The CSI Journal on
Computer Science and Engineering
Vol. 17, No. 2, 2020
Pages 34-45
Regular Paper

Authorship identification from unstructured texts: A stylometric approach

Reyhaneh Ameri and Hamid Beigy

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Abstract

With the increasing use of the Internet, a considerable volume of texts is exchanged in cyberspace in which individuals can hide their true identities. Abuses that may occur in online communities due to unknown identities reduce the confidence of cyberspace and create many challenges. Hence the importance of maintaining the security of the space by controlling the user-generated content and identifying the authors of documents increases day by day. Author Identification is a method of finding the author of the anonymous document. Since there would not be any standard corpus for the Persian language, we created a standard Persian corpus for the authorship analysis applications in this language. In this paper, we propose an approach based on modeling the authors' writing style with the extracted stylometric features from their writing documents. Performance of author identification is also improved by applying pre-processing of the documents and reducing the dimensionality of the feature space by selecting the features with higher discriminative capability. The proposed approach is evaluated in terms of performance measures in data mining by designing and conducting experiments on the benchmark datasets of standard documents in Persian and English languages. The effect of different factors on the accuracy of the author's identification has also been investigated by designing and performing experiments. The results of these experiments have shown that the proposed method has a higher performance than the related state-of-the-art methods.

Keywords: Authorship identification, feature selection, classification method, writing styles, stylometric.

1. Introduction

The rapid changes in human life in the 21st century are a product of fast information transfer; that's why this era is called the age of the explosion of information. Fast and easy access to information has made the results of science progress quickly accessible to everyone, and in addition to scientists' life and ordinary people's life will also be affected. Information exchange techniques such as websites and social networks are being developed and widely used. As a result, millions of documents are produced every day and made available on the Internet [1].

Anonymity, which is one of the most prominent features of cyberspace, provides more possibility to commit cybercrimes and makes the prosecution of crime and criminals more difficult in cyberspace [2]. Cybercriminals provide deceptive information about their sex, place, age, and nationality and deceive people through this. Hence the use of new methods for tracking the identity of criminals is necessary, and therefore the analysis of authors has been obtained increasing importance. The authorship analysis applications have also grown in civil law, criminal law, forensic analysis, security, electronic commerce, marketing, election campaign, plagiarism detection, annd literary works [3,1,4,5]. The authorship analysis is closely related to the areas of information retrieval, natural language processing, and text mining.

The authorship analysis task can be classified into authorship profiling, similarity detection, and authorship identification [6,7]. Some demographics such as the gender and age of the author of a document are determined in the authorship profiling task [8]. The similarity detection task determines the degree of similarity between a pair of texts [9]. In the authorship identification task, an author has identified from a finite set of authors that their documents are available [10]. Also, the authorship verification task that determines whether a given text was written by a certain author A or not is a specialized task of the authorship identification task [6].

In most approaches in this context, this problem is formulated as a typical classification problem in which its

performance is dependent on discriminant features that model the writing style of authors. Therefore, selecting the appropriate features for better modeling of the author's writing style and an appropriate way to predict the author of a given document is important in this task [3]. In several works, some similarity/distance measures were defined for author identification. In addition, the combination of classifiers and predefined measures such as similarity were used for author prediction in few proposed approaches. In most of the previous approaches, there were not comprehensive discriminative features for the modeling author styles. Moreover, the document pre-processing steps have not been applied in some of the previous works. Hence some of them had low accuracy in authorship identification.

A few research papers proposed the approach for the author identification in the Persian language. One of the difficulties of the authorship identification in the Persian language is the unavailability of a good corpus for implementing and evaluating the problem. Also, natural language processing tools are more limited and less accurate for the Persian language than the English language processing tools. Moreover, to evaluate the proposed approach on multiple languages, the standard Persian corpus has been generated on Persian blogs.

This paper's main objective is to develop an efficient approach to identify the author of a given document in English and Persian languages. The proposed approach predicts authors using some classification methods. Classification methods such as KNN, SVM, LDA, and Naive Bayes are applied, and their performances are evaluated by conducting some experiments. The author's style patterns are extracted by stylometric features that belong to lexical, structural, syntactic, and semantic feature categories. Then discriminant features are selected, and redundant features are removed from the feature vectors. Also, the accuracy of author identification improves with document pre-processing. The performance of the proposed approach is evaluated using some experiments conducted on two corpora in Persian and English languages. The results of the computer experiments have shown that the proposed method obtains a higher performance compared to the state-of-the-art methods of author identification. Another objective of this paper is constructing a standard Persian corpus that can serve as the standard corpus in the authorship analysis applications in this language.

The novelty of this paper is the generation of a standard corpus in Persian and the study of the influential factors in author identification. Several research questions related to the factors influencing the author's identification are listed below, and experiments are designed to answer them. What effect do different feature sets and different authors have on identification accuracy? What is the effect of the number of authors and different sizes of training data on the author's identification? Also, how the use of different classifiers affects the identification process?

The rest of this paper is organized as follows. Section 2 reviews the literature and the previous works of author identification. Section 3 describes our proposed approach for authorship identification, and Section 4 presents experiments and analyses the practical results. Finally, the paper is concluded in Section 5.

2. Related Work

In this section, we will briefly review the related work about author identification. The first attempts for author identification were made in the late 18th century. Mendenhall was one of the first ones who made a significant attempt to identify the author of the plays of the world's famous English writer Shakespeare in 1887 [1]. Mendenhall used the lexical features in his research [6]. Also, in the early and most authorship identification researches, the lexical features were used as stylometric features. The lexical features include word frequencies [3,11,12,13,14], complexity of words [12], function of word frequencies [13,3] and vocabulary richness [15]. The concept of vocabulary richness measures is based on the reality that every author uses a specific individual vocabulary. Therefore they are about the number of different words used in their documents. N-Grams, a contiguous sequence of N words from a given set of documents, is also used in some research [16,15,17]. Languages such as English and Persian have many synonym/antonym pairs, and the authors who have a good knowledge of the language use complex words from these pairs [12].

