Journal of Computing and Security

January 2015, Volume 2, Number 1 (pp. 63–76)

http://www.jcomsec.org

A Study on Intelligent Authorship Methods in Persian Language

Zeinab Farahmandpour [a], Hooman Nikmehr [b,*]

[a] *Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran.*

[b] *Department of Computer Architecture, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.*

ARTICLE INFO.

Article history:
Received: 24 September 2014
Revised: 26 January 2015
Accepted: 15 May 2015
Published Online: 7 February 2016

Keywords:
Authorship Attribution, Writing Style, Writeprint, Human Stylome

ABSTRACT

Author identification is an attempt to demonstrate the characteristics of the author of a piece of language information so that in the end, it would be possible to significantly distinguish the difference between various texts written by different people. The rapid development of Internet communication has caused Internet tools with anonymous identity, such as emails and weblogs, to become popular communication methods for the perpetrators of illegal acts and has raised some security concerns. Persian language is of interest to a great number of different individuals and organizations for various reasons such as political, social, artistic, cultural and religious issues. In this paper, a number of intelligent writeprint methods which help automatic identification of a Persian writer based on his/her writing style are studied and compared. For this purpose, after collecting two different databases, five feature types including lexical, syntactic, semantic and application-specific features, were used for extracting stylometric characteristics. In this study KNN, Delta, Neural Networks, Decision Tree and Linear Discriminate Analysis classification methods were applied to these databases. The results and their comparison showed that Linear Discriminate Analysis and KNN methods ranked first and second, respectively, in terms of accuracy among the studied methods.

1 Introduction

Identification of the author is finding or getting close enough to the real author of a text particularly in a collection of nominated authors. To achieve this goal, the distinctive information of the authors has to be recognized and analyzed using some models. Author identification is one of the oldest issues in stylistics from one hand, and one of the newest on the other hand [1]. Author identification not only can attract the attention of scientists but also, in a more practical context it can be a favorite issue for politicians, journalists and lawyers [1, 2]. Since the creation of words and documents, there have been debates about the ownership of words and identity and characteristics of the documents' author. So author identification can be defined as an attempt to demonstrate the characteristics of the producer or the author of a piece of language information. The key assumptions in author identification consist of access to a sample text which is definitely written by a member of the set of authors and also specification of the author of that text.

Many researchers believe that people use certain language patterns in their writings which act as the fingerprints of the authors. This pattern is called writeprint [3]. In this regard, [4] calls the collection

* Corresponding author.
Email addresses: zeinab.farahmandpoor@gmail.com (Z. Farahmandpour), nikmehr@eng.ui.ac.ir (H. Nikmehr)
ISSN: 2322-4460

of measurable specific characteristics which can be used for identification of a particular author, "Human Stylome". Since every human being has unique biometric characteristics and behavior patterns, it can be said that some language and writing related characteristics such as using specific patterns of word sequences, layouts, syntactic and structural features that are called stylistic characteristics remain relatively constant in people. Learning and recognition of these characteristics with high accuracy is a debating subject in the text author identification area. Theoretically, acceptable reasons can be put forward for the hypothesis of possibility of existence of such qualities. Since everyone learns language individually and their experiences as language learners are different from each other, the language that different people learn would have little differences in various but small aspects.

Although using the document analysis science and also expert judgments can conventionally result in good conclusions in author identification, recent advances in statistical techniques and application of machine learning on accessible computer sources and objective and automatic inferences form texts, have created scientific and more reliable methods in author identification. This paper implements author identification for Persian language using all four types of features that have been used for English language with automatic machine learning methods on the two collected standard databases and compares their results. In Section 2 a background on author identification will be presented. The characteristics used for identification of the unique writing style of an author are explained in Section 3. Then in Section 4 the author identification methods are introduced and in Section 5 adopting the author identification methods for Persian language is discussed. Section 6 evaluates the recommended system and then compares different methods of machine learning used in Persian language author identification. Section 7 is dedicated to discussion.

2 Research Background

The basic idea of statistical (or computational) identification of the author is the quantitative measurement of the text's characteristics so that the differences between texts written by different people can be distinguished. The first attempt to measure writing styles was carried out when Mendenhall did some studies on the Shakespeare's plays (1887). Later, in the first half of the twentieth century, some statistical research projects were conducted on literary texts [5–7]. Then to determine the identity of the author, more detailed investigations were performed on the data collection of "The Federalist Papers" by Mosteller and Wallace [1] which is, without a doubt the strongest and most effective research ever conducted in this field. Their method basically moved away from traditional approach of author identification which was identifying the author based on human judgment. From then on and until 1990s, studies regarding author identification, focusing on defining the text attributes for assessment of the writing style (a branch of research called Stylistic), resulted in introduction of many diverse measures such as the length of the sentences, the length of the words, the number of words' occurrences, the number of repeating characters and the lexical richness functions [8, 9]. In another study, nearly 1000 different measures were introduced [10]. It is worth mentioning that during this period, i.e. up until 1990, since the idea of creating a full automated system of author identification was hardly considered, the presented identification methods were only using computers rather than being computer-based.

Considering the amount of available electronic texts, the applicability of the author identification methods in variety of applications can be easily understood. Some of which are: literary researches (e.g. identification of the anonymous author of a controversial literary product from known authors) [11, 12], artificial intelligence (such as identification of terroristic messages or statements from a set of known terrorists) [13], criminal law (e.g. identification of authors of threatening messages, identification of authors of suicidal notes), civil law (the infringement of laws used in intellectual property issues) [2, 14], computer debates (e.g. identification of authors of malwares or computer viruses) [15]. In such cases, the author attempts to hide his identity for criminal purposes.

