

# User identification based on short text using recurrent deep learning

Huda Hallawi<sup>2</sup>, Huda Ragheb Kadhim<sup>1</sup>, Zahraa Najm Abdullah<sup>1</sup>,  
Noor D. AL-Shakarchy<sup>1</sup>, Dhamyaa A. Nasrawi<sup>1</sup>

<sup>1</sup>Department of Computer Science, College of Computers Science & Information Technology, University of Kerbala, Karbala, Iraq

<sup>2</sup>Department of Information Technology, College of Computers Science & Information Technology,  
University of Kerbala, Karbala, Iraq

## Article Info

### Article history:

Received Jul 8, 2022

Revised Feb 8, 2023

Accepted Mar 10, 2023

### Keywords:

Embedding layer

Identification

Long sort term memory

Tokenization

## ABSTRACT

Technological development is a revolutionary process by this time, it is mainly depending on electronic applications in our daily routines like (business management, banking, financial transfers, health, and other essential traits of life). Identification or approving identity is one of the complicated issues within online electronic applications. Person's writing style can be employed as an identifying biological characteristic in order to recognize the identity. This paper presents a new way for identifying a person in a social media group using comments and based on the Deep Neural Network. The text samples are short text comments collected from Telegram group in Arabic language (Iraqi dialect). The proposed model is able to extract the person's writing style features in group comments based on pre-saved dataset. The analysis of this information and features forms the identification decision. This model exhibits a range of prolific and favorable results, the accuracy that comes with the proposed system reach to 92.88% (+/-0.16%).

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Huda Ragheb Kadhim

Department of Computer Science, College of Computers Science & Information Technology,

University of Kerbala,

Al Wafa Street, Karbala, Iraq

Email: [huda.raghib@uokerbala.edu.iq](mailto:huda.raghib@uokerbala.edu.iq)

## 1. INTRODUCTION

User identification is an open research until now. It becomes more important with the rapid growth of Internet services which rises the value of the continuous recognition of clients that access public and private informatics resources. User identification can be useful to determine personal identification even after an authentication phase has been gone over. It would be able to handle the typing rhythms of free text, identified by users with no particular constraint.

Three main ways are often used for authentication and identification purpose: password, property authentication methods, and biometric characteristics. Three main ways are often used for authentication and identification purpose: password, property authentication methods, and biometric characteristics. It was found that keystroke governs an individual style of typing that considered as a biometric attribute for user identification [1]–[6]. Biometrics identification consists of a set of physiological ((like facial analysis, palm topology), or w,17teristics (such as keystrokes) that can allow identification of personal identity. The keystrokes is preferable since it comes low investments the only equipment required is just the normal keyboard, and high-performance software.

User identification can be classified into static and dynamic; Static identification is based on a structured or predefined text [5], [6], while the dynamic identification is based on free text entered by the user

in any application. The pre-stored text enables more reliable user identification in comparing with only depending on ID/password during early authentication phase. Additionally, it enables on time user identification that complement ID/password checks in suspicious cases. Free text is better to represent the real work environment and better to reveal the user behavioral characteristics [3], [4].

The main approaches for dynamic and static identification are not too variant from a recognition view. They can be classified into three main categories: statistical methods, probabilistic modeling methods, and machine learning methods. Although statistical methods are the first approach that used to deal with the recognition issue, but they are more convenient for all user identification aspects. The statistical indicators include a range of methods such as: average deviations, median deviations, standard deviations, and statistical t-criterion for similarity estimation [5], [7].

The probabilistic methods estimate the probability that a keystroke profile related to a client of a certain domain. There are a number of probabilistic methods that have been used in previous research such as: Bayesian methods, hidden Markov model, Gaussian density function, and weighted probability. The machine learning methods for user identification include a number of venerable algorithms such as decision tree, fuzzy logic, classification techniques, and evolutionary computation [4], [8]. However, the experimental results of previous research in this field are controversial. They come with deficiency of attained level of accuracy is unacceptable, or good performance is accomplished under very special conditions, which is really difficult to maintain in real application [3].

This paper proposed a type of user identification based keystroke that deals with storing the users' data then applying the proposed method which allow distinguishing between users using special identifier. The goal of this paper is to determine whether an instant message is sent by the real user indicated by the user's ID or not. In the next section, we give an overview of some of the related works that has been done on the user identification using machine learning and deep learning techniques. In section 3 the methodology and dataset are described. Section 4 describes the experiments and results. Finally, conclusion and future works explained in section 5.

## 2. RELATED WORKS

In this section, We focus on the multiple methods used in user identification were published during the past two decade using machine learning and deep learning techniques. Bolme *et al.* identify the person by using a weighted sum of similarity score for combining the the text and face steps such as term frequency inverse document frequency (TFIDF) used for information retrieval and text similarity computation and empirical Bayes geometric mean (EBGM) used for face recognition. They used dataset consists of multiple images and a pair of biographies for each person of 118 famous from the two site Yahoo Movies and Wikipedia, because the combining text classifier and face classifier, exposed 99% considerable enhancement in the classifre rate in compared with either method alone [9].

