Original Research Paper

Arabic Poetry Authorship Attribution using Machine Learning Techniques

[1]Al-Falahi Ahmed, [2]Ramdani Mohamed and [3]Bellafkih Mostafa

[1]*Department of Computer Science, FEN, IBB University, IBB Yemen*
[2]*Département d'informatique, FSTM Université Hassan II, Mohammedia, Maroc*
[3]*Institut National des Postes et Télécommunications, INPT-Rabat, Maroc*

Abstract: In this study, authorship attribution in Arabic poetry will be conducted to determine the authorship of a specified text after documents with recognized authorships have been allocated. This work also measures the impact performance of Naïve Bayes, Support Vector Machine and Linear discriminant analysis for Arabic poetry authorship attribution using text mining classification. Several features such as lexical features, character features, structural features, poetry features, syntactic features, semantic features and specific word features are utilized as the input data for text mining, using classification algorithms Linear discriminant analysis, Support Vector Machine and Naïve Bayes by Arabic Poetry Authorship Attribution Model (APAAM). The dataset of Arabic poetry is divided into two sets: known poetic in training dataset texts and anonymous poetic texts in a test dataset part. In the experiment, a set of 114 random poets from entirely different eras are used. The highest performance accuracy value is 99, 12%; the performance rate at the attribute level is 98.246%; the level of techniques is 92.836%.

Keywords: Authorship Attribution, Arabic Poetry, Text Classification, Machine Learning Techniques

Introduction

Meter, rhyme, weight and promotion are the most critical elements of ancient Arabic poetry. Ancient Arabic poetry can be classified into two sets: measured or rhymed and poem prose. Measured or rhymed poetry considerably precedes the latter in terms of origin, as it appears historically earlier. Rhymed poetry is classified by 16 different meters (Badawi, 1993).

The analysis of Arabic poetic text using machine learning is not an easy task, as the attributes of Arabic poetry differ from that of other Arabic texts. Arabic poetry has structural characteristics that differ from ordinary texts, such as shape, rhythm and weights.

The concept behind authorship attribution in the case of Arabic poetry lies in the idea that if given a text of a poem as a form of training data from a known poet, it is possible to determine the writer of the unrecognized text in the test data. It can be done by corresponding the unknown text of the known poet to the potential poet (Al-Falahi *et al*., 2017).

In this task, an anonymous text is assigned to an original author and the author is determined from a group of nominated authors using a set of features. The research on original poet identification and the performance measurement of classification techniques in Arabic poetry is new compared to that in other languages.

The research on poet identification in Arabic poetry is lacking or almost nonexistent, except for our published researches (Ahmed *et al*., 2016; Al-Falahi *et al*., 2017). To the best of knowledge, to the date of writing this paper, no published work explores such a problem as this work do. The primary research question is whether the potential poet x wrote the poem. However, a few related works are available, which make the job easier.

Related Works

The related works of Arabic poems aim to classify, recognize, or extract poems from writing texts. Alnagdawi *et al*. (2013) developed a program that can identify the meter name of a poem (called Bahar in Arabic) based on Aroud science; this science provides a methodology for classifying Arabic poems into 16 m to assist in locating the meter calls for any Arabic poem using Context-Free Grammar (CFG).

Moreover, the researcher discusses the solutions for problems, from the starting to the results phases, using regular expression and CFG. Also, the results 75% of the verse is observed its meter (Alnagdawi *et al*., 2013).

Mohammad (2009) utilized a Naïve Bayesian (NB) method to categorize Arabic poetry into their classification groups. She presented the results of the experiments performed on different literature ages and layers of classical Arabic poetry; the classified classes were sets of poems of classes like Ghazal, Medeh, Wasef, Hij authorship attribution, among others (Mohammad, 2009). Alhichri and Aldoori (2008) employed short and long poetic texts to build an expert system for classifying poems depending on the rhythmic structure. These rhythms are the seas (meters) of the poetry discovered by Alkhalil Bin Ahmed Al Farahidi. They used a rule-based algorithm, which was applied in several passes and then converted the resulting strings into a binary sequence. They calculated the distance degree among the binary pattern of a verse within an unknown poem and the binary patterns of all poetry seas. They considered their experimental results successful (Alhichri and Aldoori, 2008).

Ismail *et al*. (2004) utilized the "Expert System for Testing the Harmony of Arabic Poetry." The Expert System Harmony Test (ESHT) is designed to check the name of the meter in three steps: (1) The user enters the poetry and a heuristic rule in the knowledge base and the text is converted to the Arū form. This step involves retaining any letter that is spoken and discarding any letter that is unspoken and the heuristic knowledge base is written by Rule-Based Systems (RBS). (2) The text is converted from the Arū form to the binary form (0 s and 1 s) using deterministic knowledge, which involves the same inputs giving the same outputs and bypassing the same rules. (3) The next step involves a consultation, which is taken by the human expert knowledge stored in a knowledge base. This process is the core of the system. However, (Ismail *et al*., 2004) did not mention the experiment results. Shalabi *et al*. (2003) attempted to find Arabic poem meter names through two steps: (1) Using Arū rules to analyze the verse and define long and short sounds. (2) Using the generated string of short and long sounds to compare the rhythms of each meter one by one with the rhythms of verses to be checked. If the algorithm succeeds to find the correct rhyme in verse, then it will find the correct Arabic meter name. This algorithm was implemented using Turbo C under MS-DOS and this tool cannot be used in present operating systems (Shalabi *et al*., 2003).

