Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.034623
Article

Tech Science Press

Gender Identification Using Marginalised Stacked Denoising Autoencoders on Twitter Data

Badriyya B. Al-onazi[1], Mohamed K. Nour[2], Hassan Alshamrani[3], Mesfer Al Duhayyim[4,*], Heba Mohsen[5], Amgad Atta Abdelmageed[6], Gouse Pasha Mohammed[6] and Abu Sarwar Zamani[6]

[1]Department of Language Preparation, Arabic Language Teaching Institute, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[2]Department of Computer Sciences, College of Computing and Information System, Umm Al-Qura University, Makkah, 24211, Saudi Arabia
[3]Department of Teachers Training, Arabic Linguistics Institute, King Saud University, P.O. BOX 145111, Riyadh, ZIP 4545, Saudi Arabia
[4]Department of Computer Science, College of Sciences and Humanities-Aflaj, Prince Sattam bin Abdulaziz University, Al-Aflaj, 16733, Saudi Arabia
[5]Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, New Cairo, 11835, 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: 22 July 2022; Accepted: 22 September 2022

Abstract: Gender analysis of Twitter could reveal significant socio-cultural differences between female and male users. Efforts had been made to analyze and automatically infer gender formerly for more commonly spoken languages' content, but, as we now know that limited work is being undertaken for Arabic. Most of the research works are done mainly for English and least amount of effort for non-English language. The study for Arabic demographic inference like gender is relatively uncommon for social networking users, especially for Twitter. Therefore, this study aims to design an optimal marginalized stacked denoising autoencoder for gender identification on Arabic Twitter (OMSDAE-GIAT) model. The presented OMSDAE-GIAR technique mainly concentrates on the identification and classification of gender exist in the Twitter data. To attain this, the OMSDAE-GIAT model derives initial stages of data pre-processing and word embedding. Next, the MSDAE model is exploited for the identification of gender into two classes namely male and female. In the final stage, the OMSDAE-GIAT technique uses enhanced bat optimization algorithm (EBOA) for parameter tuning process, showing the novelty of our work. The performance validation of the OMSDAE-GIAT model is inspected against an Arabic corpus dataset and the results are measured under distinct metrics. The comparison study reported the enhanced performance of the OMSDAE-GIAT model over other recent approaches.

Keywords: Arabic twitter; gender identification; bat algorithm; hybrid deep learning; social media; arabic corpus

1 Introduction

Nowadays, the world is emerging around social networking platforms and its consequence for our day to day lives. Even though large number of people use social networking platform for connecting with individuals, other uses bogus account to commit violence, cyberstalking, and intimidation [1]. Nearly thirty percent of cyber harassers' gender is not known. Though English text has been widely inspected in this study, Arabic based author profiling, and specifically Gender Identification (GI), is in the course of the investigation. Arabic is the authorized language of the United Nation and the native language for 420 billion speakers in 22 Arab nations. In addition, the 1.8 million individuals following the Islam religion use Arabic in their day to day devotion, thus raising the prominence of Arabic Grammar Error Correction (GEC) [2]. Since automated GEC has become increasingly important, languages like Chinese and English have gained considerable interest from researchers. But some investigations were carried out on Arabic Grammar Error Correction (AGEC) owing to the insufficient learning dataset, the only accessible AGEC similar data is Qatar Arabic Language Bank (QALB) [3]. Besides, the complication of grammar and morphologically rich features of Arabic language are another concern. Arabic is a morphologically rich language that has many morphological characteristics like affixes, suffixes, and prefixes, demonstrating features of language like gender, numbers, and persons [4]. Those characteristics are up to sixteen morphemes, thereby raising the language uncertainty. For example, the word (-fasayashrabonahA) has a collection of morphological characteristics that are transformed into English as "They will drink it" [5].

Contextual word presentation is of considerable importance for the application of Natural Language Processing (NLP) like automatic summarization, text classification, plagiarism, data retrieval, and query suggestions [6]. Its prominence is associated with the fact that it simplifies the procedure of discovery relations among two terms and computes their resemblances. We employed word2vec method to compute the contextual presentation of the word. To calculate the vector presentation of words through neural network with single linear hidden layer on larger data set, Word2vec model used deep learning (DL) method [7]. Furthermore, word2vec trains the module on the basis of sliding window, the neighbor words within the window are considered for computing the probability of word occurrence, and the window keeps sliding over the entire corpus frequently. DL algorithm has proven its achievement in various NLP challenges with dissimilar architectures and models [8]. The model was developed on the basis of combination of structural units that accomplished better outcomes in text classification tasks. For GI problems, many structures like convolutional neural network (CNN) and long short term memory (LSTM) were investigated under the study and shows greater promise in resolving the challenge. Initially, this study begins with the basic neural network (NN) model and exploits them to resolve the GI challenge [9]. Then, add a degree of complexity to the NN model by integrating dissimilar layers from different NN methods for achieving high GI performance. In this study, a set of NN structures that are examined [10].

This study aims to design an optimal marginalized stacked denoising autoencoder for gender identification on Arabic Twitter (OMSDAE-GIAT) model. The presented OMSDAE-GIAR technique mainly concentrates on the identification and classification of gender exist in the Twitter data. To attain this, the OMSDAE-GIAT model derives initial stages of data pre-processing and word embedding. Next, the MSDAE model is exploited for the identification of gender into two classes namely male and female. In the final stage, the OMSDAE-GIAT technique involves the design of enhanced bat optimization algorithm (EBOA). The performance validation of the OMSDAE-GIAT model is inspected against an Arabic corpus dataset and the results are measured under distinct metrics.

The rest of the paper is organized as follows. Section 2 offers a brief survey of existing gender classification models and Section 3 introduces the proposed model. Section 4 draws experimental validation and Section 4 concludes the study.

2 Existing Gender Identification Models

ElSayed et al. [11] examine GI (male or female) of author posting Egyptian dialect twitters with NN model. Different structures of NN are discovered with wide-ranging parameter selection like LSTM, artificial neural network (ANN), CNN, Convolution Bi-directional gated recurrent unit (C-Bi-GRU), and Convolution Bidirectional-LSTM (C-Bi-LSTM) that is tuned for the GI problems. Ali et al. [12] make use of the customer profile name related to the submitted analyses to identify the customer's gender. Firstly, we construct dataset of profile names extracted from the customer review. Then, present a dynamic pruned n-gram models to identify gender of the user. Feature selection via a dynamic pruned n-gram method is the following step using the recurrent misspelling correction with fuzzy matching.

