Tech Science Press

Optimal Machine Learning Driven Sentiment Analysis on COVID-19 Twitter Data

Bahjat Fakieh[1], Abdullah S. AL-Malaise AL-Ghamdi[1,2,3], Farrukh Saleem[1] and Mahmoud Ragab[2,4,5,6,*]

[1]Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[2]Center of Excellence in Smart Environment Research, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[3]Information Systems Department, HECI School, Dar Alhekma University, Jeddah, Saudi Arabia
[4]Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[5]Department of Mathematics, Faculty of Science, Al-Azhar University, Naser City, 11884, Cairo, Egypt
[6]Centre for Artificial Intelligence in Precision Medicines, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
*Corresponding Author: Mahmoud Ragab. Email: mragab@kau.edu.sa

Abstract: The outbreak of the pandemic, caused by Coronavirus Disease 2019 (COVID-19), has affected the daily activities of people across the globe. During COVID-19 outbreak and the successive lockdowns, Twitter was heavily used and the number of tweets regarding COVID-19 increased tremendously. Several studies used Sentiment Analysis (SA) to analyze the emotions expressed through tweets upon COVID-19. Therefore, in current study, a new Artificial Bee Colony (ABC) with Machine Learning-driven SA (ABCML-SA) model is developed for conducting Sentiment Analysis of COVID-19 Twitter data. The prime focus of the presented ABCML-SA model is to recognize the sentiments expressed in tweets made upon COVID-19. It involves data pre-processing at the initial stage followed by n-gram based feature extraction to derive the feature vectors. For identification and classification of the sentiments, the Support Vector Machine (SVM) model is exploited. At last, the ABC algorithm is applied to fine tune the parameters involved in SVM. To demonstrate the improved performance of the proposed ABCML-SA model, a sequence of simulations was conducted. The comparative assessment results confirmed the effectual performance of the proposed ABCML-SA model over other approaches.

1 Introduction

The impact of Coronavirus Disease 2019 (COVID-19) upon the daily routines of people across the globe was disastrous. People utilized social media to express their thoughts and feelings regarding

the unprecedented events that occurred after the outbreak of COVID-19. Social Networking Sites like Twitter experienced a tremendous increase in the number of users and their usage. In other terms, the number of tweets increased in relation to the pandemic within a short period of time [1]. Twitter is one of the most typically used social media platforms. It offers real information about the ongoing events in brief and seizures the thoughts and emotions of the people. During pandemic lockdown, people used social media platforms like Twitter to express their emotions, feelings, thoughts and opinions about the pandemic worldwide [2]. COVID-19 rapidly spread across the globe and the number of COVID-19 positive cases increased within a short span of time. Various nations got affected since the infection is an air-borne one. The disease tend to affect the person who shares close space with an infected individual. Global biomedical professionals made persistent efforts to contain the spread of disease and help the patients using the available medicines [3]. Vaccination was found to be the most effectual strategy to prevent the spread of coronavirus infection globally. Those individuals affected with coronavirus were advised to isolate themselves and were prescribed to test themselves from time to time, whereas the rest of the people were advised to get vaccinated [4].

Individuals exhibited their feelings through social networking sites when they were in isolation. Because, social media had realistic and valuable information regarding coronavirus infection; but, the information from mass media was considered to be misleading or useless in some cases [5]. Miseries increased when the data from online media was found to be misleading. Work from home, isolation time and stay-at-home became the new norm and were frequently used in social media to share the emotions, news, advice and opinions [6]. Misinformation is defined as an attempt made to misdirect others with fake or inappropriate data, for instance, 'consuming bananas was a preventative measure against this viral infection.' An individual, affected by COVID-19, undergoes multiple levels of mental and physical health issues. So, the disease condition demands to apply logical policies so as to comprehend informative data streams [7]. Social networking sites like Facebook and Twitter have too much of noisy information. Hence, it is challenging to find a useful content from such huge volumes of noisy data. However, if the noisy data is cleaned, it is possible to capture different types of emotions, human feelings, thoughts and expressions. Through complete analysis, the data can provide more information regarding the present attitude, nature and mood of a large human society [8].

Sentiment Analysis (SA) is also termed as emotion or opinion Artificial Intelligence (AI). It uses Machine Learning (ML), computational linguistics, Natural Language Processing (NLP) and text analysis processes to methodically examine the texts written by people [9]. The results from SA help in determining the sentiments, opinions, attitudes and emotions. SA is broadly implemented to gain insights about online media posts. It intends to differentiate the polarity of a text as neutral, positive and negative [10]. Thus, the main topic in SA is to realize the responses for the tweets made at first. Till now, several researchers have scrutinized the effects of a tweet's Sentiment Polarity (SP) based on the responses received for the tweet. The current research work intends to fill this research gap.

The current study presents an Artificial Bee Colony (ABC) with Machine Learning-driven SA (ABCML-SA) model for conducting Sentiment Analysis of COVID-19 Twitter data. The prime focus of the presented ABCML-SA model is to recognize the sentiments present in COVID-19 tweets. The model involves data preprocessing at the initial stage followed by n-gram feature extraction to derive the feature vectors. For identification and classification of the sentiments, the Support Vector Machine (SVM) model is exploited. At last, the ABC algorithm is applied to fine tune the parameters involved in SVM. In order to demonstrate the improved performance of the proposed ABCML-SA model, a sequence of simulations was conducted.

2 Related Works

Naseem et al. [11] developed a policy to be implemented in social networking sites. For instance, the policy decides the level of moderation required to curtail the misinformation on social networking sites. This work also examined the views, with regards to COVID-19, by concentrating on individuals who communicated and shared their opinions in Twitter. In the study conducted earlier [12], SA was conducted for the tweets made by individuals from the top-10 infected nations. The experimentations were carried out using the collected data i.e., individuals' tweets from the top-10 infected nations along with the Sultanate of Oman. In literature [13], Sentiment Analysis was conducted upon the tweets made during the first wave of COVID-19 pandemic in Europe. This study used Neural Network (NN) for SA using multilingual sentence embeddings.

Pastor [14] focused on how to expose the sentiments of Filipinos, especially from Luzon region, during the regulation of extreme community quarantine at the time of COVID-19 outbreak. The study collected a twitter user's tweets from the Philippines. It examined the impact of the pandemic upon a person's lifestyle changes based on their tweets. NLP is one of the techniques that can be applied in determining a user's sentiment from the extracted tweets. In this study, opinions were assumed as data to conduct the analysis. Qualitative method was utilized in this study to determine the impact of extreme community quarantine in Luzon region, the Philippines.

