Machine Learning for Authorship Attribution in Arabic Poetry

Al-Falahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa

***Abstract*—This paper presented an authorship attribution in Arabic poetry using machine learning. Public features in poetry such as Characters, Poetry Sentence length; Word length, Rhyme, Meter and First word in the sentence are used as input data for text mining classification algorithms Na ïve Bayes NB and Support Vector Machine SVM. The main problem: Can we automatically determine who poet wrote an unknown text, to solve this problem we use style markers to identify the author. The dataset of this work was divided into two groups: training dataset with known Poets and test dataset with unknown Poets. In this work, a group of 73 poets from completely different eras are used. The Experiment shows interesting results with classification precision of 98.63%.**

***Index Terms*—Authorship attribution, Arabic poetry, text classification, NB, SVM.**

I. Introduction

The Arabic poems are the earliest type of Arabic literature traditionally, these poems are classified into two groups: rhymed or measured, and prose. The rhymed or measured poems are greatly preceding the latter since they seem traditionally eeliest. The rhymed poem is classed by sixteen completely different meters. Such meters of the measured poetry also are famed in Arabic as meters (buḥūr). As mentioned before, the syllables area unit measuring a block of meters "tafilah". every meter contains a definite number of taf'ilah that the author should observe in each verse (bayt) of the verse form. The procedure of reckoning variety of taf'ilah in a very verse form is extremely strict since adding or removing a consonant or a vowel letter "Harakh" will shift the bayt from one meter to a different. Another feature of measured poems is that each bayt (the second a part of the verse) should rhyme poetry, every verse ought to end with an identical rhyme (qāfiyah) throughout the verse [1].

The task of analysis the text content in order to identify its original author among a set of candidate authors is called Authorship attribution. The idea behind the authorship attribution is as follows: given a set of poem texts as training data of known poet, the author of the unchecked text (texts in the test data) is determined by matching the anonymous text to one poet of the candidate set. In the context of Old Arabic Poetry, the current task can be re-formalized as follows:

Given a poetry with an anonymous author, find to whom this poem is belonging known set of features of each candidate authors. Authorship attribution research in Arabic poems is considered new and is not tackled as much as in other languages [2].

Before this research, no published works and researches about authorship attribution in Arabic poems. The majority of founding works deal with Arabic poems as a classification task. Al Hichri and Al Doori, in[3] used the distance-based method to classify Arabic poems depending on the rhythmic structure of short and long syllables. Na ïve Bayesian classifier was used by Iqbal AbdulBaki [4] to classify the poems into classification sets known in Arabic as " meters" (buḥūr).

Alnagdawi and Rashideh [5] proposed a context-free grammar-based tool for finding the poem meter name. The proposed tool was worked only with trimmed Arabic poems (words with diacritics "Tashkeel"). On the other hand, there are little works deal with authorship attribution of Arabic language [6]–[11]. Among of them, Altheneyan and Menai's work [10] is attractive. In their work, four different models naive Bayes classifiers: simple naıve Bayes, multinomial naıve Bayes, multi-variant Bernoulli na ïve Bayes, and multi-variant Poisson naıve Bayes were used. The experiment was mainly dependent on feature frequency which is extracted from a large corpus of four different datasets. The overall results showed that the multi-variant Bernoulli na ïve Bayes model provides the best results among all used models since it was able to find the author of a text with an average accuracy of 97.43%.

Rhymed in Arabic Poetry

وَ فَمُ الزَّمَانِ تَبَسُّمٌ وَ ثَنَاءُ			وُلِدَ الْهُدَى فَالْكَائِنَاتُ ضِيَاءُ			**Verse**
وَثَنَاؤُا	نِتَبَسْسُمُنْ	وَفَمُزْزَمَا	تُضِيَاؤُا	فَلْكَائِنَـا	وُلِدَلْهُدَىْ	**Verbal**
o/o///	**o//o///**	**o//o///**	**o/o///**	**o//o/o/**	**o//o///**	**split**
مُتَفَاعِلْ	مُتَفَاعِلُنْ	مُتَفَاعِلُنْ	مُتَفَاعِلْ	مُسْتَفْعِلُنْ	مُتَفَاعِلُنْ	**Tafilah**
Rhyme الهمزة			الكامل			**Metre**

Fig. 1. Rhymed in Arabic poetry.