Alsharhan et al. [13] analyzed the communication among traditional corpus-compensation approaches (data selection, gender-dependent acoustic, feature selection models) and the register-dependent/dialect-dependent variations amongst Arabic corpus. The initial communication analyzed in the study is between discrete pronunciation variation and acoustic recording quality. The next aspect of register and dialect variations to be taken into consideration is difference in the fine-grained acoustic pronunciation of every phoneme in the language. Bsir et al. [14] characterize an addition of recurrent neural network (RNN) which applies variants of gated recurrent unit (GRU). This work presents a gender identification based on Facebook texts and Arabic Twitter by examining the inspected text features. The presented method makes use of the combination of unsupervised and supervised approaches for learning word vectors that capture the semantic and syntactic words.

Hussein et al. [15] present an Egyptian Dialect Gender Annotated Dataset (EDGAD) attained from Tweet and a presented text classifier solution for the GI problems. The data set involves seventy thousand tweets for each gender. During the classification of text, a Mixed Feature Vector (MFV) with dissimilar stylometric and Egyptian Arabic Dialect (EAD) language-specific features is developed, along with N-Gram Feature Vector (NFV). Ensemble weighted average is employed by the random forest (RF) and logistic regression (LR). Lokala et al. [16] aim is to evaluate and design a scheme for capturing the symptoms of mental health (MH) related to CVD that are diversely shown with gender on social networking platforms. We observed that the consistent recognition of MH symptoms shown by individuals with heart disease in user posts is very difficult due to the co-existence of distinct MH indications in single post and because of difference in the explanation of symptoms on the basis of based on gender. We developed GeM, a task-adaptive multi-task learning methodology to recognize the MH symptom in CVD patients according to gender.

3 The Proposed Gender Identification Model

In this manuscript, we have developed an intelligent gender identification approach called OMSDAE-GIAT model on Arabic Twitter. The presented OMSDAE-GIAR technique mainly concentrates on the identification and classification of gender exist in the Twitter data. To attain this, the OMSDAE-GIAT model derives initial stages of data pre-processing and word embedding. Next, the EBOA with MSDAE model is exploited for the identification of gender into two classes namely male and female. Fig. 1 displays the block diagram of OMSDAE-GIAT approach.

3.1 Data Preprocessing and Word Embedding

Initially, preprocessing for input dataset was executed in the following: The non-Arabic characters (special characters, English characters, links, and digits) were eliminated. Whitespace between emojis can be included for helping the tokenizer in extracting and separating emojis counts and types. Tweets below 5 words were discarded with the tweets which are duplicated. The Glove model is an alternative word embedding method [17]. This method makes use of an unsupervised learning technique to construct word

embeddings. The aim of these models is related to Word2Vec with respect to repelling different words and clustering similar words. But the mechanism is dissimilar from Word2Vec. Considering the contexts of the word that is surrounding, however, it inspects the occurrence of each word in the corpus. As a result, local and global data in the corpus are required for distributing the word vector. This method emphasizes on the non-zero value in global word to word co-occurrences matrixes. It evaluates the ratio of co-occurrence probability of both words from the input corpora. The affinity of this word is revealed once the ratio is larger, and vice-versa. The study presents a window size of 5 for each model, and the minimal amount of words was equivalent to five. As well, we applied three distinct dimensions for these two approaches that are: 100, 200, and 300.

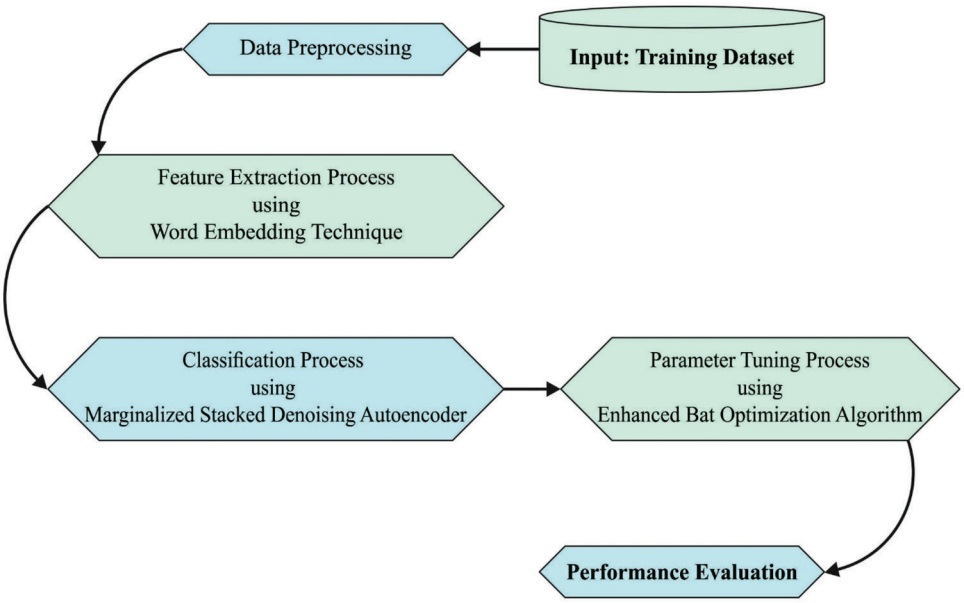

Figure 1: Block diagram of OMSDAE-GIAT approach

3.2 Gender Identification Using MSDAE Model

Here, the MSDAE model is exploited for the identification of gender into two classes namely male and female. Deep neural network (DNN) architecture (NN with multiple hidden layers) is extremely prevalent in ML because of its higher ability for data modeling. But, devising multiple layers implies that multiple variables are needed to be adjusted in the training stage. Consequently, there exists an over-fitting problem with the data training and the network falls into a local minimum [18]. Furthermore, fine-tuning additional parameters give computation problems namely increased training time and memory limitation. A method to prevent this problem is to sequentially train every layer, and later stack them on topmost of others while retaining the weight of static trained layer. This method is called a Stacked Denoising Autoencoder (SDA); every latent representation characterizes an abstract level that is utilized for a regression or classification task 20. The phases to train an SDA are displayed in Algorithm 1. Generally, the demonstration at the concluding layer is deliberated for additional investigation 6, but, concatenate each abstract representation and the new vector dataset is employed 21 (to utilize data in each abstract presentation). It should be noted that in the network construction, when the layer is trained, it obtains the uncorrupted output of the preceding layer. The initialization of weight of a deep network performs a significant part in preventing local minima. Stacked Sparse Autoencoder (SSAE) was utilized for initializing the weight layer-wise, and later tuning the weight of SDA. It is demonstrated that this

