with previous studies, although the comparison is not fair because of the use of varying datasets. The dataset used in this study is unique compared to other research works. The study in [16] achieved 69% accuracy when trained on 1000 tweets in UniGram with negation features. The study set only one rule to specify the sentiment from text by counting only positive and negative words from tweets. The limitation of [16] is that the use of only one rule which cannot properly detect the whole sentiment form the text.

In Islam *et al.* [18], a precision value of 77%, a recall/TPR value of 68%, and a F1-socre of 72% were achieved. The authors in [18] manually normalized the Bangla text with the help of valence shifting words by detecting one adjective in a sentence. However, the study did not consider the complex and compound sentences. The study in [20] trained only 850 and tested 200 texts in RF classification, achieving 85% accuracy for positive and negative data, however, the volume of training dataset is small. The study determined sentiments through only assigning feature words to positive and negative tags without considering POS tagger. In a recent paper [31], 80.48% accuracy was attained during 6-fold cross validation approach in multinomial Naive Bayes classification. The authors used polarity from given dataset as a target output without generating any sentiment from texts. This means the study did not apply any semantic connections to text and polarity.

However, our UniGram and BiGram features have higher accuracy with precision, recall/TPR and F1-socre than the abovementioned previous works. Moreover, our Unigram and Bigram both feature matrix have included stemming, normalization, POS tagger process. The dataset used in our study is much larger than the other studies shown in Figure 6. Still our results are comparable with others and thus acceptable.



(a) Performance of different classifier in Restaurant Data

(b) Performance of different classifier in Cricket Data

Figure 4 Visualization performance of different classifier in restaurant and cricket dataset.

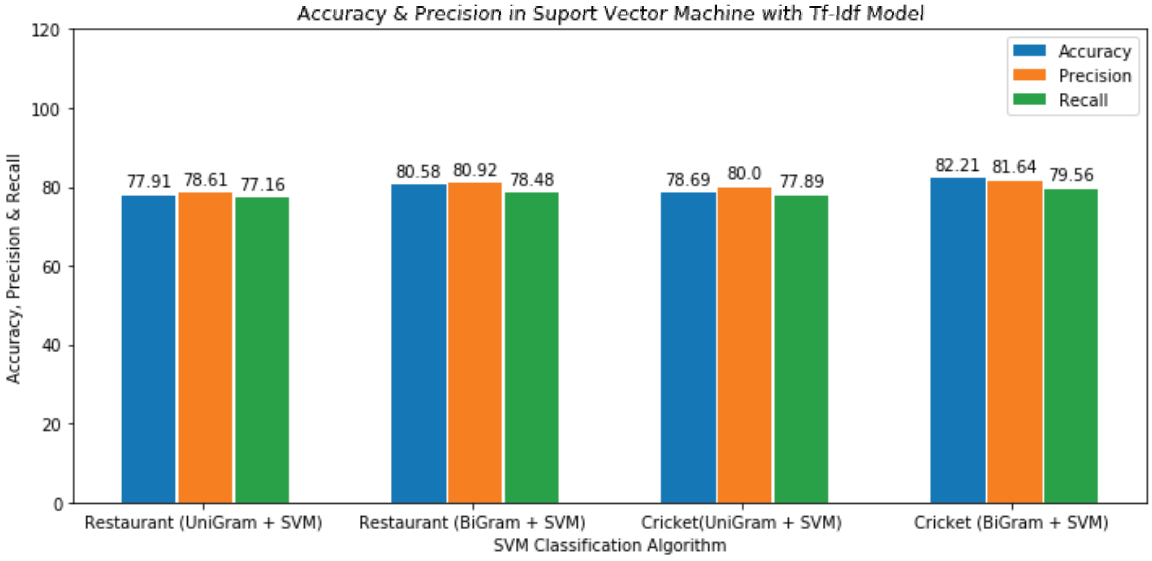


Figure 5 Visualization performance of UniGram and BiGram with SVM classifier.

