IJCAT - International Journal of Computing and Technology, Volume 7, Issue 6, June 2020
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

Machine Learning Based Sentiment Analysis for Text Messages

[1]Abhishek Bhagat; [2]Akash Sharma; [3] Sarat Kr. Chettri

[1, 2] Department of Computer Science & Engineering
Assam Don Bosco University, Assam, India

[3] Department of Computer Applications
Assam Don Bosco University, Assam, India

Abstract - People use online platforms such as Facebook, Twitter, etc. for social networking and share their opinions, feelings or beliefs with others. Sharing is done by posts on these platforms. Sentiment analysis or opinion mining of these posts using machine learning techniques is of great significance. The analysis is generally carried out with sentiment, subjectivity analysis or polarity calculations. In this study, we performed a sentiment analysis of text messages using supervised machine learning techniques. They are mainly online product reviews, general tweets in Tweeter and movie reviews. Messages are pre-processed and then used three different machine learning techniques, namely Naïve Bayes, Decision Tree and Support Vector Machine (SVM) for sentiment analysis.

Keywords - *Machine Learning, Natural Language Processing, Sentiment Analysis, Twitter*

1. Introduction

In an online micro-blogging platform like Twitter, millions of tweets on various topics are created every day. These tweets are very important for understanding the trending themes. Classification of Twitter messages is one of the most important fields of research related to tweets. It is important and appropriate to identify these topics in different categories with high accuracy for better information retrieval. Sentiment analysis is a method to determine the polarity of a piece of writing as positive, negative or neutral. A text sentiment analysis program incorporates natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to persons, subjects, themes and categories within a sentence or expression.

Machine learning techniques which are generally used for sentiment analysis, one is unsupervised and the other is supervised. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering. Supervised learning is based on labeled dataset and thus the labels are provided to the model during the process.

In this study, we focused on supervised machine learning algorithms where the labeled datasets are used to train and check to generate appropriate outputs when encountered during decision-making. The rest of the paper proceeds as follows; in Section 2, we discuss about related work in this area. Section 3 is about the proposed system of our approach towards sentiment analysis in opinions and tweets. Section 4 discusses about machine learning techniques followed by implementation in Section 5 and finally we conclude with future prospects.

2. Related Work

Several studies have been conducted in the field of sentiment analysis and it can be concluded that sentiment analysis of text messages can be carried out in a variety of ways. The work [1] shows the types and techniques of sentiment analysis used to extract sentiments from tweets and have taken a comparative study of the different techniques and approaches of sentiment analysis using twitter as data. Various supervised learning methods on pre-processing and information extraction of tweets from twitter have also been explored [2]. The author concludes that Support Vector Machine (SVM) for text categorization can be used to detect the polarity of a text tweet. It was inferred that SVM recognizes some text properties such as High Dimensional Feature Space, few irrelevant features, and a sparse instance vector. The performance of SVM can be evaluated using precision and recall. Different results show that SVM delivers strong text categorization efficiency compared to artificial neural network (ANN).

Table 1. Summary of some of the attempts made by authors in sentiment analysis of text data

References (Year)	Data	Techniques	Approaches
[8] (2002)	IMDB dataset	Naïve Bayes, SVM	The effectiveness of applying Naïve Bayes and Support Vector Machine are explored to classify sentiments based on the reviews made in movies.
[9] (2016)	Online hotel reviews	Fuzzy ontology with SVM	Proposed a system based on SVM and Fuzzy domain ontology for opinion mining based on the collection of online reviews about hotels. The system computes the polarity term of each feature.
[10] (2013)	Online product reviews	Semi-supervised learning technique	The main approach followed here is to summarize and classify online user reviews of products in order to guide consumers in making decisions while making online purchases.
[11] (2016)	Multi-domain sentiment data extracted from Amazon.com	Rule based and statistical method	Proposed a three-stage cascade model to address the polarity shift problem in the context of document-level sentiment classification Three classification algorithms were also evaluated; SVM, logistic regression and Naïve Bayes.
[12] (2013)	Twitter dataset	Domain Ontology	Proposed a model based on the original ontology-based techniques for efficient analysis of Twitter posts sentiments. The Twitter posts are not simply characterized by a sentiment score but sentiment grade is assigned to each distinct notion available in the post.
[13] (2013)	Movie reviews	SVM, Naive Bayes and KNN	Sentiment classification techniques were applied to movie reviews and comparison of three supervised machine learning techniques has been made namely, SVM, Naïve Bayes, and kNN for classifying the sentiments.
[14] (2012)	Online product reviews	SVM	Opinion mining approach has been used to summarize user reviews, which are mainly unstructured and non-grammatical. The author attempts to perform the classification of features and polarity for each class of features.
[15] (2018)	Hotel reviews	Naïve Bayes multinomial	Proposed a framework that automatically prepares training and testing datasets for sentiment mining in order to analyze the online reviews made by hotel customers of their services.
[16] (2019)	Tourism reviews in social media	Domain specific ontology	The authors attempt to do a sentiment analysis of comments from people on Oman tourism on social media. The sentiment analysis is performed using a domain specific ontology and a POS tagger.