Nowadays, statistical and machine learning methods are very common methods for author identification. Burrows was the first to use Principal Component Analysis in 1987 that attracted attention of researchers due to its high differentiation ability [16]. Other multi-variable methods such as Cluster Analysis and Discriminant Analysis [17, 18] have produced good results in author identification as well.

Powerful machine learning techniques have had a lot of applications these years. Tweedie, Lowe, Mathews et al. used Neural Network for this purpose [18, 19]. While Diedrich, de Vel et al. have successfully applied SVM method for author identification [20, 21].

The expansion of the Internet, its borderless nature and the increase of online communications has resulted in creation of security issues in author identification. For this reason, implementation and use of authors identification in different languages is becom-

ing increasingly more important. So far, studies have been conducted on English, Greek, Chinese and Arabic. Peng et al. performed tests on English texts, Chinese novels and Greek newspapers [22]. Zheng et al. implemented author identification on Internet messages in English and Chinese, etc. [23–26]. Feature extraction for author identification does not have the same difficulty in different languages. Most features of writing style have been designed for English and it is possible that these features are not equally applicable for other languages. The structural and linguistic differences in different languages can possibly make the extraction of the features difficult.

Persian language is the official language in Iran, Afghanistan and Tajikistan, and tens of millions of people speak in this language. Different groups of people are related to this language for political, social, artistic, historical, cultural and religious reasons. The morphological differences of this language with English have resulted in negligence in implementation and design of natural language processing tools consistent with Persian language. This might have partly caused the negligence of author identification in this language. In this paper, some of the well-known techniques in author identification are implemented for Persian texts.

3 Stylistic Features

Previous studies on author identification have defined and classified some features for recognition of writing styles [8, 26, 27]. Writing style features extracted from the text, show the style of people's writings and simplify the author identification process. The writing styles of individuals are different in various aspects. The main classification of these features consists of lexical, characteristic, syntactic, semantic and application-specific features.

3.1 Lexical Features

The lexical features of a text are known as sets of clauses that are divided into components called tokens. Each token can be a word, a figure or a punctuation mark. The main advantage of these features is that they are applicable in any language and with any set of texts by using the tokenizer tool (which divides a text into a set of tokens).

The lexical richness function, models the distribution of words in a text. The size of the lexis is another measure which strongly depends on the length of the text. In the way that at the beginning of the text, the size of the lexis increases rapidly and by reaching the end the increasing trend of the lexis size slows down. In another method, regardless of the content information, the text is considered as a set of words each of which is repeated a certain number of times in the text [28].

3.2 Characteristic Features

The characteristic features consider the text as a set of character strings. Some of the other features of this category can be the number of characters in the text, numbers in the text, capital and small letters, repetitious and punctuation characters [21, 26]. It has been proven that this group of information, which can be easily obtained for every natural language, is relatively suitable for determination of the writing style [29]. Cavnar and Trenkle [30], introduced n-gram as an n-character piece of a longer character string. This feature which can extract subtle style differences (including lexical information) is remarkably noise (spelling errors) tolerant. The Character n-gram extraction method is very simple in terms of computation. For example Character 4-gram for the current paragraph would be as follows: |The-|, |char|, |acte|, |rist|, |ic-f|, |eatu|, ... In a study [31], it was found that in an author identification process, the Character n-gram with different "n"s is a more suitable feature than a lexical feature. A diverse number of authors have shown that a number of Character n-gram repeats might be better than even the grammatical features for obtaining lexical information. They are useful without the need for language information background. It is noteworthy that the Character n-gram feature is conducted on words related to both the content and the subject. This feature generates useful information only if all of the texts in our database are on the same subject.

The important issue in the Character n-gram method is defining n as the length of the desired string. A large n not only provides better lexical and background information but also presents better thematic information. However, a large n would increase the dimensions of the feature representation very much by providing hundreds of thousands features for a text. On the other hand, a small n (2 or 3) would be able to display the information of sub-strings (such as syllable), however this extracted attribute is not sufficient by itself for displaying the text and background information. The procedure of choosing an optimum n depends on the natural language specifications and defining constant n, will prevent application of n-grams with different lengths [31, 32]. For instance, since Greek and German languages have longer words in comparison with English, using larger n for them is more suitable.

3.3 Syntactic Features

Since an author unconsciously tends to use the same syntactic patterns, using syntactic features provides a more accurate display of the text [24]. This provides a more reliable method for identification of the author than the lexical features. Extracting syntactic information requires powerful and accurate tools for processing the natural language (NLP) which are able to syntactically analyze texts. Extraction of syntactic features is a text-related procedure and requires a parser which can analyze a particular natural language with high precision. Moreover, the syntactic attributes generate a set of features with low precision due to inevitable errors of the parsers. Baayen, Halteren and Tweedie [33] were the first people to use syntactic information for author identification. Using the complete tree decomposition procedure of every sentence, they analyzed the corpus set syntactically and extracted the number of occurrences of the re-writing rules. Each re-writing rule expresses one part of a syntactic analysis. For example an adverb preposition phrase consists of the preposition and a noun phrase followed by a prepositional predicate. This detailed information shows the syntactic group of each word and the composition of words to make phrases and other structures. The reported results suggest that this attribute is better than lexical and lexical richness features. The extracted attribute includes the number and the length of noun phrases and clauses.