technique enhances the performance through arbitrary weight initialization on the IMDb dataset 23. Also, SDA was applied in domain adaptation for sentimental detection of review of dissimilar products 24. In the study, each corpus is existing to the SDA for extracting a shared presentation of each source; then classification is implemented on the representation. Furthermore, SDA presented an important development on sentimental significance recognition for cross corpus analysis. Even though SDA mediate specific problems of AE, it still they suffer from two limits: (i) insufficient scalability to higher dimension feature; and (ii) higher computation cost (stochastic gradient descent learning). In addition, it is essential to produce a greater amount of noisy instances that are corrupted on distinct characteristics, and deliver them to the network in training. Marginalized SDA (MSDAE) assists to overcome the limitation through a closed-form solution to evaluate the network weight and indirectly employ Denoising without producing any noisy samples. Fig. 2 showcases the infrastructure of SDAE.

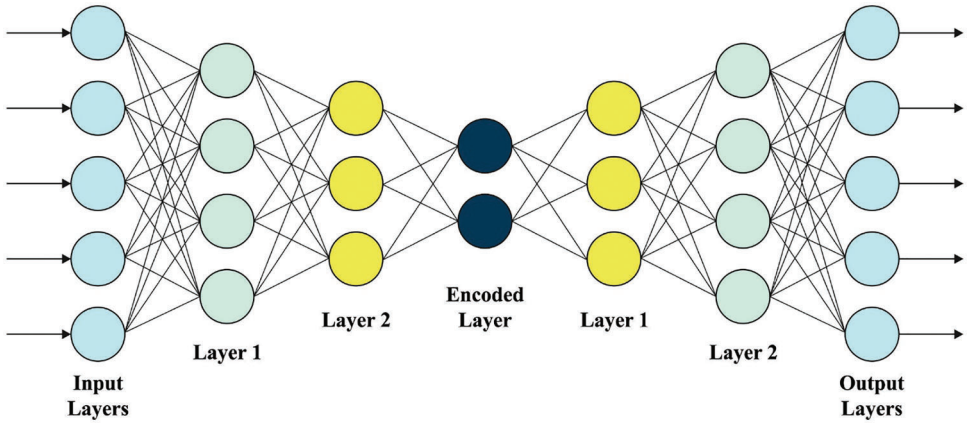

Figure 2: Structure of SDAE

A typical MSDAE comprises several layers of linear denoiser, whereas the objective function is:

$$\mathcal{L}(W) = \frac{1}{m}\sum_{i=1}^{m} \|X_i - W\breve{X}_i\|^2, \tag{1}$$

whereas $W: R^{N+1} \rightarrow R^{N+1}$, and it considers that $X_i = [X_i;\ 1]$. The solution to this formula is formulated as a closed-form ordinary least squares:

$$W = PQ^{-1}, \qquad where\ \ Q = \breve{X}\breve{X}^T\ and\ P = X\breve{X}^T. \tag{2}$$

Taking $k \rightarrow \infty$ copies of $X \in R^{N\times M}$ corrupted by noise, $P \in R^{N\times N}$ and $Q \in R^{N\times N}$ converge to its expected values $E[P]$ and $E[Q]$. Thus:

$$W = E[P]E[Q]^{-1} \tag{3}$$

Assume that $q = [1-p.\ 1-p,\ \ 1]^T \in R^{N+1}$, whereas q_α and q_β implies the probability of α^{th} and β^{th} features survive the corruption (that takes place with probability p). Afterward, 2 features m_α and m_β surviving corruption with probability $(1-p)^2$. When it can be determined $S = XX^T$ as the scatter matrix of uncorrupted inputs, it can then represent the expected value of Q as:

$$E[Q]_{\alpha,\beta} = \begin{cases} S_{\alpha,\beta}q_\alpha q_\beta\ if\ m_\alpha \neq m_\beta \\ S_{\alpha,\beta}q_\alpha\ if\ \alpha = \beta \end{cases} \tag{4}$$

and, $E[P]_{\alpha,\beta} = S_{\alpha,\beta} q_{\beta}$. For embedding non-linearity as MSDAE, afterward computing W to all the layers, the non-linear 'squashing' function was executed to the resultant of layers. The benefits contain: (1) only one pass with trained data was needed; (2) the convex optimum solution was guaranteed; and (3) optimization is from the closed procedure. Moreover, the MSDAEs demonstrated a massive speedup (×450) in the trained with comparable performance to $A_s^{27\,28}$.

Noticeable, the calculation of $E[Q]^{-1}$ was computationally costly to dataset with higher dimensionality (for instance, if the demonstrating text as Bag of Words (BoW)). For coping with this problem, the author decreased the dimensionality of input dataset to only five thousand frequent terms.

They employed an MSDAE depiction of features as the input to Domain Adversarial Neural Networks (DANNs). In the DANNs $G_f(.)$, 2 other deep networks were augmented to predict a class label, $G_y(.)$ and a domain label, $G_d(.)$. In the trained stage, the main function was planned for minimizing the label forecast loss, and concurrently, maximizing the domain forecast loss. The latter guarantees that the 2 domains were mapped to everyone. Noticeable, the MSDAE representation is a concatenation of the abstract feature of every layer of MSDAE.

