Evaluating the Effects of Textual Features on Authorship Attribution Accuracy

Reza Ramezani
Department of Computer Engineering
Ferdowsi University of Mashhad, Iran
DDEmS Lab
reza.ramezani@stu.um.ac.ir

Navid Sheydaei
Electrical & Computer Engineering
Isfahan University of Technology, Iran
DM Lab
n.sheydaei@ec.iut.ac.ir

Mohsen Kahani
Department of Computer Engineering
Ferdowsi University of Mashhad, Iran
WT Lab
kahani@um.ac.ir

***Abstract*- Authorship attribution (AA) or author identification refers to the problem of identifying the author of an unseen text. From the machine learning point of view, AA can be viewed as a multiclass, single-label text-categorization task. This task is based on this assumption that the author of an unseen text can be discriminated by comparing some textual features extracted from that unseen text with those of texts with known authors. In this paper the effects of 29 different textual features on the accuracy of author identification on Persian corpora in 30 different scenarios are evaluated. Several classification algorithms have been used on corpora with 2, 5, 10, 20 and 40 different authors and a comparison is performed. The evaluation results show that the information about the used words and verbs are the most reliable criteria for AA tasks and also NLP based features are more reliable than BOW based features.**

Keywords- Authorship Attribution, Author Identification, Textual Features, Persian Corpus, Data Mining, Classification.

I. INTRODUCTION

In the problem of authorship attribution (AA), a text with unknown author is assigned to one of the candidate authors. Each candidate author possesses a set of quantitative textual features that denotes his/her unconscious writing styles. From the machine learning point of view, AA is a multiclass, single-label text-categorization task. This problem is supported by statistical or computational methods and can utilize the results of researches done in natural language processing, information retrieval, and machine learning areas [1]. Some research topics, such as author verification, author profiling or characterization, detection of stylistic inconsistencies and plagiarism detection can be defined as special cases of AA problem [2].

AA is based on text representation. Before text documents being processed by machine learning algorithms, they should be converted to vectors of numbers or metrics of quantitative textual features in some ways. This is called text representation and is essential for text categorization. Vector space model (VSM) is the most widely used representation model in AA tasks that was first proposed by [3]. In this model, documents are represented as vectors in the space of features in a way that the vectors' entries contain some numeric values of one or more textual features. Vector construction that is known as *authorship attribution* is built in two steps: selecting a textual feature as the measurement and comparison criterion and assigning numeric values to the selected textual feature in different states. Finally machine learning algorithms perform the text-categorization task using these quantitative vectors.

The accuracy of the text-categorization task highly depends on the constructed quantitative vectors. As it was mentioned before, in VSM, textual features construct vectors' values, hence the selected textual feature and its assigned weights have a great effect on the author identification accuracy. Selecting a textual feature for quantifying writing style, also is known as *stylometry* or *style markers* [4]. Sentence length, word and character frequencies, the number of verbs and punctuation marks in the sentences and different styles of the used verbs are examples of textual features. By the late 1997 about 1,000 different textual features have been proposed in [5].

Textual features are categorized into three main groups:

- *Lexical and Character Features*: a text is considered as a mere sequence of word-tokens or characters. These features are also known as *Bag of Words (BOW)* [6] and are almost independent of the host natural language.
- *Syntactic and Semantic Features*: to extract these features, deep linguistic analyses and complex NLP techniques are required and they involve methods that highly depend on the host natural language.
- *Application-Specific Features*: these features can be defined only in the certain languages or text domains (Such as e-mail messages and online-forum messages) and are not general purposes.

In this paper, a set of 29 different textual features (belonging to the first and the second groups of textual features) along with their impacts on the accuracy of text categorization for the AA task on some Persian corpora are investigated. Most of the employed textual features look at documents as BOW, which is the most widely used text representation technique in prior researches [7]. Other textual features need some NLP techniques to be extracted. The investigated features are extracted from text documents, stemmed documents, or tagged documents and finally are experimented on 5 Persian corpora containing text documents of 2, 5, 10, 20 and 40 famous and contemporary Iranian writers. These experiments are done by SVM, K-NN and C5 classification algorithms. The obtained results show how different textual features, in different situations, affect the AA accuracy.

The organization of the paper is as follows. Section 2 investigates several textual features to be evaluated. Section 3 discusses two different kinds of learning processes employed in the experiments. Section 4 introduces the used corpus and investigates their attributes. Section 5 describes different situations of the performed experiments and shows the results of applying three different classification algorithms to different textual features. Finally Section 6 concludes the paper and presents future work.

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

II. TEXTUAL FEATURES

In order to conclude which author is the writer of an anonymous document, a text-categorization algorithm compares the values of textual features of the anonymous document with the values of the corresponding features of the candidate authors' documents. As it was mentioned earlier, different textual features have different effects on the accuracy of the AA task. To do a fair comparison among the selected textual features in quantitative AA, it's necessary that they are extracted from identical datasets in different situations and are evaluated by more than one classification algorithm. This section investigates some textual features that their effects on AA accuracy on Persian corpora will be evaluated in next sections.

