# Towards Author Identification of Arabic Text Articles

Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amer
The Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University
Zarqa, Jordan

***Abstract*—we target the problem of identifying the author of an Arabic text article. Our main aim is to develop an intelligent system that is capable of classifying a new article into one of seven classes that belong to seven different authors. For this purpose, we propose a novel dataset consisting of 12 features and 456 instances belonging to the 7 authors. In addition, we combine the proposed feature set with strong classification algorithms to assist in distinguishing between the different authors. Our results show that the proposed dataset has proved successful with a classification performance accuracy of 82% with hold-out test.**

***Keywords—authorship identification, arabic text features, fuctional trees, support vector machines.***

## I. INTRODUCTION

The area of authorship analysis has been investigated for long years going back to the early 60s of works such as [4], where the authors were studying the famous *Federalist Papers* case for solving an authorship claim by different authors. Recently, there has been growing interest in developing practical applications for the purpose of authorship analysis. These applications focus on different fields such as: email authorship [1], plagiarism detection [7] and forensic cases [10].

The area of authorship analysis includes interesting fields such as authorship identification or categorization, author characterization and similarity detection. In authorship identification, the main aim is to identify the author of a piece of writing given a number of existing writings by the same author. Author characterization intends to generate the author profile based on his/her writing style. The profile gives knowledge about the characteristics of the author such as gender and educational and cultural background. Another interesting area of authorship analysis is similarity detection, which is applied for the purpose of identifying if there is any kind of plagiarism in a given piece of writing, including a complete or partial replication of a piece of work without the permission of the original author. It is carried out by comparing multiple pieces of writings and determining whether they were produced by the same author or not [10,11].

Another linked area to authorship analysis is text categorization which is closely related to authorship identification but clearly different. In text categorization, the main aim is to categorize a set of text documents based on its content. Applications of text categorization include document filtering and document retrieval [10].

In this paper, we target an interesting area of authorship analysis, which is authorship identification or categorization or attribution. The main aim is to develop an intelligent classifier that is capable of predicting the author of a piece of writing given a predefined set of candidate authors and a number of samples for each author. To build such a classier, it is very important to identify a set of characteristics or features that is strong enough to distinguish between the different authors. Moreover, an intelligent learning algorithm must be built in an automatic way to be able to identify the authorship of a piece of writing.

In the literature, different sets of features have been implemented for the abovementioned purpose. These features can be divided into four main types: lexical, syntactic, structural, and content- specific features. Lexical features are character or word based features, examples of these features include vocabulary richness or word length...etc [2,10,11]. Syntactic features are more related to the language pattern and they determine how to syntactically form a sentence, examples of these features include the use of function words such as "the", "if", "to", "while", "upon" etc. Use of punctuation is also part of these features. In the work of [6], it was shown that including these features can improve the accuracy of author identification. Structural features are related to the style the author follows in his/her writing. It is believed that each author has a certain habit or structure in the way of writing. Examples of these features are the use indentation and paragraph length [10,11]. The last set of features is content-specific features which are content related features or keywords frequency and incorporating them can improve the performance of the classifier. In works like [12], content-specific features contributed positively in improving the performance of the author identification classifier.

To build a robust classifier, it is very important to have the right feature set and a strong learning algorithm. Recently, there has been great interest in the application of machine learning methods for the purpose of author identification and these methods proved to be useful and successful for this purpose. Examples of these methods are Support Vector Machines (SVM) [9,10], neural networks [5], Markov chains [3] and others. Each of these methods had its positives and negatives and there is no unique algorithm that can be generalized as the best among the others.

