INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Supervised Machine Learning Algorithms for Sentiment Analysis of Bangla Newspaper

Sabrina Jahan Maisha
Department of Computer Science and Engineering
BGC Trust University Bangladesh
"BGC Biddyanagar" Chandanaish Chattogram, Bangladesh
Email: sjbm1996@gmail.com

Nuren Nafisa
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
Raozan, Chattogram, Bangladesh
Email: nurennafisa@gmail.com

Abdul Kadar Muhammad Masum
Department of Computer Science and Engineering
International Islamic University Chittagong
Kumira, Chattogram, Bangladesh
Email: akmmasum@yahoo.com

Submitted: 22/4/2021. Revised edition: 18/7/2021. Accepted: 19/7/2021. Published online: 15/11/2021
DOI: https://doi.org/10.11113/ijic.v11n2.321

Abstract*—We can state undoubtedly that Bangla language is rich enough to work with and implement various Natural Language Processing (NLP) tasks. Though it needs proper attention, hardly NLP field has been explored with it. In this age of digitalization, large amount of Bangla news contents are generated in online platforms. Some of the contents are inappropriate for the children or aged people. With the motivation to filter out news contents easily, the aim of this work is to perform document level sentiment analysis (SA) on Bangla online news. In this respect, the dataset is created by collecting news from online Bangla newspaper archive. Further, the documents are manually annotated into positive and negative classes. Composite process technique of "Pipeline" class including Count Vectorizer, transformer (TF-IDF) and machine learning (ML) classifiers are employed to extract features and to train the dataset. Six supervised ML classifiers (i.e. Multinomial Naive Bayes (MNB), K-Nearest Neighbor (K-NN), Random Forest (RF), (C4.5) Decision Tree (DT), Logistic Regression (LR) and Linear Support Vector Machine (LSVM)) are used to analyze the best classifier for the proposed model. There has been very few works on SA of Bangla news. So, this work is a small attempt to contribute in this field. This model showed remarkable efficiency through better results in both the validation process of percentage split method and 10-fold cross validation. Among all six classifiers, RF has outperformed others by 99% accuracy. Even though LSVM has shown lowest accuracy of 80%, it is also considered as good output. However, this work has also exhibited surpassing outcome for recent and critical Bangla news indicating proper feature extraction to build up the model.

***Keywords*—Bangla Newspaper, Document level, Natural Language Processing, Sentiment Analysis, Supervised Machine Learning Algorithm**

I. INTRODUCTION

To humans with the blooming technology, computer is more like embellishing as the magical lamp of Aladdin. With this intent, human efforts are directed towards computers being capable of perceiving the mechanism of natural language spoken or written just the way human beings does. The area of study dealing with such sort of topics is known as Natural Language Processing (NLP). Sentiment analysis (SA) has been evolved as one of the most dominating and pivotal field of researches in this epoch of information arte-fact (data-driven age of technology).

Literally, SA refers to the summing up of underlying fruitful information exploring through the millions and millions of documents such as reviews, opinions, news, interviews and so on [1]. Mining down this huge lump of documents manually and identifying manifested opinions by organizing data systematically can be both tiresome and labor intensive [2]. With a view to solving this problem, diving through one of the most important branches of NLP is SA [3].This area is a large-scale problem domain. The basic challenging task in this field is to understand the complex semantic structure of languages. Critical case arises to inspect through the ambiguous statements in which positive words may point out negative meaning or vice-versa.

Previously, organizations used to spend a lot of time and sometimes it has seemed to be intractable for analyzing the market demand or current trend. But in this era of technology, Bangla contents are generated in huge volume where opinions of intellectuals, mass people and the happenings around the globe

are banging our minds and thought process. Such kind of phenomenon is characterizing people's concepts and ideas. Extraction of these sentiments is indispensable for understanding the implied meanings of human compositions.

At present, starting from marketing to customer service in organizations, social media monitoring, political views analysis and many more realms of human lives are surpassed by the dole of SA. Due to the presence of lexical resources like stemmer, POS tagger and other required elements in well-resourced languages, enormous efforts rendering outstanding results has been witnessed. Whilst research works in Bangla language is still going on and a lot of prospects are yet to be unveiled.

Online news reaches each and every type of people which puts an impact in their minds and thoughts. Some news can make people anxious, happy and some may have in appropriate contents. The kind of sentiment being expressed in the news needs to be extracted so that future predictions of events, removal of inappropriate contents, trend analysis and many more applications can be done easily. This triggered to build up a system for one of the most important NLP tasks i.e.; sentiment extraction implementing on dataset of Bangla online news.

The proposed system aims to establish an approach to automatically extract sentiments illustrated in those texts on large number of online Bangla news. The key contributions in this work are to introduce SA for Bangla News with new online Bangla dataset. This will help to extract inappropriate news as positive and negative news will be extracted for different category of readers such as children, aged people, etc. Several supervised ML algorithms is applied to examine behavior of dataset by comparative analyses.

As far as our knowledge goes there is scarcity of dataset in Bangla having Bangla online news portal as the source for analyzing sentiments[4]. In this proposed work, the main objective is to contribute for developing a large volume of dataset for Bangla language and work forward to create a large volume of this dataset being making it publicly available for future use in other researches in this field. With a view to building up a Bangla news dataset of 7000 documents, news from online are collected and manually annotated into positive and negative sentiments.