The occurrence of the common words, such as prepositions, articles and pronouns, which are called function words, is one of the very good features for author identification [16, 34]. The main advantage of this attribute is its independence to the text's subject since these words do not carry any semantic information. Another important advantage of this attribute is that the author uses the function words unconsciously; this means that the writing style of the writer of different texts with different subjects can be reflected in the function words used by him/her. The choice of these words is usually based on the criteria that are introduced by linguistics. A simple yet successful method to define function words for author identification is to extract words with high occurrences in the corpus of the nominated authors. In studies conducted by Burrows [16, 35] a set of 100 function words have been recognized suitable for displaying the author's style.

3.4 Semantic Features

Semantic features more precisely analyze the text and provide more applicable features. This is done by obtaining the semantic dependency graph including the semantic features and modification relations.

The tense and the mode of the verbs used by the author and the semantic similarity between the words in the text are another example of such features. The words and phrases that join clauses together are known as conjunctions. Different patterns of using conjunctions [36] result in remarkable differences in writing styles. According to Hildi and Mason [37], conjunctions are categorized into three groups of "descriptive", "exponential" and "elaborative". The descriptive conjunctions deepen the content of the text by illustrating and re-focusing. Using descriptive conjunctions can have a good effect on the text and provide solidarity throughout the whole text. The exponential conjunctions would add more information into the text, even sometimes antithetical to the current concepts. Repetitive usage of these conjunctions aggregates the text information but if these conjunctions are not used enough it is possible for the reader to get lost among the multiple and various concepts. The elaborative conjunctions describe the text with details or logical relationships.

McCarty et al. [38] defined another approach for semantic measure extraction. According to the WordNet [39], they have well recognized the information about the word's synonyms. Moreover, they used the Latent Semantic Analysis for lexical features in order to automatically recognize the semantic similarities between words [40].

3.5 Application-Specific Features

The application-specific features can be used for a more delicate presentation of differences in people's styles in a specific confine. The application-specific features, in fields such as electronic messages, are the application of structural measures which includes using greetings in messages, different types of signatures, dips, the length of the paragraph, etc. [3, 21, 26, 41]. In addition, when the text is in the form of HTML, the HTML Tag distribution-specific features [21] such as the colors and the sizes of the fonts can be used as application-specific features [13]. These features are very important in short texts, where the stylistic features of the text content are not enough, and accurate tools are required for extracting them.

4 Data Classification Methods

There are different methods for data classification such as Statistic Pattern Recognition, which works based on statistical models, Neural Pattern Recognition, which is based on neural networks and Syntactic Pattern Recognition, which works based on element structures (clustering).

4.1 Density Estimation Using K Nearest Neighbor Method

The KNN method is a probabilistic model in which to identify the author of an anonymous text, K samples of the training texts, in the feature space, which have the least distance from the anonymous text, are taken and the author who has the most number of texts among these K texts is selected as the author of the unknown text [42]. This method estimates the probability of the anonymous text belonging to the i^{th} author through

$$\widehat{p}\,(x|i) = \frac{n_i}{k} \tag{1}$$

where, n_i is the number of texts that belong to the i^{th} author in the set of K texts which have the most similarity to the anonymous text. K is constant in every probability calculation.

Suppose that there are n training texts with their authors known and we are going to classify a text with an anonymous author, x, in one of the above c classes (author). To use the KNN method, first a scale should be selected for measuring the distance (such as the Euclidean distance). Then the following steps should be taken:

A. The distance of each training text with the text x is calculated; K texts which are the closest to the text x are selected regardless of their class (author).
B. The number of each author's texts in this K texts is determined and called n_i where $i = 1, 2, \ldots, c$ and $\sum_i n_i = k$.
C. The text x is assigned to the author who has the most number of texts among the K texts, i.e. the largest n_i.

4.2 Delta Method

In 2002, a method was presented by Burrows [11] exclusively for author identification namely Delta. In this method, the normalized difference of the number of word occurrences (a set of n desired words, usually the most repetitious words) in the text D of the known author and the number of occurrences of the same words in text D' with anonymous author is calculated using Equation (2). In this equation, σ_i is the standard deviation of the i^{th} word in the comparative set and $f_i(D)$ is the number of occurrences of the word w_i (the number of occurrence of the i^{th} desired word) [43]. This method calculates the distance between the two texts, D' being the test anonymous text and D the written training text by a nominated author, as follows:

$$\Delta_B^{(n)}\,(D - D') = \sum_1^n \frac{1}{\sigma_i}\,|f_i(D) - f_i(D')| \tag{2}$$

This method classifies the nominated authors of the texts D' considering their distances from the test text D, while after each difference, the number of words occurrences reduces by a factor of $\frac{1}{\sigma_i}$. In this equation the test document is classified into the group closest to the specified training document. This method uses the function word distribution feature for classification.

4.3 Neural Networks

The neural network algorithm is one of the most widely used and most practical modeling methods for large and complex problems. In 1958, Rosenblatt [44] introduced the concept of single-layer perceptron as a useful tool for classification of a set of data into two classes and he offered a proof for stability of the perceptron learning rule. Each multilayer perceptron neural network consists of an input layer. Each node in this layer is equal to one of the predictor variables. Each input node attaches to all nodes of the hidden layer. The nodes of the hidden layer can be attached to another hidden layer or the output layer. The output layer can consist of one or more output variables depending on the problem [45, 46]. In this study, a three-layer neural network is used for the implementation of the author identification system.