Goldstein-Stewart *et al.* identify the person by applied correlated corpus on samples from text and audio and classified the samples by four classifier of Weka workbench (Naïve-Bayes, J48, sequential minimal optimization (SMO), random forest (RF)) [10]. This wok implemented on groups containg samples of communication of 21 individuals in six genres across six topics. The accuracy of identify the perons with number cases: first, 71% when the samples of their communication are availabled across 6 genres, second, 81% when the samples of their communication are availabled for specific genres (train on 5 genres, test on one), third, 94% when the communicating is applied on a new topic [10].

Mohtasseb and Ahmed identify the author (person) by text collection, features extraction by converting very blog post to a features vector and finally applied support vector machine (SVM) as the classification algorithm [11]. They used text that collected from LiveJournal 80,000 blog posts that contained 565 authors(person) with 140 posts as a rate for any user. The overall number of words is 20,172,275. Many steps of preprocessing wrere done such as: removes images, videos, elicit text in tables, and delete empty posts, all that to produce group from 63,167 posts that represent purely dataset for this work. The classification percentageis high blooger for larger number posts is greater than 90% [11].

Layton *et al.* determined the author for a given tweet by using the source code author profile (SCAP) that is applied for the raw messages of tweet [12]. Their dataset is (tweet dataset) which is a collection of 14,000 twitter users and for each of those user up to a maximum of 200 hundred tweets. The accuracy of this work > 70% for determining the author of tweet by the SCAP method [12].

Iqbal *et al.* used the National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE) framework for e-mail authorship identification by building two models from a two-class classification that is produced a one e-mails of the potential suspect person and the other from building a big e-mail dataset from multiple person called universal background model by applied a vector of feature extraction for each e-mail, applied number of classification techniques (Adaboost.M1, discriminative multinomial Naive Bayes (DMNB), and Bayesian network) and finnally applied number of regression techniques (linear

regression, SVM with sequential minimum optimization (SMO), and SVM with radial basis function (RBF) kernel [13]. Their dataset was build from a big e-mail dataset for diverse persons entitled as universal background model. As well as used normalization to reduce all the numerical values as preprocessing. The proposed method verified the writer of a virulent e-mail with the following results: an average EER= 15-20% , minDCF =0.0671 (with 10-fold cross validation technique) [13].

Poignant *et al.* identified the user in vedio based on the text that is taken from part of a newscast (sush as person name, or a position) which is done by text detection [14]. The process of text detection is performed by text features including : texture, color, contrast, geometry, and temporal information. While the recognition of text is implemented using free software called Google Tesseract. The dataset are 59 videos from the France 2 French TV channel that contains a group of a broadcast news. In this work, 91% recall are represented the performance of the detection text system in the frame that is extracted by segmentation of videos. The person is present in 96.8% of cases,when a name is written single on a box text [14].

Poignant *et al.* identified the person by text detection (coarse detection, refine detection) and temporal tracking (to correct the text boxes) are performed, adapting images of to an optical character recognition (OCR) [15]. It combines multiple transcriptions of the same text box and a person recognition is revealed by OCR and audio information. The dataset is a combination of 59 videos of France 2 TV; it contains 29,166 key frames that are extracted automatically and focused on person name and person in the screen. From the 29k images, 4,414 frames contain 9,256 text boxes. Their result was F- measure is 77.3% of this work [15].

Al-Maadeed performed the identification of text-dependent writers for arabic handwriting by features extraction are taken from writers' word images and a k-nearest neighbor classifier are performed for recognition operations of Arabic text writers [16]. A new database are built based on offline arabic handwriting text that are collected from 100 writers using the same pen for identification Arabic writer, by this time the dataset is being produced at Qatar University. The normalization operations was performed to the word in number of steps:Page scanning, document segmentation, removing background, and edge detection. It is shown that the recognition rate of top ten writers is exceeding 90% of certain words [16].

Macleod and Grant, identified the writer of short messages (one or more) by using automate process from forensic linguistics [17]. Firstly,identification of distinctive textual features were extracted from short messages for the development of a taxonomy. Secondly, compute the 'distance' between messages that contain instances of these feature types. Dataset was created using data from United Kingdom (UK) online groups in Twitter of around 18,500 tweets with 12 words long for a tweet [17].