In the works [3, 6, 7], authors find keywords in tweet and predict whether they have positive or negative weights by applying machine-leaning algorithms. Multi-classification algorithms such as SVM, Naïve Bayes, Logistic Regression for classification, KNN and Decision Tree have been used for trend analysis prediction. In another work [4], the authors discuss methods for the automatic extraction of sentiments or opinions of the text unit. The proposed model deals with mining emotions of tweets about Indian politicians using Support Vector Machine.

They have used unigram and TF-IDF as feature vectors measured the performance of the proposed system in terms of classification accuracy, and F-measure. Y. Luo et. al [5] has made a comparison of positive and negative sentences. Their proposed model extracts information from the web and manually labels a word set that requires a lot of effort. The main focus of the sentiment analysis is to categorize people's opinions by analyzing their views on social media. Table 1 provides a summary of the works done by different authors in the area of sentiment analysis or opinion mining available in the literature.

3. Proposed System

We present a framework where there are multiple layers for sentiment analysis (see Fig.1). The first layer is the initialization layer where the data collection and pre-processing of messages is performed. The second layer is the learning layer where the pre-processed data will be split into training (70% data) and test dataset (30% data). The training dataset will be used to develop three different models using supervised machine learning techniques namely Naïve Bayes, Decision Tree and Support Vector Machine.

Again, the run-time behavior of three trained models using model-based testing techniques will be used to check the model's predictions. Finally, in the third layer, comparison of the performance of the models will be made on the basis of the evaluative metrics , i.e. precision, recall, macro average F1-score, polarity classification accuracy to determine how the effects of the statistical analysis can be generalized to an independent dataset.

The task of sentiment classification can usually be seen as a two-class classification problem. This type of work mainly involves sentiment analysis as a text classification problem, where feature selection has significant effect on the performance of developed classifier models.