Almuhareb *et al*. (2013) used a method for extracting and recognizing the old Arabic poetry found in poetic texts.

Authorship attribution has been carried out in a few Arabic texts (Stamatatos, 2008; Shaker and Corne, 2010; Ouamour and Sayoud, 2012). However, certain shortcomings can be attributed to those works. For instance, they used corpora with 10 or fewer authors. Also, the forms of testing and trying out files were both extracted from the same supply (Ouamour and Sayoud, 2012) or were manipulated by a technique that influenced the style of the authors (Stamatatos, 2008). Authorship attribution is similarly carried out on Arabic messages in the study by Abbasi and Chen (2006). The kinds of stylometric functions that may be carried out for Arabic authorship attribution are lexical (Ouamour and Sayoud, 2012), individual functions (Stamatatos, 2008; Ouamour and Sayoud, 2012) and syntactic capabilities (Abbasi and Chen, 2006; Shaker, 2012) extensively utilized phrase root capabilities and mentioned that the first-class accuracy rate of 85.43% had received the use of Support Vector Machine (SVM).

The present study uses NB, SVM and Linear Discriminant Analysis (LDA) to determine the authors of unknown Arabic Poetry texts by Model (APAAM) and undertake the impact evaluation of the performance techniques.

This paper is divided into several sections: Section II discusses the Arabic poetry dataset used in this study. Section III presents the identification methodology. Section IV discusses the test and outcomes. The paper culminates with a conclusion and future recommendations.

Arabic Poetry Dataset

The dataset of the Arabic poetry collection was stored in the form of poems in relation to a particular poet. The poems of 114 poets were collected from various websites.

The dataset included 21929 poems of 114 poets. The total number of words was 12311402, with 1673465 words in the training phase and 89456 words in the testing phase, as shown in Table 1.

Table 1: The dataset

	N. Poets	N. Poems	N. Words
Training dataset	126	15145	1673465
Testing dataset	126	6784	89456
Total	252	21929	1762921

Poet Identification Methodology

Building the poet identification model for this study (Fig. 5) involved the following stages: (1) Text pre-processing feature extraction and (2) feature selection for poetic text author-style detection. In this study, we give the author identity, challenge as a category procedure as a classification task. The technique we implemented evolved from a category of the pre-processed dataset. The dataset was divided into a teaching dataset and checks dataset. The dataset was classified into a testing dataset and training dataset. Step 1: From the data category in the training/testing datasets, the features that might indicate the potential author were extracted; prescient. Step 2: The model was constructed from the training dataset and an unknown test dataset was tested. Training and test cases determined the number of vector attributes that represented the frequency of each selected attribute and were once taken by a particular author name (poet). We executed controlled classification in a situation where called training dataset was utilized for developing machine learning, as it approved the analysis of the classification case and was consequently the best technique for testing the efficiency of the text category approach (Al-Falahi *et al.*, 2017).