4.4 Decision Tree

The Decision trees are used for displaying different concepts in artificial intelligence such as the structure of sentences, equations, game modes, etc. Training of a tree is a method for approximating the objective functions with continuous and discrete values. This method is resistant against the noise in the data and is able to learn the conjunction and disjunction syntax of the predicates. Due to the high efficiency of the decision tree method in problems with high volume data, this method is used in Data Mining [47]. The decision trees are made by consecutive division of data to separate groups and the purpose of this process is to increase the distances of the groups in each division.

In the decision tree ID3, a statistical value, called Information Gain, is used which determines to what extent a feature is able to divide training examples based on their classification [48]. Likewise, the features extracted from texts of each author are given to the decision tree. The decision tree method selects features with high entropy. The feature which is in the root has a higher importance than other features. This method uses entropy and interest rate for classi-

fication of selected features and in the end composes a tree of features that are classified as important. So the features which are located in the root of the tree are more discriminant than other features and as we go farther from the root, the importance and separating quality of the selected features reduces. At the end, the precision of the decision tree is also calculated.

4.5 Linear Discriminant

A common algorithm for classification and data dimension reduction is Linear Discriminant Analysis (LDA). This method easily manages the modes in which the numbers of members of the classes are not equal. This algorithm intends to maximize the ratio of the between class variance to within-class variance in every database. In-class dispersion is calculated as the covariance for each class. This method takes the data set into another space. In LDA, the form and location of the original data set do not change when converted into another space and the endeavor is to increase the separation of the classes [49]. This method seeks a series of features among the extracted features of the texts that can best divide the authors' classes and since it is a linear method, it ignores all nonlinear features.

5 Adopting Author Identification Methods for Persian Language

The automatic author identification has been implemented and used for various languages; however it has not been used to this extent for Persian language [50]. In this Section, the works conducted for the implementation of the automatic author identification in Persian language are explained. First, the stylistic features extracted from Persian texts are discussed. Then methods of feature selection for data volume reduction and elimination of additional information will be investigated. Finally, the collected database for this application in Persian language will be analyzed.

5.1 Stylistic Features Extraction

Different features have been used in different languages for author identification. In this study the most distinguishing of these features are selected. Some of the previously discussed features are discriminant by themselves but some others are applicable and provide better results in combination with other features.

In the following, a variety of selected features to be extracted from Persian texts are introduced. Some of the features mentioned in English languages are not used due to lack of similar extraction system in Persian language.

5.1.1 Lexical Features

The lexical features used in this paper include lexical richness, n-gram and other features that are explained in detail in the following sections.

Lexical Richness

The lexical richness is used to measure the lexical frequency in a text. Various metrics are used for measuring the amount of richness for the purpose of author identification. The most common metric of this group is type-token ratio which is calculated as $\frac{V}{N}$ where V is the number of words and N is the number of tokens in the text. Since these features are affected by the length of the texts, many researchers have presented functions describing these features that are claimed to be length-independent. However our investigations did not show any proof for this claim.

To achieve better results, some scholars have used a set of lexical richness functions, instead of one, in association with multi-variable statistical techniques [17]. These functions are not ponderous in terms of calculation. According to some studies [51], the majority of lexical richness functions are very much dependent to the length of the text and are relatively impermanent for texts shorter than 1000 words.

For measuring the lexical richness, Stamatatos et al. [27] and Baayen et al. [33] have used a collection of 5 functions namely K, R, W, S and D for Yule [6], Honoré [52], Sichel [53] and Simpson [54], respectively. These functions are defined through expressions

$$K = \frac{10^4 \left(\sum_{i=1}^{\infty} i^2 v_i - N\right)}{N^2} \tag{3}$$

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}} \tag{4}$$

$$W = N^{V^{-\alpha}} \tag{5}$$

$$S = \frac{V_2}{V} \tag{6}$$

$$D = \sum_{i=1}^{V} V_i \frac{i\,(i-1)}{N\,(N-1)} \tag{7}$$

where V_i is the number of words that have repeated i times and α is a constant parameter with the value of 0.17. For each Persian text in this paper, these functions are calculated and a vector is composed by these five features.

n-Gram

The n-gram feature works well in texts with noisy inputs. Since in this method, each character string is decomposed into smaller pieces, any kind of errors in the word would affect only a few pieces and other pieces of that word would remain flawless.

To create an n-gram profile, first the input text is read and then the following steps are taken:

- The text is divided into separate tokens. Numbers, control characters and punctuation marks will be spared.
- All possible n-grams of n are generated.
- The n-grams are entered in a hash table and the number of their occurrences in the text is founded.

The largest number of n-gram occurrences will occur in uni-gram ($n = 1$) which indicates the alphabetic characters in the language. In this paper, the n-gram feature is used due to its importance in the above mentioned studies. In this study, a list of all 72 possible uni-grams in the Persian language is used

{ا, ب, پ, ت, ث, ج, چ, ح, خ, د, ذ, ر, ز, ژ, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ک, گ, ل, م, ن, و, ه, ی, آ, بّ, پّ, تّ, ثّ, جّ, چّ, حّ, خّ, دّ, ذّ, رّ, زّ, ژّ, سّ, شّ, صّ, ضّ, طّ, ظّ, عّ, غّ, فّ, قّ, کّ, گّ, لّ, مّ, نّ, وّ, هّ, أ, ؤ, ۀ, ۀ, أ, ء, ئ, آ, یّ}

The number of possible bi-grams, tri-grams and forth-grams is very large and obtaining the distribution of their occurrences in the texts would result in features with too many dimensions and too many small values close to zero. Thus, in this paper only the bi-grams, tri-grams and forth-grams that the number of their occurrences is more than 10% of the most repetitious possible bi-grams, 5% of the most repetitious possible tri-grams and 10 % of the most repetitious possible forth-grams in the databases introduced in Section 5.3 are used. These threshold values are calculated considering measures such as the number of authors and the length of the texts and by experience. The number of possible and selected bi-grams, tri-grams and forth-grams are listed in Table 1.