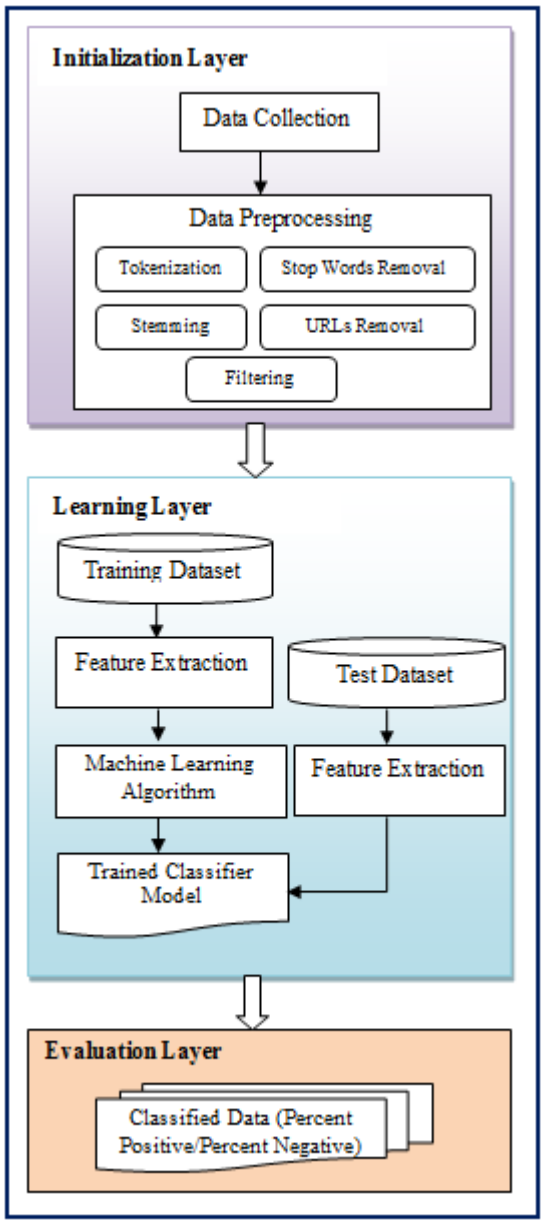

Fig. 1 Proposed framework for sentiment analysis

3.1. Data Collection

For sentiment analysis on Twitter, varieties of benchmark datasets have been released over the last few years and are available online. We choose to select five different datasets that have been widely used in Twitter sentiment analysis studies in literature. Table 2 provides a brief overview of the datasets used in this paper.

Table 2. Brief description of the datasets

Dataset	*No. of Positive Message*	*No. of Negative Message*	*Total no of Messages*	*Average no. of words*
IMDB	25,000	25000	50,000	255
Sentiment 140	800,000	800,000	1,600,000	15
SemEval-2013	2315	861	3176	23
SemEval-2014	2509	932	3441	22
STS-Gold	632	1402	2034	16

The datasets used in this paper consist of messages/reviews made in English on products, movies and general tweets made in Twitter. Tweets are extracted using Tweeter API and manually annotated by polarity (Positive, Negative, and Neutral).

The IMDB dataset is a large movie review dataset used for binary sentiment collection. It consists of 50,000 polarized movie review posts.

The Sentiment 140 dataset consists of 1.6 million Twitter messages collected using a distant supervision approach [17]. The general tweets in the dataset have been annotated (0 = negative, 4 = positive) and the large collection of data in this dataset is very useful for the development of machine learning models to be used to detect sentiment.
The SemEval-2013 dataset consists of comments taken from a variety of topics discussed on Twitter; a number of entities, items and events.

The SemEval-2014 dataset [18] deals with product reviews made by consumers on two entities or domains. They're laptops and restaurants. Both SemEval datasets are not publicly accessible but can be downloaded from the source first.

The Stanford Twitter Sentiment Gold dataset (STS-Gold) [19] is a subset of tweets from the Stanford Twitter Sentiment Corpus. In the dataset, tweets and targets (entities) are annotated separately and therefore different sentiment labels may be used to help the evaluation of sentiment classification models at both the entity and the tweet level.

3.2. Pre-processing of Data

Pre-processing is needed when building machine learning systems based on tweet data. All tweets are normalized during pre-processing. This included the following steps:

- All targets (@username) are removed from the data.
- All URLs and special characters are removed from the datasets.
- Uppercase letters are converted into lowercase.
- Stop words like articles, prepositions, conjunctions, and pronouns are filtered. Stop words provide little or no information.
- Stemming is done where the inflected words are reduced to their own word stem that affixes to suffixes and prefixes.
- Many of the texts are broken down into tokens. It is a method called tokenization. It is mainly used to eliminate the delimiters from the messages. For example, Tweets datasets are rich data sources is broken into individual tokens such as Tweets, datasets, are, rich, sources, of, data. A token is detected on encountering a space.

3.3. Feature Extraction

Feature extraction is the most important tasks in classification functions. It includes removing meaningless words or terms that do not convey any emotions. Unigram (n=1) and term frequency and inverse document frequency (TF-IDF) are the features extracted from the preprocessed datasets. The unigram features represents individual, distinct words. TF-IDF assigns a score for each term. The term-frequency is determined by counting the number of times the word or phrase appears in the document and the inverse frequency of the document is measured by dividing the total number of documents by the number of documents containing the phrase.