Texts Preprocessing

Text preprocessing is a crucial step in authorship attribution. Text files of unique shapes are not appropriate, but styles are. They ought to be transformed right into an appropriate input format. Arabic poetry texts like Arabic text classifications typically consist of certain phases as shown in Fig. 1: Normalization, text conversion, stemming, tokenization and stop-word deletion are included in the preprocessing procedure. The text modeling phase involves term selection, term weighting and vector space model construction.
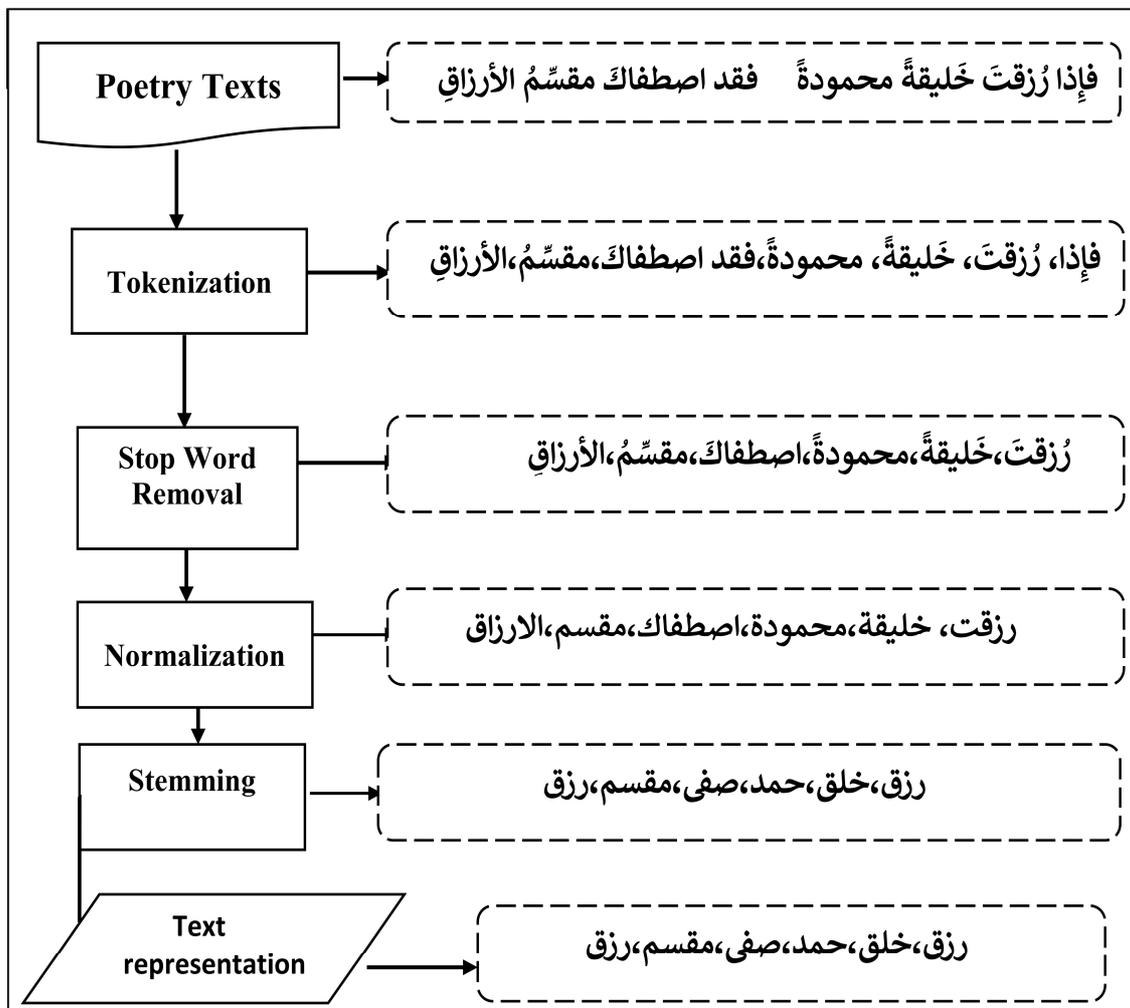

Fig. 1: Texts preprocessing phase

1014
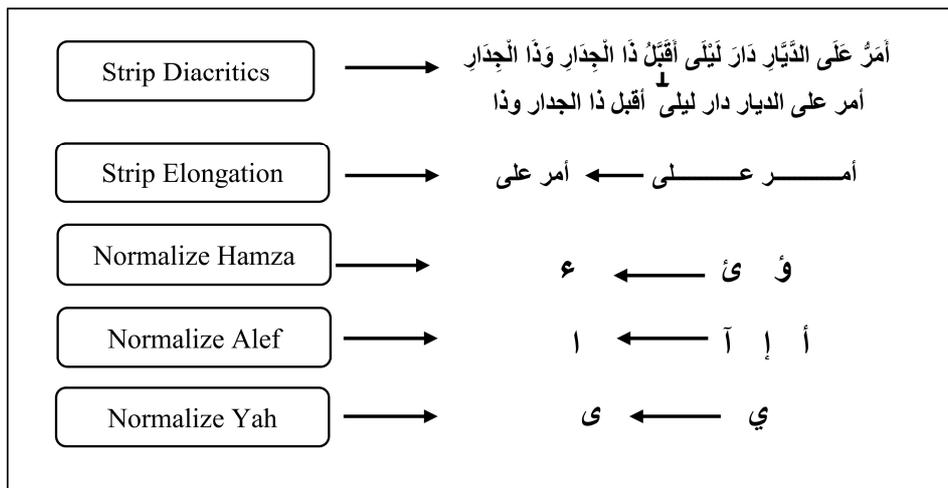
Fig. 2: Normalization Operation
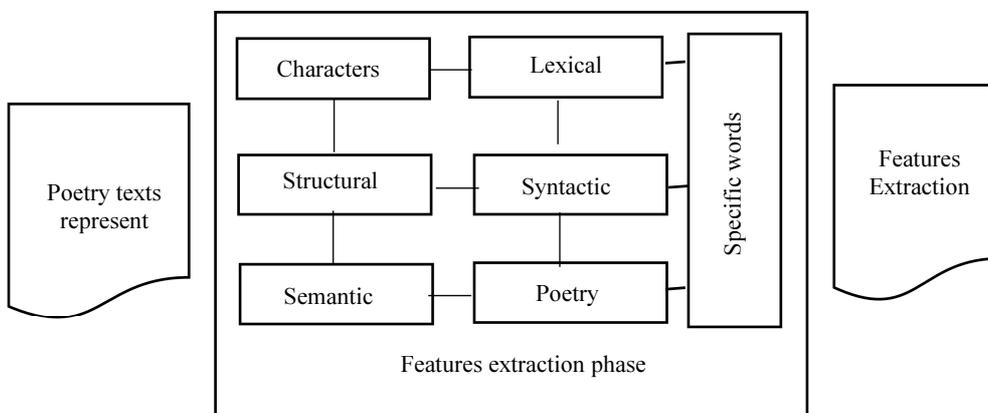
Fig. 3: Features extraction phase
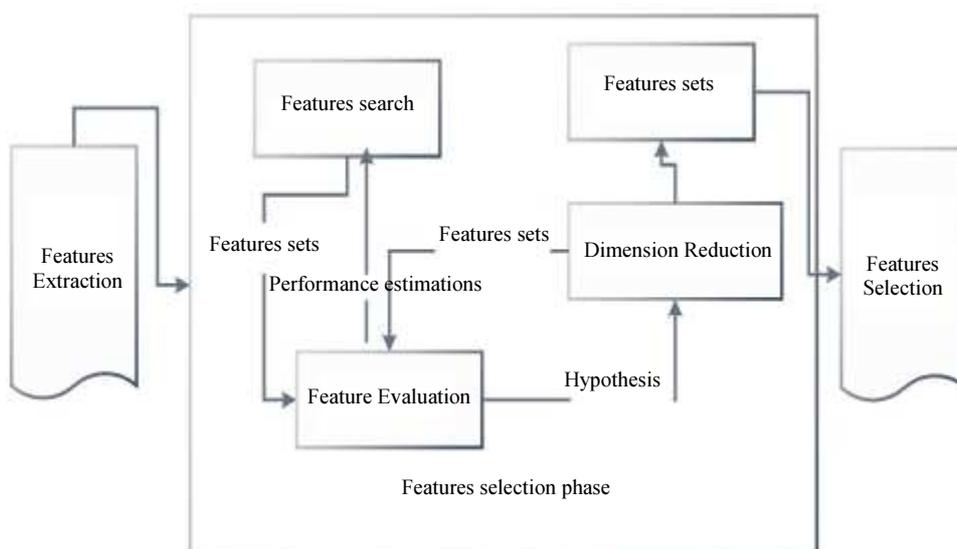
Fig. 4: Features selection phase