In the study conducted earlier [15], a rule-related method was employed to conduct SA upon the tweets made by the Honorable President, Donald Trump, during the initial spread of COVID-19 infection in the United States (US). The study found a statistically-important adverse relation between the sentiments of his messages and the sum of COVID-19 cases recorded in the US. The results inferred that his tone in the tweets made an impact and the pandemic severely affected the lives and the economy of the US. Pano et al. [16] enhanced a distinct text preprocessing strategy to correlate the sentiment scores secured by Twitter text with Bitcoin prices during COVID-19 lockdown. In addition, the authors explored the impact of several preprocessing features, time lengths and the functions of data upon the outcomes of the correlation process. Among 13 approaches, the authors found that dividing the sentences or eliminating Twitter-specific tags or both the activities enhanced the relation of volume polarity and sentiment scores with Bitcoin rates.

3 The Proposed Model

In current study, a new ABCML-SA method is proposed for conducting sentiment analysis of the COVID-19 Twitter data. The prime focus of the presented ABCML-SA is to recognize the sentiments present in COVID-19 tweets. It involves data pre-processing at the initial stage and n-gram feature extraction in the next stage to derive the feature vectors. For identification and classification of the sentiments, the SVM model is exploited. At last, the ABC algorithm is applied to fine tune the parameters involved in SVM. Fig. 1 illustrates the block diagram of ABCML-SA approach.

3.1 Data Pre-Processing

In this initial phase, ABCML-SA model performs pre-processing of the data to transform the input data into a compatible form. Data pre-processing is primarily executed to clean the raw data, by following a few steps, so as to attain superior outcomes in multiple assessments. The collected data is pre-processed using a user-defined pre-processing function related to Natural Language Toolkit (NLTK) i.e., Python library for NLP. Stemming is helpful in reducing the inflected words to its base or root form whereas Tokenization helps in dividing every sentence into small words.
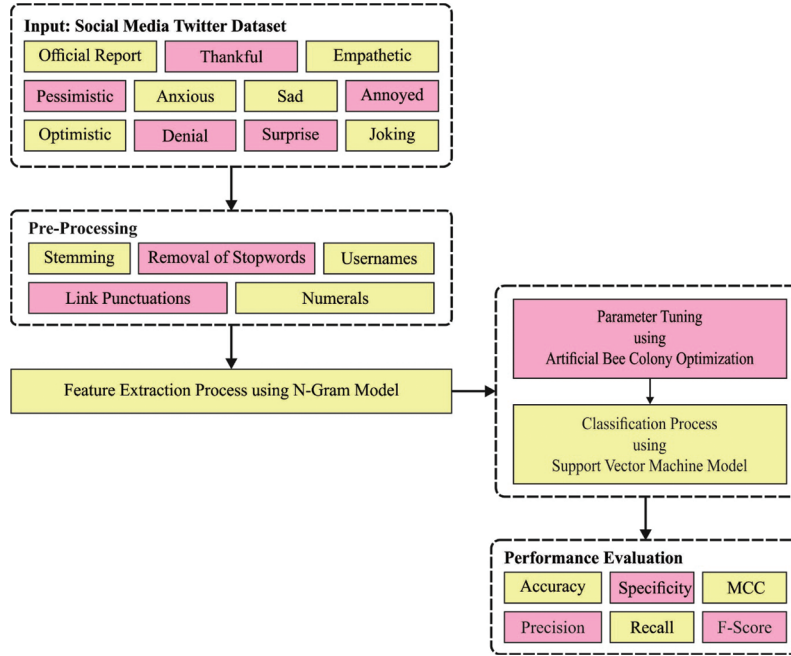
Figure 1: Block diagram of ABCML-SA approach

3.2 N-Gram Feature Extraction

Next, the n-gram model is applied to generate the feature vectors. After pre-processing, the study follows the Bag-of-Words (BOW) mechanism with a commonly-emerged word from lexicon words and attains a list of the most-commonly used words for COVID-19 [17]. Numerous words are established within the produced corpus at different times and dissimilar locations of the Twitter. Now, the reappearance of every word is calculated and the top 50 commonly-used words are proposed together with popularity. Then, word popularity is determined to evaluate the possibility of recurrence for all the words i.e., 3, 53, 704 words in total. Both probability and popularity counts of the most commonly-used words are represented as follows.

$$P\left(W_i\right) = \frac{count\left(W_i\right)}{\sum_{i=0}^{n} count\left(W_{i=0}^{n}\right)} \tag{1}$$

Lexical n-gram model is frequently applied in NLP model for syntax feature maps and statistical analysis. Therefore, the current study uses the n-gram mechanism to analyse the produced corpus with tokenized words to find a group of adjacent words or popularity of the word. Now, the occurrence probability is evaluated using the probability chain rules.

$$P\left(x_1,\ x_2,\ x_3,\ x_n\right) = P\left(x_1\right) P\left(x_2|x_1\right) P\left(x_3|x_1,\ x_2\right) \ldots P\left(x_n|x_1,\ x_2,\ x_3,\ldots x_{(n-1)}\right) \tag{2}$$

$$= \prod_{i=1}^{n} P\left(x_i|x_1^{(i-1)}\right) \tag{3}$$

For instance, a sentence is assumed to be 'Still Covid19 wave is running'. According to the probability chain rules, P("Still Covid 19 wave is running") = P("Still") x P("Covid 19" | "Still") x P("wave" | "Still Covid19") x P("is" | "Still Covid 19 wave") x P("running" | "Still Covid 19 wave is")

are expressed as follows.

$$P\left(W_1, W_2, W_3, W_n\right) = \prod_j P\left(W_j|W_1, W_2, W_3, W_{(j-1)}\right) \tag{4}$$

$$= \prod_{j=1}^{n} P\left(W_j|W_1^{(j-1)}\right) \tag{5}$$

Bigram models estimate the probability of the words using the condition probability $P\left(W_i|W_{i-1}\right)$ of the previous words. This is done based on the following condition for every preceding word $P\left(W_i|W_1^{i-1}\right)$ [6] as denoted herewith.

$$P\left(W_1, W_2\right) = \prod_{i=2} P\left(W_2|W_1\right) \tag{6}$$

The probability is determined as given below.

$$P\left(W_k|W_{(k-1)}\right) = \frac{\text{count}\left(W_{(k-1)}W_k\right)}{\text{count}\left(W_{(k-1)}\right)} \tag{7}$$

The authors found the most commonly-used unigram, bigram and trigrams within the corpus through the n-gram mechanism. Based on the study, the popularity of the trigram is established to be lesser than the bigram model, whereas the popularity of the unigram model is higher based on n-gram models.