Algorithm 1: Stacked Denoising Autoencoder training algorithm

Input : X: Data, L:#layers

Output: F: feature vector

Definitions: W: encoding weights, b: encoding bias, W': decoding weights, b':

decoding bias, q: corrupting function;

$l \leftarrow 1$;

$\tilde{X}_1 \leftarrow X$;

while $l \leq L$ do

initialize new layer with W_l, b_l, W'_l, b'_l;

$\breve{X} \Leftarrow q(\tilde{X}_l)$ //Corrupt the input;

train W_l, b_l, W'_l, b'_l with input $\breve{X}$ and output X;

$\tilde{X}_{l+1} \leftarrow f(\tilde{X}_l | W_l, \ b_l)$ //Generate features for the next layer;

$l \leftarrow l + 1$;

end

$F \leftarrow \tilde{X}_{L+1}$ OR concat $(\tilde{X}_1, \ \ldots, \ \tilde{X}_{L+1})$

3.3 Parameter Tuning Using EBOA

In the final stage, the OMSDAE-GIAT technique involves the design of EBOA. The BOA which a new metaheuristic technique that has received consideration from several researcher workers from distinct areas due to its outstanding ability of echolocation [19]. It is inspired by the echolocation capability of microbats guiding them on foraging attitude. There exist different kinds of bats that vary in weight and size however, they have analogous behaviors of hunting and navigation. The BOA was developed based on the fundamental benefits of the bat while finding the target. They have a stronger sound propagation method and are distinct from other birds named bio-sonar. While chasing the prey tends to improve the rate of sound produced by the ultrasound and reduces the loudness. The behavior of bats has been demonstrated in the following. BOA has a group of vectors signifying the frequency, location, and velocity that is upgraded as follows:

$$V_i^{(t+1)} = V_i^{(t)} + \left(X_i^{(t)} - G_{best}\right)F_i \tag{5}$$

$$X_i^{(t+1)} = X_i^{(t)} + V_i^{(t+1)} \tag{6}$$

Now, G_{best} characterizes optimal solution, F_i denotes frequency of i-th bat that is upgraded by the iteration in the following:

$$F_i = F_{\min} + (F_{\max} - F_{\min})X_{new} = X_{old} + \varepsilon A^t \tag{7}$$

In Eq. (7), β signifies a uniformly distributed random integer within [0, 1]. Note that, different frequencies of bats encourage them to tend to the solution. However, exploitation was performed by using a random walk:

$$X_{new} = X_{old} + \varepsilon A^t \tag{8}$$

In Eq. (8), ε specifies an arbitrary integer within $[-1,\ \ 1]$, and A characterizes the loudness of sound emitted. The loudness A and the pulse emission rate r are upgraded by controlling the balance:

$$A_i^{(t+1)} = \alpha A_i^{(t)} \tag{9}$$

$$r_i^{(t+1)} = r_i(0)[1 - \ \exp\ (-\gamma t)] \tag{10}$$

In Eq. (10), α and γ signify constant; t is equivalent to the cooling factor, finally, A_i is equivalent to zero and the final value of r_i is $r(O)$. The updating of the loudness and rate assurance that artificial bat moves to the optimal solution.

In this study, the EBOA is derived by the usage of Levy flight. Levy walk can be a description of diffusion paradigm of creatures that seeking is focused the location of effective solution [20]. Levy flight foraging hypothesis predicts the migration from less-resourced to the more-source environment that leads to optimal search. Animals having higher memory ability uses this method for exploring their search spaces. The theory of optimum foraging was an extension of Levy's flight foraging hypothesis, which says that creatures pay much interest to the optimum solution location than aimless search in the search spaces. Levy flights were random walks whose step length was withdrawn from Levy distributions, frequently with regard to a power-law formula

$L(\zeta) \sim \zeta^{-1-\alpha}$ whereas $0 < \alpha < 2$ was an index. Levy flight is mathematically indicated as

$$L(\zeta,\ \ \omega,\ \ \psi) = \begin{cases} \sqrt{\dfrac{\omega}{2\pi}}\exp\left[-\dfrac{\omega}{2(\zeta-\psi)}\right]\dfrac{1}{\left(\zeta - \dfrac{\psi^3}{2}\right)}, & 0 < \psi < \zeta < \infty \\ 0 & otherwise \end{cases} \tag{11}$$

Whereas $\psi > 0$ denotes a minimal step and ω refers to a scale parameter. Rather as $\zeta \to \infty$, then

$$L(\zeta,\ \ \omega,\ \ \psi) \approx \sqrt{\frac{\omega}{2\pi}}\frac{1}{\zeta^3/_2}. \tag{12}$$

Mantegna technique is used for levy flight implementation in this study. Therefore, the step length ζ is computed by

$$\zeta = \frac{\ell}{|\kappa|^1/\alpha}, \tag{13}$$

Whereas ℓ and κ were withdrawn from normal distribution. Otherwise,

$$\ell \sim N\left(0,\ \rho_\ell^2\right),\ \kappa \sim N(0,\ \rho_\ell^2), \tag{14}$$

4 Results and Discussion

The experimental analysis of the OMSDAE-GIAT model is tested using the EDGAD dataset [15]. The dataset comprises a total of 70 accounts under each gender. Table 1 offers the details of dataset.

Table 1: Dataset details

Metric	Female	Male
Number of accounts	70	70
Number of tweets	1000	1000
Average tweet length	14	14
Vocabulary per gender	9687	10233

Table 2 and Fig. 3 report the gender identification output of the OMSDAE-GIAT model under run-1. The results implied that the OMSDAE-GIAT model has accomplished enhanced outcomes with classifier results. For example, with 50 lengths, the OMSDAE-GIAT technique has gained $accu_y$ of 75.67%, 82.49%, 84.05%, and 88.97% under 1, 4, 8, and 12 tweets correspondingly. Also, with 80 lengths, the OMSDAE-GIAT algorithm has attained $accu_y$ of 77.42%, 84.18%, 84.52%, and 90.37% under 1, 4, 8, and 12 tweets correspondingly. Parallelly, with 100 lengths, the OMSDAE-GIAT approach has acquired $accu_y$ of 75.63%, 75.18%, 77.62%, and 84.25% under 1, 4, 8, and 12 tweets correspondingly.