Ragel *et al.* Identify the author by Short Message Service(SMS) messages by used unigram method for train data, used around ten SMSes piled together could yield a good data for test [18]. Finally, Cosine similarity is used that is given best accuracy from Euclidean distance in comparing between two vectors. Therefore, the author is identified as one having the most similarity. The National University of Singapore SMS Corpus (NUS) SMS message corpus contains more than 50,000 messages written in English from multiple cultures in Asia. The pre-processing steps involve: removed all SMSes that < 50 SMSes, removed from the database all types of messages, remove the repeated messages. In this work, Cosine distances gives an accuracy =40% for only one SMS 90% for the large number of SMSes stacked [18].

Brocardo *et al.* identified the authorship for short online messages by combines supervised learning and n-gram (with size 3-5 characters) analysis to comparing sample writing of an person against the profile associated with the identity by that individual at login time [19]. The real-life e-mail dataset from Enron contains more than 200,000 messages from about 150 users, the average number of words per e-mail is 200. The preprocessing steps to the data: removed all duplicate e-mails for each user's folder, used avaMail application programmable interface (API) to parse each e-mail and extract the body of the message. The result of this work is energy efficiency ratio (EER) =14.35% [19].

Nirkhi *et al.* identify the authors of short message by using SVM classifier and word Uni-gram [20]. The dataset is are used C50 corpus(contain 50 authors with 50 documents each) and Enron corpus (contains 619,446 messages for each 158 users). The best accuracy is 93.3% for 5 author of Enron corpus and 100% for 7 or 25 author [20].

Nirkhi *et al.* used two steps for authorship identification of online messages that are helped to investigator from the results visualizion by firstly hierarchical clustering for the cluster formation operation and secondly multidimensional scaling for visualization of clusters [21]. The dataset is Enron corpus contains 619,446 messages belonging to 158 users. While the accuracy of this work from 70-90% [21].

Buza used dynamic time warping (DTW) and Nearest neighbour regression with error correction (EckNN) for identifying the person by keystroke dynamics. The dataset is contained 500 typing sessions from 12 users. The accuracy of EckNN is 86.6% [2]. Ye and Zhang are find out the influential users by two steps; firstly, directed graph is used for modelling short message. Secondly, scoring policy is used for evaluating users. The data set is contaned greater than 20 millions short message of china mobile communication corporation [22].

Alsultan *et al.* are examined user authentication by used decision tree, SVMs and Ant Colony Optimization (ACO) for classify extracted non-conventional keystroke features from the users' typing [7]. The dataset are extracted from the Guardian newspaper with 8 typing tasks from the user and include 170 characters for each task [7]. Akimushkin *et al.* are used a radial basis function network that is one of the supervised learning method for classify correctly 68 from 80 text by 8 authors with author matching success rate is equal 85% [23]. Kim *et al.* used a user-adaptive feature extraction algorithm for enhancing keystroke dynamics of user authentication [6]. The methods are used 150 user with 13,000 keystrokes for each user in Korean and English languages and are gave 0.44% of EER [6].

Vijayakumar and Fuad identify short-text authors by the combinations of NLP techniques (stemming, lemmatization, stop words removal) for adding context to the reviews and machine learning algorithms SVM, multinomial Naïve Bayes (MNB), maximum entropy (ME) to classify the author of a review [8]. Vijayakumar and Fuad used yelp review dataset that consist of restaurant and hotel reviews. And dataset contains 6,685,900 unique reviews, written by 1,637,138 different authors, across 192,609 businesses. The best accuracy of author identification is 90.5% that is resulted from the combination of SVM classifier with unigram and bigram vectorization model, and lemmatization [8].

Salomatin *et al.* identify and classify the user based on browser fingerprints by using K-nearest neighbors classification algorithm and probabilistic-statistical method [4]. The data set are used 9 sets of browser attributes for 7 users. The experimental results are excited on real data and calculate the average time for attributes of browser fingerprints of users. *Selection of significant attributes method* was implemented which ignore features with a high average duration of their calculation (features >100 ms are not considered) [4].

Recent trends in artificial intelligence are deep learning. Deep learning is a part of machine learning that related to artificial neural networks. As the neural network is aimed to imitate the human brain then deep learning is also considered as a kind of imitate for human brain. Deep learning has been used in many applications, such as biometric system [24], abusive comment identification [25], skin cancers detection [26], automatic text generation [27], [28], healthcare [29], image recognition [30], and video [31].

In our work, deep neural network used to identify a person in a social media groups comment. Long-short term memory (LSTM) was used with multiple hidden layers. The dataset collected from Telegram group with Arabic language (Iraqi dialect) for 14 members. More details were explained in next section.

### 3. METHOD

The present proposed system is mainly setup by employing LSTM models for person identification via its group comments. A new LSTM architecture was established to take out the experiments on the Telegram group Arabic comments dataset (Iraqi dialect). The three stages of the proposed system was explained in Figure 1.

