Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.033945
Article

Convolutional Deep Belief Network Based Short Text Classification on Arabic Corpus

Abdelwahed Motwakel[1,*], Badriyya B. Al-onazi[2], Jaber S. Alzahrani[3], Radwa Marzouk[4], Amira Sayed A. Aziz[5], Abu Sarwar Zamani[1], Ishfaq Yaseen[1] and Amgad Atta Abdelmageed[1]

[1]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
[2]Department of Language Preparation, Arabic Language Teaching Institute, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[3]Department of Industrial Engineering, College of Engineering at Alqunfudah, Umm Al-Qura University, Saudi Arabia
[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[5]Department of Digital Media, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, 11835, Egypt
*Corresponding Author: Abdelwahed Motwakel. Email: a.ismaeil@psau.edu.sa
Received: 02 July 2022; Accepted: 18 August 2022

Abstract: With a population of 440 million, Arabic language users form the rapidly growing language group on the web in terms of the number of Internet users. 11 million monthly Twitter users were active and posted nearly 27.4 million tweets every day. In order to develop a classification system for the Arabic language there comes a need of understanding the syntactic framework of the words thereby manipulating and representing the words for making their classification effective. In this view, this article introduces a Dolphin Swarm Optimization with Convolutional Deep Belief Network for Short Text Classification (DSOCDBN-STC) model on Arabic Corpus. The presented DSOCDBN-STC model majorly aims to classify Arabic short text in social media. The presented DSOCDBN-STC model encompasses preprocessing and word2vec word embedding at the preliminary stage. Besides, the DSOCDBN-STC model involves CDBN based classification model for Arabic short text. At last, the DSO technique can be exploited for optimal modification of the hyperparameters related to the CDBN method. To establish the enhanced performance of the DSOCDBN-STC model, a wide range of simulations have been performed. The simulation results confirmed the supremacy of the DSOCDBN-STC model over existing models with improved accuracy of 99.26%.

Keywords: Arabic text; short text classification; dolphin swarm optimization; deep learning

1 Introduction

Arabic was the mother tongue of almost 300 billion persons roughly twenty-two nations; and the target language of this study, it is the 5th largest national language in the list of top 100 natural languages spoken all

over the world [1]. The writing orientation of Arabic language will be from right to left. Any Arabic word was a grouping of twenty-eight letters which belonging to Arabic alphabetic sets. The twenty-eight letters were extensible to 90 letters because of additional marks, vowels, and writing shapes. The Arabic language contains 2 main formats one is Dialectal (colloquial) Arabic and another one is Standard Arabic [2,3]. In Standard Modern Standard Arabic (MSA) and Arabic Classical Arabic (CA) will be presented and CA becomes the historical linguistic utilized in the Hadith and Quran [4]. MSA was the formal format and utilized in newspapers, books, TV, formal speeches, and education. But Arabic speakers employed the dialectal format in everyday communication and whenever it expresses their views regarding distinct aspects of life on mass media [5,6]. There exist several Arabic dialects, but 6 are considered as main such as Moroccan, Egyptian, Iraqi, Levantine, Yemini, and Gulf [7]. The dialects were one such cause for introducing several new words to any language specifically stop words [8]. Even though Arabic becomes a broadly utilized language, OM studies have only held recently, and this domain needs more research because of the exceptional nature of Arabic linguistic morphology values [9]. Arabic opinion mining has trouble because of the poorness of language sources and Arabic-specific language characteristics [10].

Arabic language users frame the rapidly growing linguistic group over the web in terms of number of Internet users [11]. 11 billion people were active on Twitter posted nearly 27.4 million tweets every day. In accordance with the survey, most of the youngest Arabs receive news from Twitter and Facebook, not television [12,13]. Like others, the region was rife with rumours and fake news. Advancing a classifier mechanism for Arabic linguistic needs understanding of syntactic framework of words so it could represent and manipulate the words for making their categorization very accurate [14]. The research into Arabic text classifiers can be confined when compared with the research volume on English textual classifiers. There exists less research volume on Arabic short textual classifications [15]. Nevertheless, the language features, the inaccessibility of free accessibility to Arabic short text corpus were other reasons [16].

Hawalah [17] devise an improved Arabic topic-discovery architecture (EATA) that uses ontology for offering an effectual Arabic topic classifier system. And presented a semantic enhancing method for enhancing Arabic text classifier and topic discovery method by using rich semantic data in Arabic ontology. Then relies in the paper on vector space method term frequency-inverse document frequency (TF-IDF) and the cosine similarity technique for classifying new Arabic text files. Ibrahim et al. [18] assess Arabic short text classifier utilizing 3 standard NB classifiers. In this technique, the classifications are made with the dissertations and thesis utilizing their titles for performing the classifier procedure. The collected data set was gathered from distinct sources by utilizing standard scrapping algorithms. This algorithm categorizes the document on the basis of their titles and is positioned in the wanted specialization. Numerous pre-processing methods were implied, like (space vectorization, punctuation removal, and stop words removal).

Beseiso et al. [19] modelled a new structure for hand Arabic words classifier and comprehends depending on recurrent neural network (RNN) and convolutional neural network (CNN). In addition, CNN method was very influential for scrutiny of social network analysis and Arabic tweets. The predominant method employed in this article was character level CNN and an RNN stacked on top of each other as the classifier structure. Najadat et al. [20] project a keyword-related algorithm to detect Arabic spam reviews. Features or Keywords were words subsets from original text which were labelled as significant. A weight of term, TF-IDF matrix, and filter approaches (like correlation, information gain, uncertainty, deviation, and chi-squared) were employed for extracting keywords from Arabic text. Albalawi et al. [21] offer a complete assessment of data preprocessing and word embedded algorithms with regard to Arabic document classifiers in the field of health-based transmission on mass media. And assess twenty-six text preprocessing implied to Arabic tweets in the training processes a classifier for identifying health-based tweets. For this one uses Logistic Regression, traditional machine learning (ML)

classifiers, etc. Additionally, reported experimental outcomes with the deep learning (DL) structures bidirectional long short term memory (BLSTM) and CNN for similar textual classifier issues.

This article introduces a Dolphin Swarm Optimization with Convolutional Deep Belief Network for Short Text Classification (DSOCDBN-STC) model on Arabic Corpus. The presented DSOCDBN-STC model majorly aims to classify Arabic short text in social media. The presented DSOCDBN-STC model encompasses preprocessing and word2vec word embedding at the preliminary stage. Besides, the DSOCDBN-STC model involves CDBN based classification model for Arabic short text. At last, the DSO approach can be exploited for optimal modification of the hyperparameters related to the CDBN technique. To accomplish the enhanced performance of the DSOCDBN-STC model, a wide range of simulations have been performed. The simulation results confirmed the supremacy of the DSOCDBN-STC model over existing models.