Lexical Features

In order to benefit from the lexical information as much as possible, the following lexical features are used

- The average number of characters used in words
- The average number of characters used in sentences
- The average number of words used in sentences
- The number of all short words equal or shorter than 3 characters
- Distribution of Persian characters in the text
- Distribution of paragraphs in the text
- Distribution of words with the length of 1 to 30 in the text
- Distribution of English characters in the text
- Distribution of numeric characters in the text
- Distribution of half space character in the text

5.1.2 Syntactic Features

These features find the rules of sentence generation. Features that are extracted from this structure include noun, adjective and adverbial phrases. In this paper, a POS tagger (Part-Of-Speech tagger) is used for extraction of the number of nouns, adjectives and adverbial phrases. This POS tagger relates every word to one of these tags.

POS Tagger for Persian Language

POS taggers are language related tools that recognize and label words based on their role in the sentence. Due to lack of a suitable software tool for automatic identification of nouns, adjectives and adverbs in a Persian text, such tool is designed to complete the research.

According to Tabatabaee [55], the structure of noun, adjective and adverbial phrases can be simple or complex. If the structure is a simple one, there is no other way of identification than listing all possible nouns, adjectives and adverbs and then comparing the words with this list. However, if the structure is complex, the structural methods mentioned By Tabatabaee [55] will be used. In this paper, the method recommended by Sadeghi and Zandi-Moghadam [56] for dividing the words is followed. For instance, an adjective like "زیبا" (beautiful) has a simple structure and it should be put in a list to be identified; but a noun like "کارگر" (worker) has a complex structure and it consists of "کار" and "گر" whereas "کار" is a noun and "گر" is a noun suffix. Some of the composing structures of noun, adverb and adjective are as follows:

Noun: (noun±گر) / (noun±بان) / (noun±کار) ...
Adjective: (noun±s) / (noun±ناک) / (noun±وار) ...
Adverb: (adjective±انه) / (بلا+noun) ...

There are some nouns, adverbs and adjectives that are neither simple nor complex, so they cannot be categorized in these structures. A complete list of those words based on Tabatabaee [55] is made and used as well. For labeling, when come across a word, first the lists of simple and irregular nouns, adverbs and adjectives are searched for that word. If the word is found in those lists, it is labeled accordingly. Otherwise, its composing structure is studied. If it is consistent with

Table 1. The number of possible applicable n-grams used in each database.

n-gram	Number		
	Possible applicable features	Selected features in *University* database	Selected features in *Writers* database
Uni-gram	72	72	72
Bi-gram	$72 \times 72 = 5184$	375	71
Tri-gram	$72 \times 72 \times 72 = 373248$	560	117
Forth-gram	$72 \times 72 \times 72 \times 72 = 26873856$	257	97
Total	27252360	1264	357

one of the known and collected structures for nouns, adverbs and adjectives, then it will be labeled appropriately.

Function Words

In Persian language, a set of 922 words are introduced by Davarpanah et al. [57] as the function words which include the most repetitious pronouns, verbs, conjunctions, prepositions and articles and they are used in this paper. The number of occurrences of each type of function words, "است" (is), "من" (me/I) and ... in each text is found and used as a feature.

Punctuation Characters

The number of punctuation characters in the text is one of the syntactic features that can differentiate people's writing styles. In Persian language, the number of usage of punctuation marks, such as "?" and "!", by each author in each text is extracted and utilized.

5.1.3 Semantic Features

The way authors use conjunctions can differentiate the texts written by them. According to classification of Persian conjunctive adjuncts and based on the theoretical framework of Hildi and Mason [37], the adjuncts are categorized in three groups. In this study, based on this classification, the type of adjuncts in each text was determined and the number of occurrences of each was calculated and the results were used as semantic features. In any case, to have more prepositions, each group was expanded by adding the synonyms so that each preposition would fit in one of the categories.

5.1.4 Application-Specific Features

Considering the type of texts in the available databases, including books and articles, some of the application-specific features used in this paper, were the distribution of tab, enter, space and line feed characters in the text.

5.1.5 Summary of Extracted Features

More than 1500 features introduced in Sections 3–5 along with the number of their occurrences collected from the databases, are listed in Table 2 (Features extracted from texts).

5.2 Pre-Processing of Information Collected from Texts

Data pre-processing includes all conversions that are done on the preliminary data and changing them to a form simpler and more effective for the next processes such as classification.

There are different tools and methods for data pre-processing for classification such as normalization, which converts data to a new one with suitable changing of confine and distribution, and dimension reduction, which is used for elimination of the repetitive, overflow or irreverent data [58, 59]. Features with larger values have greater effect on the classification cost function, although it does not necessarily mean they are more important, so it is considered a diverse effect. In this study the normalization and correlation reduction methods introduced in the next section, are used.

5.2.1 Normalization of Features' Values

Normalization maps data to $[-1, 1]$ confine through a linear conversion. For this purpose, if μ_i is the estimation of the average of the i^{th} feature and σ_i is the variance estimation of that feature on n samples,

Table 2. Features extracted from the texts.