1015
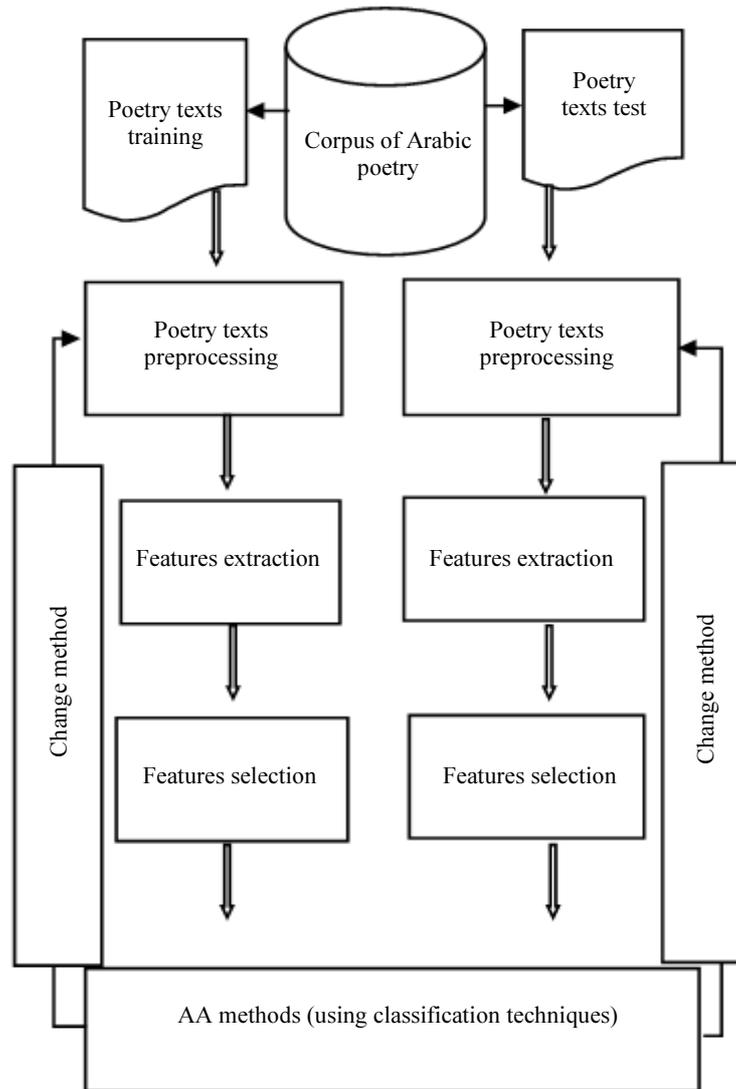
Fig. 5: APAAM for poet identification detection

Text Tokenization

Tokenization is the procedure of segmenting a poetic text into tiny units called tokens, which usually results in some word sets having a useful linguistic meaning. The words are often demarcated from each other by white space, semicolons, commas, quotes and periods by removing the extra symbols such as colons, exclamation marks, hyphens, bullets, parenthesis, numbers, among others (Nagaprasad, 2014). These tokens may be separate words that are born again, which may deter one in understanding the words while not relationships.