In this research, written documents have three different aspects as *Text*, *Stem* and *Tag*. Documents related to *Stem* and *Tag* aspects need NLP techniques to be constructed.

- *Text*: this aspect contains normal written documents of authors, without any changes.
- *Stem*: this aspect contains written documents except that all tokens' value are stemmed.
- *Tag*: this aspect contains written documents except that all tokens' value are replaced with their type (e.g. verb, adjective, adverb, noun, conjunctive, pronoun and etc.).

The proposed textual features could be extracted from one, two or all aspects of the written documents set. These textual features are tabulated in Table 1:

TABLE 1 - SOME TEXTUAL FEATURES

Feature Name	Feature Description
Words Length	As the first measurement criterion, the words length is used. This feature indicates the number of words with different lengths. For example how many of words have length 2, how many of words have length 3 and so on. This feature leads to represent a document as a numeric vector in a way that its entries indicate the number of words with special lengths. This feature is extracted only from *Text* documents.
Words Frequency	Based on this assumption that authors would tend to use special words in their writing, words frequency could be employed as a measurement criterion for AA tasks. This feature as the most widely used feature in the researches, leads to represent a document as a numeric vector in a way that its entries correspond to words and indicate the number of occurrences of the corresponding words in the document. This feature is extracted from *Text*, *Stemmed* and *Tagged* documents.
Characters Frequency	This Feature is similar to the *Words Frequency* except that the entries of the constructed vectors correspond to alphabetic characters and indicate the number of occurrences of the corresponding characters in the document. This feature is extracted only from *Text* documents.
Sentences Length	This feature is similar to the *Words Length* except instead of length of words, the length of sentences is counted. The length of a sentence is counted in two different cases: the number of words in the sentence and the number of characters in the sentence. This feature is extracted only from *Text* documents.
Verbs Frequency (Value)	This feature is similar to the *Words Frequency*, but based on this assumption that authors tend to use special verbs in their writing, instead of frequency of all words, the frequency of verbs is only counted. This feature needs NLP techniques to be extracted and is extracted from *Text* and *Stemmed* documents.
Verbs Count in Sentence	Some authors tend to use a few verbs in each sentence and in fact they would rather to use small sentences. In contrast some other authors prefer to use sentences with many verbs. Hence similar to the *Sentence Length*, the number of verbs in sentences could be used as a measurement criterion for AA. This feature needs NLP techniques to be extracted and is extracted only from *Text* documents.
Commas Count in Sentence	This feature is similar to the *Verbs Count in Sentence*, except instead of the number of verbs, the number of commas in the sentences is counted. This feature is based on this assumption that some authors tend to use integrated sentences and some others tend to use fragmented sentences using commas. This feature is extracted only from *Text* documents.
N-Gram Word	The *Words Frequency* feature looks at the written documents as a bag of words and never consider the order and the collocation of words that may lead to noisy calculation. To decrease this noise, the collocation order of *N* words could be taken into account which is called *N-Gram Word* and is based on this assumption that the authors usually tend to use special words together. This feature is extracted only from *Text* documents
N-Gram Character	This feature is similar to the *N-Gram Word* except instead of words, the collocation and the order of *N* characters is considered. This feature is extracted only from *Text* documents.
Verbs Info (Structure)	Applying complex NLP techniques to written documents leads to extract sophisticated textual features which would be beneficial for AA. *Verbs Info* is one of such features that tries to identify the structure of the verbs. After analyzing the used verbs, four different structures of verbs were extracted which can be used as textual feature individually. These structures are as follows: • *Verbs Type*: indicates the type of the used verbs. Six different verbs types are considered, such as simple verbs, auxiliary verbs, imperative verbs, subjunctive verbs and etc. • *Verbs Number*: determines whether the used verb is singular or is plural. • *Verbs Person*: indicates the person facet of the used verbs. These facets are as follows: 1st person singular, 1st person plural, 2nd person singular, 2nd person plural, 3rd person singular, and 3rd person plural. • *Verbs Tense Mood*: indicates the tense mood of the used verbs. Simple_Past, Past_Perfect, Past_Continues, Present_Prefect, Future and etc. are examples of tense moods. In these features only *Verbs Type* is extracted from *Text* and *Tagged* documents and the others are

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

	extracted only from *Text* documents.
Adjectives Info	Similar to the *Verbs Info*, complex NLP techniques could be employed to extract the structure of the used adjectives. Five different tags were used to specify the type of an adjective. For example simple/positive adjective, comparative adjective, superlative adjective, order adjective and etc. This feature is extracted from *Text* and *Tagged* documents.
Adverbs Info	As well as *Adjectives Info*, using information about the used adverbs would help to identify the author of a written document. Five different tags were used to specify the type of an adverb. For example mood, time, place, cause, degree and etc. This feature is extracted from *Text* and *Tagged* documents.
Sentence Start Token	Based on this assumption that authors usually tend to start sentences with special words or special token types (e.g. noun, conjunctive, pronoun and etc.), employing information about the first word/token of sentences would be useful for AA. This feature is extracted from *Text* and *Tagged* documents.

