Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

# Sentiment Analysis Of Bengali Text Using Machine Learning: Novel Approach

**[1]Ms.Moumita Pal , [2]Dr. Rajesh S. Prasad**

[1]Asst. Professor, NMIMS. Navi Mumbai, India.

[2]Professor and Associate Dean, MIT ADT University, Pune, India.

---

**Abstract:** In the machine learning domain, sentiment analysis has emerged as a key framework for scientific and commercial market research. As there are few research works on sentiment analysis for this language, it is currently a more significant research field of Bangla language processing system. Sentiment analysis is essentially an automated text mining procedure that determines the emotion of a given text. A given text can be classified into many emotions using sentiment analysis. This paper focuses on sentiment analysis in the context of Bangla language. Supervised machine learning classifiers such as logical regressions, K-Nearest Neighbour (K-NN), linear supervised vector machine and random forest are applied to the feature matrices. Using the Linear SVM technique, the Unigram model had a precision of 83.26%at the dataset. While the Bigram model approaches accuracy at 72.04% and precision at 85.2%, the Trigram model has the highest precision score of all, at 92%.

**Keywords:** K-NN, sentiment analysis, SVM, RF, LR, unigram, bigram, trigram.

## Introduction:

Sentiment analysis, also called opinion mining, is a field of study that predicts polarity in public opinion or textual data from microblogging sites on a well-publicized topic by extracting people emotions, attitudes, emotions, etc. As, sentiment analysis 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. Now-a-days people in social media sites, newspaper, blogs, etc., express their reviews on a specific product or items. There is also forum discussion, opinion on a specific post, 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 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 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. This matrix feature is language-dependent. Bangla, an ancient Indo-European language, is the spoken language of over 250 million people. So, extracting sentiment in Bangla language will surely

be significant for NLP researchers to make substantive progress in machine learning. [1]Among the three levels of sentiment analysis, 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. In this paper, machine learning algorithms such as logistic regression, SVM, random forest and K-NN are used.

By providing a method for automated sentiment analysis, this research opens up the path for further development of sentiment analysis methods of Bangla language in other sectors too. Such as the people having the reading disability will provide assistive technology for students with learning disabilities. [2] [3]

**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]. In Alshari et al. [4] 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 [5] pre-processed 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 [6] algorithm. In Chowdhury and Chowdhury [7], 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. Besides, In Islam et al. [8], 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 [9] 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 [10], authors designed a framework for SA by counting only positive and negative words form their feature word list dictionary. In Zhang et al. [11], 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 [12] authors described a lexicon-based dictionary model by checking the occurrences of a sentimental feature word in tagging each sentence.

**Dataset Pre-processing and Document Representation**

This experiment's dataset was compiled by hand from people's remarks on social networking websites. It has roughly 4000 samples, each of which is labelled as positive or negative. Further classification of the positive and negative classes was not possible with this dataset due to the lack of data. The test set was made up of 10% of the data and 90% of the data was utilised for training.

K-fold cross validation was used to validate the performance even more. The following are some examples from the dataset:

neg আমি অনেক অ্যাকশন মুভি দেখেছি, কিন্তু তাদের কোনটিই ভালো লাগেনি।

neg আমি মনে করি এই সিনেমাটি বছরের সবচেয়ে খারাপ সিনেমা।

neg পরের কয়েক asonsতু কয়েকটি পর্ব ছিল।

pos শেষবার যখন আমি একটি সিনেমা দেখেছিলাম, এটি আমার নজরে এসেছিল, আমি মনে করি না। কিন্তু এই সিনেমায় শেষ যে জিনিসটি আমি দেখতে চেয়েছিলাম তা হল আমার ছেলে।

pos হাজারা বাজ সিনেমা একটি মানসম্মত সিনেমা।

pos অসাধারণ, ভয়েস অভিনয় মনোমুগ্ধকর এবং চরিত্রের বিকাশ খুব সুন্দর।

**A. Pre-processing**

The raw data gathered isn't appropriate for classification on its own. Many punctuation marks, emoji's, etc. are included in the message but have no bearing on how the sentiment analysis method. To improve accuracy, the dataset must be pre-processed before beginning the classification process. There are a variety of pre-processing methods that are commonly used on datasets, depending on the dataset's language. Pre-processing is a crucial step before the classification process begins. The success of the pre-processing procedures determines the classification outcome. Tokenization, punctuation and emoticon removal, stemming, stop-word removal, and other techniques are applied.[13]