3.3 SVM Based Sentiment Classification

In order to identify and classify the sentiments, SVM model is exploited. SVM is a two-class classification model and is defined herewith. Consider that (x_i, y_i), $1 < i < N$ indicates a dataset that has N number of trained instances [18]. Every instance should fit the condition , $x_i \in R^d$. Here, y_i establishes the class of the respective instance, x_i. Hence $y_i \in \{-1, 1\}$ whereas d denotes the number of dimensions in the input dataset.

$$w \cdot x_i + b = 0, \ 1 \leq i \leq N. \tag{8}$$

In case of a hyperplane, linear separation is achieved. The instance near one to the separating hyperplane is named as support vector. In boundary (support vector), Eq. (8) is rewritten as follows.

$$w \cdot x_i + b = \pm 1. \tag{9}$$

Based on Eq. (9):

$$y_i \cdot (w \cdot x_i + b) \geq 1. \tag{10}$$

Hence, the issue is to find w and b. Various hyperplanes are available to split the two-class dataset. However, SVM generates the optimum hyperplane. This hyperplane has the maximal distance to assist the vector. The margin of the splitting hyperplane is $2/\|w\|$. Therefore, to detect the optimum hyperplane, $\|w\|$ needs to be reduced whereas $(1/2)\|w\|^2$ should be substituted with $\|w\|$. Therefore, the optimization issue is handled successfully. It implies that $(1/2)\|w\|^2$ should be minimized, subjected to Eq. (10).

Though the instances are linearly separable during most of the instances, it cannot be detached easily. For non-linear problems, a positive slack parameter ζ_i is presented.

$$\text{Min } \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{n} \zeta_i$$

$$s.t \ y_i \cdot (w \cdot x_i + b) \geq 1 - \zeta_i, \tag{11}$$

$$\zeta_i \geq 0,$$

$$1 \leq i \leq N.$$

In Eq. (11), C is named after penalty factor. It is familiarized to regulate the trade-off between error minimization and margin maximization. This issue is resolved using Lagrange multiplier. Fig. 2 illustrates the SVM hyperplane.
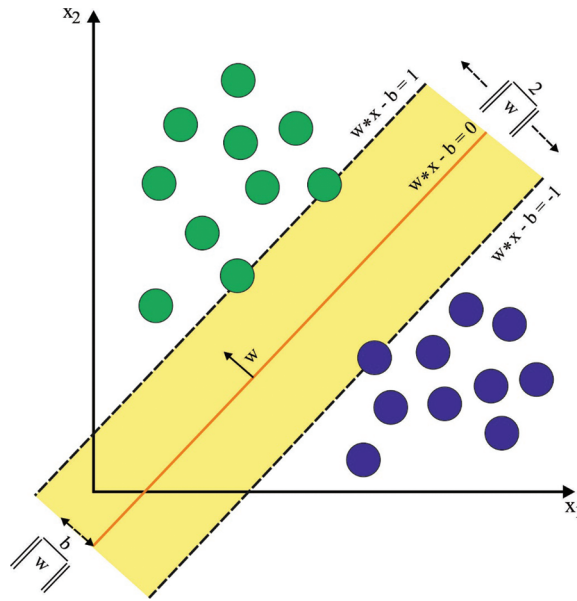

Figure 2: SVM hyperplane

$$F(x) = sign \left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot K(x_i, x_j) + b \right), \tag{12}$$

In Eq. (12), α_i denotes the Lagrange multiplier. $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ refers to a kernel function via mapping function, $\phi(x)$. Quadratic Programming (QP) solver is utilized to find the α_i.

$$w = \sum_{i=0}^{N} \alpha_i \cdot y_i \cdot \phi(x_i), \tag{13}$$

$$b = \frac{1}{N_{SV}} \sum_i \left(y_i - \sum_j \alpha_j \cdot y_j \cdot K(x_j, x) \right). \tag{14}$$

Here, N_{SV} indicates the support vector count and x represents the input unknown instance. Some kernel function is

linear: $k(x, y) = x \cdot y + 1$,

polynomial: $k(x, y) = (x \cdot y + 1)^\sigma$,

RBF: $k(x, y) = \exp(-\|x - y\|/(2 \cdot \sigma^2))$

and quadratic: $1 - \|x - y\|^2/(\|x - y\| + \sigma)$,

In each function, σ needs to be fine-tuned optimally with C.

3.4 ABC Based Parameter Tuning

Finally, ABC algorithm is applied to fine tune the parameters related to SVM. ABC algorithm is a popular algorithm that was developed to overcome continuous optimization problems. The algorithm was inspired from the natural foraging behaviour of honey bees. ABC algorithm makes use of three control variables and three types of artificial bees such as the scouts, the employed bees and the onlooker bees. The employed bee makes half of a colony. As far as the algorithm is concerned, the food sources signify a feasible solution. A single employed bee is allocated for every food source. The employed bee implements the searching method by investigating a solution in its neighbourhood. The onlooker bees select the food sources for exploitation on the basis of the data gained from the employed bees. In case, if a food source does not increase for a predefined number of cycles, the scout bees replace the food source with a novel one that is selected on a random basis. As an iterative process, ABC algorithm initiates again by relating every employed bee with an arbitrarily-produced food source. Every bee x_i ($i = 1, 2, \ldots N$) is a D dimensional vector in which N indicates the population size. The initialized population of the candidate solution is generated using the subsequent formula (15) in which $x_{i,j}$ refers to the *j-th* variable of ith bee in the population, *rand* $(0, 1)$ denotes a random integer that lies in the range of 0 and 1 and ub_j and lb_j indicate the upper and lower limits of the jth variable, correspondingly. In general, x signifies a different element in comparison with the trained instances in Eq. (15).

$$x_{i,j} = lb_j + rand(0, 1) * (ub_j - lb_j) \tag{15}$$

Various formulas are used for fitness function. However, in most of the implementation works, especially for maximization problems, fitness is proportionate to the objective function value. If the issue to be resolved targets the minimization of the function and is represented by *objFun*, then the task is transformed for maximization by means of modification as given herewith.

$$fitness_i = \begin{cases} \dfrac{1}{objFun_i}, & if \ objFun_i > 0 \\ 1 + |objFun_i|, & otherwise \end{cases} \tag{16}$$

Every employed bee finds a food source in the neighbourhood and computes the fitness. The location of the new neighbourhood solution is formulated using the Eq. (17) in which $x_{i,j}$ denotes the jth dimension of the older solution i, $x_{k,j}$ denotes the jth dimension of the neighbourhood solution k, φ represents an arbitrary value that lies in the range of 0 and 1, and *MR* corresponds to the modification rate. *MR* is a control variable of ABC algorithm.