Table 2: Result analysis of OMSDAE-GIAT approach with distinct lengths under run-1

	Run-1			
Maximum length	1 tweet	4 tweet	8 tweet	12 tweet
50	75.67	82.49	84.05	88.97
80	77.42	84.18	84.52	90.37
100	75.63	75.18	77.62	84.25
120	76.00	79.60	83.97	85.47
140	76.60	83.48	84.04	88.87
Average	**76.26**	**80.99**	**82.84**	**87.59**

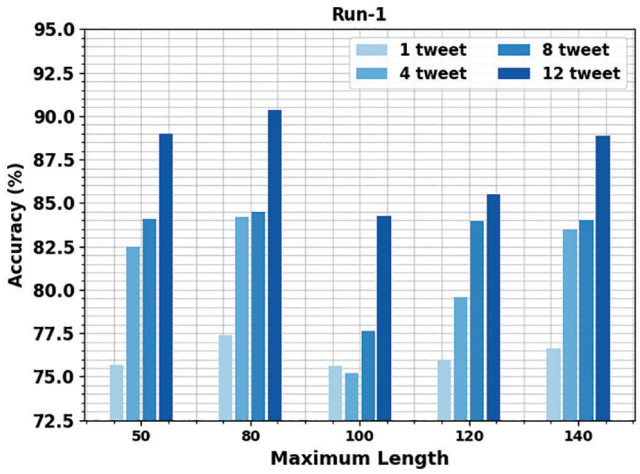

Figure 3: Result analysis of OMSDAE-GIAT approach under run-1

Fig. 4 depicts the average gender identification results of the OMSDAE-GIAT model under varying numbers of tweets. The figure represented that the OMSDAE-GIAT model has shown enhanced performance with average $accu_y$ of 76.26%, 80.99%, 82.84%, and 87.59% under 1, 4, 8, and 12 tweets respectively.

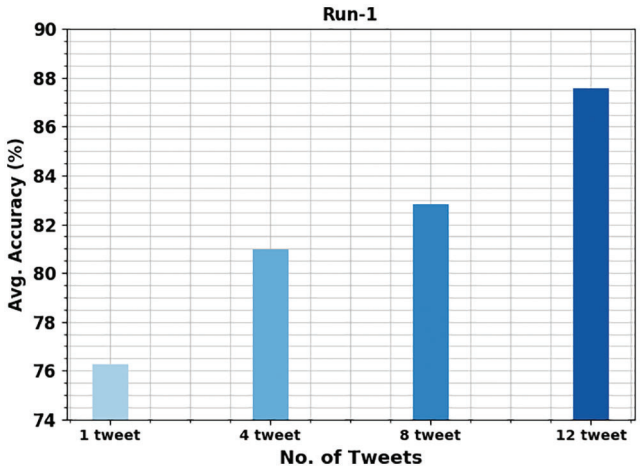

Figure 4: Average analysis of OMSDAE-GIAT approach under run-1

Table 3 and Fig. 5 report the gender identification output of the OMSDAE-GIAT technique under run-2. The results indicate the OMSDAE-GIAT algorithm has exhibited enhanced outcomes with classifier results. For example, with 50 lengths, the OMSDAE-GIAT method has reached $accu_y$ of 76.24%, 78.03%, 77.30%, and 78.69% under 1, 4, 8, and 12 tweets correspondingly. Along with that, with 80 lengths, the OMSDAE-GIAT model has obtained $accu_y$ of 76.94%, 76.70%, 82.35%, and 81.95% under 1, 4, 8, and 12 tweets correspondingly. Simultaneously, with 100 lengths, the OMSDAE-GIAT approach has attained $accu_y$ of 77.04%, 81.32%, 83.84%, and 83.82% under 1, 4, 8, and 12 tweets correspondingly.

Fig. 6 portrays the average gender identification results of the OMSDAE-GIAT approach under varying numbers of tweets. The figure indicated the OMSDAE-GIAT approach has shown enhanced performance with average $accu_y$ of 76.42%, 79.41%, 83.57%, and 85.55% under 1, 4, 8, and 12 tweets correspondingly.

Table 3: Result analysis of OMSDAE-GIAT approach with distinct lengths under run-2

Run-2				
Maximum length	1 tweet	4 tweet	8 tweet	12 tweet
50	76.24	78.03	77.30	78.69
80	76.94	76.70	82.35	81.95
100	77.04	81.32	83.84	83.82
120	76.84	83.91	89.82	93.17
140	75.02	77.11	84.55	90.13
Average	**76.42**	**79.41**	**83.57**	**85.55**