III. Learning Process

As it was mentioned earlier, AA task can be viewed as a common classification problem and hence it is done in two different phases:

At the first phase, a textual feature is selected as the measurement criterion and then some numeric values extracted from documents with known authors are assigned to the textual feature (these documents are called *training set*). This phase is referred to *training phase*.

In this work, documents are represented as quantitative vectors in which vectors' entries contain numeric values corresponding to different situations of a textual feature. For example if *Words Length* is used as measurement criterion, the first entry of a vector would contain the number of words with length 2, the second entry would contain the number of words with length 3, the third entry would contain the number of words with length 4 and so on. Finally these vectors are used by a classification algorithm at the second phase.

At the second phase, a classification algorithm after receiving the feature vectors of some documents with unknown authors (*test set*), predicts the most likely author of them by comparing the test set's feature vectors with those of documents in training set.

In this paper, in order to evaluate the effects of different textual features on AA accuracy in different situations, the process of feature extraction and weighting, and author identification, is done in two different manners as follows:

A. Integrated Training Set and Test Set

In this manner, training set and test set are integrated and in fact there is no test set. In this approach after selecting a textual feature as measurement criterion, all documents are represented as numeric vectors of the selected feature in training phase in a way that vectors' entries indicate different values of the selected feature (e.g. the number of different lengths of words). Then by n-fold technique [8] some documents' feature vectors are used as training set and the others are used as test set. This process is repeated until each document's feature vector is used in the test set at least once. In each iteration, a classification algorithm is applied to the training set and the test set feature vectors and the classification accuracy is returned as the result. Finally the average of the acquired accuracies of all iterations is considered as AA accuracy.

Using *FF-IFF* (feature frequency-inverse feature frequency) improves the accuracies acquired by the *Integrated Training Set and Test Set* approach significantly. *FF-IFF* is an adaptation of the TF-IDF model which is widely used for document categorization [9]. In this scheme a vector based representation is used where the value of each entry is given by the *FF-IFF* of the corresponding feature value.

If f is a feature and f_i is feature f with value i (e.g. f_i denotes words with length i), then *FF-IFF* is calculated by (1):

$$FF - IFF(f_i, d) = FF(f_i, d) * IDF(f_i, A_d) \quad (1)$$

The frequency of feature f with value i in document d, is denoted by $FF(f_i, d)$ which is the number of times that feature f_i occurs in document d. The higher value of $FF(f_i, d)$, indicates the more the feature f with value i is representative of document d (e.g. if words length is used as textual feature and $FF(f_4, d)$ has the maximum value, it indicates that the author of document d more uses words with length 4).

If A_d is the *author* of document d and D is the collection of all documents, the inverse feature frequency of a feature f_i, denoted $IFF(f_i, A_d)$ is given by (2):

$$IFF(f_i, A_d) = \log\left\{\left(\sum_{c \in D} FF(f_i, c)\right) \div \left(1 + \sum_{c \in D \text{ and } A_c \neq A_d} FF(f_i, c)\right)\right\} \quad (2)$$

That is $IFF(f_i, A_d)$ is calculated by taking logarithm from the total number of times that feature f_i occurs in all documents divided by the total number of times that feature f_i occurs in documents that their *author* is not A_d.

Hence, the $IFF(f_i, A_d)$ of feature f with value i for author A_d is low if f_i is used by many authors, indicating that this feature has little author discriminating power. On the other hand, the $IFF(f_i, A_d)$ of feature f with value i for author A_d is high if f_i is used by few authors, indicating that the feature has a great author discriminating power. Of course, features with a high *FF* and a high *IFF* are desirable to be used in AA tasks.

Integrated training set and test set approach, based on the textual feature f, represents each document d as vector $DV(d, f)$ as defined in (3):

$$DV(d, f) = \{FF - IFF(f_i, d) \mid i \in \mathbb{N}\} \quad (3)$$

B. Disjointed Training Set and Test Set

In this manner which compared to integrated training set and test set approach is nearer to real world applications, some documents are used as training set and the rest are used as test set, and the training phase is applied only to the training set. That is, in contrast to previous manner, the feature vectors' values of the test set are not involved in the training phase. In this approach because the test set is disjointed from the train set and the system supposes that the test set's documents have unknown authors and hence calculating *IFF* is impossible, the feature vectors' values are assigned only using $FF(f_i, d)$, and the value of $IFF(f_i, A_d)$ is not involved in the calculations.

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

Then in the test phase, regardless to the feature vectors' values of training set, for each document of the test set a feature vector is constructed using $FF(f_i, d)$. Finally the feature vectors constructed from both training set and test set are fed to a classification algorithm (such as SVM) and classification accuracy is returned as the result. This process is repeated until each document is used in the test set at least once. Finally the average of the acquired accuracies of all iterations is considered as AA accuracy.

Similar to previous manner, vector construction is based on a textual feature, and feature vectors' entries indicate values corresponding to the textual feature in different counts (e.g. the number of words with different lengths).