Here, the work is intended to present SA on the Bangla online news portals through applying various prevalent supervised ML classifiers which are MNB, LSVM, LR, DT, K-NN and RF. The comparative analysis in this manner using different classifiers will illustrate the trend and behavior of dataset as well as underlying causes for such response can be perceived apparently.

This work has been structured in the successive manner into five sections. Section II represents the literature review related to our research work which embarks on giving an overview of the works that has been done so far on SA in different languages and also goes through the research works that has been done in Bangla language. Section III comes up with brief exploration of the methodology implemented in this work comprising discussion on dataset, data pre-processing techniques, training and testing dataset and feature extraction processes. Section IV depicts the experimental results with performance metrics along with graphical and tabular representation of obtained results. Section V shows the conclusion and future work summarizing the contributions of this proposed work and suggests some potential future works.

II. LITERATURE REVIEW

This section gives a comprehensible outline of the exploitation done so far with clarification about studies and works that paved the way for more future scopes and invoked the urge of revolutionary progress in this area of NLP.

Basically, SA research follows two types of techniques: (a) lexicon based or corpus based and (b) ML based approaches. Lexicon or corpus based technique is implemented creating a dictionary manually or automatically taking the document orientation, words and phrases into account [5]. Methodologies such as Hidden Markov Model (HMM), Single Dimensional Classification (SDC), Sequential Minimal Optimization (SMO), Conditional Random Field (CRF) and many other classification approaches are usually used for this technique. On the contrary, ML approach is the most popular and advanced methods for classification discipline. Algorithms such as Multi Layer Perceptron (MLP), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Maximum Entropy (ME) and many other traditional classifiers have been implemented which has exhibited promising results [6]. To understand and analyze the works done till now in English and other languages, we can look over models evaluated to extract sentiments in domain of document level analysis such as news, reviews or blogs.

A. Sentiment Analysis in English and Other Languages

Even though English and Bangla languages structurally differs from each other in terms of sentence formation and words [7], English being a resource rich language can help the researchers to understand the outline of studying in this arena. Most of the works in this field focused on analyzing sentiments on twitter based corpus. In 2019, Taj *et al.* [8] had experimented on BBC English news dataset derived from 5 different topics consisting of 2225 documents. Using TF-IDF technique and WordNet lexical database, sentiment score of whole document had been calculated as positive, negative or neutral by summing up individual word points. As similar words can be expressed differently in each context, this approach with limited words dictionary may not yield acceptable results for every other news or reviews dataset. The work of Alshamsi *et al.* [9] in 2020 mainly focused on analyzing behavior of ML classifiers in balanced and unbalanced datasets to extract sentiments from airline tweets. Although among 6 classifiers Naïve Bayes and ID3 showed good results with balanced dataset, size can greatly impact to build up a good model with ML classifiers. In the same year, Suhasini and Srinivasu [10] employed NB and K-NN classifiers to detect sentiment from tweets and classify them into 4 emotion categories. Through different performance metrics their proposed model revealed that NB performed well with an accuracy of approximately 73%.

Whereas most of the research works are done in English, languages having complex morphological structure i.e. Chinese, Arabic and some few languages studies in this arena have been increasing rapidly. Bangla being semantically different from Chinese and Arabic is similar in terms of having own alphabets and writing style. Analyzing such works can help to perceive the style of handling such language. Among them, in one of the notable studies in 2019, Gamal *et al.* [11] had performed

extraction of Arabic sentiments from in total 438,931 positive and negative tweets. Generating n-gram feature set, ten different ML classifiers such as Passive Aggressive (PA), Ridge Regression (RR), NB, SVM, Bernoulli NB (BNB), MNB, Stochastic Gradient Decent (SGD), LR, Maximum Entropy (ME) and Adaptive Boosting (Ada-Boost) were applied to train the model. Finally applying performance evaluation metrics, they found that RR and PA exhibited highest accuracy of 99.96% using n-gram features. Authors Miao *et al.* [12] in 2020 studied on Chinese news text adopted from Fudan University. To classify texts into nine categories this work analyzed the dataset through three classifiers (i.e. KNN, NB and SVM) and obtained best accuracy of 95.7% through SVM with TF-IDF technique. In 2017, Lommatzsch *et al.* [13] explored this field in German language. This work was more intended to develop well annotated new dataset having two types' of corpora including German news articles and messages of telecommunication forum. With about 2369 sentences, Multinomial Naïve Bayes (MNB) classifier having 10 fold cross validation had been evaluated on the corpus. Even though the model had achieved a good accuracy of 87.1%, strong biasness of neutral category has been observed.