$$v_{i,j} = \begin{cases} x_{i,j} + \varphi * (x_{i,j} - x_{k,j}), & R_j < MR \\ x_{i,j}, & or \ else \end{cases} \tag{17}$$

Once the fitness of the novel solution is higher than the fitness of the previous one, then the employed bees continue to execute exploitation method using novel food sources, or else it retains

the older one. The employed bees share the data regarding the fitness of the food source to onlooker bees whereas the onlooker bees select a food source *i* with probability which is proportionate to the solution fitness.

$$p_i = \frac{fitness_i}{\sum_{i=1}^{N} fitness_i}$$

(18)

4 Experimental Validation

The proposed ABCML-SA model was experimentally validated for sentiment analysis performance using COVID-19 tweets dataset (available at https://www.kaggle.com/competitions/sentiment-analysisof-COVID-19-related-tweets/data?select=validation.csv). The dataset has a total of 2,750 sample tweets under 11 class labels as depicted in Table 1. Each class holds a total of 250 samples.

Table 1: Dataset details

Labels	Description	No. of instances
C1	Optimistic	250
C2	Thankful	250
C3	Empathetic	250
C4	Pessimistic	250
C5	Anxious	250
C6	Sad	250
C7	Annoyed	250
C8	Denial	250
C9	Surprise	250
C10	Official report	250
C11	Joking	250
Total Number of Instances		2750

The confusion matrices, generated by the proposed ABCML-SA model, on the applied dataset under different Training (TR) and Testing (TS) datasets, are shown in Fig. 3. On the entire dataset, the proposed ABCML-SA model recognized 249 samples as C1, 247 samples as C2, 245 samples as C3, 246 samples as C4, 241 samples as C5, 250 samples as C6, 246 samples as C7, 244 samples as C8, 247 samples as C9, 246 samples as C10, and 247 samples as C11 class respectively. Also, on 70% of TR data, ABCML-SA methodology categorized 181 samples under C1, 177 samples under C2, 168 samples under C3, 173 samples under C4, 168 samples under C5, 177 samples under C6, 187 samples under C7, 174 samples under C8, 170 samples under C9, 170 samples under C10 and 160 samples under C11 class correspondingly. Besides, on 30% of TS data, the presented ABCML-SA technique classified 68 samples under C1, 70 samples under C2, 77 samples under C3, 73 samples under C4, 73 samples under C5, 73 samples under C6, 59 samples under C7, 70 samples under C8, 77 samples under C9, 84 samples under C10 and 87 samples under C11 class respectively.
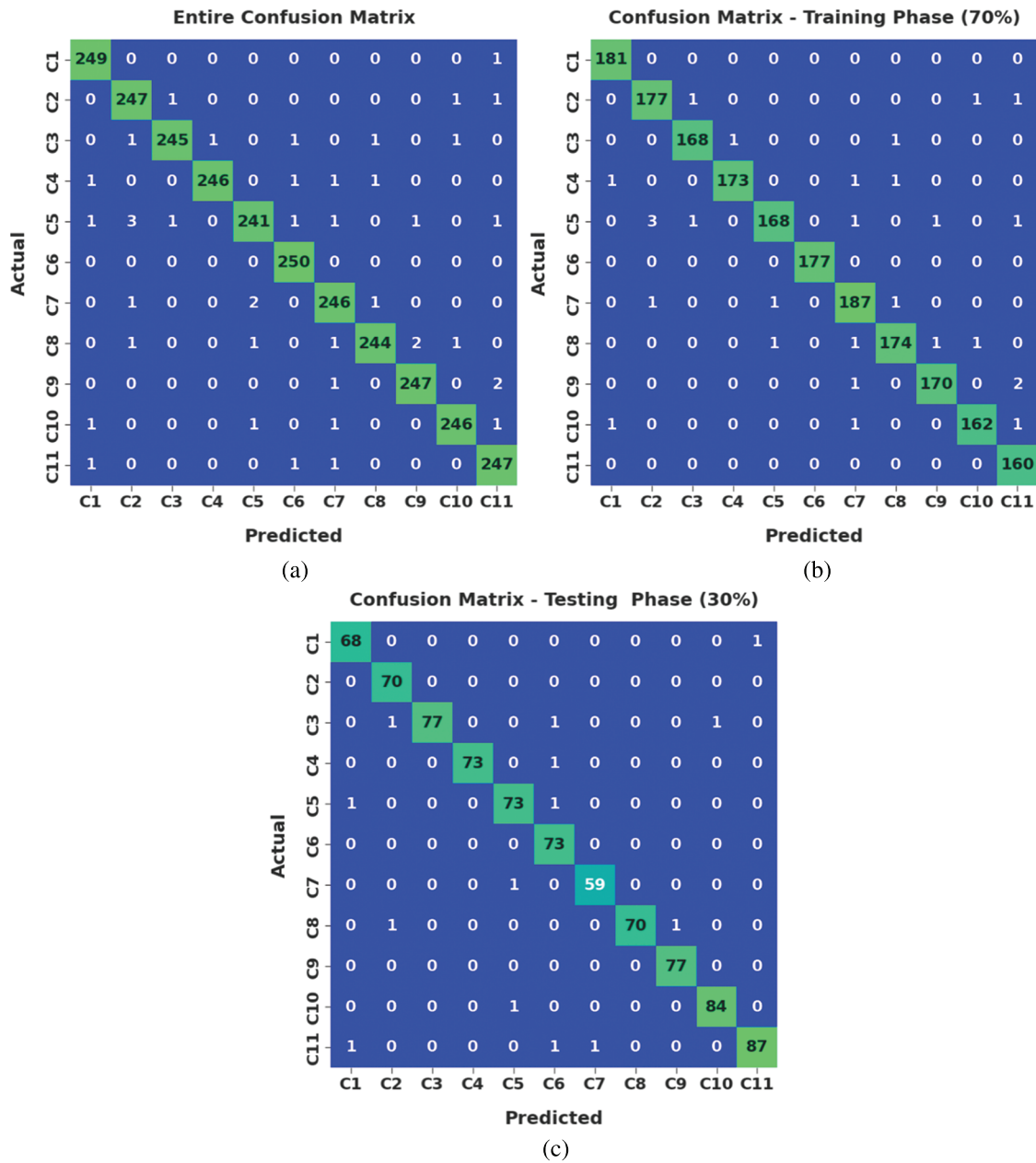
Figure 3: Confusion matrices of ABCML-SA approach (a) entire dataset, (b) 70% of TR data, and (c) 30% of TS data