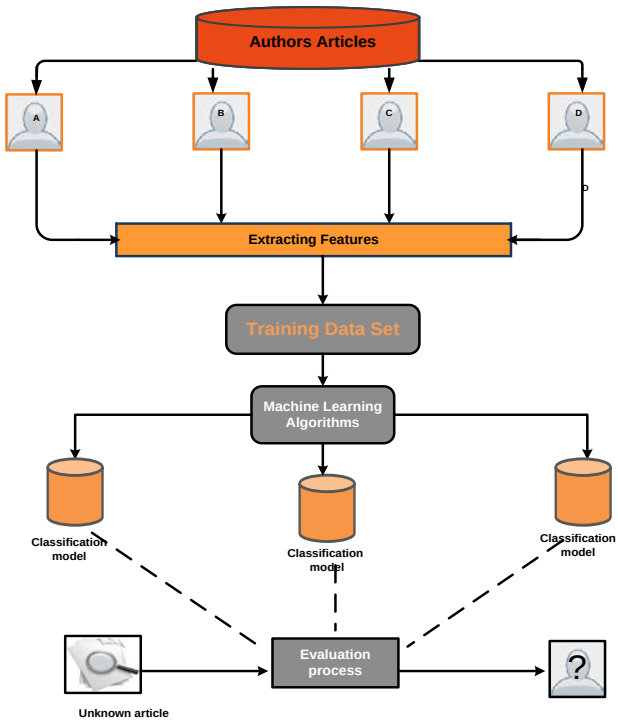


Fig. 1. The classification process.

In this paper, we propose the application of machine algorithms to the purpose of Arabic text author identification. The novelty of our work is in the proposal of a novel feature set for Arabic articles and the proof of the ability of the proposed feature set to distinguish between 7 different classes (authors) with a high percentage of accuracy.

The rest of this paper is organized as follows: In section 2, we explain the feature extraction part. Section 3 gives an overview about the classification algorithms used in our experiments. In section 4, we present the experimental results on the proposed feature set. Finally, section5 concludes the paper and suggests future work.

## II. FEATURE EXTRACTION

The first step in any classification problem is feature extraction where features are extracted for the Arabic articles based on different information. After extracting features, a classifier is trained using a classification algorithm and then a model is built that can be used for prediction. This process is explained in detail in Fig. 1.

For this purpose, we collected articles for 7 Arabic writers (for simplicity, we refer to writer as W) where the number of articles (instances) per writer is explained in Table I. Moreover, Fig. 2 gives examples of these articles for two writers: W1 and W2.

After collecting the articles for the 7 authors, we analyzed these articles and ran our feature extractor to extract important

TABLE I. # OF INSTANCES PER CLASS

| Writer name | # of instances |
|---|---|
| W1 | 49 Articles |
| W2 | 75 Articles |
| W3 | 61 Articles |
| W4 | 62 Articles |
| W5 | 74 Articles |
| W6 | 78 Articles |
| W7 | 57 Articles |

TABLE II. FEATURE TYPES

| Feature ` | Feature Type |
|---|---|
| Big Words | Lexical |
| Long sentences | Lexical |
| Short Sentences | Lexical |
| Small Statement | Lexical |
| Contain-Digit | Syntactic |
| Indeed Particles (pronounced Inna in arabic) | Syntactic |
| Conditional Particles | Syntactic |
| Special Characters | Syntactic |
| Complexity | Structural |
| Negative emotions | Content |
| Sport Words | Content |
| Political Words | Content |

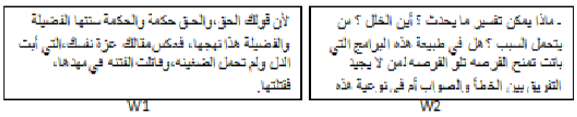


Fig. 2. Two examples of parts of articles written by two different authors.

information in these articles. As we presented in the introduction, there are four types of features that are important for identifying authors. However, in works like [6], it was proved that combining all four types of features can improve the classification performance. Thus, we ended up with a list of 27 features that include: lexical, syntactic, structural, and content- specific features.

After running an initial experiment to evaluate the impotence of these features, we ended up with 12 features that are powerful in discriminating between the seven classes, these features and there types are explained in Table II.

The *big words* feature returns the number of words that their length is greater than the average length of words related to the same author. The *long and short sentences* features return the number of short and long sentences. The *small statement* feature returns the number of small statements in the text. The *Contain-Digit* feature represents a count for the

words that contain digits in them. *Indeed particles* and *Conditional particles* are two particles that are commonly used in the Arabic language and they have a certain effect in terms of grammar in relation to the following nouns. There were other Arabic particles used in the feature set but these two were the most important. The *special characters* feature returns the number of special characters in the article (e.g. punctuations, question mark, etc..).

Regarding the *complexity* feature, it reflects how hard the readability of the text written by the author and it is calculated as the number of different words divided by the total number of words.