Some works using Markov chains is not new in this direction [12], [13], however, using NB, SVM is not new in this direction the originality is our attempt to apply them together in Old Arabic Poetry context. Current paper proposes to use NB, SVM to solve authorship attribution in Arabic poetry.

Thus, this paper is organized as follows: Section II showed a general overview of characteristics of Arabic poems.

Manuscript received December 23, 2016; revised April 10, 2017.

Al-Falahi Ahmed is with Computer Science Department in FEN, IBB University, IBB, Yemen (e-mail: flahi79@gmail.com).

Ramdani Mohamed is with Département d'informatique -FSTM Université Hassan II Casablanca, Mohammediah, Morocco (e-mail: moha@fstm.ac.ma).

Bellafkih Mostafa is with Institut National des Postes et Télécommunications, INPT-Rabat Rabat, Morocco (e-mail: bellafki@inpt.ac.ma).

doi: 10.18178/ijfcc.2017.6.2.486

Section III introduces Arabic Poetry Corpus. Section IV presents Authorship attribution methodology. Section V discusses results of experimental. Finally, a general conclusion of this work is presented in Section VI.

II. Characteristics of Arabic Poems

Old Arab poetry that includes some of the characteristics that distinguish it from the rest of literary, it's called Meter and Rhyme.

A. *Meter (wazn)*

Old Arabic poem has restricted structure which is mainly based on the length of syllables. This structure formulates, as said before, the meter. Traditionally, there are sixteen meters described by the grammarian al-Khalili in the 8th century. Each meter is constructed from two basic units called watid ('peg') and sabab ('cord'). Each unit consists of either short or long syllable [14].

B. *Rhyme (qafiya)*

The process of finding the rhyme of an Arabic poem is basically easy. In Old Arabic poetry, poems follow a very strict but simple rhyme [15]. Since the last letter of each verse in OAP must be the same. The rhyme is the last letter of the second part of any verse. In the case of vowel letter, then the second last letter of each verse must be the same as well. The basic vowel sounds in Arabic are a "ا" alif, i"ي" yaa , and o "و"wow. Each vowel sound has two versions: a long and a short version. Short vowels are written as diacritical marks below or above the letter that precedes them while long vowels are written as whole letters [16].

To build our authorship attribution model, we go through a few stages: text pre-processing features extraction and features selection for Poetry Author fingerprint detection. In this paper, we present the Authorship Attribution task as a classification process. The methodology we applied starts from a classification of pre-processed dataset. The dataset is partition into train dataset and test datasets. In the first step, prescient features are extracted from the data, then the training and test sets are made, on the premise of these features. In the second step, the model is built from training data, then it is tested on unknown test data. The training and test cases are numerical features vectors that represent term frequency of each chose features, taking after by the author's name. We perform administered classification, the circumstance in which named training data are utilized to train a machine learner, as it permits the evaluation of classification, and accordingly is the best method for examining the adaptability of the Text Classification [17].

III. Arabic Poetry Corpus

TABLE I: The Corpus of Arabic Poetry

	N. poets	N.Qasidah	N. words
Training dataset	54	18646	1856436
Testing dataset	54		106546
Total	54	18646	1962982

Arabic poetry corpus may be a store having an assortment of poems related to a specific poet. The poetry of 73 totally different poets is collected from numerous websites.

The poetry corpus includes seventy-three poets with 18646 Qasidah, the full words are equal 1235402 this words divisions into 1856436 words for training dataset and 106546 words for the testing dataset is shown in Table I.