Bangla is basically one of the languages originated from Indo-Aryan language. Other languages such as Nepali, Urdu, Hindi, Malayalam, Sanskrit and so on are also the various dialects derived from the same family language group as Bangla [7]. Very scarce amount of contributions exists in analyzing the sentiments of such under resourced languages. Going through a research study on Nepali language by Thapa and Bal [14], implementation of supervised ML classifiers on a size of 384 book and movie reviews has been observed. Extracting features by TF-IDF and Bag-of Words (BOW) technique, results unmasked that MNB performed well in comparison with SVM and LR. A study on Urdu blogs collecting resources from 14 different subjects has been done in 2017 by Mukhtar and Khan [15] . In this work, different techniques such as feature extraction, creation of balanced dataset and 10-fold cross validation are implemented on 151 blogs having 6025 sentences. Model showed that K-NN classifier outperformed among SVM and DT with 67% accuracy but it required more computational time. In a recent work of 2020 on 3184 Malayalam tweets, Soumya and Pramod [16] proposed a model by extracting features through TF-IDF, BOW, Unigram using Sentiwordnet along with and without negation words. Training the model with RF, NB and SVM, the experimentation concluded that RF obtained better results with 95.6% accuracy through Unigram including negation words with Sentiwordnet extraction technique. To interpret sentiments on the basis of sarcastic and non-sarcastic, Pawar and Bhingarkar [17] conducted a research on a dataset containing English as well as Hindi Tweets. Considering four sets of features, the model achieved highest results (81%) by RF among RF, SVM and KNN classifiers. Tweets contain raw unstructured text but it could not address all the possible patterns in the adopted work.

B. Sentiment Analysis in Bangla Language

With availability of a very few standard datasets and resources, more scope for contributing in Bangla language has been created. For this reason, more and more efforts are being put up in developing a standard dataset. Small attempts in this aspect have been observed in this field. In the paper of classification on sentence level, Hoque *et al.* [18] has been performed using doc2vec feature extraction technique with supervised machine learning classifiers. However, this approach on sentence level has achieved good accuracy of 77.72% by employing comparative analysis. A research work on Bangla document level classification by Islam *et al.* [19] has been on 31908 documents. Exploring the dataset with three calssifiers(i.e. SVM, NB and Stochastic Gradient Descent(SGD)), an accuracy of 92.56% with SVM and TF-IDF procedure had been obtained. Sarkar and Bhowmick [20] executed two popular machine learning techniques i.e. MNB and SVM with polynomial kernel on Bangla tweets of available dataset. This work basically achieved higher accuracy than the existing approach on this dataset. Al-Amin *et al.* [21] conducted six different approaches including some feature extraction techniques with machine learning classifiers to comparatively analyze performance. In a recent attempt made by Amin *et al.* [22] on 7500 Bangla sentences, two supervised ML techniques which are NB and Topical approach were applied. Though topical approach showed better accuracy with 90% in sentence level and document level clasiification, lackings of dataset smoothness were observed in the study.Another study of 2019 by Tabassum and Khan [23], 1050 Bangla texts were collected from Facebook and Twitter. Through classifying dataset into positive and negative sentiments by RF classifier and techniques such as unigram, pos tagging and negation, 87% accuracy had been acquired. Going through the previous researches, it can be seen that ML classifiers has shown promising results in extracting sentiments in other languages alike Bangla. So our proposed model aims to analyze sentiments from Bangla news dataset using pipeline startegy along with vectorizer, TF-IDF techniques and modeling with ML classifier.

III. METHODOLOGY

In this section, a brief overview of the overall research procedure is described beginning right from the collection of data to the implemented approach for the proposed system of extracting sentiments of Bangla news. The entire task is done using python programming language and its associated tools and techniques. The whole architecture of the work is illustrated through Fig. 1.

A. Dataset

Data is the most crucial aspect in acquiring the desired result implementing different tools and techniques. It must be collected in such a precise manner so that peeping through the window by this data a realistic view of present world would be visible to us clearly. As far as it could be possibly known by us, mere amount of Bangla news dataset is publicly available to be utilized in researches. For this reason, we have to build our own Bangla news dataset from 4 different Bangla online news portals which are: prothomalo.com, jugantor.com, bdnews24.com and a2i.pipilika.com. These sources of data are piled up encompassing the two pre-determined categories i.e.

- Positive News
- Negative News

The collected news from online Bangla news archive are further stored in text file format in database. Labeling is accomplished by incorporating text files to particular categorized folder manually and naming the folders as positive and negative.

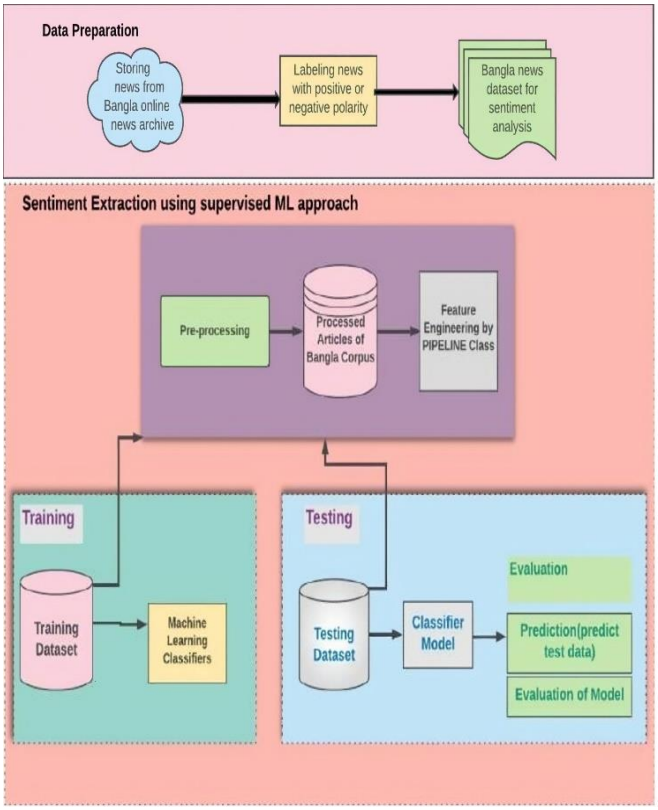

Fig. 1. System Architecture

This manual task is time consuming so we kept categorizing the news into positive and negative news gradually. For now, we have manually categorized 3500 news into positive and negative polarity. In Table I, the total statistics of positive and negative news data collection for the work are shown.