4. Machine Learning Algorithm

4.1. Naïve Bayes Algorithm

The Bayesian Classification is a supervised learning system as well as a statistical classification method. It assumes the underlying probabilistic model and allows us to capture uncertainty about the model in a principled manner by determining the probabilities of outcomes. Bayesian classification offers functional learning algorithms and prior knowledge and data can be mixed. Bayesian Classification provides a valuable viewpoint for recognizing and testing a variety of learning algorithms. It calculates explicit probabilities for hypotheses and is resilient to noise in input data. Naive Bayes conditional independence assumption can be shown in equation 1.

$$P(doc|V_j) = \prod_{i=1}^{\text{length(doc)}} P(a_i = W_k|V_j) \qquad (1)$$

Where, $P(a_i = w_k \mid v_j)$ is probability that word in position i is w_k, given v_j .. One more assumption: $P(a_i = w_k \mid v_j) = P(a_m = w_k \mid v_j) \ \forall$ i, m.

4.2. Support Vector Machine

A Support Vector Machine (SVM) is a classifier that discriminates the data in numerous planes. For a given supervised learning instance the SVM gives an optimal hyper plane as output to categorize new data records. The SVM algorithms are implemented using a kernel. The kernel is responsible for translating input data to the correct type. The SVM kernel takes input data of small dimensions and transforms it into data space of higher dimensions, thereby increasing data separability. This strategy is called kernel tricks. We used the linear kernel of the SVM, as it is mainly used to manage text data because it includes many features and is usually linearly separable.

4.3. Decision Tree

Decision Tree is a classification techniques in which the divide-n-conquer process operates. The essential aspect of the decision tree is that it breaks down the complicated decision-making process into a series of simpler decisions. In a tree where the root and the inner node are labeled with a query, and a leaf node is a prediction of an answer. We used an ID3 algorithm. It is a precursor to the C4.5 algorithm and is generally used in the area of machine learning and natural language processing. For classification of messages, information gain (see Equation 2) has been used as a split criterion.

$$G(x,y) = H(x) - \sum_{i \in value(y)} \frac{\Delta y_i}{\Delta y} H(y_i) \qquad (2)$$

Where, *H(x)* is the entropy of the training set *x* with its attribute *y*.

5. Experimental Results

The performance of sentiment classification can be evaluated by using the following metrics.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (3)$$

$$Precision = \frac{TP}{(TP+FP)} \qquad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \qquad (5)$$

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \qquad (6)$$

Where, TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances. For sentiment analysis on a tweet, high accuracy ensures that the emotions of most messages in a dataset are evaluated and correctly predicted. The F1-score is a common metric for both precision and recall. A receiver operating characteristic (ROC) curve is a graphical plot that shows the diagnostic capability of a binary classifier model as its discrimination threshold is varied. The ROC curve is developed by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings Built classifier models are evaluated on the basis of accuracy, macro-averaged F1-Scores of each model in five separate datasets (split as 70 percent training dataset and 30 percent testing dataset), and the ROC curves. The results obtained are shown in Tables 3 and 4.

Table 3. Classification accuracy (in %) of the models using the datasets

Dataset	***Naïve Bayes***	***SVM***	***ID3***
IMDB	75.43	72.5	**82.7**
Sentiment 140	73.18	75.5	**81.94**
SemEval-2013	83.56	**84.93**	82.23
SemEval-2014	75.12	**78.25**	69.56
STS-Gold	**87.78**	84.34	73.6

Table 4. Macro-Averaged F1-Score of the classifier models using the datasets

Dataset	***Naïve Bayes***	***SVM***	***ID3***
IMDB	0.75	**0.75**	0.73
Sentiment 140	0.73	0.76	**0.82**
SemEval-2013	0.85	**0.862**	0.72
SemEval-2014	0.76	**0.78**	0.63
STS-Gold	**0.862**	0.827	0.682

The plotted ROC curves of the three different classifier models developed using the Sentiment 140 dataset are shown (see Figures 2, 3, and 4). The reason for showing the ROC curve only for Sentiment 140 dataset is because of its large size which really helps in developing better classifier models. Figures 2, 3, 4 reflect ROC curves of the Naïve Bayes classifier model, SVM and decision tree model using ID3 algorithm respectively.