IV. Authorship Attribution Methodology

A. *Text Preprocessing*

We collected the poets' texts of the study sample from poetry encyclopedias and websites, for a number of famous poets. The poets were selected randomly from different eras. The bulk of the poetry texts are used in training data, however, the remaining used as testing data. We have introduced a range of fifty-four unknown poetry texts varying the number of verses (abyat/ابيات) of the test. The collected texts were classical poems which rely on weight and rhyme and not pure. They were consisting of some alphanumeric and punctuation that are usually rare in such poetry type. Thus, most of the poetic texts are subject to initialization process of strip punctuation, strip numbers, and strip alphanumeric. In the process of normalization, some Arabic letters have different forms such as (أ, إ, آ) to (ا) since they do not give any indication of discriminatory classical poetry texts, but it may play an important role in some structured texts such as the web [18].

TABLE II: The Corpus of Arabic Poetry Details

	Name of poet	N qasidah	N verse	N meter	N words training	N word test
1	**3amir iben tofil**	62	366	6	3440	267
2	**3ntarah**	156	1760	9	18916	397
3	**Abdljabar ben hamd îs**	365	4050	12	67110	570
73	**Abu Tamm âm**	313	7256	13	17168	682

B. *Extracting Features*

Extracting features for authorship attribution is a critical stage since it aims to discover unmistakable features. For each author, we assume that he or she has specific style features. We can recognize four principle sorts of features that convey possible indicators for authorship: character, lexical, syntactic, and semantic features. In this paper, we give an account of trials utilizing lexical and character features, since they are more solid than semantic features, considering the cutting edge in the semantic investigation; and are most regularly connected behind syntactic features. The features we utilized are recorded as a part of Table3. Characters, sentence length, word length, rhyme, meter and the first word in a sentence have been exhibited that they find themselves able to dependably handle restricted data [17].

C. *Feature Selection*

After features are extracted, feature selection is applied to limit the dimension of doubtless relevant options. Feature selection could be an important a part of every authorship attribution study, that aims to spot the foremost relevant ones for the task. The frequency of a feature is that the most used criterion for choosing options for authorship attribution. the best means is to limit the set to the $\boldsymbol{n}$ most frequent terms within the dataset[19]. The point of features choice routines is to diminish the dimensionality of dataset by uprooting unessential features for the grouping undertaking. A few sorts of features, for example, character and lexical features can impressively expand the dimensionality of the features set [20].

TABLE III: THE ACCURACY PERCENTAGE OF GOOD ATTRIBUTION OBTAINED

Factures	Total correct		Accuracy Percent	
	NB	SVM	NB%	SVM
Character	71	72	97,26027	95, 890
word length	68	70	93,15068	98,630
Sentences length	60	60	82,19178	82,192
First word length	40	46	54,79452	63,013
Meter	38	40	52,05479	54,794
Rhyme	40	43	54,79452	58,904
Average	52,833	55,166	72,374	75,570

In such case, features determination techniques can be utilized to lessen such dimensionality of the frequency. When all in done, the more continuous features are the more elaborate variety it catches. In this paper, the features are selected by two well-known feature selection methods: chi-squared (χ^2) and information gain (IG) methods.

Information Gain (IG) perform entropy decrease given an exact feature knowing the frequency of prevalence of a term in a very document. Since immune globulin considers each feature autonomous of others, it offers a positioning of the options relying upon their (IG) score, thus a selected number of features will be select effectively [19].

We use the Chi-Squared technique because we have predefined a number of features that have x2 check score larger than 10.83 that indicates applied mathematics significance at the 0.001 level. However not least we should always note that from applied mathematics purpose the Chi-Squared feature choice is inaccurate, because of the one degree of freedom and correction ought to be used instead (which can build it more durable to achieve applied mathematics significance). so we should always expect that out of the whole elect features, a tiny low a part of them area unit freelance from the class).

D. Experiment

After process of extracting the features values (using tool), included weka library to text preprocess and applied NB, SVM), we tend to sorted them into six sets in step with following stylistic features: F1 set- character features, F2 set- word length, F3 set - sentence length, F4 set- first word in sentence, F5 set- Meter and F6 set - rhyme.

Studies have demonstrated that lexical and syntactic features are the most imperative classifications and consequently structure the establishment for auxiliary and substance particular features [20]. We connected this idea to test the importance of classification features of Arabic poetry. For the experiment, we created 73 randomly selected samples of 73 authors, which we used in all experiments.