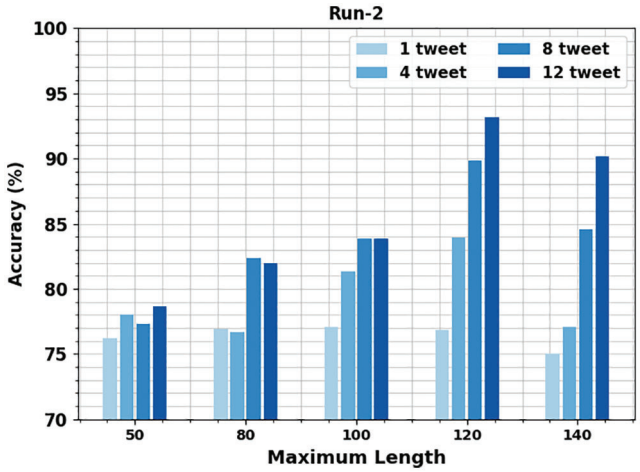

Figure 5: Result analysis of OMSDAE-GIAT approach under run-2

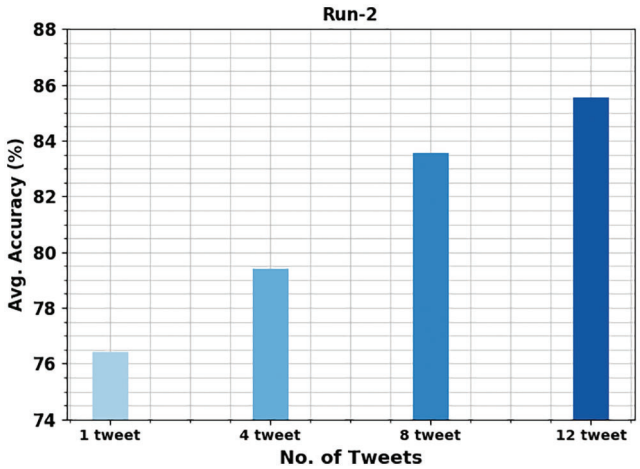

Figure 6: Average analysis of OMSDAE-GIAT approach under run-2

Table 4 and Fig. 7 report the gender identification output of the OMSDAE-GIAT algorithm under run-3. The results represented the OMSDAE-GIAT model has accomplished enhanced outcomes with classifier results. For example, with 50 lengths, the OMSDAE-GIAT model has obtained $accu_y$ of 75.47%, 81.99%, 81.45%, and 88.34% under 1, 4, 8, and 12 tweets respectively. In addition, with 80 lengths, the OMSDAE-GIAT technique has acquired $accu_y$ of 75.73%, 78.66%, 84.18%, and 84.21% under 1, 4, 8, and 12 tweets correspondingly. Alongside, 100 lengths, the OMSDAE-GIAT approach has attained $accu_y$ of 77.15%, 78.40%, 79.44%, and 79.92% under 1, 4, 8, and 12 tweets correspondingly.

Table 4: Result analysis of OMSDAE-GIAT approach with distinct lengths under run-3

Run-3				
Maximum length	1 tweet	4 tweet	8 tweet	12 tweet
50	75.47	81.99	81.45	88.34
80	75.73	78.66	84.18	84.21
100	77.15	78.40	79.44	79.92
120	75.21	79.57	79.23	80.45
140	77.15	82.44	86.38	92.38
Average	**76.14**	**80.21**	**82.14**	**85.06**

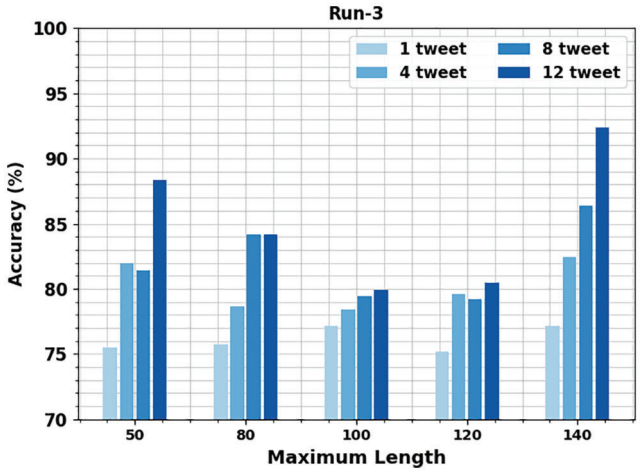

Figure 7: Result analysis of OMSDAE-GIAT approach under run-3

Fig. 8 shows the average gender identification results of the OMSDAE-GIAT technique under varying numbers of tweets. The figure signifies the OMSDAE-GIAT approach has displayed enhanced performance with average $accu_y$ of 76.14%, 80.21%, 82.14%, and 85.06% under 1, 4, 8, and 12 tweets correspondingly.

Table 5 and Fig. 9 illustrate the gender identification output of the OMSDAE-GIAT technique under run-4. The results implied that the OMSDAE-GIAT algorithm has outperformed enhanced outcomes with classifier results. For instance, with 50 lengths, the OMSDAE-GIAT approach has reached $accu_y$ of 76.44%, 78.89%, 82.68%, and 85.90% under 1, 4, 8, and 12 tweets correspondingly. Also, with 80 lengths, the OMSDAE-GIAT technique has attained $accu_y$ of 75.86%, 79.73%, 86.47%, and 86.90% under 1, 4, 8, and 12 tweets correspondingly. Synchronously, with 100 lengths, the OMSDAE-GIAT

methodology has acquired $accu_y$ of 76.51%, 81.94%, 81.05%, and 87.80% under 1, 4, 8, and 12 tweets correspondingly.

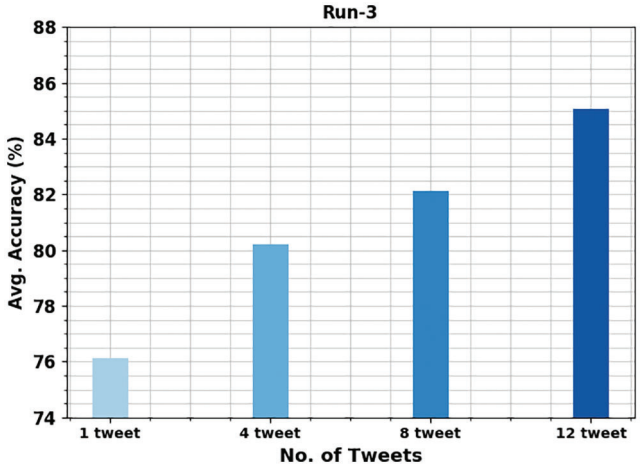

Figure 8: Average analysis of OMSDAE-GIAT approach under run-3

Table 5: Result analysis of OMSDAE-GIAT approach with distinct lengths under run-4

Run-4				
Maximum length	1 tweet	4 tweet	8 tweet	12 tweet
50	76.44	78.89	82.68	85.90
80	75.86	79.73	86.47	86.90
100	76.51	81.94	81.05	87.80
120	77.68	80.48	86.74	86.58
140	75.29	76.25	80.38	84.37
Average	**76.36**	**79.46**	**83.46**	**86.31**

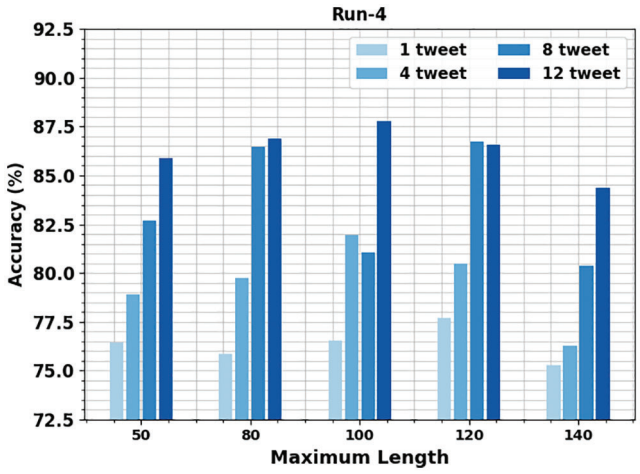

Figure 9: Result analysis of OMSDAE-GIAT approach under run-4

Fig. 10 portrays the average gender identification results of the OMSDAE-GIAT algorithm, under varying numbers of tweets. The figure denoted that the OMSDAE-GIAT technique has exhibited enhanced performance with average $accu_y$ of 76.36%, 79.46%, 83.46%, and 86.31% under 1, 4, 8, and 12 tweets correspondingly.