The structural features contain measures about the character count, word length, sentence length [3,15,18], and phrase count [12]. Some researchers use syntactic features to identify authors. To extract syntactic features, we need to use natural language processing tools for extracting part-of speeches from documents. Halvani et al. proposed the method for authorship verification that generates the expandable and scalable model for different languages, genres, and topics. Furthermore, natural language processing techniques were not used in this method that had lowered the running time [6]. The frequency distribution of types of part-of-speeches was used as syntactic features in some approaches [3,15,18]. Punctuation counts only from all kinds of part-of-speeches are applied as structural features [12,13,15]. In this algorithm that is a compression technique, after extracting the parts-of-speeches, the probability of authorship was focused on the syntactic features. They demonstrated the great validity of syntactic features through achieving competitive performance in author identification with their feature set [20,21].

Zhang et al. used frequency distribution of types of tense, voice, and nonsubject stylistic words as the semantic feature set [3]. Taxonomies of various semantic functions of different lexical items could be considered as semantic features given by Argamon et al. [22]. The method for author identification in interactive communications proposed by Villar-Rodriguez et al. had been used as the feature selection algorithm for exploiting the essential feature sets [23]. NUS SMS Dataset [24] as actual SMS data with short-length textual contents had been used in this method.

Classifiers were used as prediction method in the most approaches introduced for author identification [12,3,14,15,16,46,21,20,23,26,17]. The performance of classification algorithms had been evaluated for web author identification on imbalanced datasets by Vorobeva et al. [25]. The reported experiment results of this research showed that the Random Forest algorithm performs better than the Support vector machine (SVM), Decision trees (DT), and Naive Bayes(NB). In addition, Al-Ayyoub et al. compared the performance of different classifier methods such as NB,

DT, and SVM and various feature selection techniques such as SubEval, CorrEval, ReliefEval, PCA, and InfoG for authorship identification of Arabic tweets [26].
In some works, the similarity/distance measures were defined for author prediction [19,13,18,27,28,6,34]. Also, classifiers with predefined similarity/distance measures were combined for author prediction in some approaches [29,30,31]. Castilo et al. compared the similarity between a given document with an unknown author against the documents with the known author by similarity measures as latent semantic analysis (LSA), Jaccard similarity, Euclidean distance, Chebyshev Distance and cosine similarity measures [13]. In addition, cosine similarity was used as a similarity measure in some previous works [27,28]. Ferilli translated a collection of documents to First-Order Logic descriptions and then constructed the model of authors from the result of clustering descriptions [32]. Also, in the approach presented by Li et al., classification and ranking models are used for author identification [33]. The literature review on author identification is summarized in Table 1.

Table 1. Summary of Literature Review on the author identification.

Reference	Year	Language	lexical	structural	syntactic	semantic	feature selection	data mining techniques
[44]	2013	Persian	✓	✓	✓	×	Genetic Algorithm	KNN and Delta
[19]	2014	English	×	×	✓	×	×	PPM algorithm
[3]	2014	English	✓	✓	✓	✓	PCA	SVM, LDA and KNN
[18]	2015	English	×	✓	✓	×	×	A unique formula called CUSUM method to compare feature vectors
[45]	2015	Persian	✓	×	✓	✓	Eliminating Highly Correlated Features	KNN, Delta, Neural Networks, Decision Tree and LDA classification methods
[16]	2015	English, Spanish, Dutch, and Greek	✓	×	✓	×	Merge all features into a single meta feature space	SVM
[14]	2015	English	✓	×	×	×	×	SVM
[28]	2015	English, Spanish, Dutch, and Greek	✓	×	✓	×	×	Combination of Cosine, Dice, and MinMax as similarity/distance measure
[30]	2015	English, Spanish, Dutch, and Greek	✓	✓	✓	×	Extra tree classifier and the SVM classifier	SVM and combination of Cosine, City block, and MinMax as similarity/distance measure
[15]	2015	English, Spanish, Dutch, and Greek	✓	✓	✓	×	×	Random Forest and SVM
[32]	2016	English, Greek, and Spanish	✓	✓	✓	×	Text pre-processing	translate sentences into a set of First-Order Logic descriptions and clustering these descriptions
[50]	2019	English	✓	✓	✓	×	×	C4.5, the fuzzy and the Ada boost classifiers
[48]	2020	Arabic	×	×	×	✓	×	Using an ontology as a semantic feature and the similarity methods
[49]	2020	Thai	✓	✓	✓	×	×	Probabilistic Nearest Neighbors Classification

3. Authorship identification

In this section, we present the proposed approach for author identification. By studying different authors' documents, we concluded that those authors unconsciously use particular patterns in their writing. Therefore, we can identify authors by modeling their writing styles using extracted features from the documents written by every author. In this paper, we assume that each author has a personal style in the writing document, and the number of authors is limited. We also use the vector space model to represent the writing styles of every author. In the rest of this section, we first describe the proposed approach and then describe different features used for author identification.

3.1. Overview of the research method

The proposed approach for author identification consists of five phases: document pre-processing, extracting stylometric features, selecting stylometric features, modeling the authors' writing style, and the author identification phases.

The goal of the first phase is document pre-processing for reducing the computational complexity of the identification algorithm and increasing the efficiency of author identification system and then extracting some useful features applicable for author identification. Document pre-processing includes removing stop-words and stemming of words. Stop-words are a group of words that have no content and no relevant/useful information for discrimination of documents. Prepositions that have high frequencies in documents belong to this category of words. In documents, the verbs with different tenses are often used and the names and attributes are used with different structures. Therefore, using words in various shapes does not affect the final discrimination because all content of this family of words has no significant difference for author identification. So stemmed form of words will be used as representative of the words of the same family in the vocabulary set. It is worth noting that, pre-processed documents have been used just for extraction of lexical features.

The second phase is extracting stylometric features from documents. Documents that are available for each author are analyzed as representative of her/his writing style. The authors style pattern can be extracted through stylometric features [34]. These features can be divided into four types: lexical, structural, syntactic and semantic features. We extract these features in English language documents by Stanford library tools [35,36] and Persian language documents by JHazm and FarsNet libraries [37,38]. Later in Section 3.2, we will explain the stylometric features in more detail.