Features	Sub Features	Number of Used Features
Lexical	Lexical Richness	7
	n-gram	Depends on text's length and number of authors
	average length or words	1
	average number of characters in sentences	1
	average number of words in sentences	6
	number of short words with 3 characters or less	3
	distribution of alphabet characters in texts	72
	distribution of Persian characters in texts	1
	distribution of the number of paragraphs in texts	1
	distribution of words with the length of 1 to 30 in texts	30
	distribution of English characters in texts	1
	distribution of numerical characters in texts	1
	distribution of half space character in texts	1
Syntactic	distribution of occurrences of noun, adjective and adverbial phrases	4
	distribution of function words	922
	distribution of punctuation characters	1
Semantic	distribution of adjuncts in Persian	36
Application-Specific	distribution of tab character	1
	distribution of enter character	1
	distribution of space character	1
	distribution of empty lines	1

then the i^{th} feature is normalized [60] using

$$y_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{8}$$

where the mean of converted data in each dimension (feature) is 0 and its variance is 1.

When normalizing the data, it should be taken into consideration that the data extracted from the training texts (the classification design data) and the data extracted from the test texts must be normalized through the same method. Also, the normalization of test data should be done with the mean and variance resulted from the training data.

The dimension of the data must be reduced after normalization. The reason for this dimension reduction can be considered as simplification of the next analyses, performance improvement of the classification methods functionalities based on better display of features in feature space, elimination of repetitive information or an attempt for detection of basic structure by obtaining the graphical display of the data.

5.2.2 Eliminating Highly Correlated Features

If one or more features have a high correlation with another feature, they will be considered repetitive. A

subset of features is considered suitable if they have a low correlation. In this study, in order to reduce the correlation and thus reducing the repetitive features, the correlations were calculated and one of two features that had a correlation equal or more than 95 % with each other was eliminated.

5.3 Databases

Author identification is often used when the author of an anonymous text is searched in a small set of authors. For example, to identify the poet of a poem that is supposed to belong to Hafez.

A suitable training and test sets for an author identification case shall be studied with regard to type and subject so that the identity of the author is the only or the most important dividing factor of the texts [61]. Ideally, all the texts of the training set shall be exactly related to one subject; however there are very few sets with this characteristic. Age, education and nationality are other factors that should be closely studied in preparation of ideal evaluation sets. This would reduce the probability of selection of the style of a wide group of people to the writing style of the desired author. In addition, all the texts must be written during the same period of time so that the probability of style changes is eliminated [62]. Due to the lack of a standard database [61] for Persian writers that can meet these requirements, a database from Bu-Ali Sina University (*University* for abbreviation) was constructed. To prepare the database, texts with the exact length of 2009 words, (1500 words for training and 509 words for test,) were collected from 20 junior undergraduate engineering students from the university on a specific subject.

Since the texts of the *University* database were informal, (this would result in inaccuracy of many features such as the distribution of function words,) and also because these texts were short (this would reduce the validity of some measuring features [63], another database, called *contemporary writers* database, consisted of texts from books and articles written by eight recent Persian authors (Ali Ashraf Sadeghi, Mohammad Ali Forughi, Mojtaba Minovi, Abolhassan Najafi, Mohammad Amin Riahi, Ahmad Samiee, Fatholah Mojtabaee and Hussein Masumi Hamedani) on subjects in literary field and literary text analysis was constructed (*Writers* for abbreviation). For each writer, two documents were collected, with at most 7750 words from which 5000 words were assigned to training and the remaining for test.

According to the profile-based method, in this paper, for each author 70 % of the texts are selected for training and 30 % for test. In another evaluation, using a sample-based method, each writer's profile is randomly divided into K sub-samples. To make each sub-sample, that is a set of words, the value of K was increased to such extent that the pieces of the text which are used for training and testing are able to demonstrate the stylistic features. The value of K can be obtained by experience. From this number of sub-samples, one is used for test and $K-1$ sub-samples are used to train the classification models. Then the whole process is repeated K times, so that each one of these K samples are used only once for the test. Then the average of all outputs resulted from K tests is calculated. This method is called K-fold cross validation. In this study both methods are used for evaluation.

6 Evaluation and Comparison of Machine Learning Methods in Persian Author Identification

In this section, the results of implementation of author identification system in Persian language are presented and in the end, the results are compared with one another.

Table 3 indicates the precision of 5 classification methods in identification of the author of an anonymous text on each database separately.

As mentioned before, *University* database would not provide much precision due to the texts' short length and informality of the texts. This result can be observed in all classification methods. As shown in Table 3, KNN method, is displaying high precisions on both databases. As expected and shown in Table 3, the statistical method of Delta, which is specifically developed for author identification applications, provides acceptable results. However since it uses only function word feature which is a syntactic feature, it generates less precision in comparison with KNN which uses all features. Since Decision tree method trims the branches in order to prevent heightening and therefore eliminates features important in the decision making process, it does not provide a suitable precision in comparison with other methods, as indicated in Table 3.

As presented in Table 3, the precision of Neural Network method in text author identification is much lower than others. This can be due to the small number of training texts for each writer and large feature dimension of samples which can result in a phenomenon named Curse of Dimensionality. Thus, in applications such as the present study, where the number and the length of the training texts are small and the number of extracted features for each text is large, using the neural network method as the classifying method is not recommended. As it can be seen in Ta-

Table 3. Comparison of Classification Methods on two Databases.

Database	**Method**				
	KNN	Delta	Decision Tree	Neural Network	LDA
University	70 %	50 %	21.6 %	15 %	81.6 %
Writers	100 %	87 %	57.5 %	37.3 %	100 %

Table 4. Comparison of the average precision of classification methods versus features types on both databases.