2 The Proposed DSOCDBN-STC Model

In this article, a novel DSOCDBN-STC model was projected for short text classification on Arabic Corpus. The presented DSOCDBN-STC model majorly aims to classify Arabic short text in social media. The presented DSOCDBN-STC model encompasses pre-processing and word2vec word embedding at the preliminary stage. Besides, the DSOCDBN-STC model involves CDBN based classification model for Arabic short text. Fig. 1 demonstrates the overall process of DSOCDBN-STC method.

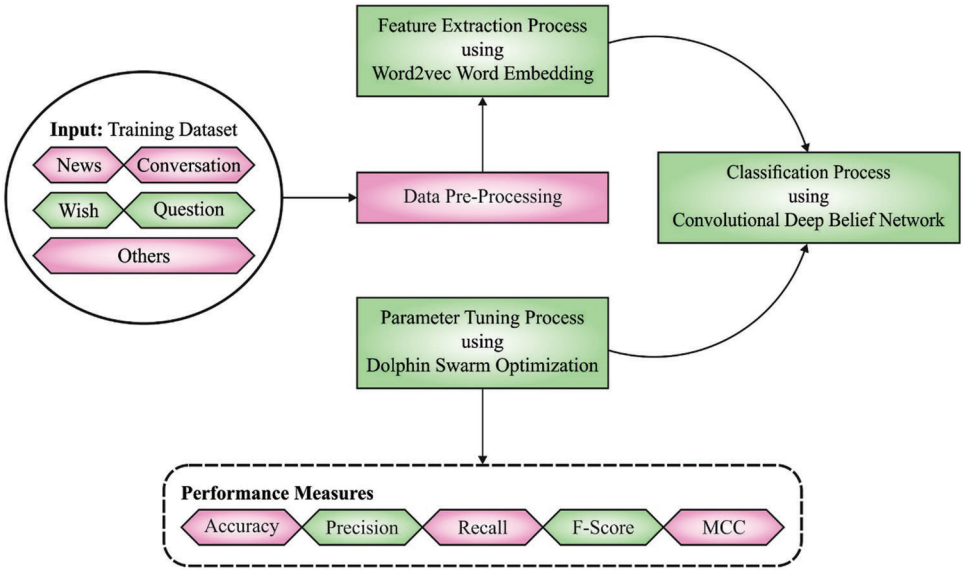

Figure 1: Overall process of DSOCDBN-STC approach

2.1 Data Pre-processing

The presented DSOCDBN-STC model encompasses pre-processing and word2vec word embedding at the preliminary stage. Raw datasets frequently needed to be pre-processed. Text preprocessing is a significant phase in text mining [22]. The pre-processing steps are given in the following:

1. Data Cleansing: As the comment contains numerous syntactic features that mayn't be beneficial for ML algorithm, the information needs to be cleaned via eliminating website link or URL (www or http). Also, comments could have repetitive letters once the user needs to highlight specific words, and the letter should be eliminated. Also, Diacritics, Emoticons, and special characters are eliminated.

2. Tokenization: they break sentences into meaningful tokens, symbols, words, and phrases by eliminating punctuation marks.
3. Stop word removal: they have shared words which don't add useful content to a file. For instance, (و, أما, على, إلى, من (in English 'from, to, on, as for, and'.
4. Normalization: The letter has multiple forms are normalized into single form. Alef in Arabic comprises various formats (ا, آ, إ, أ ,(normalized to (ا, (and Taa Almarbotah (ة, ه (was normalized to. (ه) are some of the examples.
5. POS tagging: This stage was implemented to recognize diverse POS in the text. For word embedding, we utilize Word2vec, a common prediction-based approach viz. effective interms of time and space. Word2vec is a two-layer NN, whereby input was the document, and output can be a set of real-valued feature vectors–one vector for each word–of a pre-set dimension.

2.2 Short Text Classification Using CDBN Model

Next to data pre-processing, the DSOCDBN-STC model involves CDBN based classification model for Arabic short text. A DBN is a generative graphical method which consists of a stack of Restricted Boltzmann Machine (RBM), composed by single layer of visible unit (input dataset) and multiple layers of hidden unit. The relation between the topmost 2 layers of the DBN is undirected whereas the remaining connection is directed but there is no correlation for the node in similar layer [23].

An RBM is a two layer undirected graphical method comprised of single layer of hidden unit "h", and single layer of visible unit "v", having a complete set of connections amongst them. The probabilistic semantics and the energy function for an RBM are described below:

$$E(v,h) = -\sum_{i,j} v_i w_{ij} h_j - \sum_j b_j h_j - \sum_j c_j v_j \tag{1}$$

$$P(v_t h) = \frac{1}{Z} \exp\left(-E(v,\ \ h)\right) \tag{2}$$

From the expression, $\cdot w_{ij}$ characterizes the weights among hidden units h_j and visible units v_i. b_j refers to the hidden unit bias, c_i indicates visible unit bias, Z represents the partition function. Once the visible unit is real valued, then the energy function is evaluated by:

$$E(v,\ h) = \frac{1}{2}\sum_i v_i^2 - \sum_{i,j} v_i w_{i,j} h_j - \sum_j b_j h_j - \sum_j c_i v_i \tag{3}$$

RBM is trained to learn a generative mechanism in an unsupervised way. One effective learning model for RBM is contrastive divergence "CD" that is a form of contrastive optimization that approximates the gradient of log probability of the learning dataset.

Building hierarchical structure of feature is a challenge and CDBN amongst the prominent feature extractors are widely employed in the region of pattern detection.

Convolution DBN is a hierarchical generative mechanism: supports effective top-down and bottom-up probabilistic inferences. Similar to typical DBN, this framework comprises multiple layers of max pooling CRBM stacked on top of each other, besides training can be attained through the greedy layer-by-layer process. Constructing convolution DBN, the algorithm learns higher level features namely object-part and groups of the strokes. In this study, we trained CDBN with two layers of CRBM, and for inference, apply feedforward estimation [24]. The CRBM is the basis of CDBN. We trained the CDBN method through learning a stack of CRBM whereby the output of single CRBM is the input of following CRBM. CRBM encompasses a hidden layer H and a visible layer V that are interconnected by sets of shared and local parameters. The hidden unit is binary-valued, as well as visible unit is real-valued or binary-valued.

Assume the hidden unit H is separated into K groups (maps), whereby every group is an $N_H \times N_H$ array of binary units and is related to a $N_W \times N_W$ convolution filter ($N_W \Delta N_V - N_H + l$). The visible input layer encompasses L images (with arbitrary aspect ratio), and every image comprises of $N_V \times N_V$ real unit (intensity pixel images).