The third phase is selecting the stylometric features. The feature vector is a high dimensional vector and has a damaging effect on the performance of the author identification. The reason is that all features do not have the equal capability for representing the author's writing style, and discarding irrelevant features increases the efficiency of the algorithm and also reduces the computational complexity of the algorithm. The main motivation for reducing dimensions of the feature vector is to improve the scalability and the efficiency of the classification method, reducing the running time and using the lower processing and memory resources. In the feature selection phase, we select a subset of the features that have a higher discriminative ability to identify authors and eliminate the less informative features for training [39]. The reason for choosing this type of dimension reduction is that the classification results can be interpreted easily and do not require much processing overhead [40]. The InfoGainAttributeEval method of Weka has been used for feature selection tasks that a good feature is a feature that reduces the most entropy [41].

In the fourth phase, we build a model for the writing style of the authors. For each author, a stylometric feature vector is constructed from averaging feature vectors of available documents and using the results of the previous stage and the selected features. Then a classifier is trained using the training data to identify the authors. This phase can be described as follows. There is a set of m authors $A = \{a_1, a_2, ..., a_m\}$, where each author a_i wrote a set of documents $D_i = \{d_1, d_2, ..., d_{S_i}\}$ and the training set, denoted by D, is the union of documents from these N authors, i.e. $D = \bigcup_{i=1}^{m} D_i$. In this model, each author a_i is represented by a feature vector $(f_1, f_2, ..., f_R)$, where R is the number of the selected features. Also, each feature of the author is calculated from averaging this feature on her/his available training phase documents. The output of this phase is a classifier that is able to identify the authors from the given documents.

In the fifth phase, a set of documents is available, and their authors should be identified. First, for the identification of the author of a document, pre-processing is performed to extract the feature vector. Then stylometric features that have been selected in previous phases are extracted from the document. Finally, by using a classification method and the writing style that we had in the previous step, the author of the document will be specified. The proposed approach to author identification can also be divided into two phases of training and testing that figure 1 demonstrates the steps related to each of these phases.

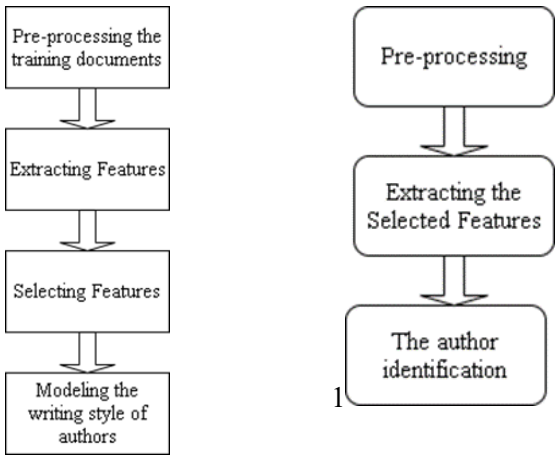


Figure 1 Steps of the proposed approach to author identification.

3.2. Stylometric features

All features used in the proposed approach can be partitioned into nine feature sets, denoted by $F_1, F_2, \ldots, F_9$ as shown in table 2. The details about what these feature sets are meant and how these feature sets are constructed are described in the rest of this subsection.

Table 2. Feature sets used in our approach.

Category	Feature set	Functionality
Lexical	F_1	Vocabulary richness rate
	F_2	Frequency rate of the most frequent words
Structural	F_{31}	Average length of sentences
	F_{32}	Percentage of sentences with longer than average
	F_{33}	Percentage of sentences with shorter than average
	F_{34}	Percentage of long sentences
	F_{35}	Percentage of short sentences
	F_{36}	Percentage of sentences with length between average and long sentences
	F_{37}	Percentage of sentences with length between short and average sentences
	F_{38}	maximum length of sentences
	F_{39}	minimum length of sentences
Syntactic	F_4	Frequency distribution rate of types of part-of-speeches
	F_5	Frequency distribution rate of bigram types of part-of-speeches
	F_6	Frequency distribution rate of types of part-of-speeches for initial sentence words
	F_7	Frequency distribution rate of types of part-of-speeches for final sentence words
Semantic	F_8	Frequency distribution rate of types of tense
	F_9	Frequency distribution rate of types of voice

3.2.1. Lexical level features

This category of features is the simplest and the most commonly used features set. In this category, a document is considered as a set of words. The main advantage of this category is applicability to any language and document corpus. Before extracting these features, pre-processing should be done. Features belonging to this feature set are vocabulary richness rate (VRR) and term frequency rate (TFR) of most frequent words. Vocabulary richness rate is the ratio between the number of words that occur only once and the total number of words in the document as given by equation (1).

$$F_1 = VRR = \frac{|\{x|\mathrm{TF}(x) = 1\}|}{W(D)}, \tag{1}$$

Where TF(x) is the term frequency of word x in the document and W(D) is the total number of words in document D. The set of the most frequent words in the corpus are represented as FW given by equation (2).

$$FW = \{x|\mathrm{TF}(x) > \alpha\}, \tag{2}$$

Where parameter α, pre-specified threshold, is a corpus-dependent parameter. The feature set F_2 is the frequency rate of the most frequent words (FW) as given by equation (3).

$$F_2(x) = TFR = \frac{\mathrm{TF}(x)}{W(D)} \qquad \forall x \in FW \tag{3}$$

3.2.2. Structural level features

Different authors often use sentences with different lengths to transfer the concept in their writings. Although the length of sentences is language dependent, while in one language this feature set could help to identify authors. The structural feature set, as denoted by F_3, includes nine features: the average length of sentences (F_{31}), the percentage of sentences that are longer than the average (F_{32}), the percentage of sentences that are shorter than the average (F_{33}), the percentage of sentences that their lengths are longer than a given maximum threshold (F_{34}), the percentage of sentences that their lengths are shorter than a given minimum threshold (F_{35}), the percentage of sentences that their lengths are between the average and a given maximum threshold (F_{36}), the percentage of sentences that their lengths are between a given minimum threshold and the average length of sentences (F_{37}), the maximum length of sentences (F_{38}) and the minimum length of sentences (F_{39}). The features in F_3 can be listed as follows:

$$F_3 = (F_{3_1}, F_{3_2}, F_{3_3}, F_{3_4}, F_{3_5}, F_{3_6}, F_{3_7}, F_{3_8}, F_{3_9}) \tag{4}$$

Also, we specify the minimum and the maximum thresholds for long and short sentences according to the corpus.