Disjointed training set and test set approach based on the textual feature f uses only $FF(f_i, d)$ and represents each document d as vector $DV(d, f)$ as defined in (4):

$$DV(d, f) = \{FF(f_i, d) \mid i \in \mathbb{N}\} \quad (4)$$

IV. Datasets

In order to evaluate the effects of textual features on AA accuracy in Persian language, several Persian written documents have been used. These documents are categorized into 5 corpora which differ in the number of authors and include literatures of 2, 5, 10, 20 and 40 famous and contemporary Iranian writers, such as: *Bozorg Alavi*[1], *Jalal Ale-Ahmad*[2], *Sadegh Hedayat*[3], *Sadeq Chubak*[4], *Forough Farrokhzad*[5] and etc. The literatures have different themes, such as romance, literary stories, social stories, satire stories and etc.

Each author has 5 documents that are used in training set or test set. The experiments we have conducted using these corpora are similar to those reported in [10]. In order to remove the side effects of imbalance classes, the size of documents concerned to authors are proportioned in a way that each document contains at least 800 and at most 1000 words.

Moreover, to evaluate the effects of some complex textual features, documents have been tagged using some NLP tools, such as *Gate* [11] and also have been stemmed by the *word and verb reduction* technique proposed by [12].

V. Experiments and Results

As it was mentioned before in Section 3, the learning process was done in two different manners: *Integrated* and *Disjointed* which differ in the involvement or not involvement of the test set in the training phase and the formula used for calculating feature vectors' values. In the constructed vectors, only 10 entries with the highest value were used and the rest entries were discarded.

In order to evaluate the effects of different textual features 30 different experiments were performed. In each experiment 29 textual features were used which are extracted from text, stemmed, or tagged documents. The accuracy of AA affected by each textual feature is calculated by three different classification algorithms as SVM, K-NN and C5.

All of the results presented here take the form of textual features and attribution algorithms accuracy table. Each table shows the results of subjecting 29 textual features to multiple experiments. The experiments differ in the used classification algorithm, in the number of possible authors per permutation and in the learning process. The obtained AA accuracy in each scenario shows the percentage of texts correctly attributed.

The first column of the presented tables identifies the tested textual features as mentioned in Section 2. The numerical values presented in other columns denote the accuracy of AA task applied by three different classification algorithms to five corpora with different authors count in two different learning processes. Finally in order to show better the effects of textual features and decrease the effects of the used classification algorithms on the obtained accuracies, for each different AA scenario, an average of the obtained accuracies of different classification algorithms are taken and presented.

Table 2 shows the results obtained by applying three different classification algorithms to two different corpora with 2 and 5 authors by two learning processes *Integrated* and *Disjointed* along with the average of the obtained accuracies. These four different combinations of corpora and learning processes are shown with different colors. Table 3 belongs only to the *Integrated* manner and depicts the accuracies of applying three different classification algorithms to corpora with 10, 20 and 40 authors and the average of the obtained accuracies.

Table 4 is similar to Table 3 except that its results are related to the *Disjointed* manner. In all tables, the maximum accuracies obtained by each scenario are **bolded**.

In order to perform a deeper analysis and identifying the most effective textual features, in each scenario, three textual features with the highest accuracy are selected and finally the number of times that each textual feature is selected as the most effective textual feature is counted. This counting was done in three cases: on all corpora and with *Integrated* manner, on all corpora and with *Disjointed* manner, and on all corpora with both manners (overall case). 10 of the most effective textual features along with the number of their effectiveness are depicted in Figure 1. Five features of these 10 features are *Lexical and Character Features (BOW)* that in contrast to 5 other features don't need NLP techniques to be extracted. Most of these 10 features are concerned with *Words* and *Verbs* information. As these results show, *Words Frequency on Tags* and *2-Gram Characters* have the most author discriminating power.

Besides as Figure 2 shows, in overall case and also in the *Integrated* manner, *BOW* textual features have more author discriminating power. But in the *Disjointed* manner which is nearer to real-world applications, textual features extracted by NLP techniques have more author discriminating power than *BOW* features.

VI. Conclusions and Future Work

This paper presents the results of evaluating the effects of a wide range of textual features on AA accuracy on a large and carefully constructed corpora of some Persian literatures. For the first time in the history of quantitative AA, Iranian inspectors now have access to reliable data about which of the presented textual features are the most useful for author identification task in Persian language.

In order to achieve reliable results about the effects of different textual features on AA accuracy, experiments were done in 30 different scenarios. These 30 scenarios include applying 3 classification algorithms (SVM, K-NN and C5) to 5 corpora with different authors count (2, 5, 10, 20 and 40

[1] http://en.wikipedia.org/wiki/Bozorg_Alavi
[2] http://en.wikipedia.org/wiki/Jalal_Al-e-Ahmad
[3] http://en.wikipedia.org/wiki/Sadegh_Hedayat
[4] http://en.wikipedia.org/wiki/Sadeq_Chubak
[5] http://en.wikipedia.org/wiki/Forough_Farrokhzad

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

authors) in 2 different learning processes (*Integrated* and *Disjointed*). In each scenario, 29 different textual features were examined.