TABLE I. DATA STATISTICS

Types of News	Amount
Positive News	1700
Negative News	1800
Total News	3500
After augmentation	7000

To utilize the dataset of 3500 news efficiently and to introduce a new Bangla news corpus the dataset is doubled by copying the data it contained. Thus, our total we have 7000 data to use in our model. For effectiveness, we have shuffled the data using python sci-kit learn.

B. Data Pre-processing

The main objective of data pre-processing is to represent data as a vector space and removing non-informative features. Because of this procedure, feature dimensionality of vector space may reduce strikingly imparting the best possible outcome for sentiment analysis. The whole pre-processing steps are shown in Fig. 2.

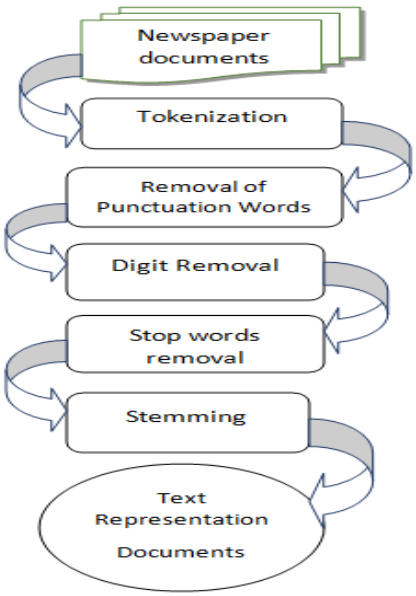

Fig. 2. Pre-processing steps

- Tokenization: At first, all the bunch of text is converted into words or texts. The words are then partitioned into linguistic segments or tokens by this process.
- Removal of Punctuation Words: Punctuation marks that makes up the contents of text documents are irrelevant and meaningless symbols in case of further analysis. (<,>, :, [,],^,&,*,(,) , | etc.) are the different types of punctuation that must be excluded in order to get a noiseless and clean dataset. White spaces, tabs and shifts are also eliminated in this step.
- Digit Removal: Bangla text documents may contain different English or Bangla numeric texts which has no
- significant meaning. This may affect the outcome negatively by increasing data redundancy so it has been removed in this process.
- Stemming: The list of words or unigrams that are separated from the input text document contains different prefixes and suffixes along with the root word or base word. This process deals with diminishing the inflected words into its root word. Noun suffixes are pointed out and analyzed to build up our corpus. In Table II, few words before stemming and after stemming has been shown in the left and right column respectively.

TABLE II. EXAMPLES OF FEW STEMMING OF WORDS

Derived Words	Root Words
দেশে	দেশ
পেঁয়াজের	পেঁয়াজ
স্থানগুলোয়	স্থানগুলো
মন্ত্রণালয়	মন্ত্রণালয়

- Stop Words Removal: Stop words basically comprise those streams of words that build up the morphology and enhance linguistic characteristics. The following Table III presents a list of stop words with their categories. It can be observed that words such as 'দিকে', 'আমি','খুব' etc. adds dimensions to the words but plays no significant role in overall sentiment extraction.

TABLE III. STOP WORDS REMOVAL PROCESS

Categories	Words Example
Postpositions	['দিকে' (dike) 'towards' 'সহ' (shôho) 'with', 'including']
Conjunctions	['এবং' (ebang) 'and', 'কিন্তু' (kintu) 'but']
Interjections	['বাহ' (bah) 'well', 'সাবাশ!' (shabash) 'bravo']
Pronouns	['আমি' (Āmi) 'i', 'তুমি' (Tumi) 'you']
Some adjectives	['ভালো' (bhalo) 'good', 'কৃষ্ণ' (krishno) 'black']
Some adverbs	['খুব' (khub) 'very', 'কালেভদ্রে' (kālēbhadrē) 'rarely']
All articles	['একটি' (ekti) 'one']
Proper nouns	['আগ্রাবাদ' (agrabad) 'Agrabad']

C. Training & Testing Data

For comparing between classification algorithms, we have used two different sampling methods: one is percentage split validation approach and the other is 10-cross fold validation.

1) *Percentage Split Validation Approach:* Subjecting to the first type of method the whole dataset is divided into two parts. One is used for the purpose of training and the other serves as the data for validation part. The model acquires knowledge by the training set of data and for testing data validation part is used. In the context of our approach, the dataset has been split into 70% training set and 30% validation set.

2) *Cross Validation Approach:* Evaluating a ML model can be quite tricky. Usually, the data set is split into training and testing sets . Then the training set and the testing set are used to train the model and to test the model respectively. Based on an error metric, model performance is evaluated to determine the accuracy of the model. This method however, is not very reliable as the accuracy obtained for one test set can be very different to the accuracy obtained for a different test set. K-fold Cross Validation (CV) provides a solution to this problem by dividing the data into folds and ensuring that each fold is used as a testing set at some point. For our proposed system, 10-fold cross validation approach has been used to split the dataset. K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point.

