Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.034609
Article

Optimal Deep Hybrid Boltzmann Machine Based Arabic Corpus Classification Model

Mesfer Al Duhayyim[1,*], Badriyya B. Al-onazi[2], Mohamed K. Nour[3], Ayman Yafoz[4], Amal S. Mehanna[5], Ishfaq Yaseen[6], Amgad Atta Abdelmageed[6] and Gouse Pasha Mohammed[6]

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, 16273, 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 Computer Sciences, College of Computing and Information System, Umm Al-Qura University, Makkah 24211, Saudi Arabia
[4]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[5]Department of Digital Media, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, 11845, Egypt
[6]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia
*Corresponding Author: Mesfer Al Duhayyim. Email: m.alduhayyim@psau.edu.sa
Received: 21 July 2022; Accepted: 11 October 2022

Abstract: Natural Language Processing (NLP) for the Arabic language has gained much significance in recent years. The most commonly-utilized NLP task is the 'Text Classification' process. Its main intention is to apply the Machine Learning (ML) approaches for automatically classifying the textual files into one or more pre-defined categories. In ML approaches, the first and foremost crucial step is identifying an appropriate large dataset to test and train the method. One of the trending ML techniques, i.e., Deep Learning (DL) technique needs huge volumes of different types of datasets for training to yield the best outcomes. The current study designs a new Dice Optimization with a Deep Hybrid Boltzmann Machine-based Arabic Corpus Classification (DODHBM-ACC) model in this background. The presented DODHBM-ACC model primarily relies upon different stages of pre-processing and the word2vec word embedding process. For Arabic text classification, the DHBM technique is utilized. This technique is a hybrid version of the Deep Boltzmann Machine (DBM) and Deep Belief Network (DBN). It has the advantage of learning the decisive intention of the classification process. To adjust the hyperparameters of the DHBM technique, the Dice Optimization Algorithm (DOA) is exploited in this study. The experimental analysis was conducted to establish the superior performance of the proposed DODHBM-ACC model. The outcomes inferred the better performance of the proposed DODHBM-ACC model over other recent approaches.

Keywords: Arabic corpus; text classification; machine learning; deep learning; dice optimization

1 Introduction

With the advancements in Natural Language Processing (NLP), the Arabic Text Categorization (ATC) process has become an active research domain since the Arabic language has several difficulties, such as highly-complicated structure, unique morphological characters and so on [1]. Indeed, the derivational and the inflectional nature of the Arabic language examine highly-complex structures and morphology. The key objective of the ATC approach is to allow pre-defined classes for the Arabic text based on its content. Text representation is a decisive stage that suggestively impacts the performance of the ATC process. In literature, an extensive array of Arabic text representation techniques was reviewed [2]. For example, a conventional text modelling related to the Bag-Of-Words (BOW) representation was attained for the existing acts in the NLP domain. But, this technique suffered from the curse of dimensionality and the non-existence of semantic relations among different text units [3,4]. Text Classification (TC) can be described as a text-mining procedure in which a category or a class is specified for the presented textual file [5]. The productivity of this procedure is measured in terms of class sets, whereas all the class sets contain a set of text files that belong to a particular kind or a topic [6].

The single-label TC presents a single label for every file, whereas the multi-label TC provides different labels for every file [7]. A dynamic classification method is required to handle the huge volumes of text generated on the web every minute. This method should categorize every file under a suitable category and simplify the tasks in other areas like information retrieval and NLP. Unsupervised and supervised Machine Learning (ML) techniques were scrutinized in the domain of TC earlier [8,9]. The unsupervised learning method varies from supervised learning in the labelled dataset. Supervised learning methods utilize labelled datasets to forecast the future, whereas these labelled datasets are known to be the knowledge repository of the models [10]. Precisely, it can be explained as a teaching process in which a method is trained with adequate information and is allowed to perform the predictions after the teaching process is over [11]. In the unsupervised learning approach, the data is not labelled. These methods are unaware of any data or its categories or classes in a dataset; such methods try to find the significant paradigms in a dataset. Both characteristics, as well as the complexities of the Arabic language, make the processing of Arabic texts [12] a challenging process. It is difficult to handle numerous complexities in Arabic like diglossia, ambiguity, etc.; at first, it is challenging to understand and read the Arabic script since the meaning conveyed by the Arabic letters changes according to their position in a word. Secondly, the language has no dedicated letter or capitalization method. Finally, the language has a complex morphology framework, while its alphabet system is not easy to understand [13]. In addition, it is also challenging to normalize the inconsistencies when using a few letters, diacritical marks and dialects. Linguistics researchers and technology developers deal with complexities in NLP tools through morphology analysis, tokenization and stemming from the Arabic language [14].

The current article designs a new Dice Optimization with a Deep Hybrid Boltzmann Machine-based Arabic Corpus Classification (DODHBM-ACC) model. The presented DODHBM-ACC model primarily relies upon different stages of pre-processing and the word2vec word embedding process. For Arabic text classification, the DHBM technique is utilized. It is a hybrid version of the Deep Boltzmann Machine (DBM) and Deep Belief Network (DBN). It has the advantage of learning the decisive intention of a classification process. To fine-tune the hyperparameters involved in the DHBM technique, the Dice Optimization Algorithm (DOA) is exploited in this study. The experimental analysis was conducted to establish the superior performance of the proposed DODHBM-ACC model.

The rest of the paper is organized as follows. Section 2 offers information about the works conducted earlier in this domain, and Section 3 explains the proposed model. Next, Section 4 provides the information on experimental validation, whereas Section 5 concludes the work.

2 Related Works

In recent times, Deep Learning (DL) methods are extensively utilized in Sentiment Analysis (SA) research. Few researchers have employed NLP or pre-processing methods to prepare the data for the classification process. El-Alami et al. [15] examined Long Short Term Memory (LSTM), Convolutional Neural Network (CNN) and a combination of these methods to accomplish the ATC process. This work further dealt with the morphological diversity of the Arabic letters by sightseeing the word embedding method using sub-word information and the position weights. This study framed a policy to refine the Arabic vector space representation to ensure an adjacent vector representation for the linked words. It was done with the help of semantic data embedding in lexical sources. The earlier study [16] devised a feature selection algorithm by integrating the Artificial Bee Colony (ABC) technique and the chi-square technique. Chi-square is a filtering technique that can perform calculations simply and rapidly. It can handle a large-dimensional feature and can be utilized as an initial level in feature selection procedures. In this study, the ABC technique, i.e., a wrapper approach, was utilized as another level, after which Naive Base was employed as a Fitness Function (FF).

Al-Anzi et al. [17] developed an innovative text classification technique that neither practised dimensionality reduction nor SA approaches. The presented technique was a space-efficient approach, i. e., it made use of an initial-order Markov method for the hierarchical ATC. A Markov chain method was arranged based on the neighbouring character series for every category and its sub-categories. Then, the preparation methods were utilized to score the files for classification. Alhaj et al. [18] developed a new TC technique to improve the performance of the ATC process utilizing ML approaches. The identification of an appropriate Feature Selection (FS) methodology along with an ideal sum of the features remains the most important step in the ATC process to achieve the finest classification outcomes. Thus, the authors devised an algorithm named Optimal Configuration Determination for ATC (OCATC). It can also be used as a Particle Swarm Optimization (PSO) technique to find the best configuration. The presented algorithm derived and transformed the attributes from the text data into an arithmetic vector with the help of Term Frequency-Inverse Document Frequency (TF–IDF) method.