Table 2 shows the SA outcomes of the proposed ABCML-SA model on the applied data. Fig. 4 shows the $accu_y$ and Area Under the Curve (A UC_{score}) assessment results accomplished by the proposed ABCML-SA model on Twitter data. The figure infers that ABCML-SA model achieved enhanced performance in all the class labels. For instance, on entire dataset-C1, ABCML-SA model achieved an $accu_y$ of 99.82% and an AUC_{score} of 99.72%. Moreover, on entire dataset-C5, ABCML-SA approach accomplished $accu_y$ and AUC_{score} values such as 99.53% and 98.12% respectively. Meanwhile, on entire dataset-C11, the proposed ABCML-SA algorithm presented $accu_y$ and AUC_{score} values such as 99.67%

and 99.28% correspondingly. Likewise, on 70% of data-C1, ABCML-SA approach offered $accu_y$ and AUC_{score} values such as 99.90% and 99.94% respectively. Similarly, on 30% of TS data-C1, the proposed ABCML-SA approach provided $accu_y$ and AUC_{score} values such as 99.64% and 99.14% respectively. At last, on 30% of TS dataset-C11, ABCML-SA methodology attained $accu_y$ and AUC_{score} values such as 99.52% and 98.27% correspondingly.

Table 2: Results of the analysis of ABCML-SA approach under different measures

Labels	Accuracy	Precision	Recall	F-measure	AUC score
Entire dataset					
C1	99.82	98.42	99.60	99.01	99.72
C2	99.67	97.63	98.80	98.21	99.28
C3	99.75	99.19	98.00	98.59	98.96
C4	99.82	99.60	98.40	98.99	99.18
C5	99.53	98.37	96.40	97.37	98.12
C6	99.85	98.43	100.00	99.21	99.92
C7	99.64	97.62	98.40	98.01	99.08
C8	99.67	98.79	97.60	98.19	98.74
C9	99.78	98.80	98.80	98.80	99.34
C10	99.75	98.80	98.40	98.60	99.14
C11	99.67	97.63	98.80	98.21	99.28
Average	99.72	98.48	98.47	98.47	99.16
Training phase (70%)					
C1	99.90	98.91	100.00	99.45	99.94
C2	99.64	97.79	98.33	98.06	99.05
C3	99.79	98.82	98.82	98.82	99.35
C4	99.79	99.43	98.30	98.86	99.12
C5	99.53	98.82	96.00	97.39	97.94
C6	100.00	100.00	100.00	100.00	100.00
C7	99.58	97.40	98.42	97.91	99.07
C8	99.64	98.31	97.75	98.03	98.79
C9	99.74	98.84	98.27	98.55	99.08
C10	99.74	98.78	98.18	98.48	99.03
C11	99.74	96.97	100.00	98.46	99.86
Average	99.74	98.55	98.55	98.55	99.20
Testing phase (30%)					
C1	99.64	97.14	98.55	97.84	99.14
C2	99.76	97.22	100.00	98.59	99.87
C3	99.64	100.00	96.25	98.09	98.12
C4	99.88	100.00	98.65	99.32	99.32
C5	99.52	97.33	97.33	97.33	98.53
C6	99.52	94.81	100.00	97.33	99.73

(Continued)

Table 2: Continued

Labels	Accuracy	Precision	Recall	F-measure	AUC score
C7	99.76	98.33	98.33	98.33	99.10
C8	99.76	100.00	97.22	98.59	98.61
C9	99.88	98.72	100.00	99.35	99.93
C10	99.76	98.82	98.82	98.82	99.34
C11	99.52	98.86	96.67	97.75	98.27
Average	99.69	98.29	98.35	98.31	99.09
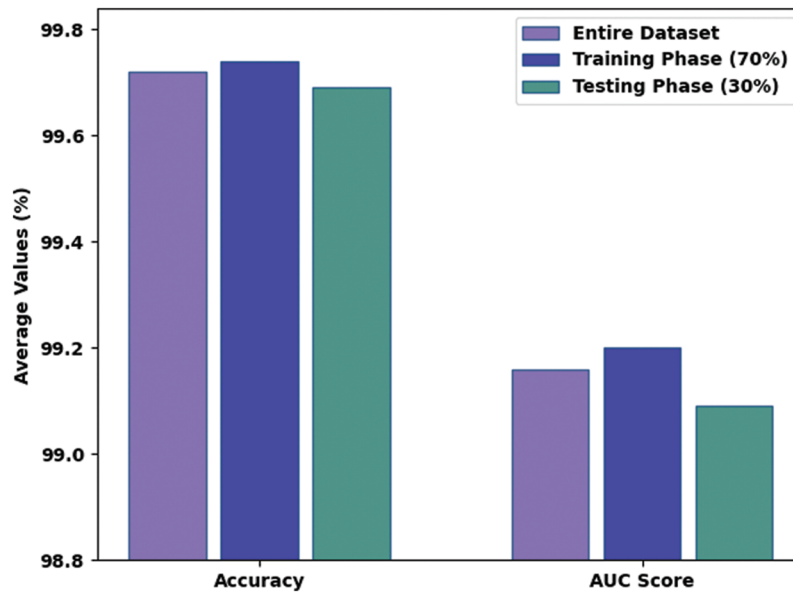
Figure 4: $Accu_y$ and AUC_{score} analyses results of ABCML-SA approach