The filter weight is shared among each location in the hidden unit within a similar map. Also, there exists a shared bias b_k for every group and a shared bias c for the visible unit.

A new function for CDBN structure termed "probabilistic max pooling" shrinks the presentation of detection layer in a probabilistic sound way. Shrinking the presentation with max -pooling enables high layer representation is invariant to local translation of input dataset, decreases the computation burden and it shows to be beneficial in visual recognition problems as described by:

$$E(v,\ h) = \frac{1}{2}\sum_{i,j=1}^{N_V} v_{i,j}^2 - \sum_{k=1}^{K}\sum_{i,j=1}^{N_H}\sum_{r,s=1}^{N_W} h_{i,j}^k W_{r,s}^k v_{i+r-1,j+s-1}$$

$$- \sum\nolimits_{k=1}^{K} b_k \sum\nolimits_{i,j=1}^{N_H} h_{i,j}^k - c\sum\nolimits_{i,j=1}^{N_V} v_{i,j} \tag{4}$$

The conditional and joint likelihood distribution of the CRBM is formulated by:

$$P(v,\ h) = \frac{1}{z}\ \exp\ (-E(,\ h)) \tag{5}$$

$$P(v_{i,j}\ = 1|h) = N(\left(\sum\nolimits_k W^k * fh^k\right)_{\mathrm{i},j} + c,\ 1\Big) \tag{6}$$

$$P\left(h_{i,j}^k = 1|v\right) = \frac{\exp\left(I\left(h_{i,j}^k\right)\right)}{1 + \sum_{(i'j')\in B\alpha}\ \exp\ \left(I\left(h_{i'j'}^k\right)\right)} \tag{7}$$

From the expression, $I\left(h_{i,j}^k\right) \Delta\, b_k + (\tilde{w} *_v v)_{i,j}$ refers to the hidden data unit in group k gained from visible layer V, $\tilde{w}$ determined by matrix filter W flipped in left-right direction and up-down side, N is a normal distribution, $*_\alpha$ is a valid convolutional layer and $*_f$ is a full convolutional layer. B_α represent a $C \times C$ block represented α whereby it is interconnected (pooled) to binary node P_α^k in pooling layer. The pooling node P_α^k is described by $P_\alpha^k = \Delta \sum_{(i,j)} h_{i,j}^k$ and the conditional probability is represented as follows:

$$P\left(p_\alpha^k = 1|v\right) = \frac{\sum_{(i'j')\in B\alpha}\ \exp\left(I\left(h_{(i'j')}^k\right)\right)}{1 + \sum_{(i'j')}\ \exp\left(I\left(h_{i'j'}^k\right)\right)} \tag{8}$$

By utilizing the operator previously determined (4),

$$E(v,\ h) = \frac{1}{2}\sum\nolimits_{i,j=1}^{N_V} v_{i,j}^2 - c\sum\nolimits_{i,j=1}^{N_V} v_{i,j} - \sum\nolimits_{k=1}^{K}\sum\nolimits_{i,j}(h_{i,j}^k(\left(\tilde{w} *_v\right)_{i,j} + b_k\Big) \tag{9}$$

Like RBM, training CRBM can be done by utilizing CD model that is an estimate of maximal likelihood approximation. Moreover, CD allows us to evaluate an estimated gradient efficiently. Learning and Inference models depend on block Gibbs sampling model. The hidden unit "activation" is exploited by the input for training the subsequent CRBM layer.

2.3 Hyperparameter Optimization Using DSO Algorithm

At last, the DSO technique can be exploited for optimal modification of the hyperparameters related to the CDBN method. DSO comprises multiple stages namely search, initialization, call, predation, and reception, involve the predatory procedure of dolphins, and the characteristics and habits assist dolphins to achieve the goal all over the predatory procedure [25]. Based on swarm intelligence, a certain amount of dolphins is required to simulate biological characteristics and living habits specified in the real predatory method of dolphins. DSO is classified into five stages in the following

1) Initialization stage: randomly and evenly producing primary dolphins swarming, Dol_i = [x1, x2, …, xD] T(i = 1, 2, …, N) whereby N represented as dolphin count. xj represent module concerning every dimension that is enhanced. Afterward, initialization, estimate fitness for every dolphin and attain Fit_k. $Fit_k = \{Fit_{k,1},\ Fit_{k,2},\ \ldots,\ Fit_{k,\,\mathrm{N}}\}$
2) Search stage: here, every dolphin implements a search of nearby region through emitting sound in M random direction. In the maximum search time Tl, sound V_j that Dol_i, $(i = 1,\ 2,\ \ldots,\ N)$ generates in time t find a novel solution X_{ijt}, as follows:

$$X_{ijt} = Dol_i + V_i * t \tag{10}$$

For the novel solution X_{ijt} that Dol_i attains, its fitness E_{ijt} is evaluated as follows:

$$E_{ijt} = Fit(X_{ijt}) \tag{11}$$

$$if\ \left(E_{iab} = \min_{j=1,2,\ldots,M;t=1,2,\ldots,T1} E_{ijt}\right) \tag{12}$$

In such cases, the individual optimal solution L_i of Dol_i is given by:

$$L_i = X_{iab} \tag{13}$$

$$if\ (Fitness\ (L_i) < Fitness\ (k_i)) \tag{14}$$

next K_i is substituted with L_i; or else, K_i doesn't alter. Afterward each Dol_i $(i = 1,\ 2,\ \ldots,\ N)$ upgrade the L_i and K_i, DSO enters call stage.

3) Call stage: here, dolphin generates sound to inform others in the searching stage that involves a better solution was attained and the position of that best solution. The matrix of transmitting time TS whereby $TS_{i,j}$ denotes the remaining duration for sound to travel from Dol_i to Dol_i and require upgraded as follows: For K_i, K_j, and $TS_{i,j}$

if (Fitness $(K_i) <$ Fitness $\left(K_j\right)$

and

$$TS_{i,j} > \left\lceil \frac{DD_{ij}}{A * speed} \right\rceil$$

then

$$TS_{i,j} = \left\lceil \frac{DD_{i,j}}{A.speed} \right\rceil \tag{15}$$

Or else $TS(i, j)$ remains its value. $(i = 1,\ 2,\ \ldots,\ N;\ j = 1,\ 2,\ \ldots,\ N)$ and $DD_{i,j}$ indicates the distance amongst Dol_i and Dol_i.

$$DD_{ij} = \|Dol_i - Dol_j\|, \quad i, j = 1, 2, \quad \ldots N, \quad i \neq j \tag{16}$$

Speed characterizes a constant corresponding to sound attributes. A denotes a constant that shows acceleration can able to make sound travel at a high speed if lower speed, next, $TS_{i,j}$ undergoes update.