Alshaer et al. [19] focused on learning the impact of the enhanced Chi (ImpCHI) square method on the performances of the six-renowned classification models. The proposed method was significant enough to enhance Arabic text classification. Further, it was considered a promising basis for the classification of the text due to its contribution in terms of pre-defined classes. Ababneh [20] attempted to find the best dataset that could offer fair evaluation and, importantly, train the method for TC. In this examination, renowned and accurate learning methods were employed. The author provided time measures and emphasized the relevance of training the methods using such datasets to enable the Arabic language authors to choose a suitable dataset and leverage a solid basis for comparison.

3 The Proposed DODHBM-ACC model

The current study has developed a new DODHBM-ACC model for automatic Arabic corpus classification. The presented DODHBM-ACC model primarily relies on four processes: data pre-processing, word embedding, Arabic text classification and hyperparameter tuning. Fig. 1 depicts the overall processes of the DODHBM-ACC technique.

3.1 Data Preprocessing

The overall steps involved in this pre-processing function are briefed herewith.

- Tokenization: This function tokenizes a text and classifies a text as either a token or a word set.

- Stop-words removal: It excludes any type of speech, neither verb nor noun. The list of different stop words in Arabic has more than 400 terminologies.
- Stemming: In this procedure, both suffixes and token prefixes are eliminated. The steaming function is a vital process and positively impacts a model's performance and efficacy.

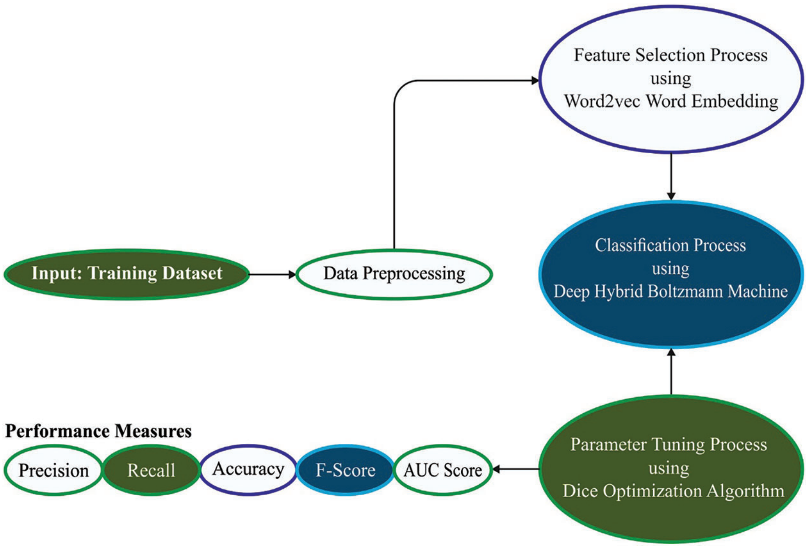

Input: Training Dataset
Data Preprocessing
Feature Selection Process using Word2vec Word Embedding
Classification Process using Deep Hybrid Boltzmann Machine
Parameter Tuning Process using Dice Optimization Algorithm
Performance Measures
Precision
Recall
Accuracy
F-Score
AUC Score

Figure 1: Overall processes of the DODHBM-ACC approach

3.2 Word Embedding

Word2Vec methodology uses neural network techniques to accomplish word representation [21]. This algorithm considers a large corpus as its input and considers each vocabulary in the corpora. Harris proposed this concept in which a word has the same meaning and is used in the same context. This technique upgrades the vector of a word based on its appearance in the external environment with the help of a pre-determined size window. The comparison amongst those words is higher than the earlier one, and the vector becomes convergent. The Word2Vec process follows two approaches such as the Skip-gram (SG) approach and the Continuous Bag-of-Words (CBOW) approach to generate a word vector. In the SG approach, the vector of an external environment within the window size is transformed based on the centre word. In the CBOW approach, the vector of a word centre is upgraded based on the external environment within the size of the window. In the current study, three dissimilar dimensions have been created for these approaches, such as 100, 200 and 300. Further, a window at 5 is also applied, whereas the minimum word appearance in the corpus is equivalent to 5.

3.3 Arabic Text Classification Using DHBM Technique

For the classification of Arabic texts, the DHBM technique is utilized. This technique is a hybrid version of DBM and DBN and is regarded as an increasingly-complicated variant of both methods [22]. The hybrid structure was created with the intention of conducting the classification process. An alternative method to consider the DHBM approach is to configure a strongly-incorporated Hybrid Restricted Boltzmann Machine (HRBM) instead of an individual module. SBEN is a stack of HRBM in which every model $p(y,\ h^l)$ is present at their corresponding level l of generalization in its general framework. To leverage the prediction power at all the generalization levels, an intermediate step of each layer-by-layer predictor is employed to compute the SBEN $p(_y|x)_{ensemble}$ at the time of inference. The first edition of these models is trained through a bottom-up greedy algorithm from which every layer learns to independently forecast the output of another layer (trained on hidden presentation). Through relative observation of the SBEN, the DHBM model is created based on the stacked HRBM expertise. Afterwards, the predictors are connected, which implies that the prediction of the total model is based on the learning of every layer about their respective abstract levels. Additionally, the relationships are built between every layer through a full Bi-directional expertise. In other terms, in order to compute the layer of the hidden parameters in the algorithm (excluding input and top hidden states), the activation from above and below layers must be integrated instantaneously. As per the discussion given above, the three layers of the DHBM model are determined. As shown in Fig. 1, the description lengthens up to L-layer model. With the help of a pattern vector input $x = (x_1, \cdots, x_D)$ and the corresponding target parameter $y \in \{1, \cdots, C\}$, two sets of the hidden parameters $h^1 = \left(h_1^1, \cdots, h_{H_1}^1\right)$ and $h^2 = \left(h_1^2, \cdots, h_{H_2}^2\right)$ and the model variables $\Theta^m = (W^1, U^1, W^2, U^2)^3$ are applied. At the same time, the energy of the DHBM method is determined as given below.

$$E\left(y, x, h^1, h^2\right) = -h^{1T} W^1 x - h^{1T} U^1 e_y - h^{2T} W^2 h^1 - h^{2T} U^2 e_y. \tag{1}$$

Eq. (1) notes that $e_y = (1_{i=y})_{i=1}^{C}$ refers to a one-hot vector encoder of y. It is likely that a 3-DHBM model is allocated to a 4-tuple (y, x, h^1, h^2) layer as given below.

$$p(y, x, \Theta) = \frac{1}{Z} \sum\nolimits_h e^{\left(E\left(y,x,h^1,h^2\right)\right)} \tag{2}$$

In Eq. (2), Z denotes the partition function which guarantees a valid and a likelihood distribution. This is evaluated by totalling each feasible module configuration.