D. Feature Engineering using Pipeline Class

Our feature engineering is done by using a "Pipeline" class of scikit-learn that includes vectorizer, transformer and ML classifier. This model of pipeline class works like a composite classifier. Reason of using the pipeline class is given below:

- Reproducibility can be achieved
- Cross validation and other types of model selection becomes easier
- Leaking data from training sets into test sets can be vanished

1) *Feature Extraction and Selection:* In ML, "dimensionality" simply refers to the number of features (i.e. input variables) in dataset. When the number of features are very large relative to the number of observations in the dataset, certain algorithms struggle to train effective models. This is called the "Curse of Dimensionality," and it's especially relevant for clustering algorithms that relies on distance calculations. In our case, feature extraction and selection has been done by vectorizer and transformer of "Pipeline" class.

a) Vectorizer: Count Vectorizer is a technique that provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words. But it is also applied to encode new documents utlizing that vocabulary.An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. Due to having a lot of zeros in these vectors, it is called as sparse. With this method, every column is a term from the corpus and every cell represents the frequency count of each term in each document.

*b) Transformer:*The count matrix obtained from the vectorized form is transformed by TF-IDF transformer to a normalized TF-IDF representation. It is estimated following the system given below:

- Term Frequency (TF) = (Number of times term 't' appears in a document)/(Number of terms in the document)
- Inverse Document Frequency (IDF) = log(N/n), where, 'N' is the number of documents and 'n' is the number of documents a term t has appeared in. The IDF of a rare word is high, whereas the IDF of a frequent word is likely to be low. Thus, having the effect of dealing with words that are distinct in the document.
- Finally TF-IDF is calculated by:
  TF-IDF value of a term = TF * IDF

The first thing that can be done is improvisation of the vectorization step. In fact, there are some biases attached with only looking at how many times a word occur in a text. In particular, the longer the text the higher its features (word counts) will be.

To fix this issue, Term Frequency (TF) instead of word counts can be used. In this process, calculation is done through dividing the number of occurrences by the sequence length, which can be formulated as equation (1). These frequencies have to be further downscaled so that words that occur all the time (e.g., topic-

related or stop words) have lower values. This downscaling factor is called Inverse Document Frequency (IDF) and is equal to the logarithm of the inverse word document frequency, shown by equation (2).

Putting all these estimations together, these new features are called TF-IDF features [24]. The formula is shown as equation (3).

$$TF(word, text) = \frac{number\ of\ times\ the\ word\ occurs\ in\ the\ text}{number\ of\ words\ in\ the\ text} \quad (1)$$

$$IDF(word) = log\left[\frac{number\ of\ texts}{number\ of\ texts\ where\ the\ word\ occurs}\right] \quad (2)$$

$$TF - IDF(word, text) = TF(word, text) \times IDF(word) \quad (3)$$

2) *Supervised Machine Learning Classifiers:* Six classifiers: MNB, LSVM, LR, DT, K-NN and RF are used to train our dataset for predicting sentiments.Our main focus is to analyze the results delivered by different classifiers for our dataset.

IV. Experimental Results

In this section, details of the experiments with analysis have been provided.

A. Performance Measurement Tools

For simplifying our experiment popular performance measures are adopted depicting the recall, precision and F-measure (F1) with accuracy scores. Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by the proposed algorithm. Accuracy is the general term describing the capability of any approach to determine all results correctly on total number of given documents. Characterizing the equations for precision, recall and accuracy equations (4), (5) and (6) respectively are shown.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (6)$$

In general term, the average of both precision and recall which prioritizes the performance of any model in a better way is the F-measure or F1-score. It is basically the harmonic mean of precision and recall and can be equated as follows in (7).

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

B. Experimental Outputs

Results of SA at document level implementation using supervised ML approach shows that in both splitting techniques either in percentage or 10-fold cross validation RF classifier has achieved the highest accuracy of 98% and 99% respectively. On the other hand, LSVM has revealed least accuracy of nearly about 81% in both forms. To get a clear idea about the predicted model few sample input of positive and negative documents has been presented in the Table IV. This table shows the predicted outcome of the experimental results of supervised ML approach along with cross check showing the actual results. The results depicts that the model of supervised ML learning approach is good enough as they could correctly predict the test results.

TABLE IV. Examples of some sample input and Predicted output results for our approach of sentiment analysis in ML based on news Document