Stop-Word Removal

As described by (Reddy *et al.*, 2011a; 2011b), a stop list can be a list of ordinarily perennial features that appear in a text document. Standard features such as conjunctions, prepositions and pronouns should be deleted to prevent their results from having an impact on the classification. If a feature is of a unique variety or character, then it must be deleted. Stop-word inventory is recognized using Parts-Of-Speech (POS) tags used by the Telugu Morphological Analyzer (Nagaprasad, 2014). Stop-word removal involves removing trivial words, such as مع، لأجل لذا, in sentences that do not have any sign of meaning. Different samples of such insignificant words are pronouns, articles, conjunctions (هم ,هي ،هو), prepositions (حول، في، على، عن، الى، من), demonstratives, (اولئك ،هؤلاء،هذا) and interrogatives (لمن متى ،اين). Besides, Arabic-specific nouns stating place and time (فوق، بعد، بجانب) and symbols (@, #, &, %, *) are considered insignificant and can be removed (Ayedh *et al.*, 2016). The list of Arabic stop-words formulated by Alajmi *et al.*

1016

(2012) was used and updated by preventing the removal of certain stop-words in documents. We eliminated any matches between the stop-word list and dataset words.

Normalization

The goal of normalization is to normalize several characters that have different shapes in the identical word to a single shape. As these configurations are not used to extract the Arabic roots and are not practical in classification, delete the modulation processing (́, ́ , , , ́, ̌, ̌, ̌,) in this step. Also, normalize (أ, آ ,إ, ئ, ؤ) to "ا" and (ة) to "ه"; as shown in Fig. 2.

Finally, the character that takes the symbol "" can be replaced by two duplicate characters of the same character, as these characters are used to extract the Arabic roots in order to eliminate them for preventing them from affecting the meaning of the words (Ayedh *et al.*, 2016).

Stemming

The process of removing all affixes from any words that contain infixes, suffixes and prefixes are called stemming. This task is undertaken to completely reduce the different styles of a word that replicate a similar pattern. In other words, this process limits the feature area to a single kind of root or stem. The root does not have to be compelled to be a dead ringer for the morphological root of a word. Usually, it is suitable to map connected words to a similar stem, although this stem is not a valid root itself. For example, words such as poets ("الشعراء"), poet ("شاعر"), poet feminine ("شاعرة") with the area unit derived from a similar root-poetry ("شعر")-or a similar stem-feel ("شعر")-of these words take part in a similarly abstract. Therefore, actions or movements using the stemming techniques within the text classification make the processes less obsessed with particular styles of phrases and decrease the capacity length of options, which, in turn, improve the performance of the classifier (Kreaa *et al.*, 2014).

Extract Features

A critical phase in authorship attribution is extracting features; this step aims to find distinct attributes. We assumed that every poet has a specific style and recognized seven principle features that convey potential signs of authorship as shown in Fig. 3: lexical features, character features, structural features, poetry features, syntactic features, semantic features, and specific words features. The features were organized as shown in Table 2 to be used as input data for processing and extraction as well as distinguishing their ability to identify a particular poet.

Character Features

This feature is equivalent to the word n-gram except instead of words, characters are the items (Kaur and Verma, 2015). Character features are considered the most discriminate authorship features to detect the author of a text. On the character level, n-grams work with limited text. Moreover, character n-grams have the ability to capture the differences at structural, syntactic and lexical levels. In this study, we assumed two kinds of character features: character-level n-grams and punctuation marks. Also, we considered all punctuation marks (!.: ؟؛،-" ([{) according to the Arabic-language orthography; this led to 11 punctuation marks by a feature vector. The aim of analyzing different character n-grams (n = 1-4) pertains to the detection of the authors of unattributed poems in the Arabic language (Sadam, 2014).

Lexical Features

In this study, lexical features were relied on for the precise authorship attribution of Arabic poetry. Character and word n-grams, as well as features extracted by n-gram technical, were utilized (Kaur and Verma, 2015). We, here, used them with variant word-level n-grams and concatenated word n-grams, features. Thereby, we obtained some dimensions of the word n-gram feature vector. The reliance here implied that the higher estimation of n was more compelling when the span of the corpus was substantial. Word-level n-grams perform better with a massive dataset (Ayedh *et al.*, 2016).

Specific Words Per Authors

Different poets use specific words. To extract this feature all the training dataset was concatenated into the corpus. This feature has information that uses specific words to create the authors' profiles of the training dataset, for example, the specific word w_n for the nth author if the investigated threshold values are two and three for the occurring number of specific words per poet. If a threshold value is less than three, then large feature vectors, including typos, are generated, while using a threshold greater than four leads to authors with no specific words (Sadam, 2014).

Syntactic Features

In this feature, function words are utilized. These features do not include any notice about the texts in the corpus that purport and benefit from clearing the linguistic associations with separate words internal to a sentence. Examples are conjunctions, auxiliary verbs, pronouns, prepositions, grammatical subject, or units. The benefit of using function word features abides in that they are topical-separate. However, there is no consensus on an official list of function words for authorship attribution purpose (Sadam, 2014).