It is to be noted that in the overview of top-down calculation, the hidden as well as the visible layers of the 3-DHBM model are calculated using the executable equations given below.

$$p(h^1|y, x, h^2) = \prod_j p(h_j^1|y, X, h^2), \text{ with } p(h_j^1 = 1|y, X) = \phi\left(U_{jy}^1 + \sum\nolimits_i W_{ji}^1 \sum\nolimits_k W_{kj}^2 h_k^2\right) \tag{3}$$

$$p(h^2|y, h^1) = \prod_k p(h_k^2|y, h^1), \text{ with } p(h_k^2 = 1|h^1) = \phi\left(U_{ky}^2 + \sum\nolimits_j W_{kj}^2 h_j^1\right) \tag{4}$$

$$p(x|h^1) = \prod_i p(x_i|h^1), \text{ with } p(x_i = 1|h) = \phi\left(\sum\nolimits_j W_{ji}^1 h_j^1\right) \tag{5}$$

$$p(y|h^1, h^2) = \frac{e^{\Sigma_j U_{jy}^1 h_j^1 + \Sigma_j U_{jy}^2 h_j^2}}{\sum\nolimits_{y*} e^{\Sigma_j U_{jy\star}^1 h_j^1 + \Sigma_j U_{jy\star}^2 h_j^2}} \tag{6}$$

In this expression, a logistic sigmoid or an activation function is denoted by $\phi(v) = 1/(1 + e^{-v})$ and y is utilized to access a specific class filter from U^l. In order to adapt the mechanism for distinct kinds of inputs like continuously-valued parameters, $\phi(v)$ is substituted with some other alternative functions like Rectified Linear Unit (ReLU). The subset of a formula can be used as a fixed-point formula to run the mean-field inference in the deep structure to obtain $\mu = \left\{\mu^1, \mu^2\right\}$. It is to be noted that the dependence among those conditions and a consequent bottom-up pass calculation should be weighed double to initialize the mean field cycling via the Eqs. (3)–(6) so as to achieve the reconstructed model of the input and target values (or predictive models). Fig. 2 portrays the infrastructure of the DHBM technique.

In order to accelerate the prediction process and training time, the DHBM structure is expanded by means of a separate auxiliary network or a co-model that is formerly applied to infer the state of the hidden parameters in the DBM model. Here, the aim is to exploit the individual bottom-up pass. MLP or the detection model, performs a part of the approximation function that is successfully merged with the deep structures of the concentration. Further, it is also trained based on the gradient-descent method. During the fundamental co-training of the detection model, it is anticipated that the mean-field parameter of the target model remains unchanged. Then, an individual learning step is experimentally demonstrated to study the realistic training of the DBM model. Similar principles are claimed for training an in-depth hybrid structure, i.e., DHBM too.

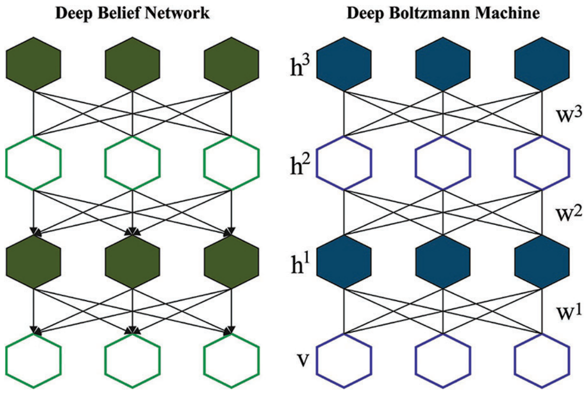


Figure 2: Structure of the DHBM technique

In this detection model, the weight is initialized to DHBM at the beginning of training and is calculated as a completely-factorized element, as given below.

$$Q^{rec}(h|v;v) = \prod_{j=1}^{H_1}\prod_{k=1}^{H_2} q^{rec}\left(h_j^1\right)q^{rec}\left(h_k^2\right) \tag{7}$$

In Eq. (7), the likelihood of $q^{rec}\left(h_i^l = 1\right) = v_i^l$ for layers $l = 1, 2$ and $v = \left\{v^1,\ v^2\right\}$ is high. Then, the detection model is run with the help of a variable $\Theta^{rec} = (R^1,\ R^2)$ (neglecting bias term for simplicity). The equation given below shows the feedforward process.

$$v_j^1 = \phi\left(\sum_{i=1}^{D} 2R_{ij}^1 v_i\right) \tag{8}$$

$$v_j^2 = \phi\left(\sum_{j=1}^{H_1} R_{jk}^2 v_j^1\right) \tag{9}$$

In these expressions, the inference network weight is doubled at all the layers, except the top-most layer, to compensate for the missing top-down feedback. In this hybrid mechanism, especially at the time of prediction, the structure might straightaway produce a suitable calculation for y with the help of the trained detection model to infer the hidden state of the DHBM model.

The detection model can be trained based on Eq. (10):

$$KL(Q^{MF}(h|v;\mu)||Q^{rec}(h|v;v)) = -\sum_i \mu_i log v_i - \sum_i (1-\mu_i) log(1-v_i) + Const \tag{10}$$

This shows a minimized Kullback-Leibler (KL) divergence between $Q^{rec}(h|v;v)$, the factorial posterior of the detection model and $Q^{MF}(h|v;\mu)$ that corresponds to the posterior of the DBM mean-field.

3.4 Hyperparameter Tuning Using DOA

In order to fine-tune the hyperparameters involved in the DHBM technique, the DOA approach is exploited in this study. DOA is a game-based optimization approach that simulates the old-age game rules i.e., dice games. In this DOA approach, the primary location of a player is randomly generated on the playing field i.e., problem description space, as expressed in the following equation [23]:

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots, x_i^n\right). \tag{11}$$

After the formation of the system, the rule is quantified. The players compete in line with the game rules set earlier and determine the winner.

Calculation of each player's score

A fitness function is applied to simulate the score of all the players. A high score is allocated to the player with the best position, calculated as follows.

$$Score_i =, \frac{f_it_i - f_it(player_{best})}{\sum_{j=1}^{N} \; f_it_j - f_it(player_{worst}))} \tag{12}$$

In Eq. (12), $Score_i$ refers to the score of a player i, fit_i denotes their fitness function value, N indicates the number of players, $player\ _{best}$ shows the location of the optimal player, and $player_{worst}$ shows the location of the worst player as given below.

$$player_{best} = \text{location of min } \left(fit_j\right) \;\&\, j \in \{1{:}N\}, \tag{13}$$

$$player_{worst} = \text{location of max } \left(fit_j\right) \;\&\, j \in \{1{:}N\}. \tag{14}$$

Tossing dice for each player

Here, every player tosses a dice. A dice count can be a discrete value between 1 and 6 that signifies the number of players guided by every player and is expressed as follows.

$$Dice_i = K\&K \in \{123456\}, \tag{15}$$

In Eq. (15) *Dice* refers to the dice count for the *i-th* player.

Selection of the Guide's players for each player

For every player, according to the count of the dice (K), a player guide is arbitrarily chosen amongst the players, as shown below

$$X^k_{Guide_i} = X_1{:}X_K, \tag{16}$$

In Eq. (16), $X^k_{Guide_i}$ refers to the location guide's player count, k.

Update the position of each player.

Here, $X^{i,d}$ is evaluated by Eq. (17)

$$X^{i,d} = X_0^{i,d} + \sum\nolimits_{k=1}^{Dice_i} \left(RK\left(X^{i,d} - X^{k,d}_{Guide_i}\right) sign(Score - Score_{Guide_k})\right), \tag{17}$$

Now, r_k refers to an arbitrary count with a standard distribution within [0,1] and $Score_{Guide_k}$ denotes the score of a player guide's count, k.

4 Experimental Validation

The proposed DODHBM-ACC method was experimentally validated using two data sets such as Waten2004 dataset (dataset 1) and Khaleej2004 dataset (dataset 2). The first dataset has a total of 4,217 samples under six classes as depicted in Table 1. The parameter settings are as follows: learning rate: 0.01, dropout: 0.5, batch size: 5, epoch count: 50, and activation: ReLU.