In additions, we used the *negative emotions* feature, which is an example of emotions features that reflect the emotions expressed in the human writing and it can be positive or negative but our feature selection algorithm selects the negative emotions feature as more important than the other one. Finally, the *sport and political words* are contents-specific features and they indicate the number of sport and political words that have been used in the article. Thus, we ended up with a dataset that contains 456 instances and each instance is represented with a normalized feature vector of a 12 – dimension space.

## III. CLASSIFICATION

The classifiers that have been used for the classification experiments in our system are the Functional Trees (FT) and sequential minimization optimization which is a variant of Support Vector Machines (SVM) which are both available as part of the WEKA package, a publicly available toolbox for automatic classification [8].

Functional Trees are classification trees that could have logistic regression functions at the inner nodes and/or leaves. On the other hand, sequential minimal optimization is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines, it breaks optimization problem into a series of smallest possible sub-problems, which are then solved analytically [8].

The performance of the classifier is evaluated in terms of classification accuracy which is calculated as the number of correctly classified samples divided by the total number of samples.

## IV. EXPEREMINTAL RESULTS AND ANALYSIS

Experiments are conducted in order to evaluate the performance of the proposed feature set. First, we run a hold-out test where the dataset is divided into two sets: training data and testing data. Almost 2/3 of the data is used to train the classifier and build the classification model using a certain classification algorithm. After that, the testing data are tested by the model and the predicted instances are compared with the original ones and the accuracy is calculated based on the number of correctly classified samples. Table III shows the number of samples used for training and testing for each of the 7 classes.

TABLE III. # OF TRAINING ANDTESTING DATA FOR HOLD-OUT TEST

| Classes | Training Data | Testing Data |
|---|---|---|
| W1 | 33 | 16 |
| W2 | 50 | 25 |
| W3 | 41 | 20 |
| W4 | 42 | 20 |
| W5 | 50 | 24 |
| W6 | 52 | 26 |
| W7 | 38 | 19 |

TABLE IV. ACCURACY RESULTS WITH HOLD-OUT TEST

| Classification Algorithm | Accuracy % |
|---|---|
| Support Vector Machines | 79.3 |
| Functional Trees | 82.0 |

TABLE V. ACCURACY RESULTS WITH 10-FOLD CROSS VALIDATION

| Classification Algorithm | Accuracy % |
|---|---|
| Functional Trees | 79.2 |

Table IV. Shows the accuracy performance of the proposed feature set using the abovementioned classification algorithms. It is clear from this table that the proposed feature set is powerful for discriminating between the 7 classes with performance accuracy of 82% by the functional trees algorithm. This accuracy is reasonably high given the high number of classes which makes it challenging for distinguishing between them. Moreover, the different writing habits for each author and the different topics they may write about especially in Arabic language makes it a not easy task to recognize the author of the text. In general, the proposed feature set proved to be effective in this problem.

Moreover, fig. 3 and fig. 4 illustrate in details the number of misclassified examples per class and among the two classification algorithms. It is clear from figure 3 that, in some cases, the proposed feature set can accurately classify two classes with 100% accuracy: W2 and W4. In figure 4, the SVM classifier is capable of recognizing class W2 with an accuracy of 100%. The reason for classifying class W2 in specific with high accuracy is because this author seems to follow a consistent style in his writings which makes the features extracted from the different articles close in its values and thus makes the recognition more accurate.

As the FT method gives better results than the other method, we decided to further experiment this algorithm with a 10-fold cross validation test to prove its robustness. So, the 456 instances of the dataset are divided into 10 disjoint groups and nine of them are used for training and the 10th one is used for testing. The algorithm runs for 10 times and the average accuracy across all the folds is calculated. Table V. shows the accuracy result of this method, it is clear from this tables that the accuracy decreased slightly compared to the results provided by the hold-out method. This is expected as the 10-fold is a result of 10 different divisions compared to a single

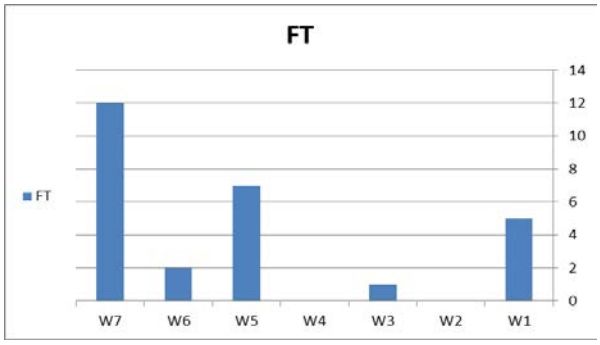


Fig. 3. Miscalssified examples across different classes with FT algorithm.