4) Reception stage: here, the procedure of exchange (including call and reception stages) is preserved with the TS, where the DSO enters the reception stage, each term $TS_{i,j}$ $(i = 1, 2, \ldots, N; j = 1, 2, \ldots, N)$, next, TS reduces by 1 to represent that sound propagates over 1. In such cases, the DSO needs to check every term $TS_{i,j}$ in a matrix

$$if \ \left(TS_{i,j} = 0\right) \tag{17}$$

Sound is transferred from $Do1_j$ to $Do1_i$ is attained by Dol_i, where is a requirement to replace the $TS_{i,j}$ by novel time term is denoted by "maximal transmitting time" (T2), to show that the corresponding sound was received.

$$if \ \left(Fitness \ (K_i) > Fitness \ \left(K_j\right)\right) \tag{18}$$

K_i will be substituted to K_j; else, K_i remain the same. Afterward, every term in the matrix TS that fulfills Eq. (17) is managed, DSO starts the predation stage.

5) Predation stage: here, dolphin is requisite to calculate the encircling radius R2, defining a distance among optimal solution of neighboring dolphin and its location succeeding the stage of predation according to the presented dataset, and later, achieves a novel location and it is evaluated as follows:

distance DK:

$$DK_i = ||Dol_i - K_i||, \quad i = 1, 2, \quad \ldots N \tag{19}$$

distance DKL:

$$DKL_i = ||L_i - K_i||, \quad i = 1, 2, \quad \ldots N \tag{20}$$

Rl: characterizes the radius of search, demonstrating the maximal searching stage is evaluated in the subsequent stage:

$$R1 = T1 \times \ speed \tag{21}$$

Generally, computing the encircling radius R2 and dolphin location update is given below.

$$\begin{pmatrix} if (DK_i \leq R1) \\ Then \ R_2 = \left(1 - \dfrac{2}{e}\right) DK_i \end{pmatrix} \tag{22}$$

$$newDol_i = K_i + \frac{Dol_{i-K_i}}{DK_i} R_2 \tag{23}$$

$$\begin{pmatrix} If (DK_i > R1) \ and \ DK_i \geq \mathrm{DKL_i}) \\ Then \ R_2 = \left(1 - \dfrac{\dfrac{DK_i}{Fitness(K_i)} + \dfrac{DK_i - DKL_i}{Fitness(L_i)}}{e.DK_i \dfrac{1}{Fitness(K_i)}}\right) DK_i \end{pmatrix} \tag{24}$$

$$newDol_i = K_i + \frac{Random}{\|Random\|} R2 \tag{25}$$

$$\left(\begin{array}{l} If\,(DKi < DKLi) \\ Then\ R_2 = \left(1 - \dfrac{\dfrac{DK_i}{Fitness(K_i)} - \dfrac{DKL_i\_DK_i}{Fitness(L_i)} \geq}{e.DK_i \dfrac{1}{Fitness(K_i)}} DK_i \right) \end{array} \right) \tag{26}$$

Compute $newDol_i$ as Eq. (25), whereby e characterizes a constant that is higher than 2. Afterward $Do1_i$ move to the new location, compare $newDol_i$ with K_i concerned fitness,

$$Fitness\ (newDol_i) < Fitness(ki) \tag{27}$$

Or else K_i is substituted with newDol1; then, K_i remain unchanged. Afterward $Dol_i\ (i = 1,\ 2,\ \ldots,\ N)$ upgrade the location and K_i, state whether the DSO fulfills the termination criteria. Once the condition is satisfied, DSO starts the termination process. Besides, DSO initiates the searching stage again. Fig. 2 showcases the steps involved in DSO technique.

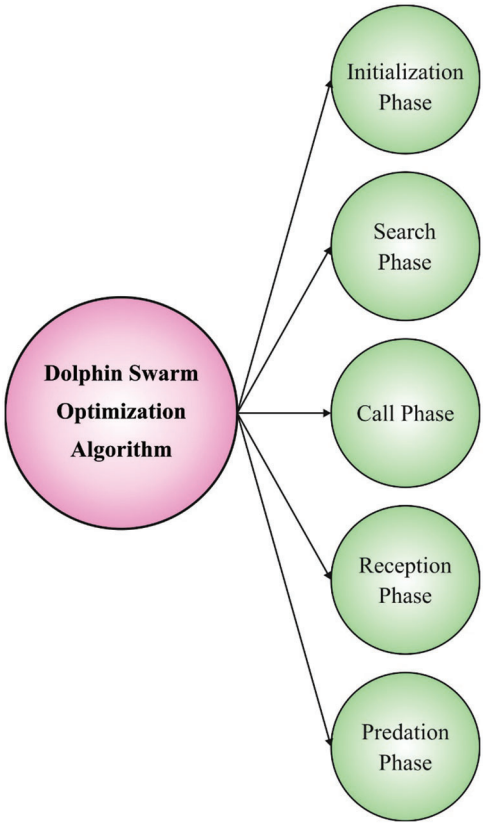

Figure 2: Steps involved in DSO

The DSO algorithm makes a derivation of fitness function for reaching improvised classifier outcome. It sets a positive numeral for indicating superior outcome of candidate solutions. In this article, the reduction of the classifier error rates can be taken as the fitness function, as presented in Eq. (28).

$$fitness(x_i) = Classifier\ Error\ Rate(x_i)$$
$$= \frac{number\ of\ misclassified\ samples}{Total\ number\ of\ samples} * 100 \quad (28)$$

3 Results and Discussion

This section examines the classification results of the DSOCDBN-STC model using a dataset comprising 5 class labels as depicted in Table 1. The proposed model is simulated using Python 3.6.5 tool.

Table 1: List of class labels

Labels	Class name	No. of tweets
News	C1	2500
Conversation	C2	2500
Question	C3	2500
Wish	C4	2500
Others	C5	2500
Total no. of tweets		12500

The proposed model is experimented on PC i5-8600k, GeForce 1050Ti 4GB, 16GB RAM, 250GB SSD, and 1TB HDD.

Fig. 3 illustrates the confusion matrices produced by the DSOCDBN-STC method on the dataset. On 80% of TR data, the DSOCDBN-STC model has recognized 1971 samples into C1, 1979 samples into C2, 1928 samples into C3, 1971 samples into C4, and 1961 samples into C5. Followed by, 20% of TS data, the DSOCDBN-STC method has recognized 487 samples into C1, 483 samples into C2, 497 samples into C3, 490 samples into C4, and 493 samples into C5. Also, on 70% of TR data, the DSOCDBN-STC technique has recognized 1728 samples into C1, 1679 samples into C2, 1729 samples into C3, 1690 samples into C4, and 1728 samples into C5. Meanwhile, on 30% of TS data, the DSOCDBN-STC algorithm has recognized 748 samples into C1, 779 samples into C2, 696 samples into C3, 727 samples into C4, and 731 samples into C5.