Table 1: Details on dataset-1

Dataset 1-Waten2004 dataset	
Class	No. of samples
Culture	656
Economy	965
Internews	415
Local	703
Religion	667
Sports	811
Total no. of samples	**4217**

The confusion matrices generated by the proposed DODHBM-ACC model on dataset-1 are portrayed in Fig. 3. The results indicate that the proposed DODHBM-ACC method achieved improved outcomes under all the aspects. With the entire dataset, the DODHBM-ACC system identified 633 samples as culture class, 941 samples as economy class, 400 samples as Internews class, 689 samples as local class, 635 samples as religion class and 775 samples as sports class. In line with this, with 70% of TR dataset, the proposed DODHBM-ACC approach categorized 435 samples under culture class, 661 samples under economy class, 280 samples under Internews class, 497 samples under local class, 443 samples under religion class and 544 samples under sports class. Similarly, with 30% of TS dataset, the proposed DODHBM-ACC method classified 198 samples under culture class, 280 samples under economy class, 120 samples under Internews class, 192 samples under local class, 192 samples under religion class and 231 samples under sports class respectively.

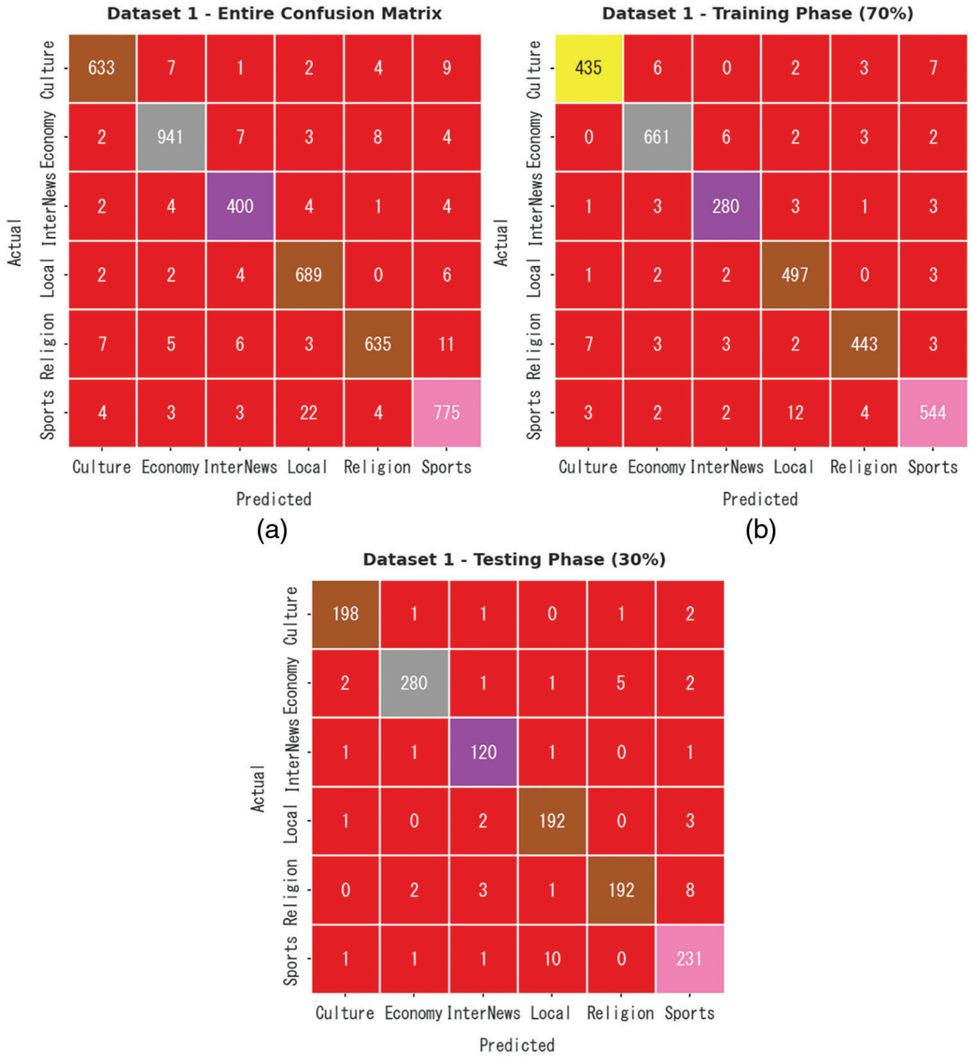
Table 2 demonstrates the overall classification results achieved by the proposed DODHBM-ACC model. With entire dataset, the DODHBM-ACC model reached average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 98.86%, 96.45%, 96.53%, 96.48% and 97.92% correspondingly. Meanwhile, with 70% of TR data, the DODHBM-ACC methodology attained average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 98.97%, 96.81%, 96.80%, 96.80% and 98.09% correspondingly. Also, with 30% of TS data, the proposed DODHBM-ACC approach gained average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 98.60%, 95.61%, 95.90%, 95.74% and 97.53% correspondingly.

Both Training Accuracy (TRA) and Validation Accuracy (VLA) values, attained by the proposed DODHBM-ACC algorithm on dataset-1, are displayed in Fig. 4. The experimental outcomes denote that the proposed DODHBM-ACC approach obtained the maximal TRA and VLA values while VLA values were higher than the TRA values.

Both Training Loss (TRL) and Validation Loss (VLL) values, obtained by the proposed DODHBM-ACC technique on dataset-1, are exhibited in Fig. 5. The experimental outcomes represent that the proposed DODHBM-ACC algorithm outperformed other methods with minimal TRL and VLL values whereas the VLL values were lower than the TRL values.

A clear precision-recall analysis was conducted upon the proposed DODHBM-ACC methodology using dataset-1, and the results are shown in Fig. 6. The figure signifies that the proposed DODHBM-ACC algorithm produced enhanced precision-recall values under all the classes.

A detailed ROC analysis was conducted upon the presented DODHBM-ACC methodology using dataset-1, and the results are presented in Fig. 7. The results indicate that the proposed DODHBM-ACC technique showcased its ability in categorizing the dataset-1 under distinct classes.



(a) (b) (c)

Figure 3: Confusion matrices of the DODHBM-ACC approach under dataset-1 (a) Entire dataset, (b) 70% of TR data, and (c) 30% of TS data

Table 2: Analytical results of the DODHBM-ACC approach upon dataset-1 under distinct class labels

Dataset-1					
Labels	Accuracy	Precision	Recall	F-score	AUC score
Entire dataset					
Culture	99.05	97.38	96.49	96.94	98.01
Economy	98.93	97.82	97.51	97.66	98.43
InterNews	99.15	95.01	96.39	95.69	97.92
Local	98.86	95.30	98.01	96.63	98.52
Religion	98.84	97.39	95.20	96.29	97.36

(Continued)

Table 2 **(continued)**

			Dataset-1		
Sports	98.34	95.80	95.56	95.68	97.28
Average	**98.86**	**96.45**	**96.53**	**96.48**	**97.92**
Training phase (70%)					
Culture	98.98	97.32	96.03	96.67	97.77
Economy	99.02	97.64	98.07	97.85	98.68
InterNews	99.19	95.56	96.22	95.89	97.87
Local	99.02	95.95	98.42	97.17	98.78
Religion	99.02	97.58	96.10	96.83	97.83
Sports	98.61	96.80	95.94	96.37	97.59
Average	**98.97**	**96.81**	**96.80**	**96.80**	**98.09**
Testing phase (30%)					
Culture	99.21	97.54	97.54	97.54	98.53
Economy	98.74	98.25	96.22	97.22	97.85
InterNews	99.05	93.75	96.77	95.24	98.04
Local	98.50	93.66	96.97	95.29	97.88
Religion	98.42	96.97	93.20	95.05	96.32
Sports	97.71	93.52	94.67	94.09	96.55
Average	**98.60**	**95.61**	**95.90**	**95.74**	**97.53**