Fig. 5 depicts the $prec_n$, $reca_l$, and $F_{measure}$ assessment results accomplished by the proposed ABCML-SA methodology on Twitter data. The figure infers that the proposed ABCML-SA technique attained increased performance in all the class labels. For example, on entire dataset-C1, the proposed ABCML-SA algorithm achieved $prec_n$, $reca_l$ and $F_{measure}$ values such as 98.42%, 99.60%, and 99.01% respectively. Furthermore, on entire dataset-C5, ABCML-SA approach attained $prec_n$, $reca_l$ and $F_{measure}$ values such as 98.37%, 96.40%, and 97.37% correspondingly. At the same time, on entire dataset-C11, ABCML-SA method yielded $prec_n$, $reca_l$, and $F_{measure}$ values such as 97.63%, 98.80%, and 98.21% correspondingly. Similarly, on 70% of data-C1, ABCML-SA approach achieved $prec_n$, $reca_l$ and $F_{measure}$ values such as 98.91%, 100%, and 99.45% respectively. Likewise, on 30% of TS data-C1, the proposed ABCML-SA model established $prec_n$, $reca_l$, and $F_{measure}$ values such as 97.14%, 98.55%, and 97.84% correspondingly. Finally, on 30% of TS dataset-C11, ABCML-SA algorithm accomplished $prec_n$, $reca_l$ and $F_{measure}$ values such as 98.86%, 96.67%, and 97.75% respectively.
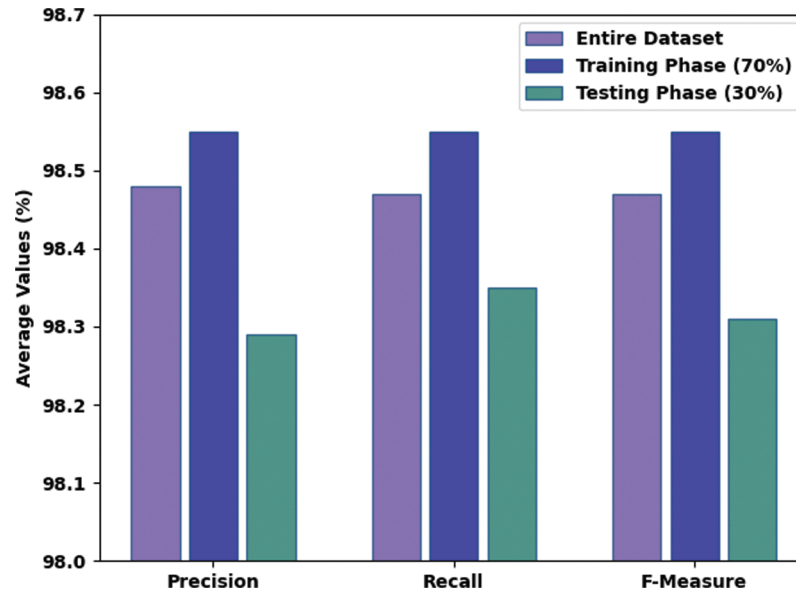
Figure 5: $Prec_n$, $reca_l$, and $F_{measure}$ analyses results of ABCML-SA approach

Both Training Accuracy (TA) and Validation Accuracy (VA) values, acquired by the proposed ABCML-SA method on test dataset, are demonstrated in Fig. 6. The experimental outcomes infer that the proposed ABCML-SA approach attained the maximum TA and VA values whereas VA values were higher than TA.
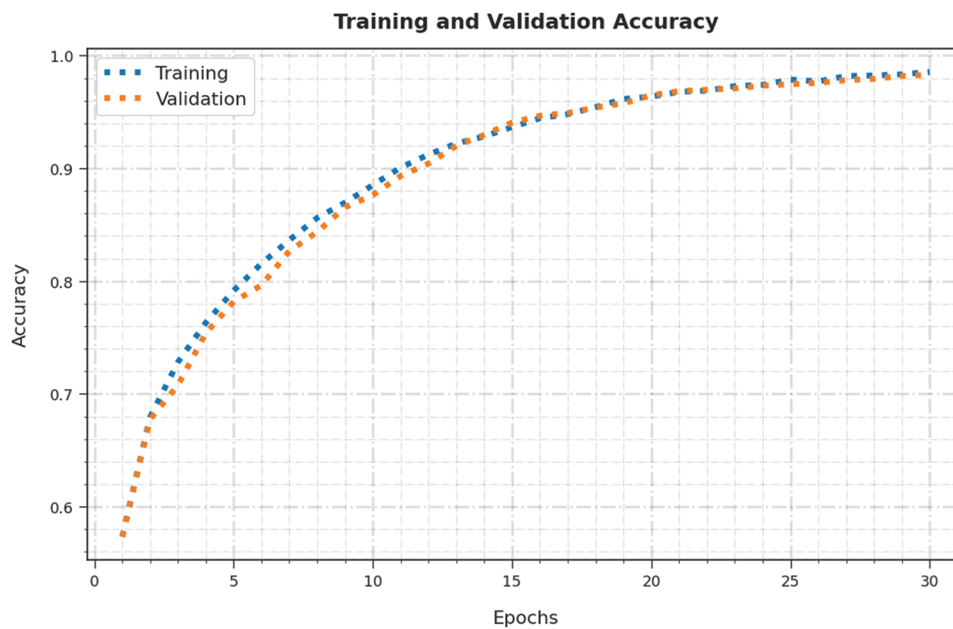

Figure 6: TA and VA analyses results of ABCML-SA approach

Both Training Loss (TL) and Validation Loss (VL) values, obtained by the proposed ABCML-SA approach on test dataset, are shown in Fig. 7. The experimental outcomes imply that the proposed ABCML-SA algorithm established the minimal TL and VL values whereas VL values were lesser than TL.
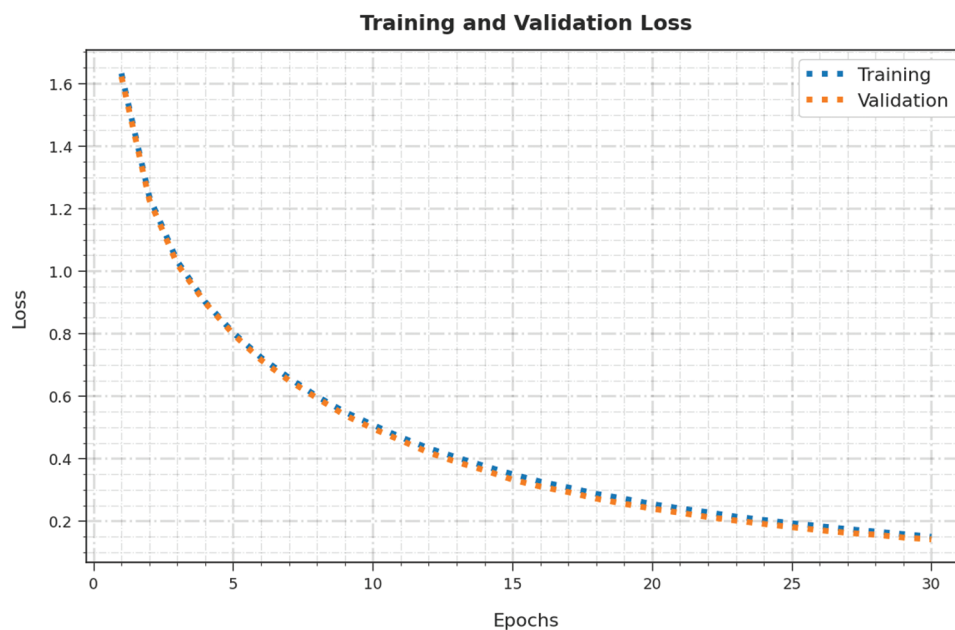

Figure 7: TL and VL analyses results of ABCML-SA approach

A comparative assessment was conducted between ABCML-SA model and other recent models and the results are shown in Table 3 [19]. Fig. 8 portrays the comparison study results accomplished by ABCML-SA model and other existing models in terms of $accu_y$. The figure denotes that the SVM model achieved poor results with the least $accu_y$ of 88.26%. Along with that, Random Forest (RF), Extreme Gradient boosting (XGBoost) and Extra Trees models demonstrated slightly enhanced $accu_y$ values such as 90.87%, 91.26%, and 91.56% respectively. Followed by, Decision Tree (DT) model gained a reasonable performance with an $accu_y$ of 92.35%. However, the proposed ABCML-SA model accomplished the maximum classification results with the highest $accu_y$ of 99.69%.