Table 2 offers a brief result analysis of the DSOCDBN-STC model on 80% of TR data and 20% of TS data. With 80% of TR data, the DSOCDBN-STC model has shown average $accu_y$ of 99.24%, $prec_n$ of 98.10%, $reca_l$ of 98.10%, F_{score} of 98.10%, and MCC of 97.62%. At the same time, with 20% of TS data, the DSOCDBN-STC approach has exhibited average $accu_y$ of 99.20%, $prec_n$ of 98%, $reca_l$ of 98.10%, F_{score} of 98%, and MCC of 97.50%.

Table 3 offers a detailed result analysis of the DSOCDBN-STC method on 70% of TR data and 30% of TS data. With 70% of TR data, the DSOCDBN-STC model has exhibited average $accu_y$ of 99.10%, $prec_n$ of 97.77%, $reca_l$ of 97.76%, F_{score} of 97.76%, and MCC of 97.12%. Meanwhile, with 30% of TS data, the DSOCDBN-STC model has shown an average $accu_y$ of 99.26%, $prec_n$ of 98.15%, $reca_l$ of 98.14%, F_{score} of 98.14%, and MCC of 97.68%.

The training accuracy (TA) and validation accuracy (VA) acquired by the DSOCDBN-STC method on test dataset is shown in Fig. 4. The experimental outcome implicit the DSOCDBN-STC technique has attained maximal values of TA and VA. In specific, the VA is greater than TA.

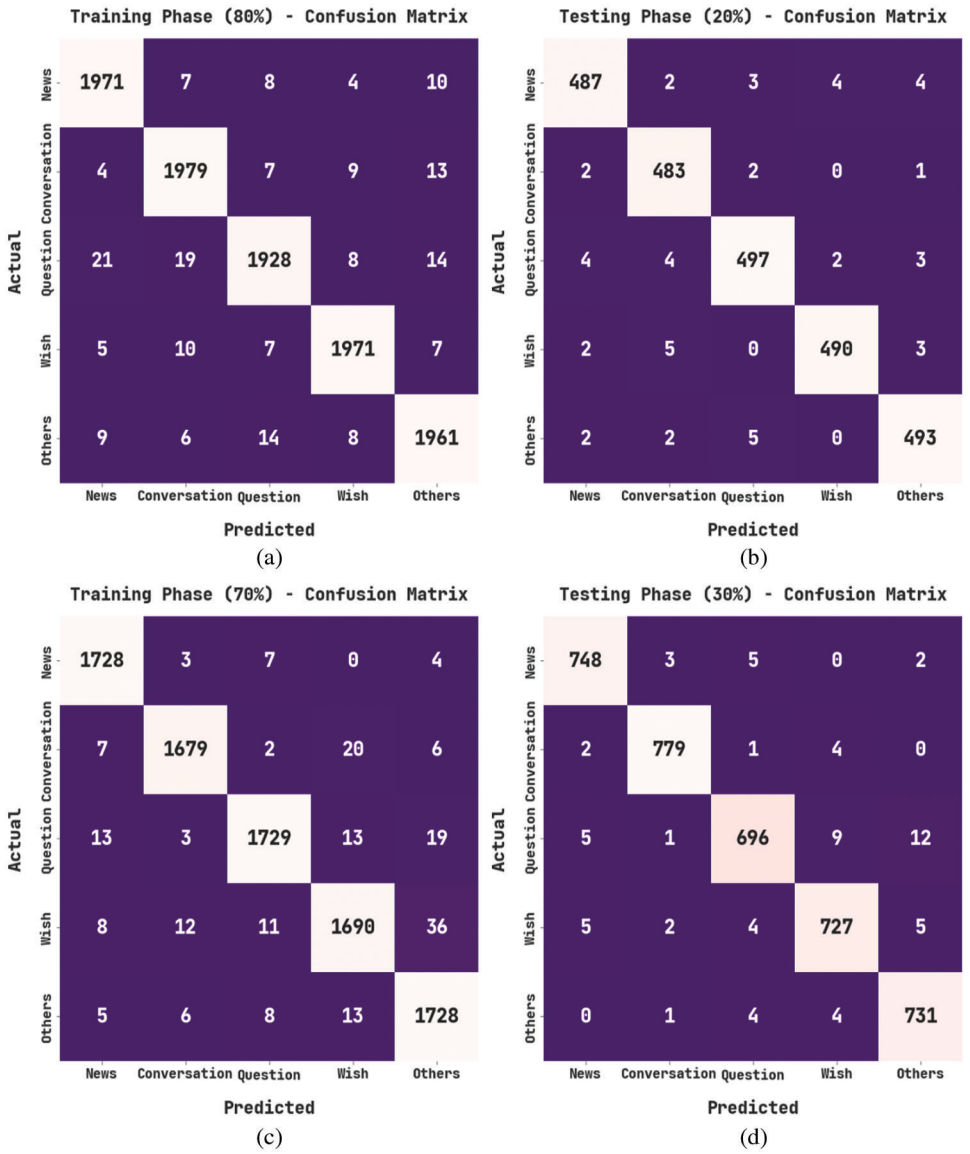

Figure 3: Confusion matrices of DSOCDBN-STC approach (a) 80% of TR data, (b) 20% of TS data, (c) 70% of TR data, and (d) 30% of TS data

Table 2: Result analysis of DSOCDBN-STC approach under 80:20 of TR/TS data

Labels	$Accu_y$	$Prec_n$	$Reca_l$	F_{score}	MCC
Training phase (80%)					
News	99.32	98.06	98.55	98.30	97.88
Conversation	99.25	97.92	98.36	98.14	97.67
Question	99.02	98.17	96.88	97.52	96.91
Wish	99.42	98.55	98.55	98.55	98.19
Others	99.19	97.81	98.15	97.98	97.47
Average	99.24	98.10	98.10	98.10	97.62

(Continued)

Table 2 (continued)

Labels	$Accu_y$	$Prec_n$	$Reca_l$	F_{score}	MCC
Testing phase (20%)					
News	99.08	97.99	97.40	97.69	97.12
Conversation	99.28	97.38	98.98	98.17	97.73
Question	99.08	98.03	97.45	97.74	97.16
Wish	99.36	98.79	98.00	98.39	98.00
Others	99.20	97.82	98.21	98.01	97.51
Average	99.20	98.00	98.01	98.00	97.50