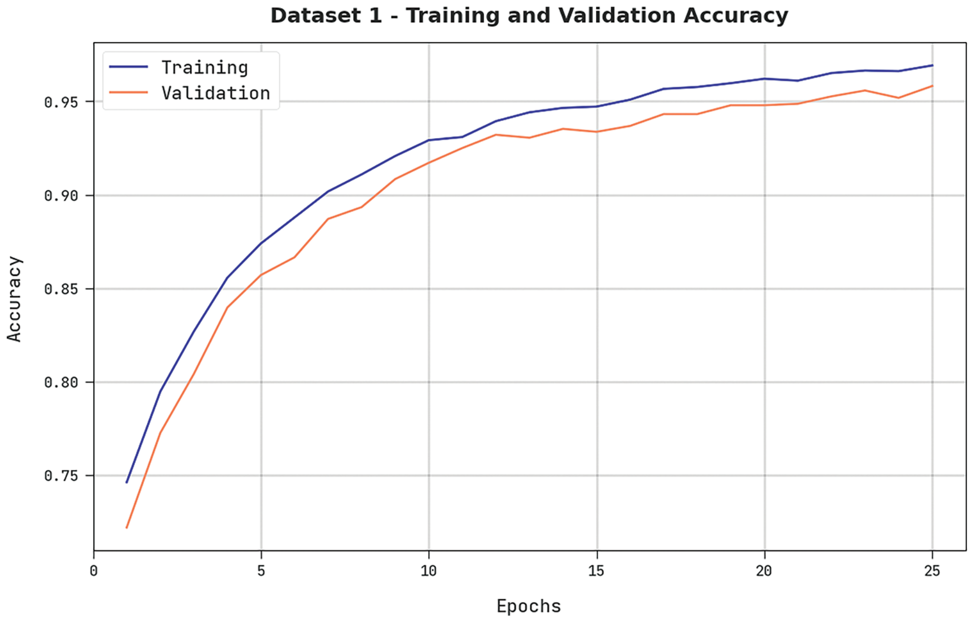


Figure 4: TRA and VLA analyses results of the DODHBM-ACC approach on dataset-1

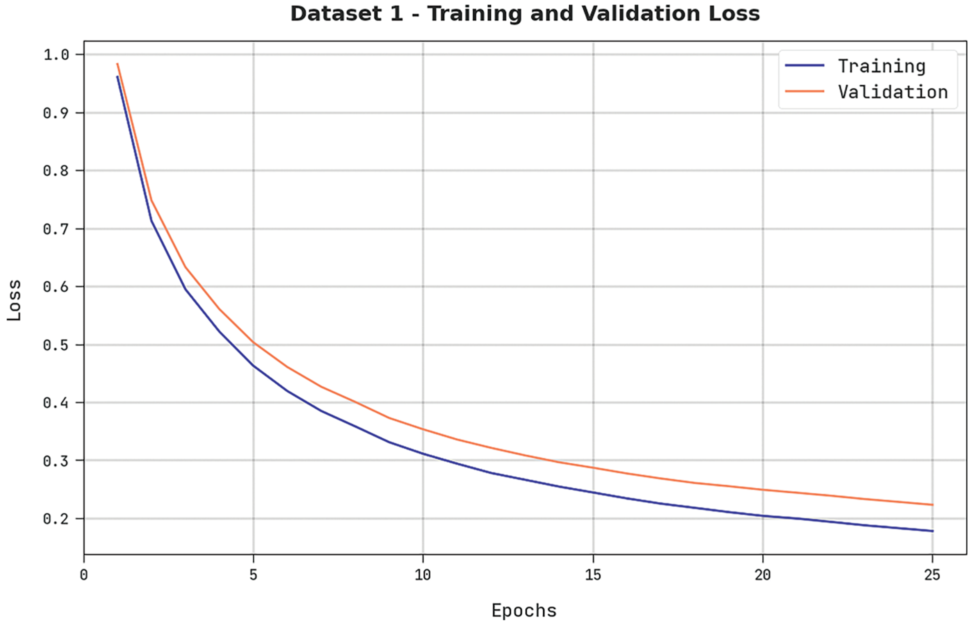


Figure 5: TRL and VLL analyses results of the DODHBM-ACC methodology on dataset-1

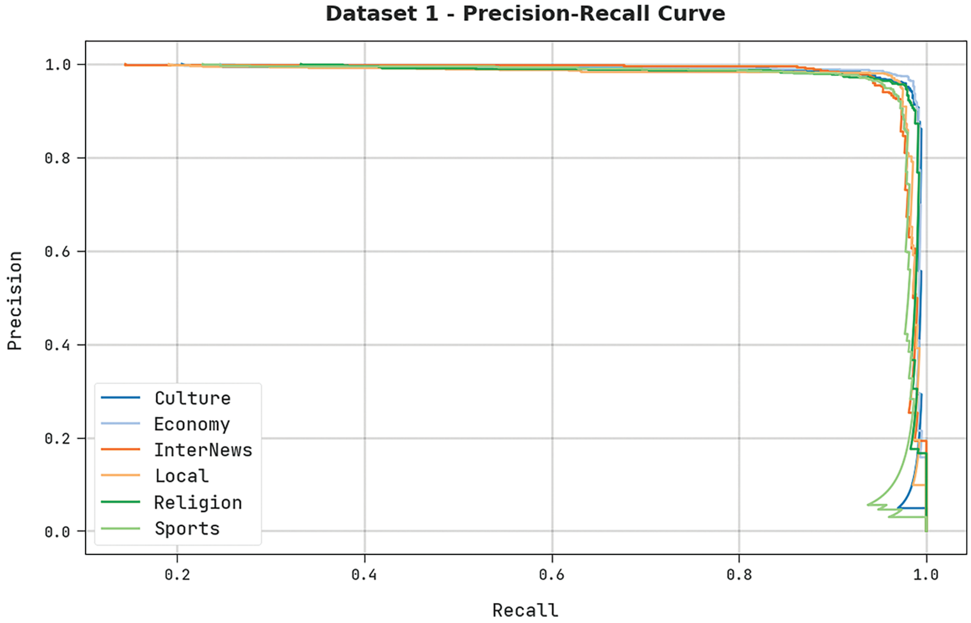


Figure 6: Precision-recall analyses results of the DODHBM-ACC approach on dataset-1

The proposed DODHBM-ACC algorithm was experimentally validated using Khaleej2004 dataset (dataset 2). The dataset holds 1,498 samples under four classes and is depicted in Table 3.

The confusion matrices generated by the proposed DODHBM-ACC method on dataset-2 are shown in Fig. 8. The results indicate that the proposed DODHBM-ACC system displayed improved outcomes under all the aspects. With the entire dataset, the DODHBM-ACC technique identified 214 samples as Economy, 210 samples as Internews, 600 samples as Local class and 438 samples as Sports class respectively. Further, upon 70% of TR dataset, the proposed DODHBM-ACC approach classified 142 samples under Economy, 146 samples under Internews, 430 samples under Local class and 302 samples under Sports class. Meanwhile, with 30% of TS, the presented DODHBM-ACC algorithm categorized 72 samples under Economy, 64 samples under Internews, 170 samples under Local class and 136 samples under Sports class.