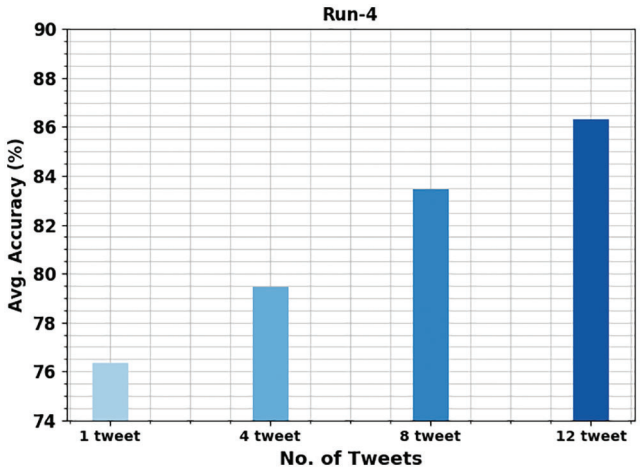

Figure 10: Average analysis of OMSDAE-GIAT approach under run-4

Table 6 and Fig. 11 demonstrate the gender identification output of the OMSDAE-GIAT algorithm under run-5. The results implied that the OMSDAE-GIAT technique has exhibited enhanced outcomes with classifier results. For example, with 50 lengths, the OMSDAE-GIAT method has reached $accu_y$ of 75.57%, 80.35%, 80.17%, and 90.18% under 1, 4, 8, and 12 tweets correspondingly. Also, with 80 lengths, the OMSDAE-GIAT approach has attained $accu_y$ of 77.61%, 84.90%, 87.65%, and 89.37% under 1, 4, 8, and 12 tweets correspondingly. Synchronously, with 100 lengths, the OMSDAE-GIAT technique has gained $accu_y$ of 76.01%, 76.02%, 82.58%, and 87.19% under 1, 4, 8, and 12 tweets correspondingly.

Table 6: Result analysis of OMSDAE-GIAT approach with distinct lengths under run-5

Run-5				
Maximum length	1 tweet	4 tweet	8 tweet	12 tweet
50	75.57	80.35	80.17	90.18
80	77.61	84.90	87.65	89.37
100	76.01	76.02	82.58	87.19
120	78.23	82.37	88.45	89.65
140	79.74	81.59	81.91	91.66
Average	**77.43**	**81.05**	**84.15**	**89.61**

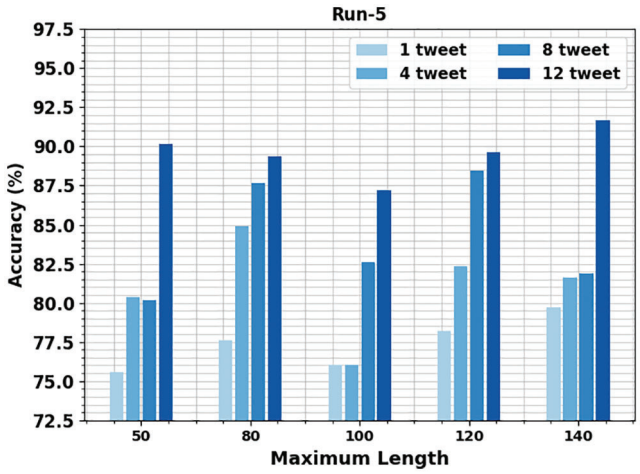

Figure 11: Result analysis of OMSDAE-GIAT approach under run-5

Fig. 12 describes the average gender identification results of the OMSDAE-GIAT technique under varying numbers of tweets. The figure denotes the OMSDAE-GIAT algorithm has shown enhanced performance with average $accu_y$ of 77.43%, 81.05%, 84.15%, and 89.61% under 1, 4, 8, and 12 tweets correspondingly.

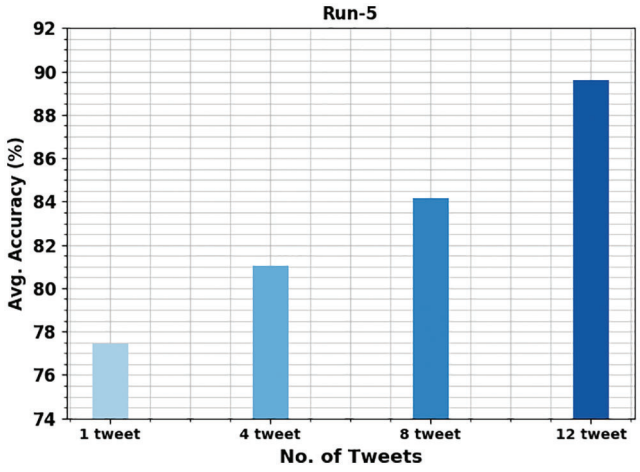

Figure 12: Average analysis of OMSDAE-GIAT approach under run-5

For illustrating the improvements of the OMSDAE-GIAT model over other techniques, a brief comparison study is carried out in Table 7 [11]. The obtained values inferred that the OMSDAE-GIAT model has shown improved values with maximum classification results under all tweets. For instance, with 1 tweet, the OMSDAE-GIAT model has offered increased $accu_y$ of 77.43% whereas the ANN, CNN, multichannel CNN, LSTM-NN, CBi-LSTM NN, and multichannel CBi-GRU NN models have reported reduced $accu_y$ of 61.96%, 56.52%, 65.10%, and 53.23% respectively. Also, with 4 tweets, the OMSDAE-GIAT algorithm has provided increased $accu_y$ of 81.05% whereas the ANN, CNN, multichannel CNN, LSTM-NN, CBi-LSTM NN, and multichannel CBi-GRU NN models have reported reduced $accu_y$ of 71.88%, 72.05%, 71.45%, 53.99%, 70.48% and 72.41% correspondingly.