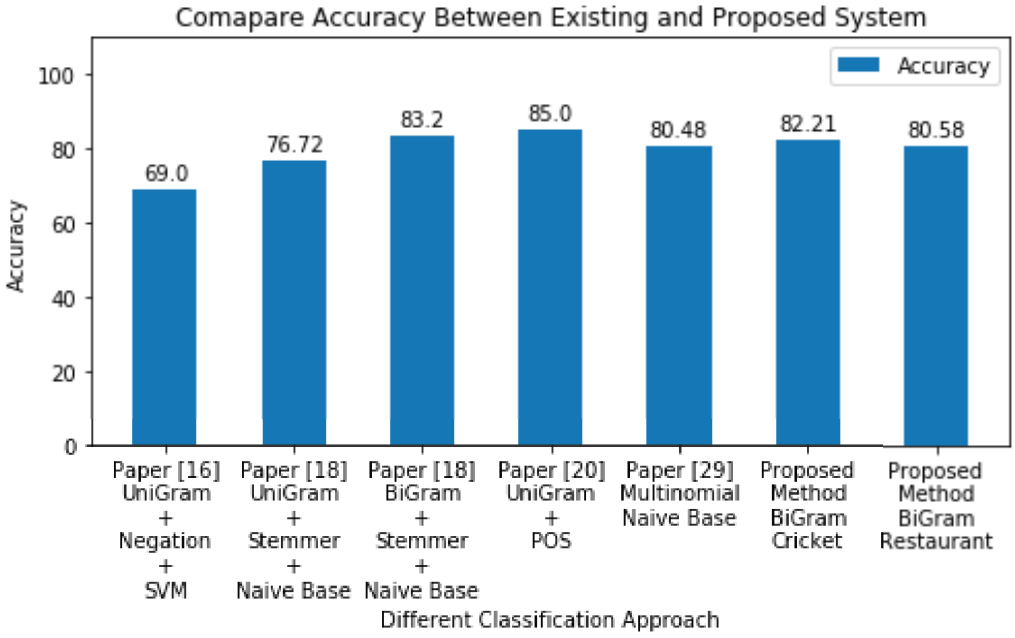


Figure 6 | Comparing accuracy between the existing and proposed systems.

6. CONCLUSION

With the rapidly growing of Internet users, SA depends on the dataset of particular content. Lexicon-based extended data dictionary is developed based on a specific domain restaurant and cricket. Manual construction of positive and negative dictionary with weighted value still a complex while mining data form Bangla dataset. However, precise observation on these data will give more accurate result in classifying polarity. In this paper, the BTSC algorithm detects the three types of polarity from the sentences using Bangla extended dictionary approach. Since a document belongs to more than one category, any rule-based algorithm is required for categorical specific domain-based data to detect text category and classification. We achieved the highest result of 82.21% accuracy in cricket on BiGram feature matrix. In CM, identifying neutral data has been less performed than the other two polarities. Every dataset has its owned variabilities. If we use i.e., fifty (50) thousand dataset in our machine learning process, our result will predict more accuracy than the obtained accuracy with the current dataset. For this, we need to construct a huge volume of sentimental dictionary. In future we will apply more datasets on our method. In our approach, approximately five thousand data is used as sentimental dictionary. Moreover, there is still a scope to redefine the weights of the dictionary. To make this approach even more significant, we introduce a categorical based data dictionary which will play a very pioneering role in further research.

REFERENCES

[1] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers LLC, 2012.

[2] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of the International Conference on Language Resources and Evaluation (LREc), Valletta, Malta, 2010, vol. 10, pp. 1320–1326.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2011), 267–307.

[4] E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12 (2009), 526–558.

[5] X. Liu, W.B. Croft, Statistical language modeling for information retrieval, Ann. Rev. Inf. Sci. Technol. 39 (2005), 1–31.