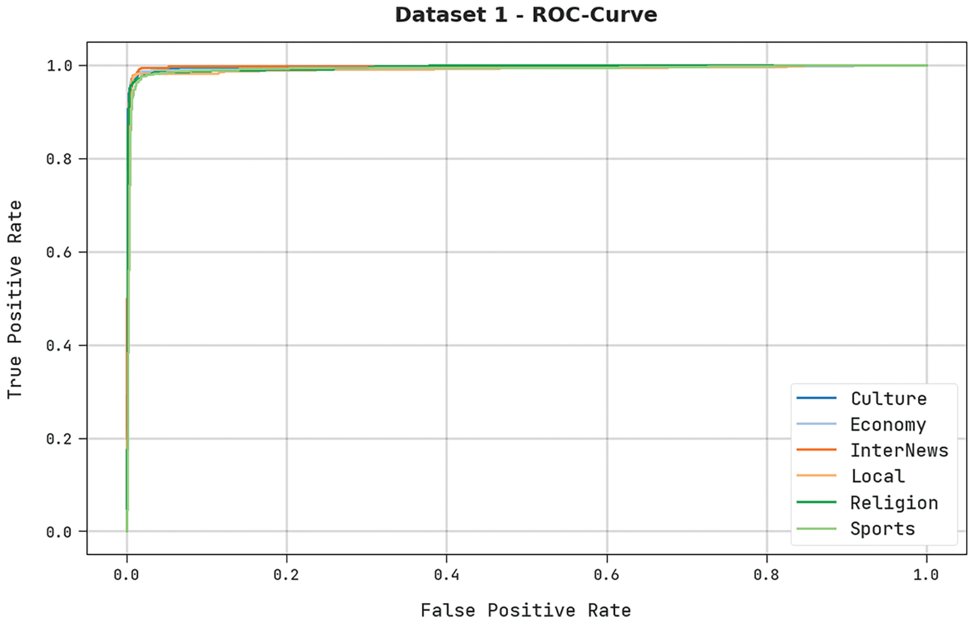


Figure 7: ROC analysis results of the DODHBM-ACC approach on dataset-1

Table 3: Details on dataset-2

Dataset 2-Khaleej2004 dataset	
Class	No. of samples
Economy	217
Internews	217
Local	613
Sports	451
Total no. of samples	**1498**

Table 4 demonstrates the overall classification results achieved by the proposed DODHBM-ACC methodology. With entire dataset, the DODHBM-ACC approach produced average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 98.80%, 97.12%, 97.60%, 97.34% and 98.39% respectively. Eventually, with 70% of TR, the proposed DODHBM-ACC method achieved average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 98.66%, 96.71%, 97.27%, 96.96% and 98.18% correspondingly. Also, with 30% of TS, the presented DODHBM-ACC approach attained average $accu_y$, $prec_n$, $reca_l$, F_{score} and AUC_{score} values such as 99.11%, 98.03%, 98.34%, 98.17% and 98.86% correspondingly.

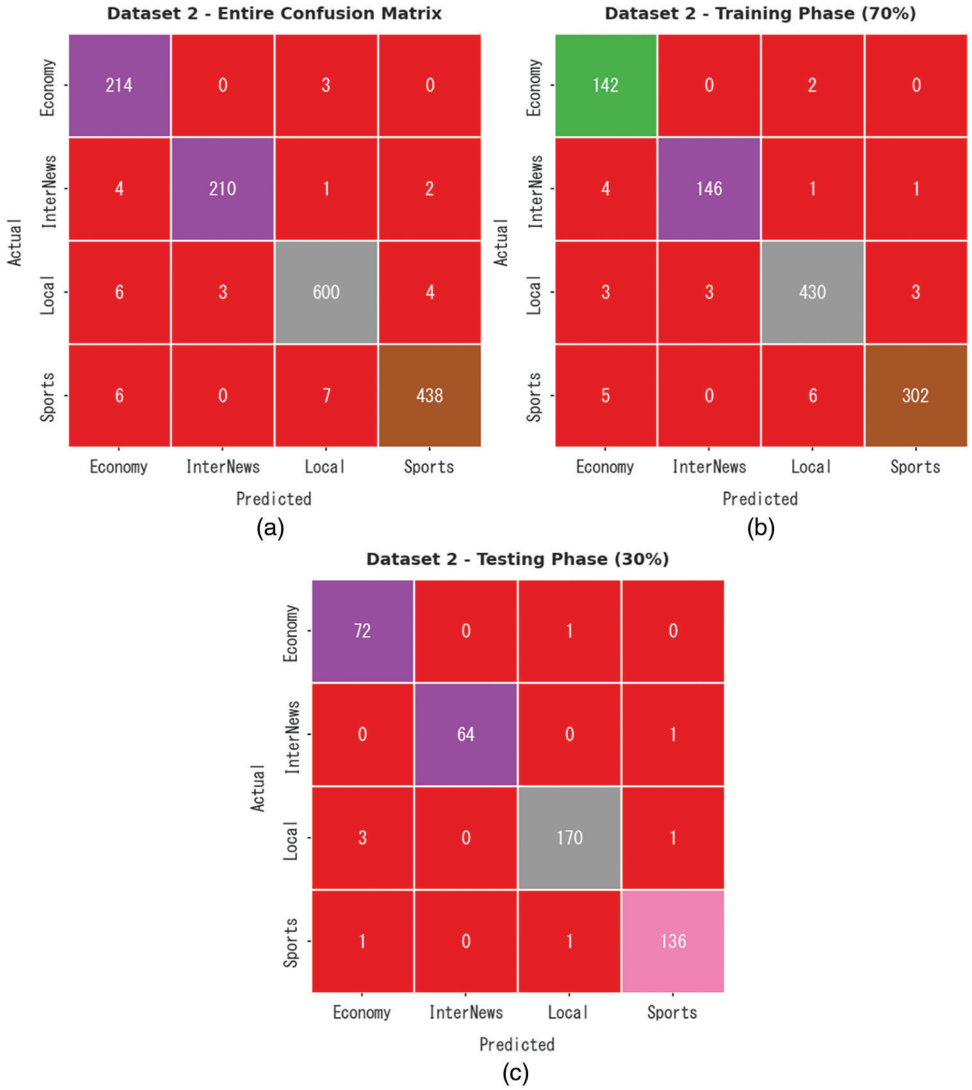


Figure 8: Confusion matrices of the DODHBM-ACC methodology under dataset-2 (a) Entire dataset, (b) 70% of TR data, and (c) 30% of TS dataset

Table 4: Analytical results of the DODHBM-ACC approach upon dataset-2 under distinct class labels

Dataset-2					
Labels	Accuracy	Precision	Recall	F-score	AUC score
Entire dataset					
Economy	98.73	93.04	98.62	95.75	98.68
InterNews	99.33	98.59	96.77	97.67	98.27
Local	98.40	98.20	97.88	98.04	98.32
Sports	98.73	98.65	97.12	97.88	98.27
Average	**98.80**	**97.12**	**97.60**	**97.34**	**98.39**

(Continued)

Table 4 (continued)

		Dataset-2			
Training phase (70%)					
Economy	98.66	92.21	98.61	95.30	98.64
InterNews	99.14	97.99	96.05	97.01	97.86
Local	98.28	97.95	97.95	97.95	98.24
Sports	98.57	98.69	96.49	97.58	97.97
Average	**98.66**	**96.71**	**97.27**	**96.96**	**98.18**
Testing phase (30%)					
Economy	98.89	94.74	98.63	96.64	98.78
Internews	99.78	100.00	98.46	99.22	99.23
Local	98.67	98.84	97.70	98.27	98.49
Sports	99.11	98.55	98.55	98.55	98.95
Average	**99.11**	**98.03**	**98.34**	**98.17**	**98.86**

Both TRA and VLA values, acquired by the DODHBM-ACC methodology on dataset-2, are illustrated in Fig. 9. The experimental outcomes denote that the proposed DODHBM-ACC approach gained the maximal TRA and VLA values while the VLA values were higher than the TRA values.