Classification Methods	**Feature**			
	Lexical and Characteristic	Syntactic	Semantic	Application-Specific
KNN	80 %	66.5 %	49.5 %	49.5 %
Decision Tree	25 %	21.88 %	18.75 %	12.5 %
Neural Network	11 %	12.5 %	16 %	8.75 %
LDA	38.12 %	29.37 %	30 %	39.37 %
Average Precision	38.53 %	32.56 %	28.56 %	27.53 %

ble 3, LDA method which uses all kinds of features is showing high precisions on both databases. The reason can be the high linear coherency of the extracted features of the texts. The studies conducted in this paper suggest that this method generates better result comparing to other methods if only this kind of features is eliminated from it.

The features in the root of decision tree which was implemented on our databases included distribution of adverbs in the texts which is a syntactic feature, the average of character distribution in the sentences and the number of desired bi-gram occurrences in the texts that are the lexical features. The level one features included the feature of distribution of adjective phrases from syntactic features and distribution of desired conjunctions in texts which is a semantic feature. The selected feature for the second level is also distribution of the punctuation characters, which is a syntactic feature. This concludes that the mentioned features have the most separating ability among all features and can be used as the important features in author identification applications.

In order to study the effect and the role of different types of features in author identification more closely, in another implementation, the mean of precision of all kinds of features in author identification in two databases of *Writers* and *University* are calculated and categorized based on the classification methods. The results are shown in rows 3 to 5 of Table 4 (Comparison of the average precision of classification methods versus feature types on both databases). Delta method that only uses function words is not addressed in this table. The last row of the table presents the mean of the precision of all four classification methods on all features. This table expresses that the set of lexical features has the higher precision and greater separating ability than the other sets. Table 3 also indicates that the next features in the ranking belong to the syntactic, semantic and application-specific features respectively. The result of this comparison is consistent with the result obtained from the decision tree as well.

This is also noteworthy that it is possible that one feature out of one “kind” of features has a better separating quality compared to the kind of feature that it belongs to. Thus, the precision of separating quality of a type of feature being higher does not mean every feature in that group would have a higher precision. Therefore, it is possible that a feature belonging to the syntactic features group has a higher precision in author identification in comparison to other features. The most discriminant features in the collected databases in Persian language include the distribution of adverbs and adjectives in the texts, the average of distribution of characters in sentences, the number of desired bi-gram occurrences in the text, distribution of conjunctions and distribution of punctuation characters in the text.

7 Discussion and Conclusion

In this study, in addition to design and implementing a system for automatic Persian author identification of an anonymous text, a comparison study is conducted on machine learning methods for identification of a Persian author using two databases.

In this research, five classification methods including K-Nearest Neighbors (KNN), Delta, Linear Discriminant Analysis, Decision Tree and Neural Networks are implemented and compared. The results show that implementation of LDA method on the databases, results in more precise results compared to other methods.

The results of implementation of these methods show that the shorter the training and test texts are, the less accurate the author identification will be. However, for short texts, the extracted features shall be selected carefully in order to present the writing style of the author [64]. Thus, texts from *University* database do not generate acceptable precision due to the texts' short length and being informal.

As mentioned in Section 6, the most discriminant features in the collected databases in Persian language include the distribution of adverbs and adjectives in the texts (the output of the software designed for this paper to recognize nouns, adverbs and adjectives in the text), the average of distribution of characters in sentences, the number of desired bi-gram occurrences in the text, distribution of conjunctions and distribution of punctuation characters in the text.

The success rate of the system designed in this study on Persian literature is 90 % in average. Considering the fact that the current study is the first attempt for presenting an automatic/machine learning based method for identification of an anonymous Persian author, the obtained precision seems to be quite promising and encouraging.

Due to lack of accurate available processing tools for Persian language, we encountered some problems in the author identification and thus we designed and implemented a number of these tools such as a POS tagger. However, some other tools were not available and so we were not able to use them. The usage of those tools would definitely increase the precision of the system.

References

[1] F Mosteller and D L Wallace. *Inference and Disputed Authorship: The Federalist*. Springer-Verlag, New York, 1964.

[2] C E Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005.

[3] J Li, R Zheng, and H Chen. From fingerprint to writeprint. *Communications of the ACM*, 49(4): 76–82, 2006.

[4] H Van Halteren, H Baayen, F Tweedie, M Haverkort, and A Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12 (1):65–77, 2005.

[5] G U Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390, 1938.

[6] G U Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.

[7] G Kingsley Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, 1932. ISBN 9780674432048.

[8] D I Holmes. Authorship attribution. *Computers and the Humanities*, 28(2):87–106, 1994.

[9] D I Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[10] J Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1997.

[11] J Burrows. 'delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

[12] D L Hoover. Delta prime? *Literary and Linguistic Computing*, 19(4):477–495, 2004.

[13] A Abbasi and H Chen. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75, 2005.

[14] T D Grant. Quantifying evidence for forensic authorship analysis. *International Journal of Speech, Language and the Law*, 14(1):1–25, 2007.

[15] G Frantzeskou, E Stamatatos, S Gritzalis, and S Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, pages 893–896. ACM, 2006.

[16] J F Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2(2):61–70, 1987.

[17] D I Holmes. A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 91–120, 1992.

[18] D Lowe and R Matthews. Shakespeare vs. fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, 29(6): 449–461, 1995.

[19] F J Tweedie, S Singh, and D I Holmes. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30 (1):1–10, 1996.