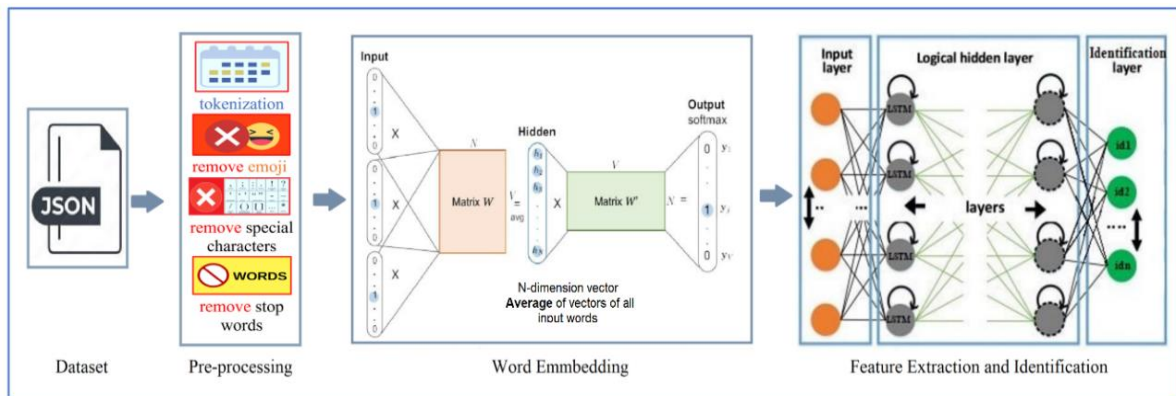


Figure 1. Proposed system block diagram

It can be notably recognized in Figure 1 the main stages of pre-processing which comprising word embedding, feature extraction and identification stages. The main goal of the proposed architecture is to predicting a person's identity from her/his comments writing style; which it in Arabic. This section will be focused on the dataset that applied in the experiments and all model details.

### 3.1. DATASET

Recently, Telegram brought an easy way to save conversations and exports some (or all) of chats, including photos and other media they contain. This paper uses this feature by used Export Telegram data tool; as a result; all data accessible can be get offline in JavaScript object notation (JSON) format. The group that used here contains 14 members which implemented in proposed model as row data. The created dataset contains (253,564) comments after preprocessing.

### 3.2. LSTM model

Long-short term memory (LSTM) is a one of recurrent neural networks. In terms of memory, LSTM is better than traditional recurrent neural networks, as it holds a good result over memorizing certain patterns. LSTM may contain multiple hidden layers. As it passes through every layer, the relevant information is kept, and all the irrelevant information is discarded in every single cell. As well as the input layer the proposed model comprises three additional layers; preprocessing, word embedding, LSTM, and output layers (feature extraction and identification) respectively. The recap description of the proposed architecture is illustrated in Table 1 that depicts all related information comprising the layers, output style, parameters values (weights) for every layer, and ultimately the overall number of parameters (weights) of the proposed system.

Table 1. The recap description of the proposed system

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 100)	0
embedding_1 (Embedding)	(None, 100, 300)	32,862,900
lstm_1 (LSTM)	(None, 10)	12,440
activation (Dense)	(None, 14)	154
Total params: 32,875,494		
Trainable params: 12,594		
Non-trainable params: 32,862,900		

*Pre-processing stage:* this stage implements all preprocessing on the input data to be suitable form on the proposed model. It consists of tokenization, remove emoji, remove special characters, remove non Arabic symbols, remove Arabic stop words (a group of unimportant words that do not affect the meaning or context of the comment and which in turn do not affect the person's writing style), remove words less than 2, remove unwanted words that meaningless, in addition to a number of overly frequent words such as "هههههه" and so on. In tokenization step we employed "Word Tokenization" so that the raw text splitting into small chunks of words called tokens. The 'space' is used to perform the word tokenization. During tokenization process, spaces, punctuations are ignored and omitted from the final list of tokens. The Tokenizer function is used to split the sentence into tokens and *texts\_to\_sequences* function converts word to integer number. Encoding the comments is an important stage in the deep neural network model since this model required input data to be an integer. Thus, the embedding layer takes a sequence of numbers as an input. Since sequence has a different length, pad all sequence to the given maximum length which in this work equals 100.

*Word embedding stage:* In word embedding stage all words are represented as vectors. It exerts to transform the high dimensional feature extent of words to low dimensional feature vectors by preserving the contextual resemblance in the corpus. Briefly, the Word2Vec, a neural network model for word embedding in a text corpus, is used for Word representations in Vector Space. Such models are working using context which means if you need to know the embedding, you should look at nearby words; if a group of words is usually discovered closed by the same words, then they will be assigned with similar embedding. Word2Vec model is mainly comprised two preprocessing techniques: Continuous bag of words (CBOW) and skip-gram. Both of the these mechanisms are fundamentally superficial neural networks that converts word(s) to the target variable that is also a word(s). The given mechanisms learn the weights that work as word vector representations, they can be applied to perform word embedding by word2vec. In this work Arabic word to vector (*AraVec*) (an open-source) is used for pre-trained word2vec embedding, which trained on a big Arabic corpus of text that contains more than 1,169,075,128 tokens and implemented with two techniques: Continuous bag of words (CBOW), the 300-dimensional Twitter-CBOW was chosen; and skip-gram (SG), 300-dimensional Twitter Skip-gram version 3 was used.