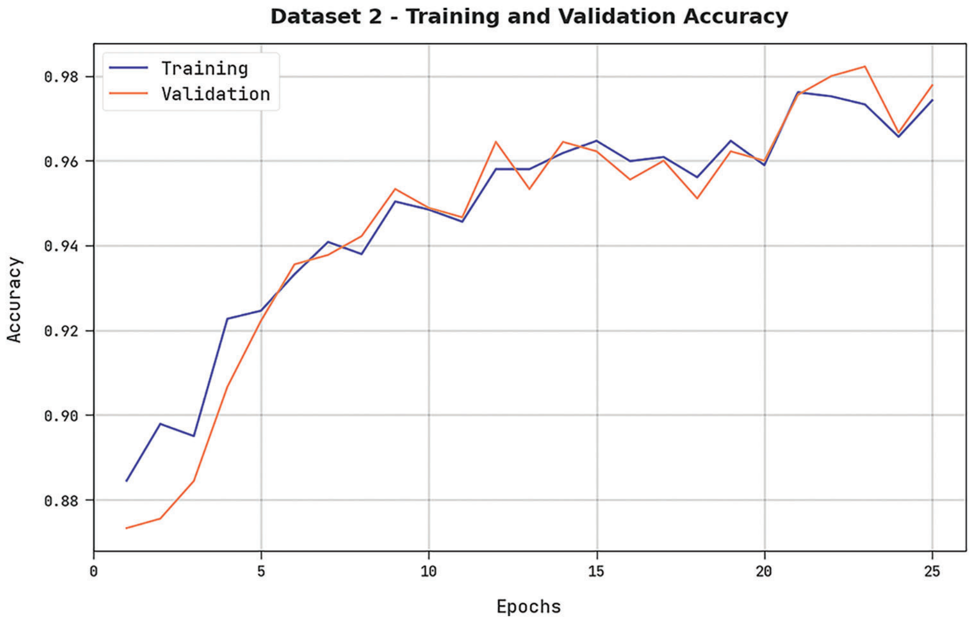


Figure 9: TRA and VLA analyses values of the DODHBM-ACC approach on dataset-2

Both TRL and VLL values, attained by the proposed DODHBM-ACC method on dataset-2, are shown in Fig. 10. The experimental outcomes imply that the proposed DODHBM-ACC technique exhibited the least TRL and VRL values while the VLL values were lesser than the TRL values.

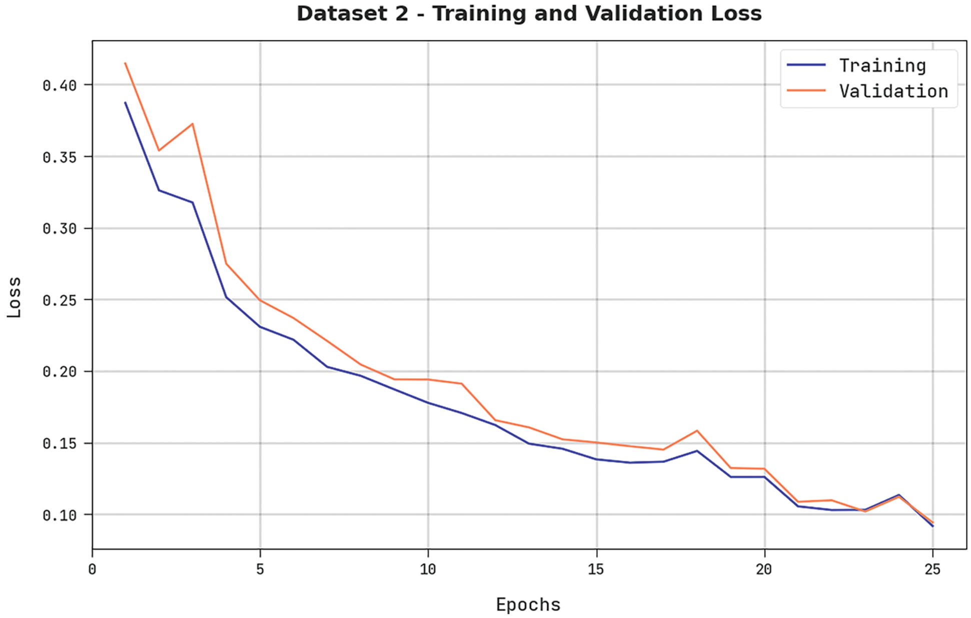


Figure 10: TRL and VLL analyses results of the DODHBM-ACC approach on dataset-2

A clear precision-recall analysis was conducted upon the proposed DODHBM-ACC approach on dataset-2 and the results are portrayed in Fig. 11. The figure represents that the proposed DODHBM-ACC algorithm produced enhanced precision-recall values under all classes.

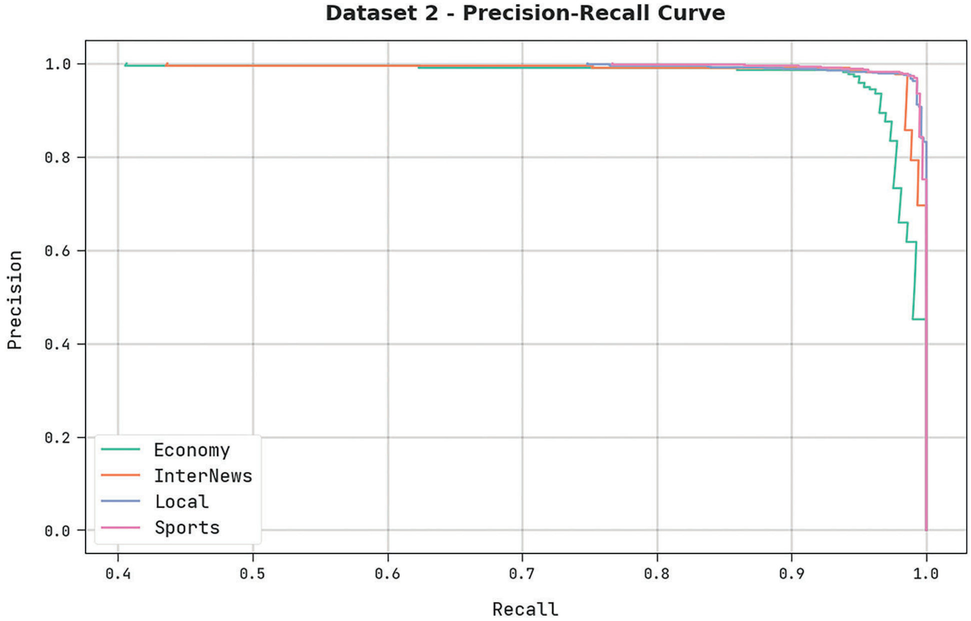


Figure 11: Precision-recall analysis results of the DODHBM-ACC approach on dataset-2

A brief ROC analysis was conducted upon the proposed DODHBM-ACC method using dataset-2, and the results are shown in Fig. 12. The results denote that the proposed DODHBM-ACC methodology established its ability in categorizing the dataset-2 under distinct classes.