The results shows a close competion between SVM and ID3, as they are found to be better than Naïve Bayes

classifaction algorithm. However, Naïve Bayes performs better with the STS-Gold dataset. In case of SemEval-2014, the best accuracy achieved is 78.25% which is quite low.

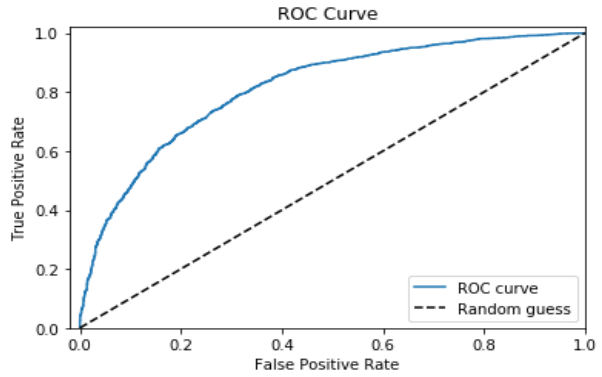

Fig. 2 ROC curve of Naïve Bayes classifier model

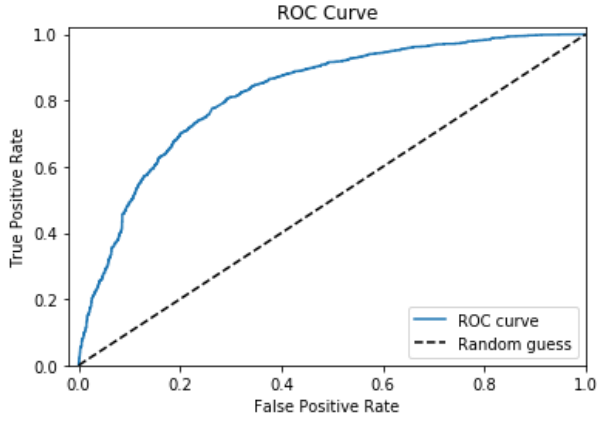

Fig. 3 ROC curve of SVM classifier model

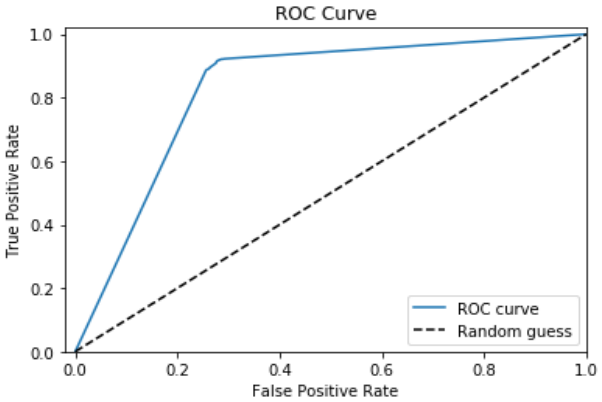

Fig. 4 ROC curve of decision tree classifier model

6. Conclusion and Future Works

In this paper, text data from product reviews, general tweets and movie reviews are taken into account to assess the polarity (positive or negative) of messages or tweets. We used the classification algorithms namely SVM, Naïve Bayes and decision tree. We evaluated our models on the basis of metrics; classification accuracy, precision, recall, F1-score, and ROC curve. After evaluating the developed classifiers, we find that the results obtained from the Decision Tree and SVM have a lower mean square error or a higher accuracy with most of the datasets and are considered to be good classifiers. We find our work to be unique, as we have attempted in our study to provide an overview of the various methods used in the sentiment analysis of text data. We have also built and compared three different classifiers using machine learning techniques to five different datasets of varying sizes and domains.

Future prospects in this area include the development of techniques for aspect-based sentiment analysis, which will take into account the various aspects of the text. Social networking data can be accessible in various languages, making it an obstacle to sentiment analysis. The development of a multilingual approach to sentiment analysis is therefore a major challenge. Also, the main drawback of using a machine learning approach to opinion mining is that each opinion is treated as a single uniform statement and assigns a sentiment score to the post as a whole. Efficient methods must therefore be developed to identify the subjects discussed in the message and to convert each message into subject-level aspects.

References

[1] A. Mittal and S. Patidar, Sentiment Analysis on Twitter Data: A Survey, In: Proceedings of the 7th International Conference on Computer and Communications Management, pp. 91--95, 2019.