*Feature extraction and identification (LSTM and output layers):* Proposed LSTM model performs its functions based on memorizing important information. So that the classification process achieved by trained a model on multiple word not as separate inputs with no actual meaning as a sentence, and predicting the class according to statistics; instead a LSTM model trained on multiple word string each input depend on previous

input to produce real meaning as a sentence, and predicting the class according to meaning. That means, the model will be able to find out the actual meaning in input string and will give the most accurate output class.

#### 4. RESULTS AND DISCUSSION

Primarily, the training and evaluation processes performance is computed with two vital metrics the accuracy and loss functions. The proposed system trained with two Word2Vec techniques: Continuous bag of words (CBOW) and skip-gram (SG). A direct approach to understand the learning behavior of the given system on a certain data is by analyzing the training and a validation data for every epoch then plotting the resultant curve. The accuracy and loss function for training the proposed model are primarily based on CBOW Word2Vec embedding which is presented in Figure 2(a). Whereas, the accuracy and loss function of tested model based on SG Word2Vec embedding is presented in Figure 2(b).

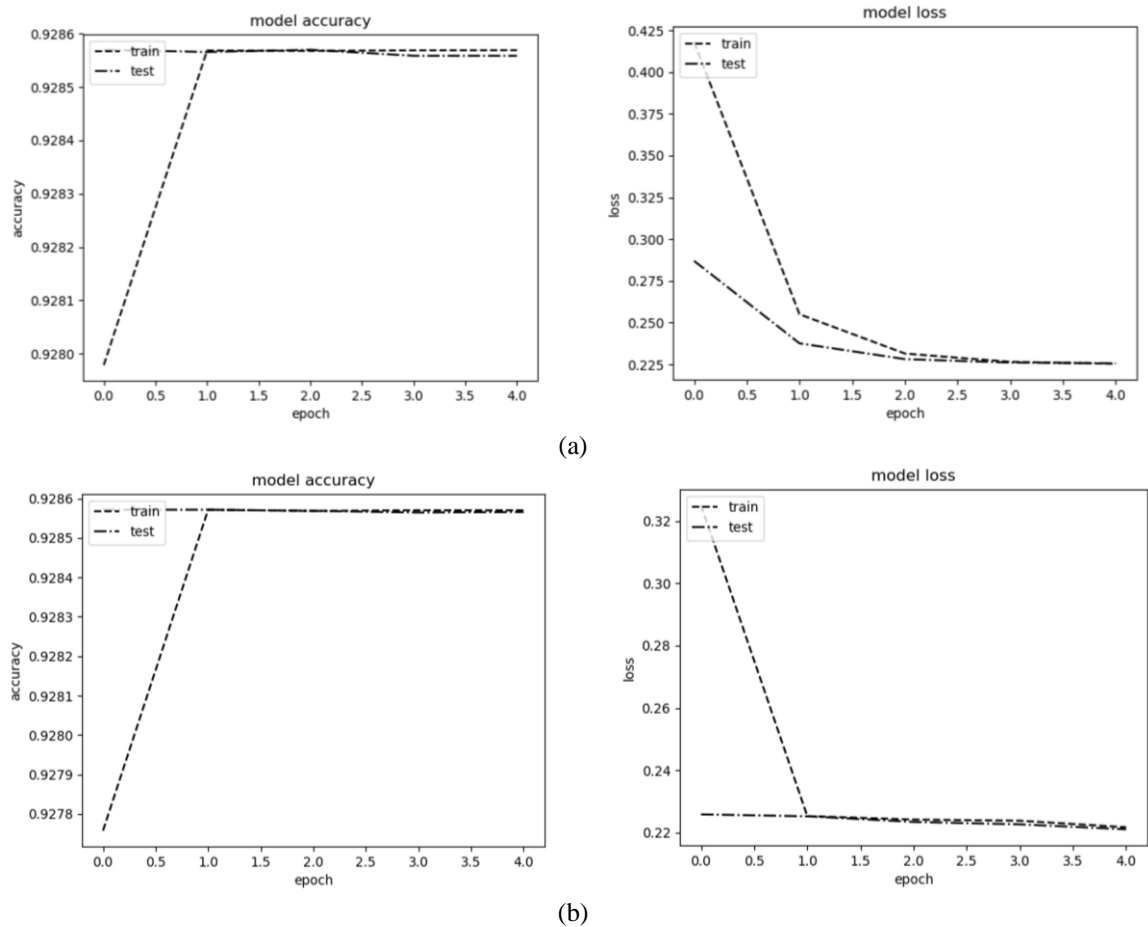


Figure 2. Accuracy and loss function of proposed model: (a) Using CBOW Word2Vec embedding  
(b) Using SG Word2Vec embedding