1017

Poetry Features

Some elements of Arabic poetry are meter, rhyme, meaning and shape. The meter is determined by the structure of the rhyming verses through the grouping of the vowel sounds (Haraka حركة) and steady sounds (Sukoun سكون) in a specific poetic line. At that point, a meter name is detected that corresponds to the verses at that point.

Determining the meter of a poem is a challenging task for someone who is not an expert, this issue urges us to construct this. To consider any verse to be rhyming, it should have a place with one of the 16 Arabic meters of the verse rhythms. To discover the meter name its sound should be compared with all grammars. At a point when any of these punctuations are substantial, it implies that the verse has a place with the meter. Presently, we explored one meter (Taweel meter) as a case for showing how this stage works.

In this step, we used the model proposed by Alnagdawi) called كاشف بحور الشعر (Alnagdawi *et al.*, 2013).

Features Selection

Reducing the dimensions of the training/testing data by removing unnecessary features for the classification is the purpose of using feature selection (Fig 4) methods. Certain types of features, such as lexical features, character features, structural features, poetry features, syntactic features, semantic features, specific words feature and other features can drastically increase the dimensionality of the feature dataset.

In this case, feature selection ways were also used to lessen such spatial properties of the representation (Nagaprasad, 2014).

While Principal Component Analysis (PCA) or other techniques extract all features, feature selection is applied to limit the pool of potentially essential features. Feature selection is an essential part of every authorship attribution study that starts from a broad set of features and aims to identify the most important ones for the task.

The frequency of a feature (character, lexical, syntactic, or semantic) is the most potent criterion for selecting features for authorship attribution (Howedi and Mohamed, 2014). The simplest way of performing feature selection is restricting the set to the most frequent *n* terms in the dataset. The purpose of features selection routines is to diminish the spatial property of the dataset by uprooting incidental features for the grouping enterprise. In this case, attributes fixing techniques may be used to minimize such spatial properties of the frequency. Once done, the additional continuous features are the additional elaborate varieties it catches.

Principal Component Analysis PCA

The primary use of PCA is to reduce the scale of the feature area while preserving as much information as feasible.

A method to find out how much information can be retained is to look at the contrast ratio explained in the main content. If we define the disparity of a dataset as $\sigma = \sum j \lambda j$, then the clarity disparity ratio of component is defined as $rj = \lambda j\sigma$ (Johannes, 2016). After feature extraction, PCA is employed to lower the dimensions of attribute scalars. PCA can reduce a potentially high correlative dimensional stylistic feature set into an unrelated lower dimensional feature set and replace the original features of the new uncorrelated features as principal components (Zhang *et al.*, 2014).

Experiment

While extracting the feature values (using APAAM and applied techniques-NB, SVM, LDA), we grouped them into seven sets according to following stylistic features: character features, lexical features set, structural features, syntactic features, semantic features set, poetry features set and specific words feature set. Studies have shown that linguistic features are the most important classifications, following the structuring of additional features and special materials (Stamatatos, 2009).

For the experiment, we selected a sample of 114 poets as the corpus (Table 2), which we used in all the experiments.

For every sample, we evaluated the usage of the most poetry text according to the author of the training texts and conducted an analysis of 114 anonymous texts to the detected creator with NB, SVM, LDA classifiers, which we applied as part of all investigations. The accuracy of (APAAM) model is outlined in the period of accuracy (total style of as it should be known textual content over the 114 total texts).

Results and Discussion

Experiments were conducted to determine the real authors of certain Arabic poetry texts. The texts dataset that was written by 114 Arabic poets were introduced. After the implementation of the three classification techniques and testing on the features, we obtained the results listed in Tables 2 and 3. As observed in Table 3, the maximum accuracy value is 99, 12% of accurate attribution by applying LDA.

The best features score acquired in utilizing NB, SVM, LDA are shown in Table 2. The best value is obtained by applying LDA on specific word attributes (99, 12%); this result means that the specific word attributes in the Arabic poetry used in different ways by the poets can be distinguished by the poets' style or

stylistic choices. The same value (99, 12%) was obtained when implementing LDA on a combination of C, L, St, Sy, Se, P, Sw features, as provided in Table 3. This end result is pleasant in terms of accuracy for all the attributes that we used in this study.

However, we received the worst result of 71, 93% by applying SVM on syntactic features (Table 2) and the same value 71, 93% by applying SVM on a combination of C, L, St, Sy, Se features (Table 3).

The low rate of 71, 93% cannot be used to identify that real author of a poetry text though syntactic features are not repeated in all poets. NB yielded the next lowest rate of 74, 56%. The rate was intended to handle data with different cases with structural features; this means that the structure is unclear to identify the authors of the poetic texts significantly. Nevertheless, when we used semantic and poetry features in this experiment, we obtained the perfect rate of 87.71%, 78.95%, as shown in Table 2.