**NGram:**

As the name suggests, Ngrams are used to break a string into a number of substrings of N length, where N=string.length. To put it in the broadest possible terms, a Ngram is just the integration of neighbouring characters of N length that can be found in N=2 and N=3 were used in our investigation for bigrams and trigrams, respectively. It's a little different to use Ngrams in the Bangla language. To create a coherent phonetic piece in English, all letters are identical and have the same significance or impact. So, for example, all of the vowels have the same meaning. Whereas, in the Bangla language we have both vowel and vowel marks like {◌া, ি◌, ◌ী, ৈ◌, ো◌ৌ, ◌ু, ◌ূ}. Vowels like {অ, আ, ই, উ, এ, ও} do not necessarily have the same significance as the aforementioned vowel marks. Such as bigrams of (Kowsher) কাওছার> ['কা', '◌াও', 'ওছ', 'ছা', '◌ার’] don’t make any clear sense, but ['কাও', 'ওছ', 'ছার'] make a clear phonetic sense. For this reason, vowel signs were not considered strong characters while creating Ngrams. With counter vector, we combined bigrams and trigrams to provide new features.[14]

**1) Tokenization and Punctuation Removal:** 'Tokenization' is the process of breaking down a text into smaller, more Vocabulary or punctuation marks may be used as tokens. As a result, the words were separated from one another by breaking the sentence based on the All punctuation marks, alphabets of other languages, emoticons etc. were deleted from each data sample during tokenization. An array containing sub-arrays of tokenized data was created as a result of this step. In order to hold the labels, a new array was. 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 from the sentence. The characters nad stop word will no impact on creating training, test data, and machine learning model construction.

**2)**

**3) Stop word Removal:** Stop words are words whose value in the text corpus is negligible. When it comes to document classification, these words have no meaning. Stop words in English include "a," "of," "the," "for," "my," and others. Similarly, in Bangla, the terms “অতএব”, “অথচ”, “অথবা”, “অনুযায়ী”, “এটা”, “এটাই”, “এ্রিট” etc. are considered as stop words. The list of Bangla stop words was compiled from [12]. After removing the stop word from the sentence " ছোট এ্রিপেসাড হেলও অেনক মজাদার " the following tokens are obtained: [েছাট, এ্রিপেসাড, হেলও, মজাদার]. Here, “অেনক” was deemed a stop word in this context, it was omitted.

**4) Stemming:** The phrase stemming refers to the process of reducing a word's variations to its most fundamental form. Depending on the situation, a word might take on several different forms. For instance, "করা", "করিছ", "করিছলাম", "করিছেল", "কেরেছ", "কেরিছ" etc. for all these words, "কর" is the root word. The primary goal of stemming is to reduce a word's conjugational variants to a shared fundamental form. As a result, the overall number of words with which the classifier must contend can be drastically reduced. The typical prefixes and postfixes used in Bangla words were recorded in an array for this procedure. The prefix and postfixes in the words were detected using the Python Regular Expression module, and the trimmed versions of the words were added to the newly processed corpus. In this sentence " ছোট এ্রিপেসাড হেলও অেনক মজাদার", after stemming the words become as follows (excluding stop words): [ছোট, এ্রিপেসাড, হল, মজা]. Here, "হেলও" is changed into its base form "হল".

## B. Document Representation

Document representation is a pre-processing approach that can minimise a dataset's complexity and make it easier for a machine learning model to manage. The document's existing text version must be transformed to a vector representation. One of the most extensively used document vector representations is the vector space model [13], in which documents are represented by vectors of words. In this experiment, the Count Vectorizer and the Tf-Idf Vectorizer were used for feature extraction and vector representation.

**1) Count Vectorizer:** For each corpus (collection) of words in the vocabulary, Count Vectorizer builds a lexicon of words and counts their frequencies. Use of the produced vocabulary can also be used to encrypt fresh text documents

**2) Tf-Idf Vectorizer**: TF-IDF i.e. “Term Frequency-Inverse Document Frequency”. It is a statistical metric for determining the significance of a word in a corpus of text documents. While the importance of a word increases in proportion to how often it appears in a document, that same

importance falls when that same word appears more It builds a matrix of TF-IDF features from a set of raw documents using scikit-Tf-idf learn's Vectorizer function.[15]

**Methodology:**

The main goal of this research is to analyze the sentiment from Bangla text in machine learning approach by a 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:

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

**Algorithms**

The training data that was provided for the machine learning environment in the initial phase was combined into each algorithm to determine the target variable, and the model execution was monitored by obtaining exactness. We implemented machine learning classifier algorithms to execute the most appropriate strategies for demographic attribute recognition.