3.2.3. Syntactic level features

Every author writing a document uses a specific syntactic pattern. As a result, syntactic features are selected as the stylometric features. Syntactic feature sets include F_4, F_5, F_6 and F_7. Feature sets of F_4 and F_5 record frequency distribution rates of unigrams and bigrams types of part-of-speeches, respectively as given below.

$$F_4(x) = \sum_{n=1}^{N_D} \frac{\mathrm{TF}(x)}{W(S_n(D))} \tag{5}$$

$$F_5(x,y) = \sum_{n=1}^{N_D} \frac{\mathrm{TF}(xy)}{W(S_n(D))} \tag{6}$$

Where TF(x) denotes the frequency of the part-of-speech x in each sentence of a document and $W(S_n(D))$ is the number of words in the sentence S_n and also N_D is the total number of sentences of the corresponding document. Function TF(x, y) shows the frequency of the part-of-speech x comes before the part-of-speech y in S_n as given in equation (6).

Feature sets F_6 and F_7 are frequency rates of types of part-of-speeches for initial and final sentence words, respectively as given below.

$$F_6(x) = \frac{\text{TF-B}(x)}{S(D)} \tag{7}$$

$$F_7(x) = \frac{\text{TF-E}(x)}{S(D)} \tag{8}$$

In equation (7), TF-B(x) represents the frequency of part-of-speech x as an initial sentence word and S(D) is the total number of sentences of the document. Also in equation (8), TF-E(x) shows the frequency of part-of-speech x as an end of sentence word.

3.2.4. Semantic level features

This category of features gives a semantic model to express the writing style of documents including tense and voice feature sets. The reason that this category of features is chosen for author identification lies in that they are independent of specific words, phrases, and contents of documents. Also, different authors have different preferences for using tenses and voices in their sentences of writings. The tense feature set, as denoted by F_8, contains all kinds of verb tenses in the language of documents and calculated using equation (9). The voice feature set, as denoted by F_9 includes active and passive voices as and given in equation (10).

$$F_8(x) = \frac{\mathrm{Tense}(x)}{V(D)} \tag{9}$$

$$F_9(x) = \frac{\mathrm{Voice}(x)}{V(D)} \tag{10}$$

Where Tense(x) shows the frequency of tense x in the document, V (D) is the total number of verbs of the document D and function Voice(x) represents frequency of voice x in the given document.

4. Experiments

In order to evaluate the performance of the proposed approach, computed experiments are conducted, and their performances are measured on some benchmark datasets. In Section 4.1, we explain datasets and evaluation measures that are used for evaluating the proposed approach. Then, we describe the experimental results in Section 4.2. Thereafter in Section 4.3, we study the influence of the number of features, the numbers of authors, training size, pre-processing, and classification method on the overall accuracy of the proposed approach. Finally, we analyze and discuss the results of experiments in Section 4.4.

4.1. Corpus and evaluation measures

We use the following two text corpora in English and Persian languages in our experiments: Reuter 50 50 corpus and Persian memoir blog corpus. Reuter_50_50 is a subset of RCV1 data set [42,43]. The training corpus consists of 2,500 documents written by 50 authors, where each author wrote 50 documents and the test corpus includes other 2,500 documents (50 documents per author). This corpus includes documents with CCAT (corporate and industrial) topics. We generate Persian memoir blog corpus (PMBC) automatically with the crawl on Persian blogs. PMBC includes 4977 documents that are written by 32 authors. The number of documents per author is not equal for all authors. Also, 20 percent of total authors' documents are considered as the testing corpus. PMBC is described in more detail in A. Also, table 3 reports statistical profiles of these two corpora.

Table 3. The statistical profile of datasets

Feature	PMBC	Reuter 50 50
Texts	4977	5000
Authors	32	50
Average number of documents per author	155.53	100
Minimum number of documents per author	50	100
Maximum number of documents per author	278	100
Average number of words per document	423	577.64
Average number of sentences per document	20.45	21.81
Average number of words per sentence	20.12	26.87

In order to evaluate the proposed method, we use the accuracy of classifiers as the performance measure. Accuracy shows the ability of the proposed method to identify the author of a given document. Also, cross-validation is used to evaluate the accuracy of the proposed approach.

4.2. Experimental results

In order to evaluate the performance of the proposed method, some experiments are designed. Experiments that are designed for evaluation of the proposed approach on datasets are explained in this section.

In the first experiment, we study the effect of features on the performance of author identification method and "What effect do different feature sets have on identification accuracy?" is to be answered. In this experiment, we group the features into three groups: GF_1, GF_2, and GF_3. The following three combined features sets GF1, GF2 and GF3 based on the type of features category are built for analyzing the influences of the feature sets $F_1, F_2, \ldots, F_9$ on the performance of the proposed approach.

$$GF_1 = (F_1, F_2, F_3); \quad (11)$$

$$GF_2 = (F_1, F_2, F_3, F_4, F_5, F_6, F_7); \quad (12)$$

$$GF_3 = (F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9): \quad (13)$$

The first group of experiments is executed on the C50 corpus. In these experiments, 10−folds cross-validation is used to evaluate the accuracy of the proposed approach by using training and testing documents of this corpus. Table 4 lists the identification accuracy of KNN, SVM, LDA and Naive Bayes classifiers by using the features GF_1, GF2 and GF3. The dimension of F2 is 1500. GF3 generates higher accuracy than GF1 and GF2 and LDA obtains the highest accuracy in all existing test cases as given in Table 4. The identification performance of the proposed approach using KNN and Naive Bayes classifiers is significantly lower than the performance of SVM and LDA classifiers.

Table 4. The identification accuracy with KNN, SVM, LDA and Naive Bayes on C50 corpus.