Input Data	Actual Value	Predicted Value
বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়ের (বুয়েট) শিক্ষার্থী আবরার ফাহাদকে পিটিয়ে হত্যা করা হয়েছে বলে অভিযোগ উঠেছে। গতকাল রোববার দিবাগত রাত তিনটার দিকে বুয়েটের শেরই বাংলা হলের নিচতলা থেকে তাঁর লাশ উদ্ধার করা হয়। [Abrar Fahad, a student of Bangladesh University of Engineering and Technology (BUET), has been allegedly beaten to death. His body was recovered from the ground floor of Sher-e-Bangla Hall of BUET at around 3 am on Sunday.]	Negative	Negative
২০০০ সালে টেস্ট মর্যাদা পাওয়ার পর এইপ্রথম ভারতের মাটিতে পূর্ণাঙ্গ সিরিজ, স্বাভাবিকভাবেই বিশেষ উপলক্ষ। সৌরভ গাঙ্গুলীর জন্যও এটি বিশেষ কিছু। ভারতীয় ক্রিকেট বোর্ডের বিসিসিআই) প্রধান হওয়ার পর এটিই ভারতীয় দলের প্রথম সিরিজ। দুই টেস্টেরএকটিও হবে তাঁর নিজ শহর কলকাত য়াদারুণ এই ব্যাপারটিকে স্মরণীয় করে রাখতে বাংলাদেশের প্রধানমন্ত্রী শেখ হাসিনাকে কলকাতার ইডেনের সেই টেস্টে আমন্ত্রণ জানিয়েছেন তিনি। [This is the first full series on Indian soil since gaining Test status in 2000, naturally a special occasion. It is also something special for Sourav Ganguly. This is the first series for the Indian team since becoming the head of the Board of Control for Cricket in India (BCCI). One of the two Tests will be in his hometown of Kolkata.]	Positive	Positive
বিশ্বজুড়ে করোনাভাইরাসে প্রাণহানির সংখ্যা ৯০ হাজার ছাড়িয়েছে। যুক্তরাষ্ট্রের জনস হপকিন্স ইউনিভার্সিটির তথ্য অনুযায়ী, বৃহস্পতিবার রাত ১১টা পর্যন্ত এই সংখ্যা ৯০হাজার ৫৭জন। এই সময় পর্যন্ত শনাক্ত হওয়া রোগীর সংখ্যা ১৫ লাখের বেশি। এর মধ্যে সুস্থ হয়েছেন ৩ লাখ ৪০ হাজার ৬৩০ জন। [The number of deaths from coronavirus worldwide has exceeded 90,000. As of 11 p.m. Thursday, the number was 90,056, according to Johns Hopkins University in the United States. The number of patients identified so far is more than 1.5 million. Of these, 3 lakh 40 thousand 630 people have recovered.]	Negative	Negative

Input Data	Actual Value	Predicted Value
নতুন করোনাভাইরাসের এই বিস্তারের সময়ে চারপাশ যেন দুঃসংবাদে ভরে উঠেছে।প্রতিটি মৃত্যু বুকের ওপর ভারী পাথর হয়ে চেপে বসছে।বিশ্বের যে প্রান্তেই মানুষ মারা যাকনাকেন,তা ব্যথিত করছে মানুষকে। [During this outbreak of the new coronavirus, there seems to be a flood of bad news around. Every death is a heavy stone on the chest. ? Wherever people die in the world, it hurts people.]	Positive	Positive

C. Ambiguous Input

To identify the system accurately, we need to consider some critical cases in this aspect. Both our positive and negative cases are being identified accurately. Although our dataset does not contain any neutral labeled sentiments, to understand the behavior of our model some neutral sentiments input have been tested on it. Table V gives a sample demonstration of such scenario where a sample dataset of neutral label has been used to analyze by our model. It is observed that our model has predicted the given "neutral" label document as "negative". The reason behind such kind of sentiment extraction is that our model has not considered the neutral label. So the model in such case choses the most relevant words of the document and measuring similarity classifies that document into the prevailing or considered tags. In our case, also same thing has happened. Seeing words like 'হত্যা' and other relevant negative words the whole document has been detected as negative.

TABLE V. AMBIGUOUS OUTPUT RESULT FOR SENTIMENT ANALYSIS

Input Data	Actual value	Predicted result by Supervised ML
বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়ের (বুয়েট) ছাত্র আবরার ফাহাদ হত্যা মামলায় মুজাহিদুর রহমান নামের এক আসামি আদালতে ১৬৪ ধারায় স্বীকারোক্তিমূলক জবানবন্দি দিয়েছেনাপুলিশের আবেদনের পরিপ্রেক্ষিতে ঢাকার মুখ্যমহানগর হাকিম আদালত এই আসামির জবানবন্দি রেকর্ডক রেনামুজাহিদু রইলেকট্রিক্যাল ইঞ্জিনিয়ারিং বিভাগের তৃতীয় বর্ষের ছাত্র। [Mujahidur Rahman, a student of Bangladesh University of Engineering and Technology (BUET), has given a confessional statement in the court under section 164. The statement of the accused was recorded by the Dhaka Chief Metropolitan Magistrate's Court on the plea of the police. Mujahidur is a third year student of Electrical Engineering Department.]	Neutral	Negative

D. Performance Statistics

After performing sentiment analysis on our dataset with six classifiers: RF, LR, DT, KNN, MNB and SVM, firstly the accuracy of the system is compared with respect to Precision,Recall,F1-Score and Macro Average on each case for each classifier as shown in Table VI. Secondly, our results are validated by two processes in Table VII showing that the order of accuracy in percentage split process is quite similar with 10-fold cross validation having minimal difference. Table VII exhibited that RF classifier achieved the highest accuracy of 98% among all the six classifiers. Moreover, it showed f1-measure of 97% for both positive and negative news. In terms of next highest accuracy, DT ranked in the second position with 95 % accuracy. In third ranking, LR showed an accuracy of 94%. Following this, KNN, MNB and LSVM obtained accuracies of 83%, 82% and 81% in respective order.