[20] O De Vel. Mining e-mail authorship. In *In Proceeding of KDD-2000 Workshop on Text Mining, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[21] O De Vel, A Anderson, M Corney, and G Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[22] F Peng, D Schuurmans, Sh Wang, and V Keselj. Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 267–274, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. ISBN 1-333-56789-0.

[23] G Tambouratzis, S Markantonatou, N Hairetakis, M Vassiliou, G Carayannis, and D Tambouratzis. Discriminating the registers and styles in the modern greek language-part 2: Extending the feature vector to optimize author discrimination. *Literary and Linguistic Computing*, 19(2):221–242, 2004.

[24] E Stamatatos, N Fakotakis, and G Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[25] E Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2): 790–799, 2008.

[26] 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*, 57(3):378–393, 2006.

[27] E Stamatatos, N Fakotakis, and G Kokkinakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4): 471–495, 2000.

[28] S Fabrizio. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[29] J Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[30] W B Cavnar and J M Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.

[31] R S Forsyth and D I Holmes. Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[32] J Houvardas and E Stamatatos. N-gram feature selection for authorship identification. In *AIMSA*, volume 4183 of *Lecture Notes in Computer Science*, pages 77–86. Springer, 2006. ISBN 3-540-40930-0.

[33] H Baayen, H Van Halteren, and F Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[34] Sh Argamon and Sh Levitan. Measuring the usefulness of function words for authorship attribution. In *In Proceedings of ACH/ALLC Conference 2005*, 2005.

[35] J F Burrows. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7 (2):91–109, 1992.

[36] Sh Argamon, C Whitelaw, P Chase, S R Hota, N Garg, and Sh Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

[37] A Jafari. The study of persian adjuncts based on functional and formal approaches. *Especial Issue of Name-ye Farhangestan*, 5:128–156, 2008.

[38] Ph M McCarthy, G A Lewis, D F Dufty, and D S McNamara. Analyzing writing styles with cohmetrix. In *FLAIRS Conference*, pages 764–769. AAAI Press, 2006.

[39] Ch Fellbaum. Wordnet: An electronic lexical database, 1998.

[40] S Deerwester, S T Dumais, G W Furnas, T K Landauer, and R Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[41] G-F Teng, M-Sh Lai, J-B Ma, and Y Li. E-mail authorship mining based on svm for computer forensic. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 2, pages 1204–1207. IEEE, 2004.

[42] K Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990.

[43] Sh Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

[44] F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[45] Th V N Merriam and R A J Matthews. Neural computation in stylometry ii: An application to

the works of shakespeare and marlowe. *Literary and Linguistic Computing*, 9(1):1–6, 1994.
[46] R A J Matthews and Th V N Merriam. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguistic Computing*, 8(4):203–209, 1993.
[47] J R Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
[48] Ö Uzuner and B Katz. A comparative study of language models for book and author recognition. In *IJCNLP*, volume 3651 of *Lecture Notes in Computer Science*, pages 969–980. Springer, 2005. ISBN 3-540-29172-5.
[49] S Balakrishnama and A Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and Information Processing*, 1998.
[50] R Ramezani, N Sheydaei, and M Kahani. Evaluating the effects of textual features on authorship attribution accuracy. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, pages 108–113. IEEE, 2013.
[51] F J Tweedie and R H Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.
[52] A Honoré. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.
[53] H S Sichel. On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351a):542–547, 1975.
[54] E H Simpson. Measurement of diversity. *Nature*, 163(4148):688, April 1949.
[55] A Tabatabaee. Word formation and grammatical category: The identification of grammatical categories of persian words, based on morphological characteristics, 2009.
[56] A A Sadeghi and Z Zandi-Moghadam. *Orthography of Persian Script, Based on Directions from Academy of Persian Language and Literature*. Academy of Persian Language and Literature, Tehran, Iran, 2008.
[57] M R Davarpanah, M Sanji, and M Aramideh. Farsi lexical analysis and stop word list. *Journal of Library Hi Tech*, 27:435–449, 2009.
[58] H T Bao. Introduction to knowledge discovery and data mining. Hanoi, Vietnam, 2002. Institute of Information Technology National Center for Natural Science and technology.
[59] S M M Takamy. Data preprocessing and pattern recognition methods. Technical Report KNTU-AADCL-05, Khaje Nasir Toosi University, 2005.
[60] S Theodoridis and K Koutroumbas. *Pattern Recognition*. Academic Press, New York, USA, 1999.
[61] E Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
[62] F Can and J M Patton. Change of writing style with time. *Computers and the Humanities*, 38 (1):61–82, 2004.
[63] K Luyckx and W Daelemans. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1): 35–55, 2011.
[64] G Hirst and O Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

Zeinab Farahmandpour received her BSc and MSc degrees in Software Engineering and Artificial Intelligence Engineering both from University of Bu-Ali Sina, Hamedan, Iran, in 2009 and 2012, respectively. She used to be lecturing at University of Bu-Ali Sina, University College of Omran and Toseeh and Azad University (Parand branch) for several years. Her main research interests are data mining, software and network security, artificial intelligence and natural language processing.

Hooman Nikmehr received his BSc in Electronic Engineering and MSc in Computer Architecture Engineering both from University of Tehran, Tehran, Iran, in 1992 and 1997, respectively, and PhD degree in Computer Engineering from the University of Adelaide, Adelaide, Australia, in 2005. He is an Assistant Professor with the Department of Computer Architecture, University of Isfahan, Isfahan, Iran. His current research interests include VLSI, digital arithmetic, computer architecture, reconfigurable hardware design and low-power design.