Feature sets	KNN	NB	SVM	LDA
GF1	34.93±0.63	36.8±0.74	79.12±0.79	88.01±0.84
GF2	36.01± 0:68	37.09±0.76	80.68±0.81	89.19±0.86
GF3	36.02±0.69	37.11±0.77	80.87±0.81	89.21±0.87

In this experiment, "What effect do different authors have on identification accuracy?" is to be answered. The influence of different authors on the accuracy of results is studied by dividing authors randomly into five sets A_1, A_2, A_3, A_4 and A_5. Also, these five subsets of authors have equal sizes. In addition, 5000 documents in the C50 corpus are segmented into five datasets D_1, D_2, D_3, D_4 and D_5. Thus the documents in the data set D_i (for i = 1,2,... , 5) were written by authors belonging to set Ai and Di contains 1000 documents of ten authors. Table 5 reports the identification accuracy of KNN, SVM, LDA and Naive Bayes on D_1, D_2, D_3, D_4 and D_5 by using the features GF_1, GF_2 and GF_3. In this experiment, the dimension of F2 is 1500. The results given in Table 5 show that the identification accuracy of classifier method are close together and the highest value on D_4 and the accuracy are the lowest value on D_2. Our approach models the styles of A_4 better than the styles of A_2 authors.

The second group of experiments is conducted using the PMBC corpus. Table 6 reports the identification accuracy of KNN, SVM, LDA and Naive Bayes by using the features GF_1, GF_2 and GF_3. In this experiment, the dimension of F_2 is 1000. Also, 10−folds cross-validation is used to evaluate the accuracy of results. The proposed approach with LDA and GF_3 obtains the highest identification accuracy on the PMBC corpus. As shown in Table 6, classifiers can be sorted based on accuracy in the following order: LDA, SVM, Naive Bayes, and KNN.

For investigating the influence of different authors on the accuracy of results, PMBC corpus authors are divided randomly into A1, A2, A3, A4 and all documents in the PMBC corpus are segmented into D1, D2, D3 and D4 that Di (for i = 1, 2, 3, 4) were written by authors belonging to the set of Ai and Di contains documents of eight authors. Table 7 shows the identification accuracy of KNN, SVM, LDA and Naive Bayes on D1, D2, D3, and D4 by using the feature sets GF1, GF2 and GF3. Also, the results of LDA classifier on D2 by using the feature set GF3 is shown in terms of confusion matrix in Fig. 1. As seen in this matrix, the numbers of documents of the first author (a1) that not correctly detected are more than others. It may be because of that the number of training documents of a1 is fewer than others.

The third experiment is executed to compare the results presented in [3]. The results are donated in Figure 3. As shown in Figure 3, the proposed method's classification accuracy is 4.6% higher than the method in Article [3] on the C50 test corpus. One of the reasons why the proposed approach is more efficient is that it has richer syntactic features than [3].

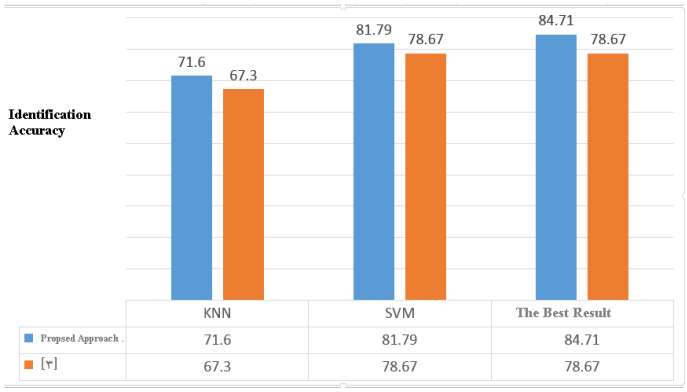


Fig 2 Compare the proposed method with the method presented in the [3].

Table 5 The identification accuracy on five datasets split from the C50 corpus.

Datasets	Feature sets	KNN	NB	SVM	LDA
D_1	GF1	55.12±0.71	59 .01±0 .79	89 .65±0 .88	93 .04±0 .92
D_1	GF2	56.29±0.73	59.52±0.81	89.96±0.90	94.13±0.95
D_1	GF3	56.31±0.74	59.55±0.81	89.98±0.92	94.14±0.97
D_2	GF1	51.01±0.62	54.78±0.70	81.12±0.85	92.76±0.89
D_2	GF2	52.16±0.64	55.21±0.71	82.43±0.89	93.34±0.93
D_2	GF3	52.88 0.65	55.98±0.72	82.77±0.88	93.91±0.95
D_3	GF1	52.76±0.66	56.11±0.71	83.59±0.86	94.56±0.96
D_3	GF2	53.54±0.67	57.32±0.73	83.88±0.89	94.93±0.98
D_3	GF3	53.91±0.67	57.93±0.74	83.92±0.90	94.95±1.01
D_4	GF1	61.42±0.82	65.32±0.82	88.03±0.89	94.91±0.99
D_4	GF2	62.85±0.84	66.89±0.84	89.54±0.91	95.44±1.03
D_4	GF3	63.11±0.85	67.22±0.86	89.85±0.91	95.76±1.03
D_5	GF1	59.13±0.75	63.77±0.78	87.75±0.87	93.48±0.97
D_5	GF2	60:98±0.77	65.02±0.79	88.42±0.90	94.57±0.99
D_5	GF3	61.06±0.78	65.46±0.80	88.96±0.91	94.95±1.00

Table 6 The identification accuracy with KNN, SVM, LDA and Naïve Bayes on PMBC corpus.

Feature sets	KNN	NB	SVM	LDA
GF1	71.18±0.47	77.45±0.52	89.51±0.71	93.11±0.93
GF2	72.01±0.49	78.11±0.56	89.88±0.78	93.87±0.95
GF3	72.12±0.51	78.15±0.58	89.93±0.80	93.91±0.96

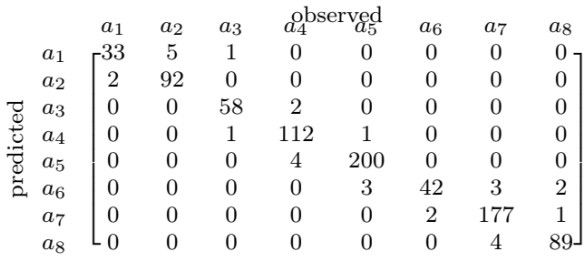


Fig 3 The confusion matrix from the PMBC corpus.

4.3. The influence of parameter

To find the appropriate values for the parameters of the proposed approach, we study the effect of each parameter on the performance of the method. For finding the impact of each parameter, we fix all parameters and change only the given parameter. After studying the effect of the selected, we choose the value that results in a good performance.