Table 3: Comparative analysis results of ABCML-SA approach and other existing methodologies

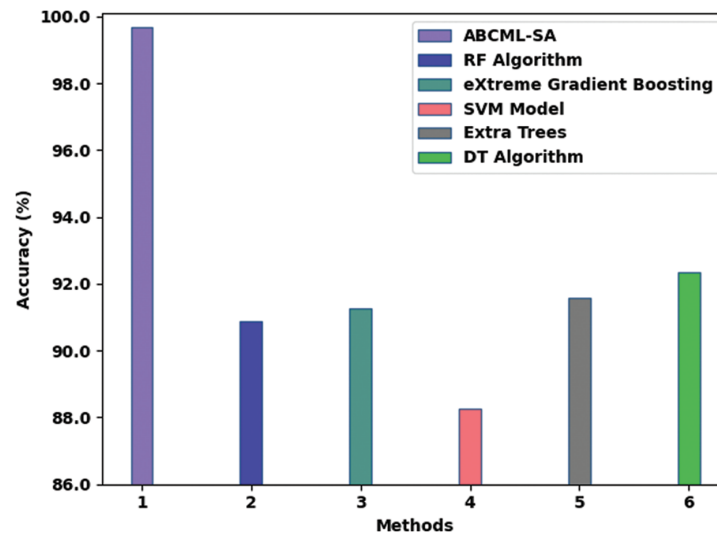| Methods	Accuracy	Precision	Recall	F-measure
ABCML-SA	99.69	98.29	98.35	98.31
RF algorithm	90.87	90.82	90.93	91.04
Extreme gradient boosting	91.26	89.84	89.31	91.28
SVM model	88.26	89.96	87.92	90.18
Extra trees	91.56	92.66	91.51	93.10
DT algorithm	92.35	93.34	95.92	94.01

Figure 8: *Accu_y* analysis results of ABCML-SA approach and other existing methodologies

Fig. 9 shows the results achieved by the proposed ABCML-SA method and other existing models in terms of *prec_n*. The figure displays that the SVM method achieved poor results with the least *prec_n* of 89.96%. Also, RF, XGBoost and Extra Trees algorithms demonstrated slightly enhanced *prec_n* values such as 90.82%, 89.84% and 89.96% correspondingly. Then, DT approach attained a reasonable performance with a *prec_n* of 93.34%. However, the proposed ABCML-SA approach established the maximum classification results with the highest *prec_n* of 98.29%.
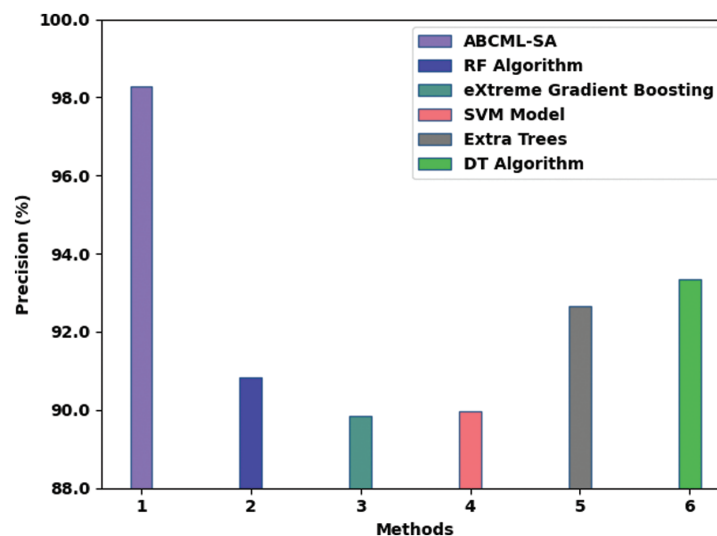

Figure 9: *Prec_n* analysis results of ABCML-SA approach and other existing methodologies

Fig. 10 indicates the detailed study outcomes achieved by ABCML-SA methodology and other existing models in terms of *reca_l*. The figure shows that the SVM technique achieved poor results with the least *reca_l* of 87.92%. Moreover, RF, XGBoost, and Extra Trees models demonstrated slightly enhanced *reca_l* values such as 90.93%, 89.31%, and 91.51% correspondingly. Then, DT model achieved

a reasonable performance with a *reca_l* of 95.92%. However, the proposed ABCML-SA approach established the maximum classification results with the highest *reca_l* of 98.35%.
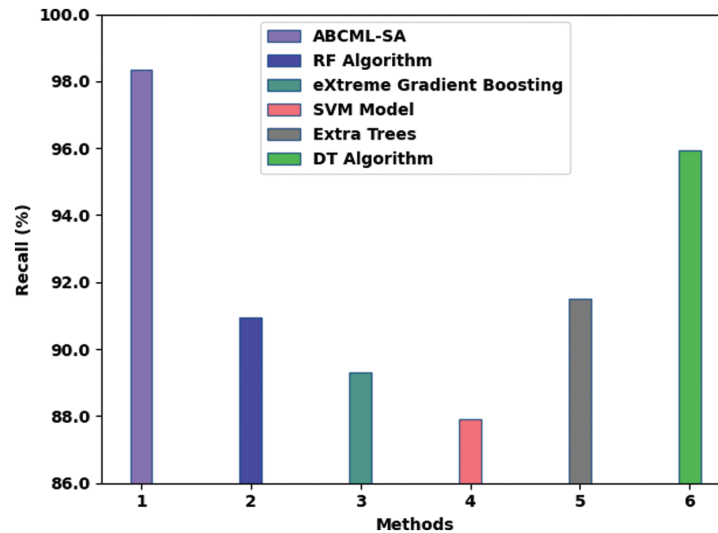

Figure 10: *Reca_l* analysis results of ABCML-SA approach and other existing methodologies