[6] Wikipedia, Bengali language. https://en.wikipedia.org/wiki/Bengali_language

[7] S. Buddeewong, W. Kreesuradej, A new association rule-based text classifier algorithm, in 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), IEEE, Hong Kong, China, 2005, p. 2.

[8] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, N. Eight Street, Stroudsburg, PA, 18360, United States, 2002, vol. 10, pp. 79–86.

[9] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in Proceedings of the 12th International Conference on World Wide Web, Association for Computing Machinery, New York, NY, United States, 2003, pp. 519–528.

[10] M.A. Karim, Technical Challenges and Design Issues in Bangla Language Processing, IGI Global, 2013.

[11] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, X. Wu, Chinese text sentiment analysis based on extended sentiment dictionary, IEEE Access. 7 (2019), 43749–43762.

[12] N.A. Abdulla, N.A. Ahmed, M.A. Shehab, M. Al-Ayyoub, Arabic sentiment analysis: lexicon-based and corpus-based, in 2013 IEEE Jordan Conference on Applied Electrical Engineering and

Computing Technologies (AEECT), IEEE, Amman, Jordan, 2013, pp. 1–6.
[13] E.M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, M. Alkeshr, Effective method for sentiment lexical dictionary enrichment based on word2vec for sentiment analysis, in 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, Kota Kinabalu, Malaysia, 2018, pp. 1–5.
[14] S.A. Mahtab, N. Islam, M.M. Rahaman, Sentiment analysis on Bangladesh cricket with support vector machine, in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, Sylhet, Bangladesh, 2018, pp. 1–4.
[15] C.J. Hutto, 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, Ann Arbor, MI, USA, 2014.
[16] S. Chowdhury, W. Chowdhury, Performing sentiment analysis in Bangla microblog posts, in 2014 International Conference on Informatics, Electronics & Vision (ICIEV), IEEE, Dhaka, Bangladesh, 2014, p. 1–6.
[17] K.M.A. Hasan, M. Rahman, Sentiment detection from Bangla text using contextual valency analysis, in 2014 17th International Conference on Computer and Information Technology (ICCIT), IEEE, Dhaka, Bangladesh, 2014, pp. 292–295.
[18] S. Islam, A. Islam, A. Hossain, J.J. Dey, Supervised approach of sentimentality extraction from Bengali Facebook status, in 2016 19th International Conference on Computer and Information Technology (ICCIT), IEEE, Dhaka, Bangladesh, 2016, pp. 383–387.
[19] R.A. Tuhin, B.K. Paul, F. Nawrine, M. Akter, A.K. Das, An automated system of sentiment analysis from Bangla text using supervised learning techniques, in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE, Singapore, 2019, pp. 360–364.
[20] N. Tabassum, M.I. Khan, Design an empirical framework for sentiment analysis from Bangla text using machine learning, in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, Cox'sBazar, Bangladesh, 2019, pp. 1–5.
[21] Z. Shunxiang, Z. Wei, Y. Wang, T. Liao, Sentiment analysis of chinese micro-blog text based on extended sentiment dictionary, Future Gener. Comput. Syst. 81 (2018), 395–403.
[22] S. Akter, 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), IEEE, Dhaka, Bangladesh, 2016, p. 1–4.
[23] A. Rahman, Bangla absa datasets for sentiment analysis. 2018. https://github.com/AtikRahman/Bangla_ABSA_Datasets
[24] M. Rahman, E.K. Dey, Datasets for aspect-based sentiment analysis in Bangla and its baseline evaluation, Data. 3 (2018), 15.
[25] G. Hub, Adjective, নাম বিশেষণ. http://www.grammarbd.com/en-grammar/adjective
[26] G. Hub, Adverb, ক্রিয়া বিশেষণ. http://www.grammarbd.com/en-grammar/adverb
[27] S. Hossain, Bltk, the bengali natural language processing toolkit. 2020. https://pypi.org/project/bltk/
[28] N.L. Ranks, Bengali stopwords - ranks nl. https://www.ranks.nl/stopwords/bengali
[29] T. Traylor, J. Straub, N. Snell, Classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator, in 2019 IEEE 13th International Conference on Semantic Computing (ICSC), IEEE, Newport Beach, CA, USA, 2019, pp. 445–449.
[30] A.G. Vural, B.B. Cambazoglu, P. Senkul, Z.O. Tokgoz, A framework for sentiment analysis in Turkish: application to polarity detection of movie reviews in Turkish, in: E. Gelenbe, R. Lent (Eds.), Computer and Information Sciences III, Springer, London, England, 2013, pp. 437–445.
[31] O. Sharif, M.M. Hoque, E. Hossain, Sentiment analysis of Bengali texts on online restaurant reviews using multinomial naive bayes, in 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), IEEE, Dhaka, Bangladesh, 2019, pp. 1–6.