TABLE VI. CLASSIFICATION REPORT OF SIX CLASSIFIERS FOR SENTIMENT ANALYSIS USING ML APPROACH

Type of News	Classifiers of ML	Precision %	Recall %	f1-measure %
Positive News	RF	98	96	97
	DT	96	95	95
	LR	95	93	94
	KNN	85	80	83
	MNB	79	86	83
	LSVM	79	85	82
Negative News	RF	96	98	97
	DT	97	98	97
	LR	95	99	97
	KNN	81	86	83
	MNB	85	77	81
	LSVM	84	78	81
Macro Average %	Random Forest (RF)	98		
	Decision Tree (DT)	95		
	Logistic Regression (LR)	94		
	K-Nearest Neighbor (KNN)	83		
	Multinomial Naive Bayes (MNB)	82		
	Linear Support Vector Machine (LSVM)	81		

Table VII and Fig. 3 and Fig. 4 shows that the comparative analysis of the six classifier's accuracy with 10-fold cross validation is almost same yielding quite similar accuracy results. This is a perfect insight that reveals that our model is efficiently

detecting the sentiments. Though the accuracy has been increased to some extent in case of Fig. 4, the difference is very small that can be clearly noticed from Fig. 3 and Fig. 4. As more than one classifier that is in total three classifiers has shown above 90% accuracy and others have shown above 80% accuracy so it reflects that the model is well trained with good amount of data.

TABLE VII. COMPARISON OF ACCURACIES IN PERCENTAGE SPLIT METHOD WITH 10-FOLD CROSS VALIDATION

Name of Classifier	Percentage Split method Accuracy %	10-Fold Cross Validation Accuracy %
RF	98	99
DT	95	95
LR	94	95
KNN	83	84
MNB	82	80
LSVM	81	80

Moreover, it also contemplates that our build model is good for sentiment analysis based on supervised ML approach for each classifier. This 10-fold cross validation approach is one of the most popular k-fold approaches that have shown less biasness in case of validation. Additionally, in our context it has also shown good results.

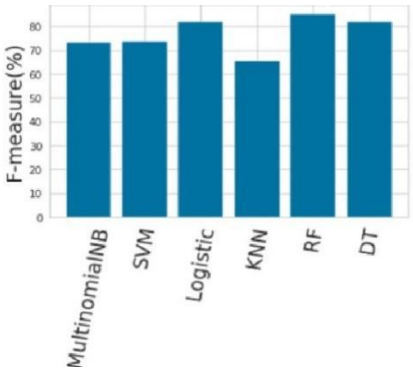

Fig. 3. Graphical plotting of f-measure versus six different classifiers (Splitting in percentage form)

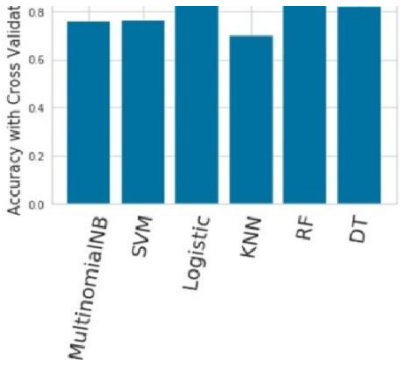

Fig. 4. Graphical plotting of f-measure versus six different classifiers (10-cross fold validation in Sentiment analysis using ML approach)

E. Comparison and Discussion

The behavior of most of the ML algorithms reveal that their performance differs from dataset to dataset which leads to a confusion that which algorithm is best suited for a specific domain's dataset. As earlier stated, very meagre amount of analysis are done on Bangla newspaper. So in Table VIII, we have compared our model with some sentiment analysis work with ML algorithm that includes SA on Bangla texts, Bangla sentences and Bangla tweets. This comparison has been quite in support of our proposed model as it achieved highest accuracy of 99% for RF algorithm being far better than other's accuracy. Another aspect is that our lowest accuracy of 80% for LSVM is also better or very close to some of the comparable results. This insight reveals that our model is very efficient and can be used as a common model for SA on Bangla document. Moreover, as our dataset is small having 3500 data and it has been eventually converted to 7000 data for better model fit. It can be said from the results that with this small dataset a good model is built up. So we hope to increase our dataset which will lead to more efficient fit of our model.

TABLE VIII. COMPARISON BETWEEN METHODS

Reference	Method	Domain	Size of Dataset	Accuracy %
Hoque, *et al.* [18] (2019)	Several supervised learning & deep learning algorithms with doc2vec features such as SM,SGD,BLSTM,LDA, SVM ,KNN, DT, Gaussian NB	Bangla Sentence	7000	77.72
Sarkar and Bhowmick [20](2017)	MNB, SVM with polynomial kernel c=10	Bangla Tweet	1000	45
Al-Amin, *et al.* [21](2017)	Parts of speech ratio, Cosine similarity using TF-IDF, Cosine similarity using custom TF-IDF, NB model using Uni-gram & stammer, NB model using Bi-gram stammer & normalizer, Word Embedding with Hellinger PCA	Bangla Sentence	2500	83.20
Tuhin, *et al.* [22] (2019)	NB and Topical approach	Bangla Sentence	7500	90%
Proposed Model	Pipeline class with vectorizer (Count Vectorizer), transformer(TF-IDF) and six supervised machine learning classifiers (MNB, K-NN, RF, DT (C4.5), LR and Linear SVM	Bangla News	7000	99%

V. Conclusion and Future work

In this paper, SA on Bangla news is done using "Pipeline" class along with six state-of-the-art supervised ML algorithms including DT, MNB, KNN, LR, RF and LSVM. As promising results have been seen using such methodologies in earlier works, in this purpose these classifiers are implemented to choose the best among them for our domain. RF algorithm outstands all other algorithms securing 98% accuracy in percentage split method and 99% accuracy in 10-fold cross validation method. Whereas LSVM has been seen to obtain an accuracy of 80%, which is the lowest.