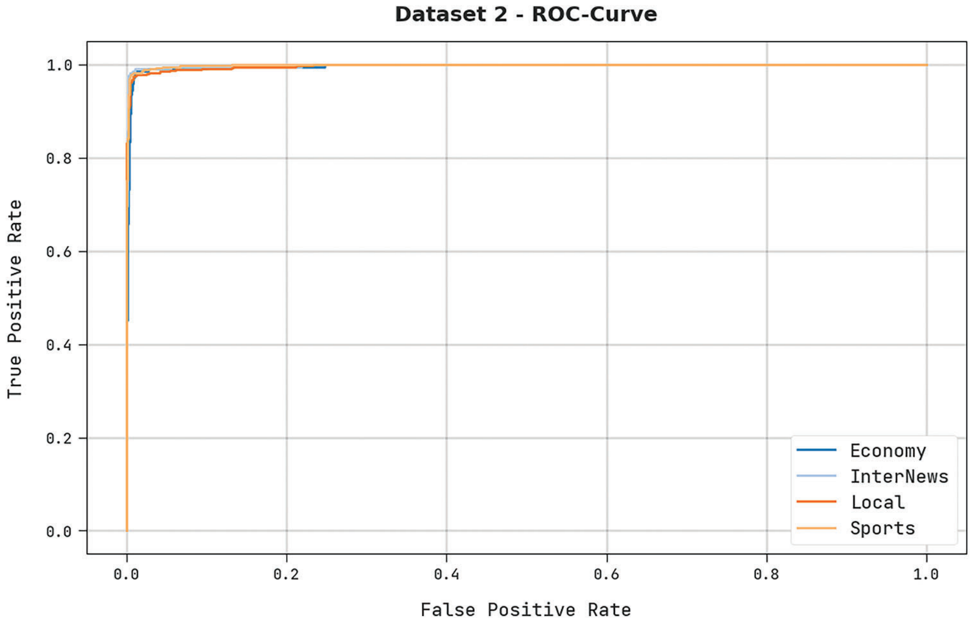


Figure 12: ROC analysis results of the DODHBM-ACC approach under dataset-2

Table 5 highlights the comparative inspection results accomplished by the proposed DODHBM-ACC model on two datasets [16]. The results imply that the DODHBM-ACC model achieved improved performance on both the datasets. For instance, on dataset-1, the proposed DODHBM-ACC model achieved an increased $accu_y$ of 98.60%, whereas the KNN, LOR, NB, RF and the SVM models obtained the least $accu_y$ values such as 94.96%, 96.07%, 95.56%, 97.26% and 96.80% respectively.

Table 5: Comparative analysis results of the DODHBM-ACC algorithm and other existing approaches on two datasets

Accuracy (%)		
Methods	Dataset-1	Dataset-2
DODHBM-ACC	98.60	99.11
KNN algorithm	94.96	95.03
LOR Model	96.07	96.14
NB Model	95.56	95.08
Random forest algorithm	97.26	95.96
SVM model	96.80	95.63

Moreover, on dataset-2, the presented DODHBM-ACC approach offered an increased $accu_y$ of 99.11%, whereas the other models such as KNN, LOR, NB, RF and SVM achieved low $accu_y$ values such as 95.03%, 96.14%, 95.08%, 95.96% and 95.63% correspondingly. Thus, the proposed DODHBM-ACC model can be utilized for the classification of the Arabic text in an effectual manner.

5 Conclusion

In the current study, a new DODHBM-ACC model has been developed for automated Arabic corpus classification. The presented DODHBM-ACC model primarily relies on different stages of pre-processing and word2vec word embedding process. In addition, the presented model uses the DHBM-based classification and DOA-based hyperparameter tuning processes. To adjust the hyperparameters of the DHBM technique, the DOA is exploited in this study. The experimental analysis was conducted to establish the supreme performance of the proposed DODHBM-ACC model. The outcomes confirmed the supremacy of the proposed DODHBM-ACC model over other recent approaches. In the future, the feature selection models can be utilized to reduce the computational complexity of the DODHBM-ACC model.

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: (22UQU4310373DSR53).

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

References

[1] M. Sayed, R. K. Salem and A. E. Khder, "A survey of Arabic text classification approaches," *International Journal of Computer Applications in Technology*, vol. 59, no. 3, pp. 236–251, 2019.

[2] 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.

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

[4] F. N. Al-Wasabi, "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.

[5] H. Chantar, M. Mafarja, H. Alsawalqah, A. A. Heidari, I. Aljarah *et al.,* "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification," *Neural Computing and Applications*, vol. 32, no. 16, pp. 12201–12220, 2020.

[6] F. N. Al-Wasabi, "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.

[7] A. Elnagar, R. Al-Debsi and O. Einea, "Arabic text classification using deep learning models," *Information Processing & Management*, vol. 57, no. 1, pp. 102121, 2020.

[8] F. N. Al-Wasabi, "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.

[9] S. Bahassine, A. Madani, M. Al-Sarem and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University—Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020.

[10] 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.

[11] K. A. Wahdan, S. 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.

[12] A. El Kah and I. Zeroual, "The effects of pre-processing techniques on Arabic text classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 1, pp. 41–48, 2021.

[13] N. Aljedani, R. Alotaibi and M. Taileb, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning," *Egyptian Informatics Journal*, vol. 22, no. 3, pp. 225–237, 2021.

[14] M. Hijazi, A. Zeki and A. Ismail, "Arabic text classification based on semantic and relation," *Computer Science*, vol. 37, no. 4, pp. 992, 2018.

[15] F. -Z. El-Alami, S. O. El Alaoui and N. En-Nahnahi, "Deep neural models and retrofitting for Arabic text categorization," *International Journal of Intelligent Information Technologies*, vol. 16, no. 2, pp. 74–86, 2020.

[16] M. Hijazi, A. Zeki and A. Ismail, "Arabic text classification using hybrid feature selection method using chi-square binary artificial bee colony algorithm," *International Journal of Mathematics and Computer Science*, vol. 16, no. 1, pp. 213–228, 2021.

[17] F. S. Al-Anzi and D. AbuZeina, "Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach," *Information Processing & Management*, vol. 54, no. 1, pp. 105–115, 2018.

[18] Y. A. Alhaj, A. Dahou, M. A. Al-qaness, L. Abualigah and A. A. Almaweri, "A novel text classification technique using improved particle swarm optimization: A case study of Arabic language," *Future Internet*, vol. 14, no. 7, pp. 194, 2022.

[19] H. N. Alshaer, M. A. Otair, L. Abualigah, M. Alshinwan and A. M. Khasawneh, "Feature selection method using improved CHI Square on Arabic text classifiers: Analysis and application," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10373–10390, 2021.

[20] A. H. Ababneh, "Investigating the relevance of Arabic text classification datasets based on supervised learning," *Journal of Electronic Science and Technology*, vol. 20, no. 2, pp. 100160, 2022.

[21] Y. Yao, X. Li, X. Liu, P. Liu, Z. Liang *et al.,* "Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model," *International Journal of Geographical Information Science*, vol. 31, no. 4, pp. 825–848, 2017.

[22] A. G. Ororbia II, C. L. Giles and D. Reitter, "Online semi-supervised learning with deep hybrid boltzmann machines and denoising autoencoders," arXiv preprint arXiv:1511.06964, 2015.

[23] M. Dehghani, Z. Montazeri and O. P. Malik, "DGO: Dice game optimizer," *Gazi University Journal of Science*, vol. 32, no. 3, pp. 871–882, 2019.