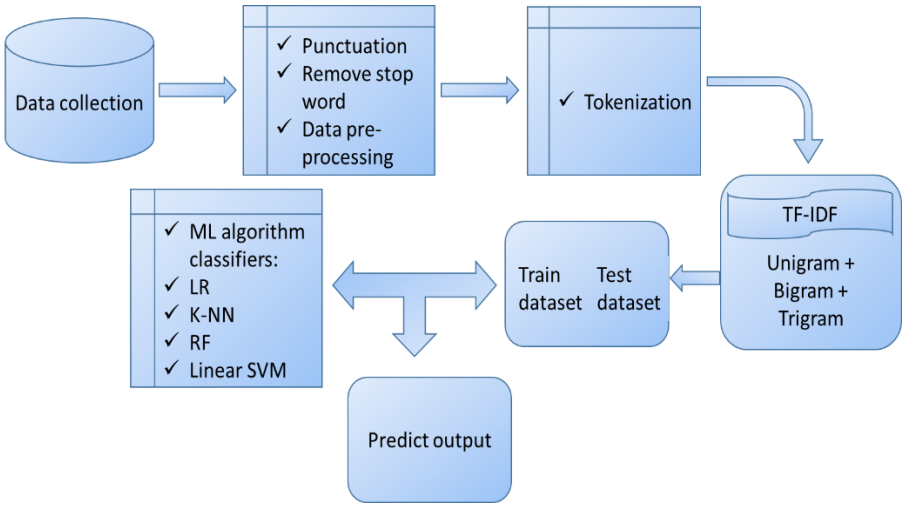


Fig 1: Architecture of the system

**Machine learning algorithms**

When dealing with classification and regression difficulties, the random forest classifier method is a great tool to have When using a support vector machine (SVM), you create hyperplanes in N-dimensional space to isolate data points based on at least two classes. The Naive Bayes Method, which is derived from Thomas Bayes' conditional probabilistic hypothesis, is another extensively used algorithm. As a machine learning technique for regression and classification, K-Nearest Neighbour (KNN) is another Our chosen neighbour value in our scenario is 5. [16]

To classify distinct features with varied prediction accuracies, we used Linear and Logistic Regression (LR).

**Result:**

Performance Table for Unigram feature:

| Sr. No. | Accuracy | Precision | Recall | F1 score | Model name |
|---|---|---|---|---|---|
| 0 | 71.26 | 73.93 | 63.93 | 68.56 | LR |
| 1 | 70.48 | 76.22 | 57.82 | 65.76 | RF |
| 2 | 64.76 | 64.40 | 62.86 | 63.62 | KNN |
| 3 | 70.61 | 83.26 | 50.13 | 62.58 | Linear SVM |

Table 1: Performance Table for Unigram feature

In case of Unigram feature:

Highest Accuracy achieved by LR at = 71.26

Highest F1-Score achieved by LR at = 68.56

Highest Precision Score achieved by Linear SVM at = 83.26

Highest Recall Score achieved by LR at = 63.93

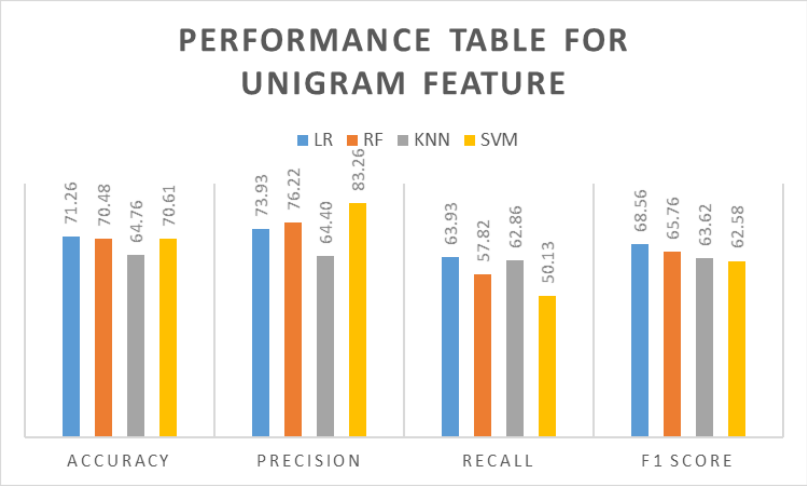


Fig 2: Visualization performance of different classifiers on the dataset with unigram feature

Performance Table for Bigram feature:

| Sr. No. | Accuracy | Precision | Recall | F1 score | Model name |
|---|---|---|---|---|---|
| 0 | 72.04 | 77.18 | 61.01 | 68.15 | LR |
| 1 | 69.31 | 79.75 | 50.13 | 61.56 | RF |
| 2 | 65.15 | 64.93 | 62.86 | 63.88 | KNN |

| | | | | | |
|---|---|---|---|---|---|
| 3 | 61.51 | 85.22 | 25.99 | 39.84 | Linear SVM |