Table 3: Result analysis of DSOCDBN-STC approach under 70:30 of TR/TS data

Labels	$Accu_y$	$Prec_n$	$Reca_l$	F_{score}	MCC
Training phase (70%)					
News	99.46	98.13	99.20	98.66	98.32
Conversation	99.33	98.59	97.96	98.27	97.86
Question	99.13	98.41	97.30	97.85	97.31
Wish	98.71	97.35	96.19	96.76	95.96
Others	98.89	96.37	98.18	97.27	96.58
Average	99.10	97.77	97.76	97.76	97.21
Testing phase (30%)					
News	99.41	98.42	98.68	98.55	98.18
Conversation	99.63	99.11	99.11	99.11	98.87
Question	98.91	98.03	96.27	97.14	96.47
Wish	99.12	97.72	97.85	97.78	97.23
Others	99.25	97.47	98.78	98.12	97.66
Average	99.26	98.15	98.14	98.14	97.68

The training loss (TL) and validation loss (VL) achieved by the DSOCDBN-STC approach on test dataset are displayed in Fig. 5. The experimental outcome denoted the DSOCDBN-STC algorithm has established least values of TL and VL. Particularly, the VL is lesser than TL.

A clear precision-recall analysis of the DSOCDBN-STC method on test dataset is displayed in Fig. 6. The figure indicated that the DSOCDBN-STC technique has resulted to enhanced values of precision-recall values under all classes.

A brief ROC analysis of the DSOCDBN-STC method on test dataset is shown in Fig. 7. The results denoted the DSOCDBN-STC algorithm has shown its ability in categorizing distinct classes on test dataset.

For confirming the improvements of the DSOCDBN-STC model, a brief comparison study is made in Table 4 [26].

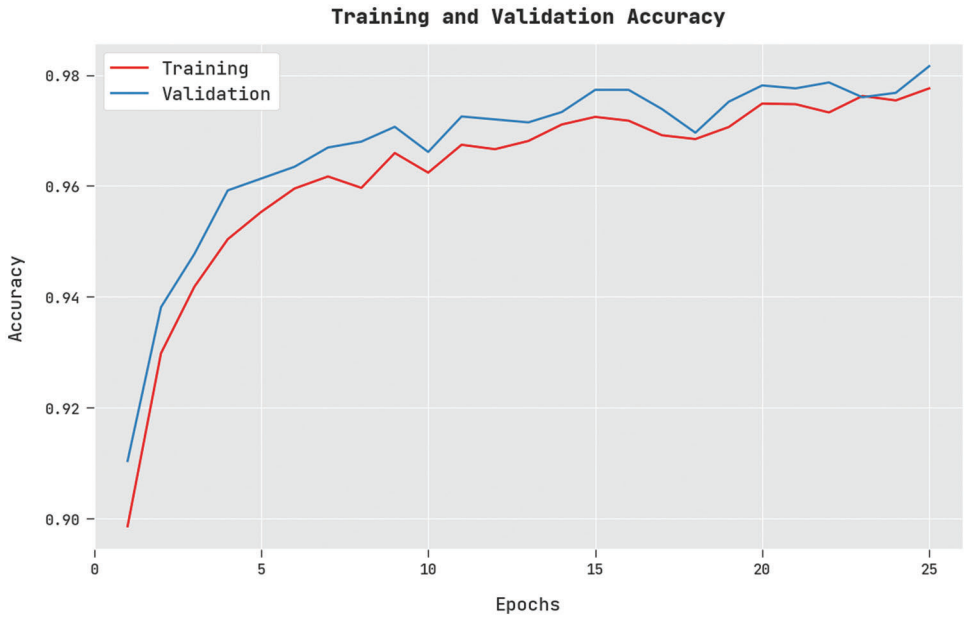

Figure 4: TA and VA analysis of DSOCDBN-STC approach

Figure 5: TL and VL analysis of DSOCDBN-STC approach

Fig. 8 illustrates a comparative $accu_y$ inspection of the DSOCDBN-STC method with other existing models. The figure represented the GoogleNet, MLP, and LOR models have shown poor performance with lower $accu_y$ values of 98.28%, 98.22%, and 98.16% respectively. Then, the RF and GNB models have reported slightly enhanced $accu_y$ values of 98.89% and 98.53% respectively. Next, the SVM model has resulted in reasonable $accu_y$ of 99.11%. However, the DSOCDBN-STC model has resulted in maximum $accu_y$ of 99.26%.