##### 4.1. Evaluation of the proposed architecture

Given that deep learning patterns are almost stochastic which means the same model is appropriate for the same data on different time. Although, it might come with different expectations therefore have variant general skill. In this work, the evaluation process is created by the procedure to estimating model skill (Controlling for model variance); which provides disparate outcomes of the same model that training on variable dataset using K-fold cross-validation. On the other hand, there is another procedure dedicated to predicting a stochastic model's skill with the aim of controlling for model stability on which variant outcomes are produced from the training of identical data. The evaluation test of a non-stochastic data is by repeated for a number of times then computing the estimating mean model skill.

## 4.2. K-folds cross validation

This procedure is based on estimating the ability of a machine learning pattern on hidden data; it is carried by assessing the machine learning patterns on pre-specified resampling data set using a unique identifier which is called k. Apparently, the restricted sample is used to assess the overall model expectation on hidden data through the training process. The performance of 10-fold cross-validation on training and testing processes is illustrated in Figure 3.

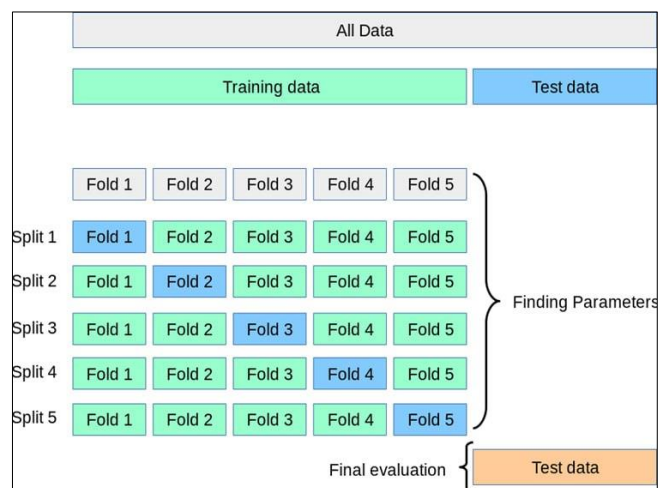


Figure 3. Cross-validation procedure with 5-folds

It can be seen that proposed models come with  $k = 10$ , for any value of  $K$ , which will divide dataset into  $Tr$  (80%) +  $Va$  (10%) = 90% and  $Te = 10\%$ . By registration of the testing performance within the per-mentioned metric (adopted accuracy). Finally, the overall performance is counted to illustrate the ultimate result; all experimental outcomes within 10-Fold CV implementation is shown in Table 2.

Table 2. 10-Fold CV for proposed model

NO OF FOLD	ACCURACY OF EACH FOLD	MODEL ACCURACY
1	92.84%	92.88% (+/- 0.16%)
2	92.45%	
3	92.83%	
4	92.88%	
5	92.91%	
6	92.98%	
7	92.99%	
8	92.96%	
9	93.01%	
10	92.98%	

## 5. CONCLUSION

Generally, deep neural networks are investigated as the best techniques for identification systems as they succeed to obtain high accuracy. Using deep neural networks shows superior outcomes in compared traditional techniques for both accuracy and loss functions. This study has come four outcomes. Firstly, the proposed model is succeeded to identify the individuals through their group comments. Secondly, the proposed model saved the time and the efforts that are required for the annoying pre-processing of the short text comment by providing the possibility of dealing with raw data directly in the same model. The third conclusion is using the combination of LSTM integrated with a rectified linear unit (ReLU) activation function can provide a pattern of feature map for each person writing style, which can be considered as the core of identification decision. The fourth conclusion is the Word2Vec Embedding method provides a successful approach for learning an Arabic word and their relative meanings from a corpus of text and representing the word as a dense vector.

## ACKNOWLEDGEMENTS

We would like to express our deep gratitude to our college (Computer Science & Information Technology) at University of Kerbala for their continued encouragement and support. We would like to thank the technical staff in our college for providing basic requirements of scientific activities.