In our approach, we have the following parameters that affect the algorithm's performance: feature sets, dimensions of feature set, the classification algorithm, the size of training data, and the number of authors.

In the rest of this section when we do not specify the classification method and feature set explicitly, the classifier is LDA, the feature group is GF_3 and the dimensions of F_2 are 1500 and 1000 in C50 corpus and PMBC corpus, respectively.

We investigate the influence of different dimensions of feature set F_2, the most frequent words. The dimensions of F_2 are set as 250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000, respectively. Fig. 2 reports the identification accuracy. The curves in Fig. 2 show that increased dimensions of feature set F_2 improves the accuracy of identification in the most test cases. However, when the dimension of F_2 is more than 1500, the improvement of accuracy stops with the increase of dimensions.

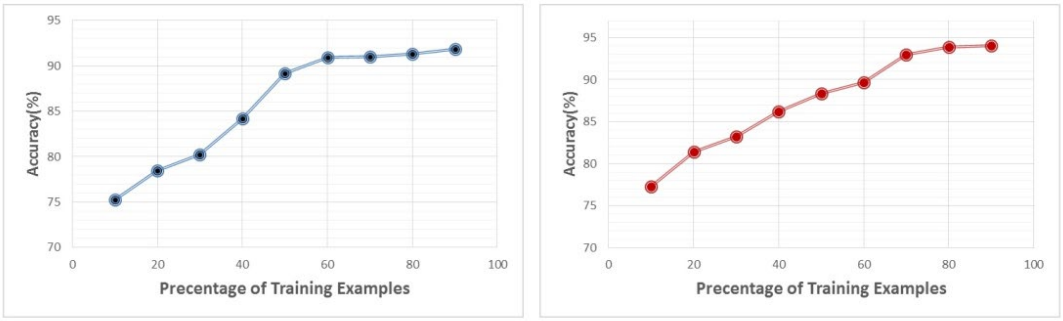
Also, the influence of different sizes of training data is investigated in the following experiments and "What is the effect of the number of different sizes of training data on the author's identification? " is to be answered. In these experiments, we use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of training documents for authorship identification. Fig. 3 shows the identification accuracy and indicates that the amount of training data has a great impact on the accuracy of the author identification.

In order to evaluate the proposed author identification, we use different classifiers and "how the use of different classifiers affects the identification process? " is to be answered. In the experiments, we use KNN, SVM, LDA and Naive Bayes classifiers and the accuracy of identification tasks is shown in Tables 4 and 6. The results show that in the author identification task, LDA as classification method has the best performance and the accuracy of identification with SVM is higher than KNN and Naive Bayes classifiers. Also, KNN and Naive Bayes are unable to solve the identification task when the number of authors increases and enough training data is not available.

We investigate the influence of the different number of authors in our experiments and "What is the effect of the number of authors on the author's identification? " is to be answered. In these experiments, we use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of authors to conduct the authorship identification. Fig. 4 shows the identification accuracy. By careful inspection of Fig. 4, it can be pointed that an increase in the number of authors decreases the accuracy of identification. The slope of the curve also decreases with increasing the number of authors.

Table 7 The identification accuracy on four datasets split from the PMBC corpus.

Datasets	Feature Sets	KNN	NB	SVM	LDA
D_1	GF1	86.17±0.85	81.01±0.77	95.85±0.87	97.24±0.96
D_1	GF2	87.01±0.87	82.58±0.82	96.37±0.90	98.01±0.99
D_1	GF3	87.08±0.87	82.91±0.84	96.38±0.93	98.20±1.03
D_2	GF1	84.01±0.80	80.71±0.75	94.94±0.95	95.52±0.92
D_2	GF2	84.91±0.85	80.73±0.76	95.80±0.98	96.27±0.98
D_2	GF3	84.93±0.88	80.73±0.80	95.81±1.01	96.28±1.02
D_3	GF1	86.25±0.84	81.25±0.69	95.19±0.91	97.01±1.00
D_3	GF2	87.60±0.88	83.02±0.74	96.51±0.93	97.79±1.09
D_3	GF3	87.61±0.88	83.23±0.76	96.56±0.95	98.81±1.11
D_4	GF1	81.56±0.71	74.84±0.64	91.07±0.73	94.91±0.89
D_4	GF2	82.27±0.77	75.39±0.69	92.51±0.86	95.49±0.98
D_4	GF3	82.30±0.79	75.42±0.71	92.83±0.89	95.98±1.01



(a) The accuracy of the methods on C50 corpus. (b) The accuracy of the methods on PMBC corpus.

Figure 4 The identification accuracy of different training sizes.

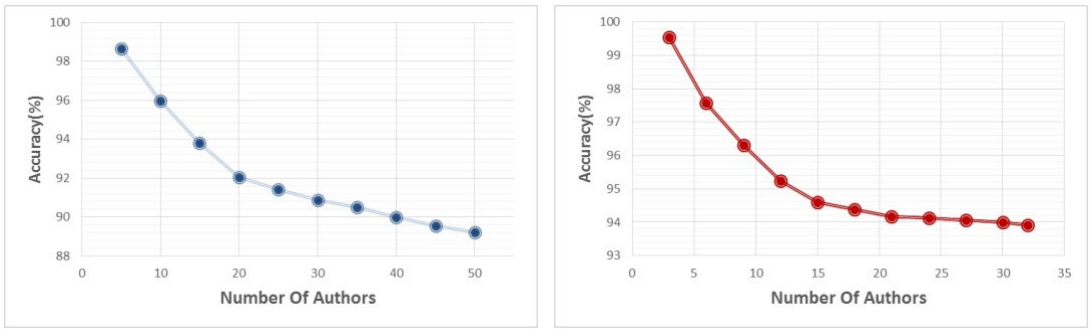


Figure 5 The identification accuracy of the different numbers of authors.

4.4. Discussions

In the previous section, we give the results of the proposed author identification approach on two corpora. By careful inspection of the results, the following points can be concluded. Results of the feature group GF_3 with more stylometric features have the highest accuracy for the author identification. The type of classification approach affects the performance of the author identification. Also, the result of experiments indicates that the approach with the LDA method identifies the author with the highest accuracy. It can be concluded that the linear combination of features creates the best discriminative capability.