As Figure 1 shows, information about the used words and verbs (such as frequency or type) can be regarded as the most reliable criteria for AA tasks and the results depicted in Figure 2 indicate that for real-world problems, NLP based features (*Syntactic and Semantic Features*) are more reliable than BOW based features (*Lexical and Character Features*) to be employed. Also the results presented in Table 2 to Table 4 show that in addition to the intrinsic impact of textual features on AA accuracy, other different factors have impact on the acquired results' accuracy too, such as: the number of authors, the number of document per author, the manner used to feature extraction and weighting them, and finally the classification algorithm used for text categorization.

For future work a combination of the proposed features (those that have higher effectiveness) to acquire more accuracy for the AA task, by considering features' scaling problem, is suggested.

TABLE 2 – THE EFFECTS OF DIFFERENT TEXTUAL FEATURES ON BOTH INTEGRATED AND DISJOINTED MANNERS WITH 2 AND 5 AUTHORS

Datasets Info		**Integrated Manner Number of Authors: 2**				**Disjointed Manner Number of Authors: 2**				**Integrated Manner Number of Authors: 5**				**Disjointed Manner Number of Authors: 5**			
Feature	**Source**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**
Words Length	Text	1	1	1	**1**	0.8	1	0.9	0.9	0.75	0.5	0.96	0.74	0.54	0.5	0.76	0.6
Words Frequency	Text	1	1	0.7	0.9	0.7	1	0.5	0.73	**1**	0.33	0.88	0.74	**0.71**	0.5	0.83	**0.68**
Words Frequency	Stem	1	1	0.7	0.9	0.7	1	0.8	0.83	**1**	0.33	0.79	0.71	0.63	0.5	0.79	0.64
Words Frequency	Tag	1	1	1	**1**	0.9	1	**1**	**0.97**	0.91	0.5	0.96	**0.79**	0.54	0.5	**0.92**	**0.65**
Characters Frequency	Text	1	1	1	**1**	0.9	1	**1**	**0.97**	0.63	0.5	**1**	0.71	0.33	0.5	**0.96**	0.6
Sentences Length (Word)	Text	1	1	0.9	0.97	0.7	1	0.8	0.83	0.66	0.5	0.38	0.51	0.25	0.33	0.67	0.42
Sentences Length (Character)	Text	1	1	0.5	0.83	0.7	1	0.5	0.73	0.45	0.33	0.21	0.33	0.38	0.33	0.21	0.31
Verbs Count In Sentence	Text	1	1	1	**1**	0.8	1	0.8	0.87	0.7	0.5	0.83	0.68	0.46	0.33	0.57	0.45
Commas Count In Sentence	Tag	1	1	0.9	0.97	0.8	1	0.7	0.83	0.75	0.5	0.38	0.54	0.5	0.33	0.75	0.53
Verbs Frequency	Text	0.9	1	0.5	0.8	0.8	1	0.8	0.87	0.41	0.33	0.5	0.41	0.54	0.33	0.31	0.39
Verbs Frequency	Text	0.8	1	0.8	0.87	0.6	1	0.9	0.83	0.42	0.5	0.42	0.44	0.25	0.33	0.67	0.42
Verbs Info (Type)	Text	0.9	1	1	0.97	**1**	1	0.7	0.9	0.79	0.5	**1**	0.76	0.5	0.5	0.88	0.63
Verbs Info (Number)	Tag	0.9	1	1	0.97	0.9	1	0.9	**0.93**	0.7	0.5	0.83	0.68	0.21	0.33	0.67	0.4
Verbs Info (Person)	Text	0.9	1	1	0.97	0.6	1	0.7	0.77	0.79	0.5	0.96	0.75	0.46	0.5	0.83	0.6
Verbs Info (Tense Mood)	Text	0.9	1	1	0.97	0.9	1	0.7	0.87	0.75	0.5	0.88	0.71	**0.67**	0.5	0.79	**0.65**
Adjectives Info (Value)	Text	1	1	0.5	0.83	0.7	1	0.5	0.73	0.79	0.5	0.21	0.5	**0.71**	0.33	0.16	0.4
Adjectives Info (Type)	Text	0.7	1	0.9	0.87	0.6	1	0.7	0.77	0.41	0.5	0.71	0.54	0.33	0.5	0.71	0.51
Adverbs Info (Value)	Tag	0.9	1	0.5	0.8	0.5	0.5	0.5	0.5	0.53	0.5	0.21	0.41	0.29	0.17	0.21	0.22
Adverbs Info (Type)	Text	0.9	1	1	0.97	0.7	1	0.8	0.83	0.33	0.5	0.67	0.5	0.38	0.33	0.63	0.44
Sentence Start Token	Tag	1	1	0.5	0.83	0.9	1	0.8	0.9	0.7	0.5	0.21	0.47	0.46	0.5	0.58	0.51
Sentence Start Token	Text	1	1	0.5	0.83	0.8	1	0.8	0.87	0.66	0.5	0.21	0.46	0.5	0.5	0.58	0.53
2-Gram Word	Tag	0.9	1	0.5	0.8	**1**	1	0.5	0.83	0.88	0.33	0.21	0.47	0.38	0.33	0.29	0.33
3-Gram Word	Text	0.8	1	0.5	0.77	**1**	0.5	0.5	0.67	0.41	0.33	0.21	0.32	0.25	0.33	0.21	0.26
4-Gram Word	Text	1	0.5	0.5	0.67	0.9	0.5	0.5	0.63	0.38	0.33	0.21	0.31	0.33	0.33	0.21	0.29
5-Gram Word	Text	1	0.5	0.5	0.67	0.8	0.5	0.5	0.6	0.29	0.33	0.21	0.28	0.17	0.33	0.21	0.24
2-Gram Character	Text	1	1	1	**1**	0.7	1	0.7	0.8	0.95	0.5	0.96	**0.8**	0.25	0.5	**0.92**	0.56
3-Gram Character	Text	1	1	0.8	0.93	**1**	1	0.8	**0.93**	0.88	0.5	0.79	0.72	0.33	0.5	0.88	0.57
4-Gram Character	Text	1	1	0.8	0.93	0.8	1	0.5	0.77	0.75	0.5	0.46	0.57	0.29	0.33	0.38	0.33
5-Gram Character	Text	0.8	1	0.8	0.87	0.5	0.5	0.5	0.5	0.91	0.33	0.33	0.53	0.29	0.17	0.38	0.28