Semantic and poetry features yielded a good result (78.95%) when LDA was applied and the good value was obtained by NB, SVM on the semantic features. Likewise, a score of significant attribution of 96, 49% by utilizing one of the accompanying two features was yielded: the character via C, L when we applied NB.

Likewise, the same results as provided in Table 3 of the features C, L together were obtained; it is a good result and is more than the result of C, L, St, Sy, which St and Sy to C, L where the value is 74.56% and 77.19% by using NB, SVM and LAD. This decline occurred because the structural and syntactic features in Arabic poetry obligates by (wazn). Also; the structural (74.56%, 74.56%, 92.10%) and lexical (88.59%, 92.98%, 98.25%) features when tested separately by NB, SVM and LDA while when added to the other features were variable

according to integrating with other features.

Conclusion

The experiments conducted separately for each feature of the Arabic poem data using NB, SVM and LDA revealed the following important points:

- The best performance average in Table 3 on features level of 97.95% for a specific word, 94.15% of character and 93.28% of lexical features, are better than all features in Table 2. Because of these features when applying classification techniques
- The poetry features secured a good performance average of 81.58%. However, the performance rate of these features is lower than expected because these features are restricted by the meter, rhyme and the
- Length of the sentences and these characteristics are similar to most poets, so those features are not ideal for identifying the real authors if texts if used separately

The best performance average in Table 2 on the techniques level is 89.72% by LDA; it is better than all classifiers techniques in Table 3:

- The best performance average in Table 3 for the features level is 98.25% that we obtain after using all attributes together (Ahmed *et al.*, 2019)
- The poetry features with other features give a visible indication for identifying the exact author, but while using it alone does not give the best result
- The best performance average result is 92.84% by LDA; it is better than the techniques in Table 4

Table 2: The accuracy of complete features

Features	Accuracy percent			
	NB%	SVM%	LDA%	Average
Character = C	96.490	91.23	94.740	94.15
Lexical = L	88.590	92.98	98.250	93.28
Structural = St	74.560	74.56	92.110	80.41
Syntactic = Sy	74.560	71.93	77.193	74.56
Semantic = Se	75.440	76.32	78.950	76.90
Poetry = P	78.070	78.95	87.720	81.58
specific W = Sw	97.370	97.37	99.120	97.95
Average	83.580	83.33	89.720	

Table 3: The accuracy features together

Features	Accuracy percent			
	NB%	SVM%	LDA%	Average
C, L	96.49	91.23	98.25	95.32
C, L, St	88.60	92.98	92.11	91.23
C, L, St, Sy	74.56	74.56	77.19	75.44
C, L, St, Sy, Se	74.56	71.93	98.25	81.58
C, L, St, Sy, Se, P	91.23	94.74	92.11	92.69
C, L, St, Sy, Se, P, Sw	97.37	98.25	99.12	98.25
Average	87.14	87.28	92.84	

Table 4: The recall of complete features

| Features | Recall | | | |
	NB%	SVM%	LDA%	Average
Character	0.9649	0.9123	0.94740	0.9415
Lexical	0.8859	0.9298	0.98250	0.9328
Structural	0.7456	0.7456	0.92110	0.8041
Syntactic	0.7456	0.7193	0.77193	0.7456
Semantic	0.7544	0.7632	0.78950	0.7690
Poetry	0.7807	0.7895	0.87720	0.8158
Specific W	0.9737	0.9737	0.99120	0.9795
Average	0.8358	0.8333	0.89720	

The LDA technique has a definite impact on the identification of the real author because of the ability of the technology to handle the features compared to other techniques.

From the above points, we find that the results match the expectations. We have got a success rate of 99.12% and a valuable performance of 92.84% for LDA efficiency and 98.246% accuracy rate for the features.

Future Work

To overcome those obstacles, we propose the introduction of different poetry attributes and using attributes such as weight, synonyms, "Medh", "Tham", "Hija".

Also, we propose increasing the investigations into more massive datasets of over 114 poets and extending the test through different algorithms and comparing the results with these new results.

Acknowledgment

Author's Contributions

Al-Falahi Ahmed: Worked on all the experiments and analyzed them and contributed to writing the article

Ramdani Mohamed and Bellafkih Mostafa: Supervising, providing the study plan and, reviewing.

Ethics

This article is original and developed from previous work. The contents of this paper are unpublished.

References

Abbasi, A. and H. Chen, 2006. Visualizing Authorship for Identification. In: Intelligence and Security Informatics, Mehrotra, S., D.D. Zeng, H. Chen, B. Thuraisingham and F.Y. Wang (Eds.), Springer, Berlin, Heidelberg, ISBN-13: 978-3-540-34478-0, pp: 60-71.

Ahmed, A., R. Mohamed and B. Mostafa, 2016. Authorship attribution in Arabic poetry using NB, SVM, SMO. Proceedings of the 11th International Conference on Intelligent Systems: Theories and Applications, Oct. 19-20, IEEE Xplore Press, Mohammedia, Morocco, pp: 1-5. DOI: 10.1109/SITA.2016.7772287

Ahmed, A., R. Mohamed and B. Mostafa, 2019. Use SMO SVM, LDA for Poet Identification in Arabic Poetry. In: Smart Data and Computational Intelligence, Khoukhi, F., M. Bahaj and M. Ezziyyani (Eds.), Springer, Cham, ISBN-13: 978-3-030-11913-3, pp: 161-169.