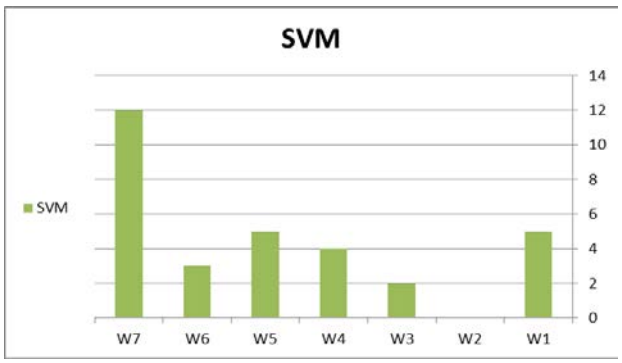


Fig. 4. Miscalssified examples across different classes with SVM algorithm.

division by the hold-out test. However, the results across the two methods remain close and are not changing dramatically.

## V. Conclusion

Authorship analysis is a long explored area that has been implemented in a number of applications including authorship identification and plagiarism detection. Recently, researchers have paid attention for developing applications for the area of authorship identification. Many of these applications use modern machine learning techniques as they proved successful for a variety of identification problems.

In this paper, we targeted the problem of identifying the author of an Arabic text articles. To have accurate author identification, it is very important to have a strong feature set that can discriminate between the different authors. For this purpose, we collected articles of 7 authors with a total number of 456 articles and applied a feature extractor to extract a total number of 27 features. Out of this feature set, we chose the most important feature and normalized their values to end up with a feature vector of 12 dimensions.

To prove the strength of the proposed feature set, we ran experiments with two popular classifiers: FT and SVM. An accuracy of 82% was achieved with the FT method and hold-out testing proving the robustness of the proposed feature set. In addition, in one of the classes, 100% accuracy has been achieved. Moreover, we tested FT using 10-fold cross validation and the method retained, to a certain degree, its accuracy result.

In the future, we plan to expand the feature set and experiment with more classifiers. In addition, we plan to develop a stand-alone application that can be commercially implemented and widely distributed.

## References

[1] A. Abbasi and H. Chen, " Applying authorship analysis to extremist-group web forum messages", *Intelligent Systems Journal*, vol. 20, no. 5, pp. 67-75, 2005.

[2] D. I. Holmes, "The evolution of stylometry in humanities scholarship", *Literary and linguistic computing*, vol. 13, no. (3), pp. 111-117, 1998.

[3] D. V. Khmelev, "Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts", *Journal of Quantitative Linguistics*, vol. 7, no. 3, pp. 201-207, 2000.

[4] F. Mosteller and D.L. Wallace, Applied Bayesian and Classical Inference: The Case of the Federalist Papers, New York: Springer-Verlag, 1964.

[5] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The Federalist Papers", *Computers and the Humanities*, vol. 30, no. 1, pp. 1-10, 1996.

[6] H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie, " An experiment in authorship attribution", *Proc. of the 6th International Conference on Statistical Analysis of Textual Data (JADT)*, pp. 29-37, 2002.

[7] H. Van Halteren, " Linguistic profiling for author recognition and verification", *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 199 -205, 2004.

[8] I. H. Witten and E. Frank, Data mining: Practical machine learning tools with java implementations, Morgan Kaufmann, San Francisco, CA, 2000.

[9] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, " Authorship attribution with support vector machines", *Applied Intelligence*, vol. 19, no (1-2). pp. 109-123, 2000.

[10] O. De Vel, A. Anderson, M. Corney, and G. Mohay, " Mining e-mail content for author identification forensics", *ACM Sigmod Record*, vol. 30, no. 4, pp. 55-64, 2001.

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

[12] R. Zheng, Y. Qin, Z. Huang, and H. Chen, " Authorship analysis in cybercrime investigation", *Proc. of Intelligence and Security Informatics*, Springer Berlin Heidelberg, pp. 59 -73, 2003.