TABLE 3 - THE EFFECTS OF DIFFERENT TEXTUAL FEATURES ON INTEGRATED MANNER WITH 10, 20 AND 40 AUTHORS

Datasets Info		**Integrated Manner Number of Authors: 10**				**Integrated Manner Number of Authors: 20**				**Integrated Manner Number of Authors: 40**			
Feature	**Source**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**
Words Length	Text	0.63	**0.33**	**0.9**	0.62	0.73	**0.25**	0.62	0.53	0.65	**0.13**	0.28	0.35
Words Frequency	Text	0.95	0.17	0.29	0.47	0.73	0.17	0.27	0.39	0.72	0.02	0.13	0.29
Words Frequency	Stem	**0.97**	0.17	0.29	0.47	0.78	0.17	0.26	0.4	0.69	0.02	0.24	0.31
Words Frequency	Tag	0.91	**0.33**	0.88	**0.71**	0.8	**0.34**	**0.77**	**0.64**	**0.76**	0.04	**0.52**	**0.44**
Characters Frequency	Text	0.71	0.25	**0.94**	0.63	0.6	0.23	**0.75**	0.53	0.65	0.06	**0.63**	**0.45**
Sentences Length (Word)	Text	0.61	**0.33**	0.1	0.35	0.45	0.13	0.15	0.24	0.42	**0.08**	0.03	0.18
Sentences Length (Character)	Text	0.6	0.17	0.1	0.29	0.3	0.08	0.16	0.18	0.25	0.04	0.08	0.12
Verbs Count In Sentence	Text	0.66	0.17	0.29	0.37	0.73	0.17	0.15	0.35	0.5	0.02	0.06	0.19
Commas Count In Sentence	Tag	0.33	0.17	0.27	0.25	0.68	0.17	0.27	0.37	0.48	0.04	0.22	0.24
Verbs Frequency	Text	0.5	0.25	0.39	0.38	0.39	0.08	0.2	0.22	0.42	0.06	0.21	0.23
Verbs Frequency	Text	0.73	0.25	0.2	0.4	0.65	0.08	0.22	0.32	0.52	0.06	0.23	0.27
Verbs Info (Type)	Text	0.53	0.27	0.65	0.48	0.69	0.08	0.23	0.33	0.65	0.04	0.21	0.3
Verbs Info (Number)	Tag	0.23	0.19	0.43	0.28	0.2	0.08	0.16	0.15	0.2	0.04	0.05	0.1
Verbs Info (Person)	Text	0.6	0.27	0.65	0.51	0.63	0.08	0.37	0.36	0.58	0.04	0.12	0.25
Verbs Info (Tense Mood)	Text	0.8	0.25	0.53	0.53	0.69	0.08	0.22	0.33	0.56	0.04	0.08	0.23
Adjectives Info (Value)	Text	0.65	0.25	0.1	0.33	0.33	0.13	0.22	0.22	0.5	0.04	0.21	0.25
Adjectives Info (Type)	Text	0.73	0.27	0.47	0.49	0.6	0.13	0.15	0.29	0.52	0.02	0.05	0.2
Adverbs Info (Value)	Tag	0.65	0.25	0.29	0.4	0.54	0.13	0.16	0.28	0.49	0.04	0.25	0.26
Adverbs Info (Type)	Text	0.4	0.25	0.1	0.25	0.46	0.08	0.05	0.2	0.33	0.04	0.03	0.13
Sentence Start Token	Tag	0.51	0.08	0.1	0.23	0.36	0.17	0.05	0.19	0.26	0.01	0.04	0.1
Sentence Start Token	Text	0.56	0.08	0.1	0.25	0.43	0.17	0.05	0.21	0.3	0.01	0.03	0.11
2-Gram Word	Tag	0.61	0.17	0.1	0.29	0.6	0.08	0.22	0.3	0.45	0.04	0.23	0.24
3-Gram Word	Text	0.3	0.17	0.1	0.19	0.43	0.08	0.19	0.23	0.29	0.02	0.28	0.2
4-Gram Word	Text	0.13	0.17	0.1	0.13	0.16	0.04	0.15	0.12	0.1	0.01	0.28	0.13
5-Gram Word	Text	0.16	0.08	0.1	0.12	0.07	0.04	0.15	0.09	0.08	0.02	0.29	0.13
2-Gram Character	Text	**0.96**	**0.33**	0.67	**0.66**	**0.94**	0.17	0.64	**0.58**	**0.84**	**0.08**	0.43	**0.45**
3-Gram Character	Text	0.83	**0.34**	0.57	0.58	**0.89**	0.17	0.32	0.46	0.68	0.04	0.27	0.33
4-Gram Character	Text	0.86	0.25	0.51	0.54	0.33	0.13	0.23	0.23	0.7	**0.08**	0.14	0.31
5-Gram Character	Text	0.93	0.17	0.2	0.43	0.69	0.17	0.15	0.34	0.56	0.04	0.13	0.25