Alajmi, A., E.M. Saad and R.R. Darwish, 2012. Toward an Arabic stop-words list generation. Int. J. Comput. Applic., 46: 975-8887.

Al-Falahi, A., M. Ramdani and M. Bellafkih, 2017. Machine learning for authorship attribution in Arabic poetry. Int. J. Future Comput. Commun., 6: 42-46.

Alhichri. A. and M. Aldoori, 2008. Expert System for Classical Arabic Poetry (ESCAP). Proceedings of International Conference on APL, (PIC' 08), Toronto, Ontario, Canada.

Almuhareb, A., I.A. Lama, A. Saud and H. Altuwaijri, 2013. Recognition of classical Arabic poems. Proceedings of the 2nd Workshop on Computational Linguistics for Literature, Jun. 14-14, Atlanta, Georgia, pp: 9-16.

Alnagdawi, M., H. Rashideh and A. Fahed, 2013. Finding Arabic poem meter using context free grammar. J. Commun. Comput. Eng., 3: 52-59. DOI: 10.20454/jcce.2013.600

Ayedh, A., G. Tan, K. Alwesabi and H. Rajeh, 2016. The effect of preprocessing on Arabic document categorization. Algorithms, 9: 27-37. DOI: 10.3390/a9020027

Badawi, M., 1993. Perennial themes in modern Arabic literature. British J. Middle Eastern Stud., 20: 3-19.

Howedi, F. and M. Mohamed, 2014. Text classification for authorship attribution using Naive Bayes classifier with limited training data. Comput. Eng. Intell. Syst., 5: 48-57.

Ismail, M.A., M.I.E. Adawy, H. Keshk and S.A. Saleh, 2004. Expert system for testing the harmony of Arabic poetry. J. Eng. Sci., 32: 401-411.

Johannes, O., 2016. Principal Component Analysis (PCA) for feature selection and some of its pitfalls· Johannes Otterbach.

Kaur, N. and A. Verma, 2015. Authorship attribution of punjabi poetry using SVM classifier. Int. J. Adv. Res. Comput. Sci. Software Eng.

Kreaa, A.H., A.S. Ahmad and K. Kabalan, 2014. Arabic words stemming approach using Arabic wordnet. Int. J. Data Min. Knowledge Manage. Process.

Mohammad, I.A., 2009. Naive Bayes for classical Arabic poetry. J. Al-Nahrain Uni., 12: 217-225.

Nagaprasad, S., 2014. Influence of machine learning techniques on Authorship attribution for Telugu text features. Int. J. Adv. Res. Comput. Eng. Technol., 3: 3633-3640.

Ouamour, S. and H. Sayoud, 2012. Authorship attribution of ancient texts written by ten Arabic travelers using a SMO-SVM classifier. Proceedings of the International Conference on Communications and Information Technology, Jun. 26-28, IEEE Xplore Press, Hammamet, Tunisia, pp: 44-47. DOI: 10.1109/ICCITechnol.2012.6285841

Reddy, P.V., B.V. Vardhan and A. Govardhan, 2011a. Analysis of BMW model for title word selection on Indic script. Int. J. Comput. Applic., 18: 21-25. DOI: 10.5120/2304-2915

Reddy, P.V., B.V. Vardhan and A. Govardhan, 2011b. Corpus-based extractive document summarization for Indic script. Proceedings of the International Conference on Asian Language Processing, Nov. 15-17, IEEE Xplore Press, Penang, Malaysia, pp: 154-157. DOI: 10.1109/IALP.2011.66

Sadam, A.A., 2014. Authorship attribution of Arabic texts. MSc Thesis, King Fahd University of Petroleum and Minerals, Saudi Arabia Deanship of Graduate Studies.

Shaker, K. and D. Corne, 2010. Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. Proceedings of the UK Workshop on Computational Intelligence, Sept. 8-10, IEEE Xplore Press, Colchester, UK. DOI: 10.1109/UKCI.2010.5625580

Shaker, K., 2012. Investigating features and techniques for Arabic authoriship attribution. PhD Theses, Mathematical and Computer Sciences.

Shalabi, R., G. Kana'an and A. AL-Jarah, 2003. Computing system for analyzing Arabic poems meter. Yarmouk Research, Irbid.

Stamatatos, E., 2008. Author identification: Using text sampling to handle the class imbalance problem. Inform. Process. Manage., 44: 790-799. DOI: 10.1016/j.ipm.2007.05.012

Stamatatos, E., 2009. A survey of modern authorship attribution methods. J. Am. Soc. Inform. Sci. Technol., 60: 538-556. DOI: 10.1002/asi.21001

Zhang, C., X. Wu, Z. Niu and W. Ding, 2014. Authorship identification from unstructured texts. Knowledge-Based Syst., 66: 99-111. DOI: 10.1016/j.knosys.2014.04.025