The pre-processing of the documents increases the accuracy of the author identification by increasing the discriminative capability of lexical features. The feature selection solves the curse of dimensionality of vector space model dimensions problem. Increasing the number of candidate authors and reducing the training size decrease the classifier accuracy and as a logical consequence decrease the accuracy of the author identification. Experimental results have indicated that the performance of identifying authors can be different in different languages.

5. Conclusions

Regarding the important role of documents in social exchanges in recent years, the problem of recognizing the author of documents has become more important. The efficient and cross-language approach has been proposed for author identification in this paper. This problem is solved as a typical classification problem. In this approach, the writing style of authors has been modeled with stylometric features that are categorized into lexical, structural, syntactic and semantic features. Pre-processing of the documents, the stylometric features richness and stylometric feature selection have improved the accuracy of identification task.

The performance of the proposed approach is evaluated on C50 and PMBC copra. Results of experiments show that LDA has a higher performance than SVM, KNN and Naive Bayes classifiers for author identification task. The amount and the type of training data and the number of authors affect the accuracy of author identification.

In future works, we will investigate the possibility to improve the efficiency of the prediction algorithm by combining other methods such as genetic algorithm. Also, we would like to develop an approach to recognize the evolution and change the style of the author's writing.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which improved this paper.

References

[1] V. Chandani, N. Deshmane, K. Buva, S. Apte, and D. R. Prasad, "Study of different methods for author identification," International Journal of Engineering Research and Technology, vol. 4, pp. 558–560, Jan. 2015.

[2] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378–393, 2006.

[3] C. Zhang, X. Wu, Z. Niu, and W. Ding, "Authorship identification from unstructured texts," Knowledge-Based Systems, vol. 66, pp. 99–111, 2014.

[4] M. Fatima, K. Hasan, S. Anwar, and R. M. A. Nawab, "Multilingual author profiling on facebook," Information Processing & Management, vol. 53, no. 4, pp. 886–904, 2017.

[5] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, and B. Stein, "Overview of the PAN/CLEF 2015 Evaluation Lab," in Experimental IR Meets Multilinguality, Multi-modality, and Interaction. 6th International Conference of the CLEF Initiative (CLEF15) (J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. SanJuan, L. Cap- pellato, and N. Ferro, eds.), (Berlin Heidelberg New York), pp. 518–538, Springer, Sept.2015.

[6] O. Halvani, C. Winter, and A. Pflug, "Authorship verification for different languages, genres and topics," Digital Investigation, vol. 16, pp. S33–S43, 2016.

[7] V. Benjamin, W. Chung, A. Abbasi, J. Chuang, C. A. Larson, and H. Chen, "Evaluating text visualization: An experiment in authorship analysis," in Proceedings of IEEE International Conference on Intelligence and Security Informatics, pp. 16–20, 2013.

[8] M. Kocher and J. Savoy, "Distance measures in author profiling," Information Processing & Management, vol. 53, no. 5, pp. 1103–1119, 2017.

[9] J.-M. Torres-Moreno, G. Sierra, and P. Peinl, "A german corpus for text similarity detection tasks," International Journal of Computational Linguistics and Applications, vol. 5, no. 2, pp. 9-24, 2014.

[10] E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. L´opez-L´opez, M. Potthast, and B. Stein, "Overview of the author identification task at pan 2014.," in Proceedings of Conference and Labs of the Evaluation Forum, pp. 877–897, 2014.

[11] C.-T. Li, Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions: Methods and Solutions. 2009.

[12] A. Gokhale, K. Borkar, and R. S. Prasad, "A proposed system for author identification using statistical method," International Journal of Engineering Research and Technology, vol. 2, pp. 1609–1611, Sept. 2013.

[13] E. Castillo, O. Cervantes, D. V. Ayala, D. Pinto, and S. Le´on, "Unsupervised method for the authorship identification task.," in Proceedings of Conference and Labs of the Evaluation Forum, vol. 1180, pp. 1035–1041, 2014.

[14] 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, vol. 3, no. 2, p. 41, 2015.

[15] A. Bartoli, A. Dagri, A. De Lorenzo, E. Medvet, and F. Tarlao, "An author verification approach based on differential features," in Proceedings of Conference and Labs of the Evaluation Forum, vol. 1391, 2015.

[16] O. Pimas, M. Kr¨oll, and R. Kern, "Know-center at PAN 2015 author identification," in Proceedings of Conference and Labs of the Evaluation Forum, vol. 1391, 2015.

[17] M. A. Sanchez-Perez, I. Markov, H. G´omez-Adorno, and G. Sidorov, "Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus," in Proceedings of International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 145–151, 2017.

[18] N. Deshmane, V. Chandani, K. Buva, S. Apte, and R. Prasad, "Author identification system using hybrid technique," International Journal of Engineering Research and Technology, vol. 4, pp. 100–102, Apr. 2015.

[19] S. Harvey, "Author verification using PPM with parts of speech tagging," in Proceedings of Conference and Labs of the Evaluation Forum, vol. 1180, pp. 1063–1068, 2014.

[20] C. Zhao, W. Song, L. Liu, C. Du, and X. Zhao, "Research on author identification based on deep syntactic features," in Proceedings of the 10th International Symposium on Computational Intelligence and Design, vol. 1, pp. 276–279, 2017.

[21] J. Soler and L. Wanner, "On the relevance of syntactic and discourse features for author profiling and identification," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 681–687, 2017.

[22] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan, "Stylistic text classification using functional lexical features," Journal of the Association for Information Science and Technology, vol. 58, no. 6, pp. 802–822, 2007.

[23] E. Villar-Rodriguez, J. Del Ser, M. N. Bilbao, and S. Salcedo-Sanz, "A feature selection method for author identification in interactive communications based on supervised learning and language typicality," Engineering Applications of Artificial Intelligence, vol. 56, pp. 175–184, 2016.

[24] T. Chen and M.-Y. Kan, "Creating a live, public short message service corpus: the NUS SMS corpus," Language Resources and Evaluation, vol. 47, no. 2, pp. 299–335, 2013.

[25] A. Vorobeva, "Examining the performance of classification algorithms for imbalanced data sets in web author identification," in Proceedings of the 18th Conference of Open Innovations Association FRUCT, pp. 385–390, 2016.