Table 2: Performance Table for Bigram feature

In case of Bigram feature:

Highest Accuracy achieved by LR at = 72.04

Highest F1-Score achieved by LR at = 68.15

Highest Precision Score achieved by Linear SVM at = 85.22

Highest Recall Score achieved by KNN at = 62.86000000000001

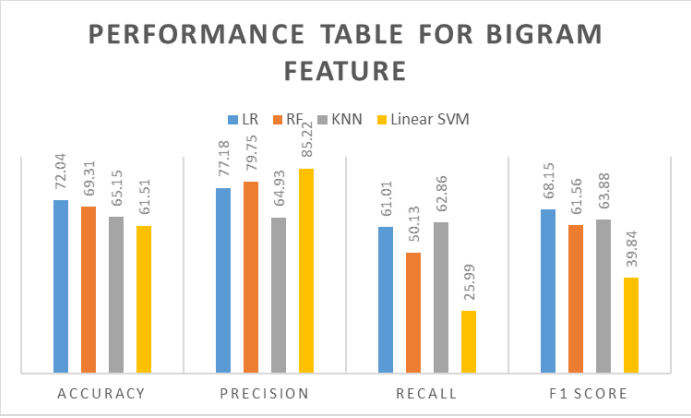


Fig 3: Visualization performance of different classifiers on the dataset with bigram feature

Performance Table for Trigram feature:

| Sr. No. | Accuracy | Precision | Recall | F1 score | Model name |
|---|---|---|---|---|---|
| 0 | 70.87 | 77.42 | 57.29 | 65.85 | LR |
| 1 | 69.7 | 83.96 | 47.21 | 60.44 | RF |
| 2 | 64.24 | 63.56 | 63.4 | 63.48 | KNN |
| 3 | 55.4 | 92.5 | 9.81 | 17.75 | Linear SVM |

Table 3: Performance Table for Trigram feature

In case of Trigram feature:

Highest Accuracy achieved by LR at = 70.87

Highest F1-Score achieved by LR at = 65.85

Highest Precision Score achieved by Linear SVM at = 92.5

Highest Recall Score achieved by KNN at = 63.4

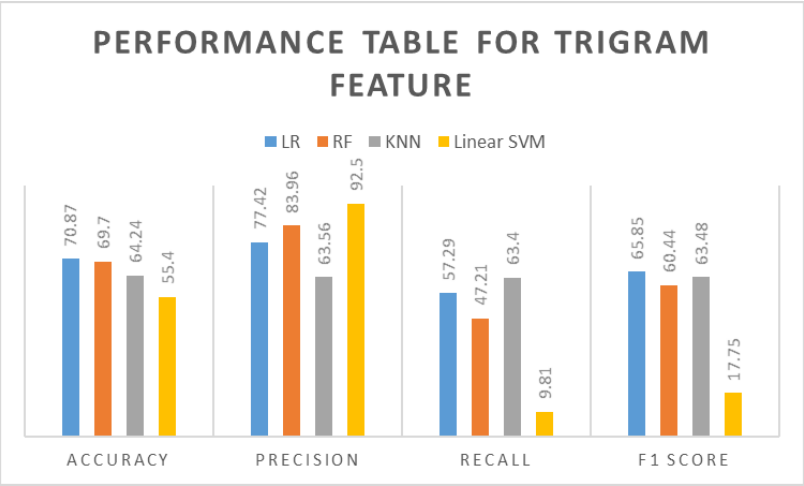


Fig 4: Visualization performance of different classifiers on the dataset with trigram feature

Besides, other classifier like LR, K-nearest neighbours (KNN), random forest (RF) algorithm is ap plied on our UniGram model. Among these classifiers, SVM shows better accuracy with 83.26%. F igure 2, 3 and 4 depicts the performance of different classifier. At Figure 4a, we have achieved bes t accuracy 71.26% and precision 83.26% at dataset. The dataset has much better accuracy and prec ision 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 LR and Linear SVM classificati on algorithm, finally accuracy is attained in dataset 72.04% which is greater than UniGram approa ch and also having precision 85.22% on the given dataset. Figure 3 shows the performance and su mmary of the sentiment analysis of Bigram model.