We evaluated each sample using most poetry text per author on training data and conducted a 73-unknown text to the detected author with NB, SVM classifiers, which we utilized as a part of all investigations. The accuracy of a classification model is outlined in term of traditional accuracy (total variety of properly known text over the fifty-four total texts).

V. RESULTS AND DISCUSSION

In our many experiments of authorship attribution on Old Arabic poetry, a set of texts that were written by 73 Arabic poets are introduced. Many features: characters, words length, sentences length, rhyme, first-word in sentences and meter are tested[15]. We observed from Table 3 after applying all algorithms that maximum accuracy value is 98.63% of true attribution with apply SVM. Accuracy is utilized to demonstrate the quantity of accurately characterized examples over the aggregate number of test cases by figuring the normal of accuracy, as in Eq.1.

$$Accuracy \quad = \frac{Number\ of\ texts\ are\ well\ attributed}{total\ number\ of\ texts} \qquad (1)$$

$$Recall \qquad = \frac{Number\ of\ correct\ ID\ Poets}{total\ number\ of\ poetrys\ in\ test} \qquad (2)$$

Table III demonstrates the best features score that acquired in utilizing NB, SVM. The maximum value is attained by applying SVM on word length features (accuracy= 98.63%), this means that the word length in Arabic Poem used in different ways by the authors and can distinguish among poets stylistic. The same value (98.63%) we obtained when implementing NB on a combination of (F1+F2) and (F1 + F2 + F3 + F4 + F5+F6) features with SVM in table 4. This result is the best for all features that have used in our experiment. However, we obtained the least result 52,055% by applying NB on F5=Meter features (table 3) and 54,794% by applying SVM on F5= Meter and we obtained the least result 65,75% by applying SVM on a combination of (F1 + F2 + F3 + F4 + F5) features (table 4). The small rate 52,055% cannot use to certify that correct Poet wrote a text. This is possible because some Poets use the same Meter, this means that the meter is not a clear sign to identify the authors of the text because of the similarity with the meter by most poets. Nevertheless, when we used meter feature of other features with rhyme feature we obtained 98.15% this mean that meter features given a good result when we applied SMO. Likewise, a score of great attribution of 96,63% by utilizing one of the accompanying two features: The F1=character via F1+F2 when we applied Na ïve Bayes.

Likewise, we obtained the same result of the experience on features (F1+F2) characters and word length together, it is a good result compared with the result of (F1+F2+F3+F4), which added sentence length (F3) and First word length (F4) to (F1+F2) where the value is 82,19%, and 82,19% by used NB and SVM. This decline because of the sentence length and first-word length in OAP obligates by a number of taf'ilahs and meters (wazn).

We also observed that the sentence length and first-word length when tested separately score was same value 82,19%

by used NB and SVM while when added to the other features was varying according to integrating with other features.

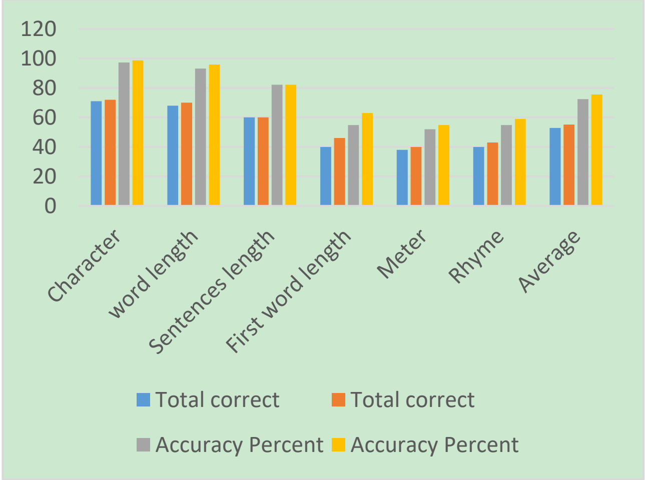

Fig. 2. Chart of accuracy result.