[26] M. Al-Ayyoub, Y. Jararweh, A. Rababah, and M. Aldwairi, "Feature extraction and selection for arabic tweets authorship authentication," Journal of Ambient Intelligence and Humanized Computing, vol. 8, no. 3, pp. 383–393, 2017.

[27] H. G´omez-Adorno, G. Sidorov, D. Pinto, and I. Markov, "A graph based authorship identification approach: Notebook for PAN," in Proceedings of Conference and Labs of the Evaluation Forum, vol. 1391, 2015.

[28] D. Castro, Y. Adame, M. Pelaez, and R. Mun˜oz, "Authorship verification, combining linguistic features and different similarity functions," in Proceedings of Conference and Labs of the Evaluation Forum, Sept. 2015.

[29] S. Mechti, M. Jaoua, R. Faiz, L. H. Belguith, and B. Bsir, "On the empirical evaluation of author identification hybrid method," in Workshop Proceedings of Conference and Labs of the Evaluation forum, vol. 1391, 2015.

[30] Y. Sari and M. Stevenson, "A machine learning-based intrinsic method for cross-topic and cross-genre authorship verification," in Workshop Proceedings of Conference and Labs of the Evaluation forum, vol. 1391, 2015.

[31] S. Jie, "Authorship identification based on extraction and combined svm of similar attribute features," Bolet´ın T´ecnico, vol. 55, no. 5, pp. 40–47, 2017.

[32] S. Ferilli, "A sentence structure-based approach to unsupervised author identification," Journal of Intelligent Information Systems, vol. 46, no. 1, pp. 1–19, 2016.

[33] C.-L. Li, Y.-C. Su, et al., "Combination of feature engineering and ranking models for paper-author identification in KDD Cup 2013," The Journal of Machine Learning Research, vol. 16, no. 1, pp. 2921–2947, 2015.

[34] C. Klaussner, J. Nerbonne, and Ç.Çöltekin, "Finding characteristic features in stylometric analysis," Digital Scholarship in the Humanities, vol. 30, no. suppl_1, pp. i114–i129, 2015.

[35] StanfordTagger, "Stanford Log-linear Part-Of-Speech Tagger," 2015. http://nlp.stanford.edu/software/tagger.html [Accessed: 08/08/2018].

[36] StanfordCoreNLP, "Stanford CoreNLP a suite of core NLP tools," 2015. http://stanfordnlp.github.io/CoreNLP/ [Accessed: 08/08/2018].

[37] mojtaba khallash, "JHazm," 2015. https://github.com/mojtaba-khallash/JHazm [Accessed: 08/08/2018].

[38] FarseNet, "Farse Net," 2015. http://dadegan.ir/catalog/farsnet [Accessed: 08/08/2018].

[39] D. Agnihotri, K. Verma, and P. Tripathi, "An automatic classification of text documents based on correlative association of words," Journal of Intelligent Information Systems, vol. 50, no. 3, pp. 549–572, 2018.

[40] S. Sadeghi and H. Beigy, "A new ensemble method for feature ranking in text mining," International Journal on Artificial Intelligence Tools, vol. 22, no. 3, 2013.

[41] L. Thomas, "Class InfoGainAttributeEval," 2018. http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html [Accessed: 10/23/2018].

[42] "Reuters rcv1 rcv2 multilingual, multiview text categorization test collection data set." https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection. Accessed: 03/03/2017.

[43] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," Journal of machine learning research, vol. 5, pp. 361–397, Apr. 2004.

[44] Z. Farahmandpoor, H. Nikmehr, M. Mansoorizade, and O. Tabibzadeh Ghamsary, "A novel intelligent persian authorship system based on writing style," Soft Computing Journal, vol.1, no.2, pp.35–26. 2013.

[45] Z. Farahmandpour, and H. Nikmehr, 2015. A Study on Intelligent Authorship Methods in Persian Language. Journal of Computing and Security, 2(1), pp.63-76.

[46] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the feasibility of internet-scale author identification," in 2012 IEEE Symposium on Security and Privacy, pp.314–300 , IEEE, 2012.

[47] Stamatatos, E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), pp.538-556.

[48] El Bakly, A.H., Darwish, N.R. and Hefny, H.A., 2020. Using Ontology for Revealing Authorship Attribution of Arabic Text. Int. J. Eng. Adv. Technol.(IJEAT), 4, pp.143-151.

[49] Sarwar, R., Porthaveepong, T., Rutherford, A., Rakthanmanon, T. and Nutanong, S., 2020. StyloThai: A scalable framework for stylometric authorship identification of thai documents. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(3), pp.1-15.

Appendix. The Persian memoir blog corpus

The Persian memoir blog corpus includes 32 authors and 4977 texts and the number and the length of the texts is different for each author, because the probability of copying from other author's documents by authors in their memoirs may be low. Therefore, memories are chosen as the content of this corpus. Also, it generated automatically with crawl on blogs.

Figure 6 shows an example of XML file with one document in PMBC corpus. For each author, we consider one XML file. The structure for this corpus is as comprehensive and expanded as possible for use in other applications in data mining.

Figure 6 An example of a document in Persian memoir blog corpus.

Reyhaneh Ameri received the BS degree in computer engineering from Iran University of Science and Technology. She also received the MS degree computer engineering from Sharif University of Technology in 2016. His research interests include data mining, blockchain and machine learning. Currently, she is a PhD student in Amirkabir university of technology.

Email: ameri@ce.sharif.edu

Hamid Beigy received the B.Sc. and M.Sc. degrees in computer engineering from the Shiraz university in Iran, in 1992 and 1995, respectively. He also received the Ph.D. degree in computer engineering from the Amirkabir university of technology, Iran, in 2004. Currently, he is an Associate Professor in department of computer engineering at the Sharif university of Technology, Tehran, Iran. His research interests include machine learning, large scale machine learning, and social networks.

Email: beigy@sharif.edu

Paper Handling Data:

Submitted: 04.20.2020
Received in revised form: 05.06.2021
Accepted: 05.22.2021
Corresponding author: Dr. Hamid Beigy
Affiliation of the corresponding author: Department of Computer Engineering, Sharif University of Technology