[2] Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta, Sentiment Analysis of Tweets using Machine Learning Approach, In proceeding of Eleventh International Conference on Contemporary Computing (IC3), IEEE, 2018.

[3] Tejal Rathod , Mehul Barot, Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events, International Journal of Computer Applications, Vol. 180 (26), 2018.

[4] Suman Rani, Jaswinder Singh Sentiment Analysis Of Tweets Using Support Vector Machine, International Journal of Computer Science and Mobile Applications, Vol.5 (10), 2017.

[5] Y. Luo, W.Huang, Product Review Information Extraction Based on Adjective Opinion Words, In: Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313, 2011.

[6] Aditya Kulkarni, Shubham Mhaske, Tweet Sentiment Analysis and Study and Comparison of Various Approaches and Classification Algorithms Used,

International Research Journal of Engineering and Technology, Vol 7(4), 2020.

[7] Apurva Dixit, Alok Kumar Pal, Shraddha Temghare and Vikas Mapari, Emotion Detection Using Decision Tree, International Journal of Advance Engineering and Research Development. 2017.

[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, vol 10, pp. 79-86. 2002.

[9] Farman Ali, Kyung-Sup Kwak and Yong-Gi Kim, Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification, Applied Soft Computing, 2016.

[10] M.K. Dalal, M.A. Zaveri, Semi supervised learning based opinion summarization and classification for online product reviews, Applied Computational Intelligence and Soft Computing, pp. 1–8, 2013.

[11] Rui Xia, Feng Xu, ‚Jianfei Yu et.al., Polarity shift detection, elimination and ensemble: A three stage model for document-level sentiment analysis, Information Processing and Management, vol 52(1), pp. 36– 45, 2016.

[12] Efstratios Kontopoulo, Christos Berberidis, Theologos Dergiades , Nick Bassiliades, Ontology- based sentiment analysis of twitter posts, Expert Systems with Applications
vol. 40, pp.4065-4074, 2013.

[13] P. Kalaivani, K. L. Shunmuganathan, Sentiment classification of movie reviews by supervised machine learning approaches, Indian Journal of Computer Sci. Eng. Vol 4, pp.285–292, 2013.

[14] T. Hassan, A. Soliman, M.A. Elmasry, A.R. Hedar, M.M. Doss, Utilizing Support Vector Machines in mining online customer reviews, In: Proceedings of 22nd International Conference on Computer Theory and Applications (ICCTA), pp.192—196, 2012.

[15] K. Zvarevashe, O.O. Olugbara, A framework for sentiment analysis with opinion mining of hotel reviews. In: conference on Information Communications Technology and Society (ICTAS), pp. 1–4, 2018.

[16] V. Ramanathan, T. Meyyappan, Twitter text mining for sentiment analysis on people's feedback about Oman tourism, In: 4th MEC International Conference on Big Data and Smart City (ICBDSC) IEEE. pp. 1--5, 2019.

[17] Alec Go, Richa Bhayani and Lei Huang, Twitter Sentiment Classification using Distant Supervision Stanford University Stanford, CS224N project report, Stanford, 2009

[18] S. Rosenthal, Semeval 2014 task 9 description http://alt.qcri.org/semeval2014/ task9/

[19] H. Saif, M. Fernandez, Y. He, and H. Alani, Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold, In: Proceedings, 1st Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM), Turin, Italy, 2013.

Authors -

Abhishek Bhagat is currently pursuing his BTech (CSE) from the Department of CSE, School of Technology, Assam Don Bosco University. Currently, he is in his final year. His areas of interest are Machine Learning, Big data analytics and Natural Language Processing.

Akash Sharma is currently pursuing his BTech (CSE) from the Department of CSE, School of Technology, Assam Don Bosco University. He is a final year student at present. His areas of interest are Machine Learning, Internet of Things (IoT) and Natural Language Processing (NLP).

Dr. Sarat Kr. Chettri is an Assistant Professor in the Department of Computer Applications, School of Technology, Assam Don Bosco University. He has made several publications in international journals and conferences. His research area includes data science, machine learning and Internet of Things (IoT).