978-1-4799-2093-8/13/$31.00 ©2013 IEEE

TABLE 4 - THE EFFECTS OF DIFFERENT TEXTUAL FEATURES ON DISJOINTED MANNER WITH 10, 20 AND 40 AUTHORS

Datasets Info		Disjointed Manner Number of Authors: 10				Disjointed Manner Number of Authors: 20				Disjointed Manner Number of Authors: 40			
Feature	**Source**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**	**SVM**	**K-NN**	**C5**	**AVG**
Words Length	Text	0.3	**0.25**	**0.76**	0.44	0.3	**0.17**	0.7	**0.39**	0.24	**0.14**	0.55	0.31
Words Frequency	Text	0.33	0.17	0.59	0.36	0.41	**0.17**	0.62	**0.4**	**0.43**	0.04	0.51	0.32
Words Frequency	Stem	0.5	0.17	0.61	0.43	**0.56**	0.08	0.34	0.33	**0.44**	0.02	0.42	0.29
Words Frequency	Tag	0.41	0.17	**0.8**	**0.46**	0.34	0.08	**0.71**	0.38	0.33	0.04	**0.62**	**0.33**
Characters Frequency	Text	0.15	**0.33**	**0.8**	0.43	0.15	0.13	**0.74**	0.34	0.09	0.06	**0.64**	0.26
Sentences Length (Word)	Text	0.15	0.17	0.18	0.17	0.13	0.08	0.36	0.19	0.08	0.01	0.26	0.11
Sentences Length (Character)	Text	0.15	0.1	0.1	0.12	0.14	0.08	0.07	0.1	0.07	0.01	0.05	0.04
Verbs Count In Sentence	Text	0.13	0.17	0.57	0.29	0.23	0.08	0.48	0.26	0.16	0.01	0.38	0.18
Commas Count In Sentence	Tag	0.26	0.18	0.57	0.34	0.23	**0.17**	0.45	0.28	0.25	0.02	0.31	0.19
Verbs Frequency	Text	0.43	0.17	0.31	0.3	0.25	0.08	0.18	0.17	0.19	0.02	0.27	0.16
Verbs Frequency	Text	0.28	0.08	0.47	0.28	0.45	0.08	0.43	0.32	0.18	0.02	0.34	0.18
Verbs Info (Type)	Text	0.26	0.18	0.63	0.36	**0.46**	0.04	0.66	**0.39**	0.27	**0.13**	**0.62**	**0.34**
Verbs Info (Number)	Tag	0.03	0.08	0.57	0.23	0.08	0.04	0.43	0.18	0.08	0.02	0.4	0.17
Verbs Info (Person)	Text	**0.56**	0.17	0.71	**0.48**	0.23	0.04	0.59	0.29	0.17	0.04	0.56	0.26
Verbs Info (Tense Mood)	Text	0.51	0.17	0.71	**0.46**	0.28	0.08	0.55	0.3	0.31	0.04	0.58	0.31
Adjectives Info (Value)	Text	**0.61**	0.17	0.16	0.31	0.07	0.04	0.19	0.1	0.23	0.01	0.17	0.14
Adjectives Info (Type)	Text	0.26	0.17	0.55	0.33	0.14	0.04	0.44	0.21	0.09	0.01	0.39	0.16
Adverbs Info (Value)	Tag	0.16	0.17	0.29	0.2	0.08	0.04	0.2	0.1	0.05	0.02	0.19	0.09
Adverbs Info (Type)	Text	0.13	0.13	0.49	0.25	0.1	0.04	0.35	0.16	0.06	0.02	0.38	0.16
Sentence Start Token	Tag	0.16	0.17	0.43	0.25	0.15	0.13	0.05	0.11	0.13	0.01	0.04	0.06
Sentence Start Token	Text	0.15	0.17	0.45	0.26	0.15	0.13	0.45	0.24	0.13	0.01	0.03	0.06
2-Gram Word	Tag	0.43	0.17	0.2	0.27	0.2	0.08	0.12	0.14	0.17	0.02	0.08	0.09
3-Gram Word	Text	0.13	0.08	0.1	0.11	0.03	0.04	0.13	0.07	0.05	0.01	0.08	0.05
4-Gram Word	Text	0.13	0.08	0.1	0.11	0.03	0.04	0.08	0.05	0.02	0.02	0.07	0.04
5-Gram Word	Text	0.15	0.08	0.1	0.11	0.07	0.04	0.06	0.06	0.03	0.01	0.07	0.04
2-Gram Character	Text	0.3	**0.25**	0.73	0.43	0.23	**0.17**	0.65	0.35	0.18	0.04	0.6	0.27
3-Gram Character	Text	0.3	**0.25**	0.67	0.41	0.26	0.08	0.52	0.29	0.18	0.04	0.44	0.22
4-Gram Character	Text	0.23	0.17	0.63	0.34	0.25	0.08	0.45	0.26	0.23	0.02	0.34	0.2
5-Gram Character	Text	0.41	0.08	0.35	0.28	0.23	0.01	0.2	0.15	0.31	0.01	0.29	0.2