Fig. 11 depicts the comparative study results accomplished by the proposed ABCML-SA algorithm and other existing models in terms of $F_{measure}$. The figure denotes that the SVM algorithm achieved poor results with the least $F_{measure}$ of 90.18%. Furthermore, RF, XGBoost, and Extra Trees models demonstrated slightly enhanced $F_{measure}$ values such as 91.04%, 91.28%, and 93.10% correspondingly. Then, DT model achieved a reasonable performance with an $F_{measure}$ of 94.01%. But, the proposed ABCML-SA approach accomplished the maximum classification results with the highest $F_{measure}$ of 98.31%. Thus, it can be confirmed that the proposed ABCML-SA model gained the maximum results compared to other models.
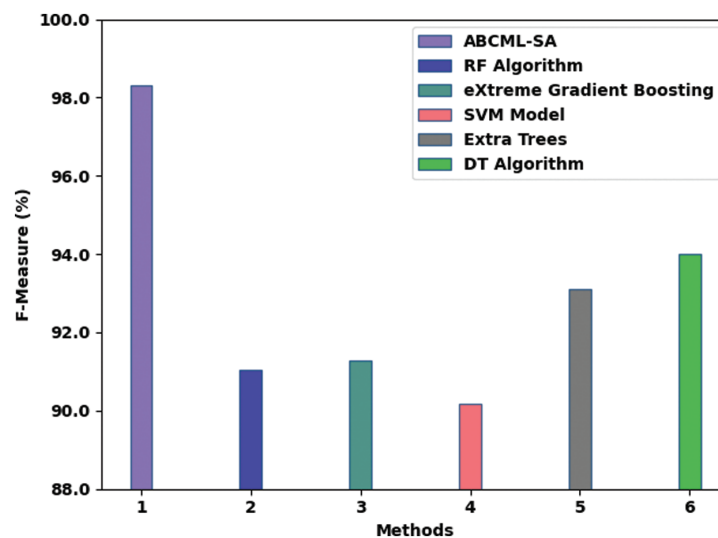

Figure 11: $F_{measure}$ analysis results of ABCML-SA approach and other existing methodologies

5 Conclusion

In this study, a new ABCML-SA method has been developed for conducting Sentiment Analysis on COVID-19 Twitter data. The key objective of the presented ABCML-SA model is to recognize the sentiments expressed in COVID-19 tweets. It involves data pre-processing at the initial stage followed by n-gram based feature extraction in the second stage to derive the feature vectors. For identification and classification of the sentiments, the SVM model is exploited. At last, the ABC algorithm is applied to fine tune the parameters involved in SVM. To demonstrate the enhanced performance of ABCML-SA model, a sequence of simulations was conducted. The comparative analysis outcomes confirmed the effectual performance of the proposed ABCML-SA model over other approaches.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1]   D. V. Cruz, V. F. Cortez, A. L. Chau and R. S. Almazán, "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the COVID-19 periods," *Cognitive Computation*, vol. 14, no. 1, pp. 372–387, 2021.

[2]   S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, pp. e21978, 2020.

[3]   K. H. Manguri, R. N. Ramadhan and P. R. Mohammed Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan Journal of Applied Research*, vol. 5, no. 3, pp. 54–65, 2020.

[4]   S. H. Sahir, R. S. A. Ramadhana, M. F. R. Marpaung, S. R. Munthe and R. Watrianthos, "Online learning sentiment analysis during the COVID-19 Indonesia pandemic using twitter data," *IOP Conference Series: Materials Science and Engineering*, vol. 1156, no. 1, pp. 012011, 2021.

[5]   M. Rahman and M. N. Islam, "Exploring the performance of ensemble machine learning classifiers for sentiment analysis of COVID-19 tweets," in *Sentimental Analysis and Deep Learning*, vol. 1408, pp. 383–39, 2022, Springer, Singapore: Advances in Intelligent Systems and Computing book series.

[6]   K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag *et al.,* "Sentiment analysis of COVID-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, vol. 97, pp. 106754, 2020.

[7]   Z. B. Nezhad and M. A. Deihimi, "Twitter sentiment analysis from Iran about COVID 19 vaccine," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 16, no. 1, pp. 102367, 2022.

[8]   C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng and J. G. Hsieh, "Twitter sentiment analysis towards COVID-19 vaccines in the Philippines using naïve Bayes," *Information*, vol. 12, no. 5, pp. 204, 2021.

[9]   N. A. Alkhaldi, Y. Asiri, A. M. Mashraqi, H. T. Halawani, S. A. Khalek *et al.,* "Leveraging tweets for artificial intelligence driven sentiment analysis on the COVID-19 pandemic," *Healthcare*, vol. 10, no. 5, pp. 910, 2022.

[10] Y. Lu and Q. Zheng, "Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic," *Current Issues in Tourism*, vol. 24, no. 7, pp. 892–898, 2021.

[11] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, 2021.

[12] M. Kausar, A. Soosaimanickam and M. Nasar, "Public sentiment analysis on twitter data during COVID-19 outbreak," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 415–422, 2021.

[13] A. Kruspe, M. Häberle, I. Kuhn and X. X. Zhu, "Cross-language sentiment analysis of european twitter messages during the COVID-19 pandemic," arXiv:2008.12172v1, pp. 1–10, 2020.

[14] C. K. Pastor, "Sentiment analysis of Filipinos and effects of extreme community quarantine due to coronavirus (COVID-19) pandemic," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3574385, Apr. 2020. 2022. [Online]. Available: https://papers.ssrn.com/abstract=3574385 (accessed on 15 June 2022).

[15] U. Yaqub, "Tweeting during the COVID-19 pandemic: Sentiment analysis of twitter messages by president trump," *Digital Government: Research and Practice*, vol. 2, no. 1, pp. 1–7, 2021.

[16] T. Pano and R. Kashef, "A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of COVID-19," *Big Data and Cognitive Computing*, vol. 4, no. 4, pp. 33, 2020.

[17] M. Ghiassi, J. Skinner and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6266–6282, 2013.

[18] D. K. Pathak, S. K. Kalita and D. K. Bhattacharya, "Hyperspectral image classification using support vector machine: A spectral spatial feature based approach," *Evolutionary Intelligence*, pp. 1–15, 2021, DOI: 10.1007/s12065-021-00591-0.

[19] K. Shankar, E. Perumal, V. G. Díaz, P. Tiwari, D. Gupta *et al.,* "An optimal cascaded recurrent neural network for intelligent COVID-19 detection using chest X-ray images," *Applied Soft Computing*, vol. 113, Part A, pp. 1–13, 2021.