Nitish Ranjan Bhowmik was born in Bangladesh in 1993. He received the B.Sc.(Honors) degree in Computer Science and Engineering (CSE) from Dhaka University (DU), at Dhaka, Bangladesh. He is currently pursuing the master's degree in Information and Communication Technology(ICT) with the Bangladesh University of Engineering and Technology (BUET). His research interests include data minig, text mining, natural language processing, and machine learning.

Dr. Mohammad Arifuzzaman received the B.Sc. degree in Computer Science and Engineering from BUET (Bangladesh University of Engineering and Technology). He received Masters in Information and Telecommunication Studies, from Waseda University, Tokyo, Japan, in 2012 and Ph.D in Information and Telecommunication from the Waseda University, Tokyo, Japan, in 2015. He worked as a postdoctoral fellow at Memorial University of Newfoundland, Canada. His research interest is in AI and Data Science, Future Internet Architecture, Green Wireless and Sensor Communication. Currently, he is an Assistant professor at Department of ECE, East West University, Dhaka, Bangladesh. Before that, he worked as a faculty member of IICT, BUET. Dr. Arif published 60+ research papers in International Journals (including *IEEE Sensors Journal, IEEE Access Journal, IEEE Transactions on Instrumentation and Measurement, Orphanet journal of rare diseases*, etc.) and conferences (including *IEEE Globecom, ITU Kaleidoscope*, etc.). He achieved best paper award at ITU Kaleidoscope Conference, Cape Town, South Africa, 12–14 December 2011. Dr. Arif worked as a reviewer of different peer reviewed journals including *IEEE Sensors Journal, IEEE Communications Letters, IEEE Communications Surveys and Tutorials, IEEE Internet of Things (IoT) Journal,* etc. For more details, visit https://fse.ewubd.edu/electronics-communications-engineering/faculty-view/mazaman

M. Rubaiyat Hossain Mondal received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. He obtained the Ph.D. degree in 2014 from the Department of electrical and computer systems engineering, Monash University, Melbourne, Australia. From 2005 to 2010, and from 2014 to date he has been working as a faculty member at the Institute of Information and Communication Technology (IICT) in BUET. His research interests include wireless communications, optical wireless communications, image processing, bioinformatics, and machine learning. For more details, visit https://iict.buet.ac.bd/?page_id=106

M. S. Islam is serving in the Institute of Information and Communication Technology (IICT) of Bangladesh University of Engineering and Technology (BUET). He obtained his B.Sc. in Electrical and Electronic Engineering from BUET, Dhaka in 1989; M. Sc. in Computer Science and Engineering from Shanghai University, China in 1997 and Ph. D in Electrical and Electronic Engineering degree from BUET in 2008. He has designed, coordinated, and implemented various IT projects at national levels and published many research articles in peer-reviewed journals. His research interest includes wireless communication, networking, and cyber security. In his long academic career, he has supervised about 40 postgraduate students which results the solution of many real-life problems. For more details, visit https://iict.buet.ac.bd/?page_id=100