In our work, we faced challenges due to insufficient amount of data to work with and the different behavior of the supervised ML algorithm in different domain and parameter setting. The main future directions of our research are pointed below:

- To add Neutral polarity as neutral news is mistreated in our model
- Enlarge our dataset effectively by integrating existing Bangla news dataset [25]
- Performing semi-supervised learning and unsupervised learning to use our large news corpus for Sentiment Analysis
- Evaluate deep learning algorithm for more efficient sentiment analysis and for more effective comparison

References

[1] E. Aydoğan and M. A. Akcayol. (2016). A Comprehensive Survey for Sentiment Analysis Tasks using Machine Learning Techniques. *2016 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 1-7: IEEE.

[2] A. Kumar and A. Jaiswal. (2020). Systematic Literature Review of Sentiment Analysis on Twitter using Soft Computing Techniques. *Concurrency and Computation: Practice and Experience*, 32(1), e5107.

[3] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu. (2020). Natural Language Processing (NLP) in Management Research: A Literature Review. *Journal of Management Analytics*, 7(2), 139-172.

[4] M. K. A. Chy, M. A. Rahman, A. K. M. Masum, S. A. Chowdhury, M. G. R. Alam, and M. S. I. Khan. (2019). Bengali Poem Generation Using Deep Learning Approach. *International Conference on Information, Communication and Computing Technology*, 148-157: Springer.

[5] H. Kaur and V. Mangat. (2017). A Survey of Sentiment Analysis Techniques. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 921-925: IEEE.

[6] P. Yang and Y. Chen. (2017). A Survey on Sentiment Analysis by using Machine Learning Methods. *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 117-121: IEEE.

[7] S. Rani and P. Kumar. (2019). A Journey of Indian Languages over Sentiment Analysis: A Systematic Review. *Artificial Intelligence Review*, 52(2), 1415-1462.

[8] S. Taj, B. B. Shaikh, and A. F. Meghji. (2019). Sentiment Analysis of News Articles: A Lexicon based Approach. *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 1-5: IEEE.

[9] A. Alshamsi, R. Bayari, and S. Salloum. (2020). Sentiment Analysis in English Texts. *Advances in Science, Technology and Engineering Systems Journal*, 5(6), 1683-1689.

[10] M. Suhasini and B. Srinivasu. (2020). Emotion Detection Framework for Twitter Data using Supervised Classifiers. *Data Engineering and Communication Technology*, Springer: 565-576.

[11] D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem. (2019). Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis using N-Gram Features. *Procedia Computer Science*,154, 332-340.

[12] F. Miao, P. Zhang, L. Jin, and H. Wu. (2018). Chinese News Text Classification based on Machine Learning Algorithm. *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2, 48-51: IEEE.

[13] A. Lommatzsch, F. Bütow, D. Ploch, and S. Albayrak. (2017). Towards the Automatic Sentiment Analysis of German News and Forum Documents. *International Conference on Innovations for Community Services*, 18-33: Springer.

[14] L. B. R. Thapa and B. K. Bal. (2016). Classifying Sentiments in Nepali Subjective Texts. *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1-6: IEEE.

[15] N. Mukhtar and M. A. Khan. (2018). Urdu Sentiment Analysis using Supervised Machine Learning Approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02), 1851001.

[16] S. Soumya and K. Pramod. (2020). Sentiment Analysis of Malayalam Tweets using Machine Learning Techniques. *ICT Express*, 6(4): 300-305.

[17] N. Pawar and S. Bhingarkar. (2020). Machine Learning based Sarcasm Detection on Twitter Data. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 957-961: IEEE.

[18] M. T. Hoque, A. Islam, E. Ahmed, K. A. Mamun, and M. N. Huda. (2019). Analyzing Performance of Different Machine Learning Approaches with Doc2vec for Classifying Sentiment of Bengali Natural Language. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1-5: IEEE.

[19] M. Islam, F. E. M. Jubayer, and S. I. Ahmed. (2017). A Comparative Study on Different Types of Approaches to Bengali Document Categorization, *arXiv preprint arXiv:1701.08694*.

[20] K. Sarkar and M. Bhowmick. (2017). Sentiment Polarity Detection in Bengali Tweets using Multinomial Naïve Bayes and Support Vector Machines.*2017 IEEE Calcutta Conference (CALCON)*, 31-36: IEEE.

[21] M. Al-Amin, M. S. Islam, and S. D. Uzzal. (2017). A Comprehensive Study on Sentiment of Bengali Text. *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 267-272: IEEE.

[22] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das. (2019). An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques. *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 360-364: IEEE.

[23] N. Tabassum and M. I. Khan. (2019). Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1-5: IEEE.

[24] S. Qaiser and R. Ali. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 181(1), 25-29.

[25] A. Khatun, A. Rahman, H. A. Chowdhury, M. S. Islam, and A. Tasnim. (2020). A Subword Level Language Model for Bangla Language. *Proceedings of International Joint Conference on Computational Intelligence*, 385-396: Springer.