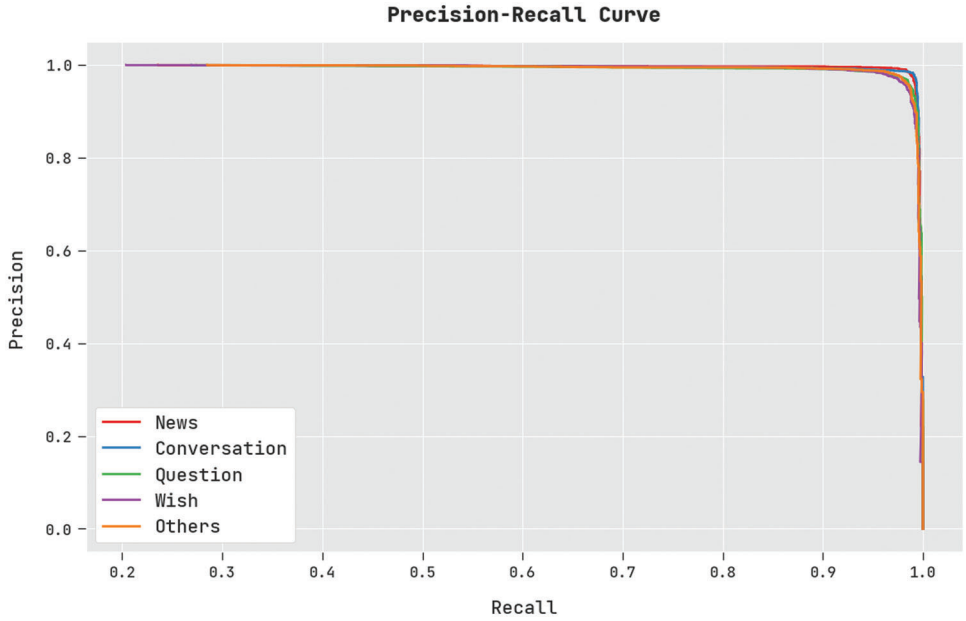

Figure 6: Precision-recall curve analysis of DSOCDBN-STC approach

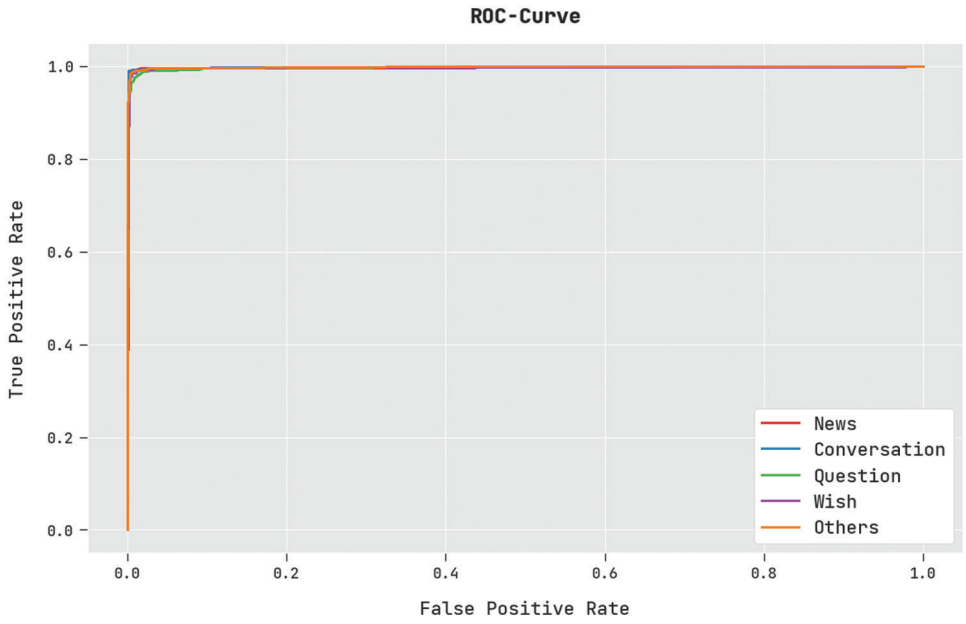

Figure 7: ROC curve analysis of DSOCDBN-STC approach

Fig. 9 demonstrates a comparative $prec_n$ analysis of the DSOCDBN-STC model with other existing models. The figure implicit the GoogleNet, MLP, and LOR models have shown poor performance with lower $prec_n$ values of 97.31%, 96.36%, and 97.99% correspondingly. Then, the RF and GNB techniques have reported slightly enhanced $prec_n$ values of 96.64% and 97.66% correspondingly. After, the SVM model has resulted in reasonable $prec_n$ of 96.50%. However, the DSOCDBN-STC model has resulted in maximum $prec_n$ of 98.15%.

Table 4: Comparative analysis of DSOCDBN-STC approach with existing methodologies

Methods	Accuracy	Precision	Recall	F1-Score
DSOCDBN-STC	99.26	98.15	98.14	98.14
GoolgeNet	98.28	97.31	96.74	96.30
MLP	98.22	96.36	96.21	96.13
LOR	98.16	97.99	97.12	97.15
RF	98.89	96.64	96.71	97.64
GNB	98.53	97.66	97.50	96.42
SVM	99.11	96.50	96.52	96.02

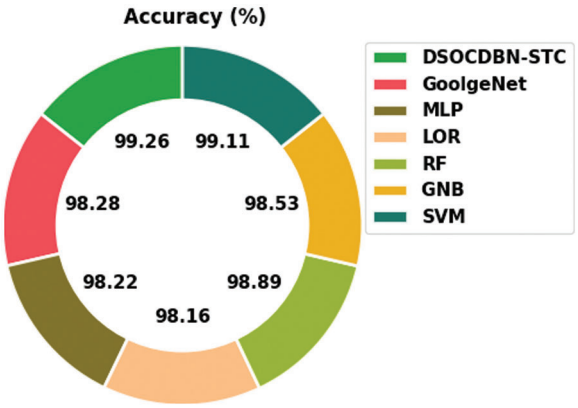

Figure 8: $Accu_y$ analysis of DSOCDBN-STC approach with existing methodologies

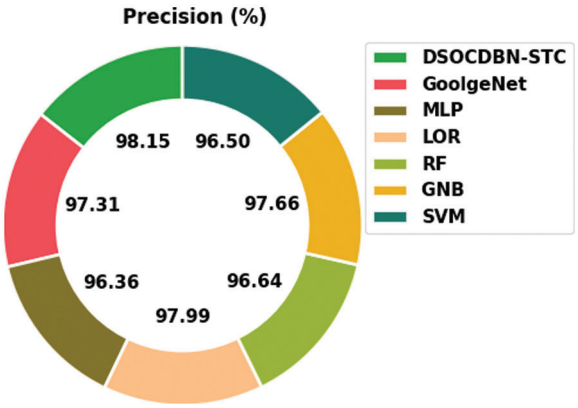

Figure 9: $Prec_n$ analysis of DSOCDBN-STC approach with existing methodologies

Fig. 10 shows a comparative $reca_l$ review of the DSOCDBN-STC model with other existing models. The figure denoted the GoogleNet, MLP, and LOR models have shown poor performance with lower $reca_l$ values of 96.74%, 96.21%, and 97.12% correspondingly. Next, the RF and GNB approaches have reported slightly enhanced $reca_l$ values of 96.71% and 97.50% correspondingly. Then, the SVM model has resulted in reasonable $reca_l$ of 96.52%. But, the DSOCDBN-STC model has resulted in maximum $reca_l$ of 98.14%.

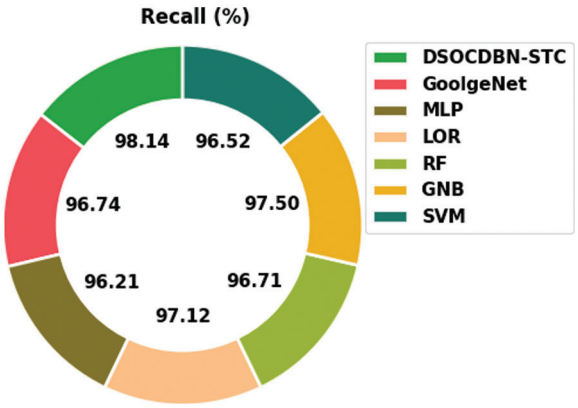

Figure 10: *Recal* analysis of DSOCDBN-STC approach with existing methodologies

Fig. 11 exemplifies a comparative $F1_{score}$ scrutiny of the DSOCDBN-STC approach with other existing models. The figure indicated the GoogleNet, MLP, and LOR models have shown poor performance with lower $F1_{score}$ values of 96.30%, 96.13%, and 97.15% correspondingly.