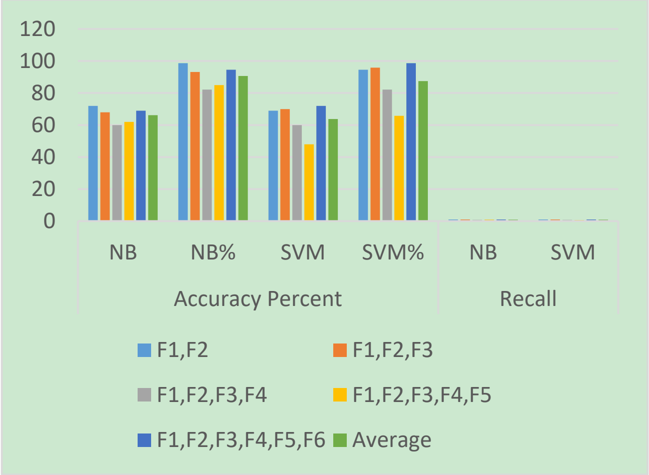

Fig. 3. Chart of accuracy and recall result.

TABLE IV: THE ACCURACY AND RECALL PERCENTAGE OF GOOD ATTRIBUTION OBTAINED THE DIFFERENT FEATURES BY APPLYING ALL ALGORITHMS

Factures	**Accuracy Percent**				**Recall**	
	NB	***NB%***	***SVM***	***SVM%***	***NB***	***SVM***
F1,F2	72	98,63	69	94,52	0,99	0,95
F1,F2,F3	68	93,15	70	95,89	0,93	0,96
F1,F2,F3,F4	60	82,19	60	82,19	0,82	0,822
F1,F2,F3,F4,F5	62	84,93	48	65,75	0,852	0,66
F1,F2,F3,F4,F5,F6	69	94,52	72	98,63	0,95	0,99
Average	66,2	90,684	63,8	87,40	0,91	0,87

VI. CONCLUSION

In this work, an Authorship Attribution task has experimented on OAP set of texts that were written by 73 Arabic poets. Each author is presented by many different texts. The algorithms classifiers are implemented using many texts validation. The experiments, which have been done separately for each, feature on the Old Arabic Poetry dataset using NB, SVM classifier shows the following remarkable points.

- The F1= Character and F2= word length features are better than all features in Table3, regarding the maximum accuracy =98.63% by SVM on word length and maximum accuracy = 97,26 % by NB on Character of all features.
- Meter without other features does not give a clear indication contributed to accurate author identify but while used within other features not giving a good result.
- F5= Meter and F6=Rhyme features are fewer results and cannot be used to certify that correct Poet wrote a text if it used separately.
- The best performance we got after it has been used all features together regarding the accuracy =98,63% when we use SVM on this features.
- Rhyme and Meter within other features gives a clear indication contributed to accurate author identify but while used alone not giving a good result.
- The best Average recall we obtained from applied NB algorithm on all features the recall is 0,91 better than SVM algorithms.

A. Future Work

- We propose the introduction of other Poetry features and used some features like weight, rare words, synonyms.
- Also, we propose a plan to extend the investigations into bigger datasets of more than 73 of Poets. In addition, we intend to extend the experiments using other algorithms and compared the results with these new results.

ACKNOWLEDGMENT

This paper was supported/partially supported by IBB University in Yemen. We tend to convey thanks our colleagues from FSTM and INPT who provided insight and skill that greatly motor-assisted the paper.

REFERENCES

[1] Wiki. (2015). Arabic poetry — Wikipedia, the free encyclopedia. [Online]. Available: http://en.wikipedia.org/wiki/Arabic_poetry

[2] K. Shaker and D. Corne, “Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis,” in *Proc. 2010 UK Workshop on Computational Intelligence (UKCI)*, 2010, pp. 1–6.

[3] A. M. A. Alhichri, “Expert system for classical Arabic poetry (ESCAP),” in *Proc. International Conference on APL*, Toronto, Ontario, Canada, 2008.

[4] I. A. Mohammad, “Naive bayes for classical Arabic poetry,” vol. 12, no. 4, pp. 217–225, 2009.