Later on, we discovered a significant improvement in the Bigram technique in the Tf-Idf model, th erefore we established another model, Trigram, in the Tf-Idf word vectorization. In this model, we used the LR and Linear SVM classification algorithms, and the final accuracy in the dataset is 70.8 %, which is higher than the UniGram approach, with precision of 92.5% on the supplied dataset. F igure 4 depicts the performance and summary of the Trigram model's sentiment analysis.

**Conclusion:**

Extraction of information from Bangla names is a major topic. As major features, n-grams approaches were used to attain this purpose. Ten machine learning classifiers, including Logistic Regression, SVM, K-NN, Support Vector Machine, and Random Forest, were used in this work. In the Bigram feature matrix, we attained the highest accuracy of 72.04 percent. Here we conclude that the we created three model viz. Unigram, Bigram and Trigram, we found that the Unigram model gave precision 83.26% at dataset by using the Linear SVM algorithm. Whereas bigram model approaches accuracy at 72.04% and precision at 85.22% and Trigram model having the highest precision score amongst all i.e. 92.5%.

**Reference:**

[1] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, “Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon

Dictionary," Nat. Lang. Process. Res., vol. 1, no. 3–4, p. 34, 2021, doi: 10.2991/nlpr.d.210316.001.

[2] O. Poobrasert, S. Luxsameevanich, S. Chompoobutr, N. Satsutthi, S. Phaykrew, and P. Meekanon, "Heuristic–based Usability Evaluation on Mobile Application for Reading Disability," Int. J. Electron. Eng. Appl., vol. VIII, no. II, p. 11, 2020, doi: 10.30696/ijeea.viii.ii.2020.11-21.

[3] R. Kumari, A. K. Sinha, and M. Banerjee, "a Comparative Study of Software Product Lines and Dynamic Software Product Lines," Int. J. Electron. Eng. Appl., vol. IX, no. II, p. 01, 2021, doi: 10.30696/ijeea.ix.ii.2021.01-10.

[4] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, and M. Alkeshr, "Effective Method for Sentiment Lexical Dictionary Enrichment Based on Word2Vec for Sentiment Analysis," Proc. - 2018 4th Int. Conf. Inf. Retr. Knowl. Manag. Diving into Data Sci. CAMP 2018, pp. 177–181, 2018, doi: 10.1109/INFRKM.2018.8464775.

[5] S. Arafin Mahtab, N. Islam, and M. Mahfuzur Rahaman, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," 2018 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2018, pp. 1–4, 2018, doi: 10.1109/ICBSLP.2018.8554585.

[6] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014, pp. 216–225, 2014.

[7] S. Chowdhury and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 Int. Conf. Informatics, Electron. Vision, ICIEV 2014, 2014, doi: 10.1109/ICIEV.2014.6850712.

[8] O. Sen et al., "Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods," vol. 4, 2021, [Online]. Available: http://arxiv.org/abs/2105.14875.

[9] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, "An automated system of sentiment analysis from Bangla text using supervised learning techniques," 2019 IEEE 4th Int. Conf. Comput. Commun. Syst. ICCCS 2019, pp. 360–364, 2019, doi: 10.1109/CCOMS.2019.8821658.

[10] N. Tabassum and M. I. Khan, "Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 1–5, 2019, doi: 10.1109/ECACE.2019.8679347.

[11] S. Zhang, Z. Wei, Y. Wang, and T. Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary," Futur. Gener. Comput. Syst., vol. 81, pp. 395–403, 2018, doi: 10.1016/j.future.2017.09.048.

[12] M. I. Siddiqi Emon, S. S. Ahmed, S. A. Milu, and S. S. Mahtab, "Sentiment Analysis of Bengali Online Reviews written with English Letter Using Machine Learning Approaches," ACM Int. Conf. Proceeding Ser., pp. 109–115, 2019, doi: 10.1145/3362966.3362977.

[13] F. Alam, S. Habib, and M. Khan, "Text normalization system for Bangla," no. May 2014, 2008.

[14] M. Kowsher, M. Z. Islam Sanjid, A. Das, M. Ahmed, and M. M. Hossain Sarker, "Machine Learning and Deep Learning based Information Extraction from Bangla

Names," Procedia Comput. Sci., vol. 178, pp. 224–233, 2020, doi: 10.1016/j.procs.2020.11.024.

[15] R. R. Chowdhury, M. S. Hossain, S. Hossain, and K. Andersson, "Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques," 2019 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2019, pp. 27–28, 2019, doi: 10.1109/ICBSLP47725.2019.201483.

[16] R. K. Tiwari, "Human age estimation using Machine Learning Techniques," Int. J. Electron. Eng. Appl., vol. VIII, no. I, p. 01, 2020, doi: 10.30696/ijeea.viii.i.2020.01-09.