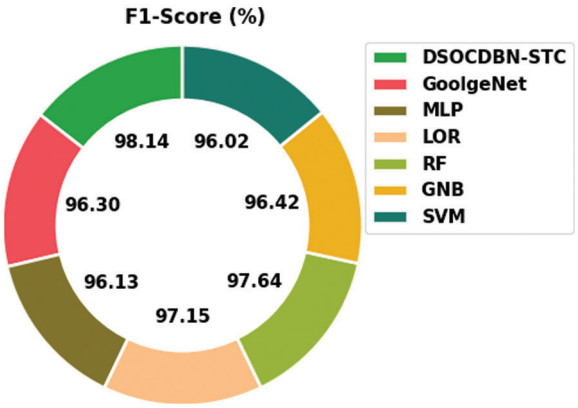

Figure 11: $F1_{score}$ analysis of DSOCDBN-STC approach with existing methodologies

Then, the RF and GNB approaches have reported slightly enhanced $F1_{score}$ values of 97.64% and 96.42% correspondingly. Next, the SVM model has resulted in reasonable $F1_{score}$ of 96.02%. However, the DSOCDBN-STC model has resulted in maximum $F1_{score}$ of 98.14%.

4 Conclusion

In this article, a new DSOCDBN-STC model was devised for short text classification on Arabic Corpus. The presented DSOCDBN-STC model majorly aims to classify Arabic short text in social media. The presented DSOCDBN-STC model encompasses pre-processing and word2vec word embedding at the preliminary stage. Besides, the DSOCDBN-STC model involves CDBN based classification model for Arabic short text. At last, the DSO technique can be exploited for optimal modification of the hyperparameters related to the CDBN method. To demonstrate the enhanced performance of the DSOCDBN-STC model, a wide range of simulations have been performed. The simulation results confirmed the supremacy of the DSOCDBN-STC model over existing models with improved accuracy of

99.26%. As a part of future scope, the performance of the presented model can be enhanced by the feature selection models.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R263), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: 22UQU4340237DSR40.

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

References

[1] A. Wahdan, S. A. Hantoobi, S. A. Salloum and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 6629, 2020.

[2] A. H. Ombabi, W. Ouarda and A. M. Alimi, "Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 53, 2020.

[3] M. Alruily, "Classification of Arabic tweets: A review," *Electronics*, vol. 10, no. 10, pp. 1143, 2021.

[4] M. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari and A. Hilal, "Preprocessing arabic text on social media," *Heliyon*, vol. 7, no. 2, pp. e06191, 2021.

[5] F. N. Al-Wesabi, "Proposing high-smart approach for content authentication and tampering detection of Arabic text transmitted via internet," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 10, pp. 2104–2112, 2020.

[6] S. A. Salloum, C. Mhamdi, M. Al-Emran and K. Shaalan, "Analysis and classification of arabic newspapers' facebook pages using text mining techniques," *International Journal of Information Technology and Language Studies*, vol. 1, no. 2, pp. 8–17, 2017.

[7] F. N. Al-Wesabi, A. Abdelmaboud, A. A. Zain, M. M. Almazah and A. Zahary, "Tampering detection approach of Arabic-text based on contents interrelationship," *Intelligent Automation & Soft Computing*, vol. 27, no. 2, pp. 483–498, 2021.

[8] A. S. Almasoud, S. B. Haj Hassine, F. N. Al-Wesabi, M. K. Nour, A. M. Hilal *et al.,* "Automated multi-document biomedical text summarization using deep learning model," *Computers, Materials & Continua*, vol. 71, no. 3, pp. 5799–5815, 2022.

[9] F. N. Al-Wesabi, "A hybrid intelligent approach for content authentication and tampering detection of Arabic text transmitted via internet," *Computers, Materials & Continua*, vol. 66, no. 1, pp. 195–211, 2021.

[10] A. Oussous, F. -Z. Benjelloun, A. A. Lahcen and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," *Journal of Information Science*, vol. 46, no. 4, pp. 544–559, 2020.

[11] F. N. Al-Wesabi, "A smart English text zero-watermarking approach based on third-level order and word mechanism of markov model," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1137–1156, 2020.

[12] B. Y. AlHarbi, M. S. AlHarbi, N. J. AlZahrani, M. M. Alsheail, J. F. Alshobaili *et al.,* "Automatic cyber bullying detection in Arabic social media," *International Journal of Engineering Research & Technology*, vol. 12, no. 12, pp. 2330–2335, 2019.

[13] F. N. Al-Wesabi, "Entropy-based watermarking approach for sensitive tamper detection of Arabic text," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 3635–3648, 2021.

[14] A. S. Alammary, "BERT models for arabic text classification: A systematic review," *Applied Sciences*, vol. 12, no. 11, pp. 5720, 2022.

[15] F. Alhaj, A. Al-Haj, A. Sharieh and R. Jabri, "Improving Arabic cognitive distortion classification in twitter using BERTopic," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 854–860, 2022.

[16] S. L. M. Sainte and N. Alalyani, "Firefly algorithm based feature selection for arabic text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 3, pp. 320–328, 2020.

[17] A. Hawalah, "Semantic ontology-based approach to enhance Arabic text classification," *Big Data and Cognitive Computing*, vol. 3, no. 4, pp. 53, 2019.

[18] M. F. Ibrahim, M. A. Alhakeem and N. A. Fadhil, "Evaluation of naïve Bayes classification in Arabic short text classification," *Al-Mustansiriyah Journal of Science*, vol. 32, no. 4, pp. 42–50, 2021.

[19] M. Beseiso and H. Elmousalami, "Subword attentive model for Arabic sentiment analysis: A deep learning approach," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 2, pp. 1–17, 2020.

[20] H. Najadat, M. A. Alzubaidi and I. Qarqaz, "Detecting Arabic spam reviews in social networks based on classification algorithms," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–13, 2022.

[21] Y. Albalawi, J. Buckley and N. Nikolov, "Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media," *Journal of Big Data*, vol. 8, no. 1, pp. 1–29, 2021.

[22] F. R. Alharbi and M. B. Khan, "Identifying comparative opinions in Arabic text in social media using machine learning techniques," *SN Applied Sciences*, vol. 1, no. 3, pp. 213, 2019.

[23] V. T. Tran, F. AlThobiani and A. Ball, "An approach to fault diagnosis of reciprocating compressor valves using teager–Kaiser energy operator and deep belief networks," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4113–4122, 2014.

[24] X. Zhao, M. Jia and Z. Liu, "Semisupervised graph convolution deep belief network for fault diagnosis of electormechanical system with limited labeled data," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5450–5460, 2021.

[25] A. K. Srivastava, D. Pandey and A. Agarwal, "Extractive multi-document text summarization using dolphin swarm optimization approach," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 11273–11290, 2021.

[26] S. M. Alzanin, A. M. Azmi and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," *Journal of King Saud University-Computer and Information Sciences*, pp. S1319157822001045, 2022, https://doi.org/10.1016/j.jksuci.2022.03.020.