Table 7: Comparative analysis of OMSDAE-GIAT approach with recent methodologies

Methods	1 tweet	4 tweet	8 tweet	12 tweet
OMSDAE-GIAT	77.43	81.05	84.15	89.61
ANN	61.96	71.88	78.15	81.23
CNN	56.52	72.05	81.11	81.81
Multichannel CNN	65.10	71.45	80.22	77.54
LSTM-NN	53.23	53.99	54.25	53.00
CBi-LSTM NN	56.87	70.48	78.86	78.73
Multichannel CBi-GRU NN	66.27	72.41	82.36	84.64

Similarly, with 8 tweet, the OMSDAE-GIAT approach has presented increased $accu_y$ of 84.15% whereas the ANN, CNN, multichannel CNN, LSTM-NN, CBi-LSTM NN, and multichannel CBi-GRU NN algorithms have reported reduced $accu_y$ of 78.15%, 81.11%, 80.22%, 54.25%, 78.86% and 82.36% correspondingly. Therefore, the results confirmed the betterment of the OMSDAE-GIAT model over the other recent approaches.

5 Conclusion

In this manuscript, we have developed an intelligent gender identification approach called OMSDAE-GIAT model on Arabic Twitter. The presented OMSDAE-GIAR technique mainly concentrates on the identification and classification of gender exist in the Twitter data. To attain this, the OMSDAE-GIAT model derives the initial stages of data pre-processing and word embedding. Next, the MSDAE model is exploited for the identification of gender into two classes namely male and female. In the final stage, the OMSDAE-GIAT technique involves the design of EBOA. The performance validation of the OMSDAE-GIAT model is inspected against an Arabic corpus dataset and the results are measured under distinct metrics. The comparison study reported the enhanced performance of the OMSDAE-GIAT model over other recent approaches. In future extensions, the presented OMSDAE-GIAT model can be extended to other languages.

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: 22UQU4310373DSR55.

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

References

[1] G. O. Adebayo and R. V. Yampolskiy, "Estimating intelligence quotient using stylometry and machine learning techniques: A review," *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 163–191, 2022.

[2] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh *et al.,* "A panoramic survey of natural language processing in the arab world," *Communications of the ACM*, vol. 64, no. 4, pp. 72–81, 2021.

[3] A. Hanani and R. Naser, "Spoken arabic dialect recognition using x-vectors," *Natural Language Engineering*, vol. 26, no. 6, pp. 691–700, 2020.

[4] S. Alzahrani, "Grammatical gender assignment on arabic nouns: Saudi dialects," *US-China Foreign Language*, vol. 17, no. 6, pp. 251–270, 2019.

[5] M. Salim, s. SAAD, and M. Aref, "Preprocessing the Egyptian arabic dialect for personality traits prediction," *International Journal of Intelligent Computing and Information Sciences*, vol. 19, no. 1, pp. 1–12, 2019.

[6] A. I. Al-Ghadir and A. M. Azmi, "A study of arabic social media users—posting behavior and author's gender prediction," *Cognitive Computation*, vol. 11, no. 1, pp. 71–86, 2019.

[7] P. Rosso, F. Rangel, I. H. Farías, L. Cagnina, W. Zaghouani *et al.,* "A survey on author profiling, deception, and irony detection for the arabic language," *Language and Linguistics Compass*, vol. 12, no. 4, pp. e12275, 2018.

[8] A. Y. Muaad, G. H. Kumar, J. Hanumanthappa, J. B. Benifa, M. N. Mourya *et al.,* "An effective approach for arabic document classification using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 267–271, 2022.

[9] M. Awais, F. Naeem, N. Rasool, and S. Mahmood, "Identification of sex from footprint dimensions using machine learning: A study on population of Punjab in Pakistan," *Egyptian Journal of Forensic Sciences*, vol. 8, no. 1, pp. 72, 2018.

[10] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 6, pp. 275–284, 2017.

[11] S. ElSayed and M. Farouk, "Gender identification for Egyptian arabic dialect in twitter using deep learning models," *Egyptian Informatics Journal*, vol. 21, no. 3, pp. 159–167, 2020.

[12] N. M. Ali, A. Alshahrani, A. M. Alghamdi, and B. Novikov, "Using dynamic pruned n-gram model for identifying the gender of the user," *Applied Sciences*, vol. 12, no. 13, pp. 6378, 2022.

[13] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975–998, 2020.

[14] B. Bsir and M. Zrigui, "Enhancing deep learning gender identification with gated recurrent units architecture in social text," *Computación y Sistemas*, vol. 22, no. 3, pp. 757–766, 2018.

[15] S. Hussein, M. Farouk, and E. Hemayed, "Gender identification of Egyptian dialect in twitter," *Egyptian Informatics Journal*, vol. 20, no. 2, pp. 109–116, 2019.

[16] U. Lokala, A. Srivastava, T. G. Dastidar, T. Chakraborty, M. S. Akhtar *et al.* "A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, Atlanta, Georgia, vol. 16, pp. 640–650, 2022.

[17] C. Zhang, L. Xu, Z. Yan, and S. Wu, "A Glove-based poi type embedding model for extracting and identifying urban functional regions," *ISPRS International Journal of Geo-Information*, vol. 10, no. 6, pp. 372, 2021.

[18] P. Wei, Y. Ke, and C. K. Goh, "Feature analysis of marginalized stacked denoising autoenconder for unsupervised domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1321–1334, 2019.

[19] O. Qasim and Z. Algamal, "Feature selection using different transfer functions for binary bat algorithm," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 5, no. 4, pp. 697–706, 2020.

[20] C. P. Igiri, Y. Singh, and D. Bhargava, "An improved African buffalo optimization algorithm using chaotic map and chaotic-levy flight," *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 4570–4576, 2018.