## REFERENCES




- [1] E. A. Kochegurova and Y. A. Martynova, "Aspects of continuous user identification based on free texts and hidden monitoring," *Programming and Computer Software*, vol. 46, no. 1, pp. 12–24, 2020, doi: 10.1134/S036176882001003X.
- [2] K. Buza, "Person identification based on keystroke dynamics: Demo and open challenge," *CEUR Workshop Proceedings*, vol. 1612, no. 6, pp. 161–168, 2016, [Online]. Available: <http://www.biointelligence.huhttp://ceur-ws.org>.
- [3] J. Poignant, F. Thollard, G. Quenot, and L. Besacier, "Keystroke analysis of free text," *Keystroke analysis of free text*, vol. 8, no. 3, pp. 312–347, 2005, doi: 10.1145/1085126.1085129.
- [4] A. A. Salomatina, A. Y. Iskhakov, and A. O. Iskhakova, "Web user identification based on browser fingerprints using machine learning methods," *IFAC-PapersOnLine*, vol. 54, no. 13, pp. 582–587, 2021, doi: 10.1016/j.ifacol.2021.10.512.
- [5] P. H. Pisani and A. C. Lorena, "A systematic review on keystroke dynamics," *Journal of the Brazilian Computer Society*, vol. 19, no. 4, pp. 573–587, 2013, doi: 10.1007/s13173-013-0117-7.
- [6] J. Kim, H. Kim, and P. Kang, "Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection," *Applied Soft Computing Journal*, vol. 62, pp. 1077–1087, 2018, doi: 10.1016/j.asoc.2017.09.045.
- [7] A. Alsultan, K. Warwick, and H. Wei, "Non-conventional keystroke dynamics for user authentication," *Pattern Recognition Letters*, vol. 89, pp. 53–59, 2017, doi: 10.1016/j.patrec.2017.02.010.
- [8] B. Vijayakumar and M. M. M. Fuad, "A new method to identify short-text authors using combinations of machine learning and natural language processing techniques," *Procedia Computer Science*, vol. 159, pp. 428–436, 2019, doi: 10.1016/j.procs.2019.09.197.
- [9] D. S. Bolme, J. R. Beveridge, and A. E. Howe, "Person identification using text and image data," *IEEE Conference on Biometrics: Theory, Applications and Systems, BTAS'07*, 2007, doi: 10.1109/BTAS.2007.4401934.
- [10] J. Goldstein-Stewart, R. Winder, and R. E. Sabin, "Person identification from text and speech genre samples," *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pp. 336–344, 2009, doi: 10.3115/1609067.1609104.
- [11] H. Mohtasseb and A. Ahmed, "More blogging features for author identification," *The 2009 International Conference on Computer Engineering and Applications*, pp. 461–466, 2009, [Online]. Available: <http://eprints.lincoln.ac.uk/1862/>.
- [12] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for Twitter in 140 characters or less," *Proceedings - 2nd Cybercrime and Trustworthy Computing Workshop, CTC 2010*, pp. 1–8, 2010, doi: 10.1109/CTC.2010.17.
- [13] F. Iqbal, L. A. Khan, B. C. M. Fung, and M. Debbabi, "E-mail authorship verification for forensic investigation," *Proceedings of the ACM Symposium on Applied Computing*, pp. 1591–1598, 2010, doi: 10.1145/1774088.1774428.
- [14] J. Poignant, F. Thollard, G. Quénot, and L. Besacier, "Text detection and recognition for person identification in videos," *Proceedings - International Workshop on Content-Based Multimedia Indexing*, pp. 245–248, 2011, doi: 10.1109/CBMI.2011.5972553.
- [15] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From text detection in videos to person identification," *Proceedings - IEEE International Conference on Multimedia and Expo*, pp. 854–859, 2012, doi: 10.1109/ICME.2012.119.
- [16] S. Al-Maadeed, "Text-dependent writer identification for arabic handwriting," *Journal of Electrical and Computer Engineering*, 2012, doi: 10.1155/2012/794106.
- [17] N. MacLeod and T. Grant, "Whose Tweet? Authorship analysis of micro-blogs and other short-form messages," *Proceedings of the International Association of Forensic Linguists' 10th Biennial Conference*, vol. 2, no. July, p. 224, 2012.
- [18] R. Ragel, P. Herath, and U. Senanayake, "Authorship detection of SMS messages using unigrams," *2013 IEEE 8th International Conference on Industrial and Information Systems, ICIIS 2013 - Conference Proceedings*, pp. 387–392, 2013, doi: 10.1109/ICIInfS.2013.6732015.
- [19] M. Luiz Brocardo, I. Traore, S. Saad, and I. Woungang, "Verifying online user identity using stylometric analysis for short messages," *Journal of Networks*, vol. 9, no. 12, 2014, doi: 10.4304/jnw.9.12.3347-3355.
- [20] S. M. Nirkhi, R. Dharaskar, and V. Thakare, "Authorship identification using generalized features and analysis of computational method," *Transactions on Machine Learning and Artificial Intelligence*, 2015, doi: 10.14738/tmlai.32.1064.
- [21] S. Nirkhi, R. V. Dharaskar, and V. M. Thakare, "Authorship verification of online messages for forensic investigation," *Physics Procedia*, vol. 78, pp. 640–645, 2016, doi: 10.1016/j.procs.2016.02.111.
- [22] Z. Ye and P. Zhang, "Identification of seed users via short messages based on Hadoop," *Proceedings of the 2016 International Conference on Education, Management, Computer and Society*, vol. 37, 2016, doi: 10.2991/emcs-16.2016.456.
- [23] C. Akimushkin, D. R. Amancio, and O. N. Oliveira, "Text authorship identified using the dynamics of word co-occurrence networks," *PLoS ONE*, vol. 12, no. 1, 2017, doi: 10.1371/journal.pone.0170527.
- [24] C. Medjahed, A. Rahmoun, C. Charrier, and F. Mezzoudj, "A deep learning-based multimodal biometric system using score fusion," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 65–80, 2022, doi: 10.11591/ijai.v11.i1.pp65-80.
- [25] T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, "Abusive comment identification on Indonesian social media data using hybrid deep learning," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 895–904, 2022, doi: 10.11591/ijai.v11.i3.pp895-904.
- [26] H. M. Ahmed and M. Y. Kashmola, "A proposed architecture for convolutional neural networks to detect skin cancers," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 485–493, 2022, doi: 10.11591/ijai.v11.i2.pp485-493.
- [27] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2515–2528, 2022, doi: 10.1016/j.jksuci.2020.04.001.
- [28] S. Zhou, "Research on the application of deep learning in text generation," *Journal of Physics: Conference Series*, vol. 1693, no. 1, 2020, doi: 10.1088/1742-6596/1693/1/012060.
- [29] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019, doi: 10.1038/s41591-018-0316-z.
- [30] W. Mo, X. Luo, Y. Zhong, and W. Jiang, "Image recognition using convolutional neural network combined with ensemble learning algorithm," *Journal of Physics: Conference Series*, vol. 1237, no. 2, 2019, doi: 10.1088/1742-6596/1237/2/022026.
- [31] N. D. Al-Shakarchy and I. H. Ali, "Detecting abnormal movement of driver's head based on spatial-temporal features of video using