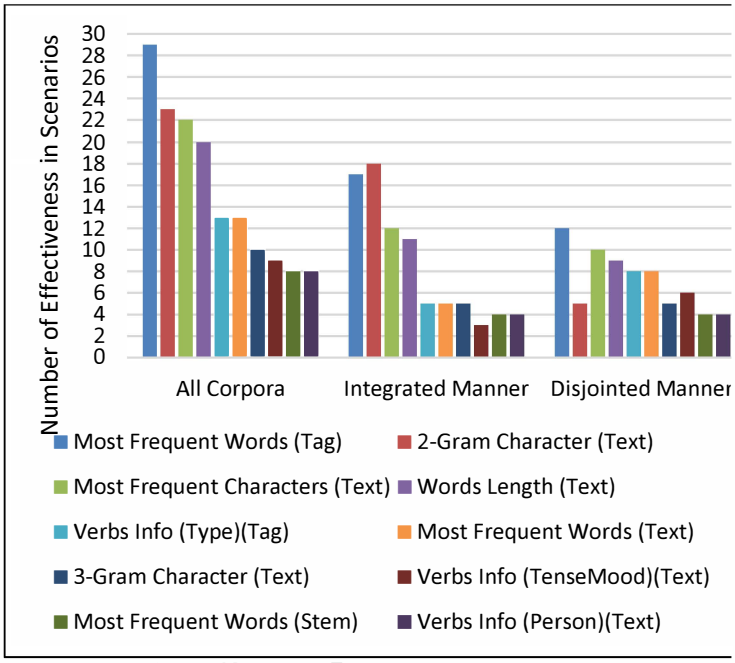

FIGURE 1 – THE NUMBER OF EFFECTIVENESS OF TEXTUAL FEATURES

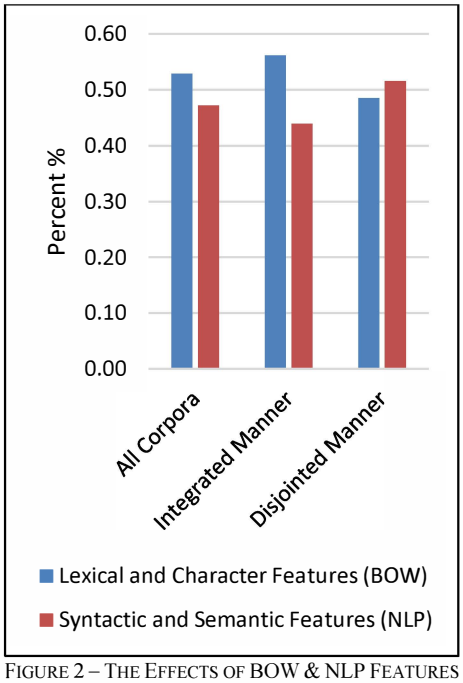

FIGURE 2 – THE EFFECTS OF BOW & NLP FEATURES

REFERENCES

[1] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 62.

[2] S. M. Zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections," in *Advances in data analysis*, ed: Springer, 2007, pp. 359-366.

[3] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. 18, pp. 613-620, 1975.

[4] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology,* vol. 60, pp. 538-556, 2009.

[5] J. Rudman, "The state of authorship attribution studies: Some problems and solutions," *Computers and the Humanities,* vol. 31, pp. 351-365, 1997.

[6] Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li, "Fast text categorization using concise semantic analysis," *Pattern Recognition Letters,* vol. 32, pp. 441-448, 2011.

[7] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 31, pp. 721-735, 2009.

[8] E. Gose, R. Johnsonbaugh, and S. Jost, *Pattern recognition and image analysis*: Prentice-Hall, Inc., 1996.

[9] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification," *Expert Systems with Applications,* vol. 38, pp. 2758-2765, 2011.

[10] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques,"*Literary and linguistic computing,* vol.22, pp.251-270,2007

[11] Gate. (2013/07/01). *General Architecture for Text Engineering,* http://GATE.ac.uk.

[12] N. Sheydaei, M. Saraee, and A. Shahgholian, "An Efficient and Powerful Preprocessing Method for Persian Corpora, " presented at the 21st Iranian Conference on Electrical Engineering (ICEE2013), 2013, (in Persian) "یک روش پیش‌پردازش کارآمد و قوی برای متون فارسی".

978-1-4799-2093-8/13/$31.00 ©2013 IEEE