[5] M. A. Alnagdawi, H. Rashideh, and A. Fahed, “Finding Arabic poem meter using context free grammar,” vol. 3, no. 1, pp. 52–59, 2013.

[6] A. Abbasi and H. Chen, “Analysis to extremist-messages,” no. October, pp. 67–75, 2005.

[7] K. Shaker and D. Corne, “Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis,” *Comput. Intell. (UKCI)*, 2010.

[8] R. Baraka, S. Salem, M. Abu, N. Nayef, and W. A. Shaban, “Arabic text author identification using support vector machines,” *J. Adv. Comput. Sci. Technol. Res.*, vol. 4, no. 1, pp. 1–11, 2014.

[9] G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski, and B. S. Howald, “Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method,” *Int. J. Digit. Evid.*, vol. 6, no. 1, pp. 1–18, 2007.

[10] A. S. Altheneyan and M. E. B. Menai, “Naïve bayes classifiers for authorship attribution of Arabic texts,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 26, no. 4, pp. 473–484, 2014.

[11] A. Abbasi and H. Chen, “Applying authorship analysis to Arabic web content,” *Intell. Secure. Informatics*, 2005.

[12] D. Khmelev and F. Tweedie, “Using markov chains for identification of writer,” *Lit. Linguist. Comput.*, vol. 16, no. 4, pp. 299–307, 2001.

[13] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, “Using literal and grammatical statistics for authorship attribution,” *Probl. Inf. Transm.*, vol. 37, no. 2, pp. 172–184.

[14] H. Scott, “Pegs, cords, and ghuls: Meter of classical Arabic poetry,” 2009.

[15] A. F. Ahmed, R. Mohamed, B. Mostafa, and A. S. Mohammed, “Authorship attribution in Arabic poetry,” in *Proc. 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2015, pp. 1–6.

[16] A. Almuhareb, I. Alkharashi, L. A. Saud, and H. Altuwaijri, “Recognition of classical Arabic poems,” in *Proc. Work. Comput. Linguist. Lit*., pp. 9–16, 2013.

[17] K. Luyckx, “Scalability issues in authorship attribution,” 2010.

[18] F. Howedi and M. Mohd, “Text classification for authorship attribution using naive bayes classifier with limited training data,” *Computer Engineering and Intelligent Systems*, vol. 5, no. 4. pp. 48–56, 2014.

[19] F. Howedi and M. Mohd, “Text classification for authorship attribution using naive bayes classifier with limited training data,” *Comput. Eng. Intell. Syst*., vol. 5, no. 4, pp. 48–57, 2014.

[20] E. Stamatatos, “A survey of modern authorship attribution methods,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.

[21] Y. Zhao and J. Zobel, “Effective and scalable authorship attribution using function words,” *Inf. Retr. Technol.*, vol. 3689, pp. 174–189, 2005.

Al-Falahi Ahmed was born in IBB-Yemen in December 1979. He received his BS degree in science at Taiz University in 1997. He got the MS degree in computer science in Iraqi Commission for Computers and Informatics (ICCI). He got his diploma in information technology at the University of Technology in Iraq. He is currently a professor in IBB University. His current research area includes authorship attribution in Arabic poetry. He has a rich professional career and possesses several journals and conference publications articles in the area of AI. Now he is a researcher with authors in FSTM at Hassan II Casablanca & INPT-Rabat Morocco.

Mohammed Ramdani received the PhD thesis in computer science from the University of Paris 6, France, in February 1994 and habilitation in computer science from the University of Paris 6, France, in June 2001. His research interests include the, knowledge management, A.I., data mining and database. He is professor in Mohammedia Faculty of Sciences and Technologies (FSTM), Morocco since 1995.

Mostafa Bellafkih received the PhD thesis in computer science from the University of Paris 6, France, in June 1994 and the doctorate Es science in computer science (option networks) from the University of Mohammed V in Rabat, Morocco, in May 2001.

His research interests include the network management, knowledge management, A.I., data mining and database. He is professor in the National Institute of Posts and Telecommunications (INPT) in Rabat, Morocco since 1995.