deep neural network DNN," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 344–352, 2020, doi: 10.11591/ijeecs.v19.i1.pp344-352.

## BIOGRAPHIES OF AUTHORS






**Huda Hallawi**    received the Ph.d degree in "Resource Allocation in Cloud Computing Using Genetic Algorithm" from Cranfield University ,UK in 2017. She got B.Sc. and M.Sc in Computer Engineering and Information Technology from University of Technology, Baghdad Iraq in 2005 and 2009 respectively. Her research interests are Evolutionary Algorithms, Cloud Computing, System Modeling and Simulation, Artificial intelligent and deep neural networks. She is currently working as a lecturer in faculty of Computer science and Information Technology College, Kerbala University, Karbala, Iraq. She can be contacted at email: huda.f@uokerbala.edu.iq.






**Huda Ragheb Kadhim**    holds a master's degree in computer science from University of Babylon, Iraq 2020. She also received her B. Sc. From University of Kerbala in 2010. She is also an Assistant Lecturer at computer science department, College of Computer Science & Information Technology, University of Kerbala. She can be contacted at email: huda.raghib@uokerbala.edu.iq.






**Zahraa Najm Abdullah**    received the M.Sc. degree in computer science from the College of Information Technology, University of Babylon, Iraq with the Dissertation "Enhancement of Association Rules Interpretability by Combining Generalization and Graph-based Visualization" since 2016. Her research interests are in Data mining, Modular Neural Networks and Pattern Recognition approaches. She is currently a lecturer in the Department of computer science, College of Computer Science and Information Technology, University of Kerbala, Iraq. She can be contacted at email: saad@um.edu.my and zahraa.najm@uokerbala.edu.iq.



**Noor D. Al-Shakarchy**    Faculty member of Computer science and Information Technology College, Kerbala University, Karbala, Iraq. She got Ph.D. From The University of Babylon. B.Sc. and M.Sc at University of Technology, Computer science and information systems/information systems Department, Bagdad, Iraq in 2000 and 2003 respectively. Her research interests include computer vision, pattern recognition, Information Security, Artificial intelligent and deep neural networks. She can be contacted at email: noor.d@uokerbala.edu.iq.



**Dr. Dhamyaa A. Nasrawi**    hold B.Sc., M.Sc., Ph.D. degrees in (Computer Science) from Babylon University, IRAQ in 1996, 2003, and 2014 respectively. she is an Asst. Prof. at College of Computer Science and Information Technology, University of Kerbala. Her research interests include Text Steganography, Evolutionary Algorithms, Graph theory, Information Retrieval and search engine, Data Mining and Machine Learning, Natural Language Processing, and text editors. She can be contacted at email: dh.alnasrawy@uokerbala.edu.iq.