African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 2, Issue 4, October-December 2023, Page No: 352-359
Website: *https://aaasjournals.com/index.php/ajapas/index*
معامل التأثير العربي 2023: (1.55) **SJIFactor 2023: 5.689** **ISI 2022-2023: 0.557**

Punctuation Marks Effect on Arabic Authorship Attribution Using a Variable Length of Character *N*-grams

Fatma Howedi [1 *], Souad Alharm [2]
[1,2] Computer Science, Information Technology Collage, Alasmarya Islamic University, Zliten, Libya

[*] *Corresponding author: f.howedi@asmarya.edu.ly*

Received: October 20, 2023	Accepted: December 10, 2023	Published: December 17, 2023

Abstract:
The problem of Authorship Attribution (AA) relies on distinguishing features to capture the writing style of the author. The models of character n-gram have been identified as the most successful features for representing the stylistic properties of a text. This study explores the use of punctuation marks within character n-grams as a feature representation of a document for Arabic AA of short texts. Starting from a variable length of character n-grams (2-, 3-, 4-, and 5-grams) used to generate feature vectors, the experiments were conducted independently for each feature condition, using *Chi*-squared selection method with varying feature set sizes. Different machine learning was trained to represent the probability of membership for certain authors. This study showed that by adding punctuation to the construction of character n-grams, the length of 5-grams and 4-grams enhanced the classification performance more than smaller lengths of 2-grams and 3-grams conditions. The results confirmed a high attribution effectiveness at 0.93% with Macro F_1- measure for AA of short texts. This method yields an improvement in the performance of AA by 7.5% with Macro F_1- measure that when punctuation marks are used within character n-grams. The punctuation therefore provides further insight into the writing style of the author. This study contributes in improving the attribution performance of the issue of text size for Arabic authorship attribution.

Keywords: Authorship Attribution, Text classification, Punctuation marks, Character n-grams, Machine Learning.

Cite this article as: F. Howedi, S. Alharm, "Punctuation Marks Effect on Arabic Authorship Attribution Using a Variable Length of Character N-grams," *African Journal of Advanced Pure and Applied Sciences (AJAPAS)*, vol. 2, no. 4, pp. 352–359, October-December 2023

تأثير علامات الترقيم على اسناد التأليف للنصوص العربية باستخدام أطوال متغيرة لأحرف *n*-grams

فاطمة سليمان هويدي[1*] سعاد إبراهيم الهرم[2]
[2،1] قسم علوم الحاسوب، كلية تقنية المعلومات، الجامعة الأسمرية الإسلامية، زليتن، ليبيا.

الملخص
تتمثل مشكلة اسناد التأليف Authorship Attribution (AA) في كيفية استخراج السمات (features) التمييزية التي من شأنها تساعد على تمثيل ومعرفة اسلوب الكتابة للمؤلف. تعتبر طريقة أحرف *n*-grams التقليدية من أكثر الطرق نجاحاً في تمثيل الخصائص الأسلوبية للنصوص. تهدف هذه الدراسة لاستخدام علامات الترقيم ضمن أحرف *n*-grams كسمة أسلوبية لإسناد التأليف للنصوص العربية القصيرة، وذلك من أجل تحسين أداء

مهمة الاسناد واستكشاف إلى أي مدى يمكن أن تأثر هذه العلامات على تحسين دقة اسناد التأليف. بدءاً من استخدام أطوال متغيرة لأحرف *n*-grams (2, 3, 4, and 5-grams) تم إجراء التجارب بشكل مستقل لكل طول من n-grams، مع علامات الترقيم في بعض التجارب وبدون هذه العلامات في تجارب أخرى. كما تم تدريب ثلاث خوارزميات تصنيف لتعلم الآلة وذلك لتعزيز أداء مهمة اسناد التأليف بشكل أفضل. أظهرت هذه الدراسة انه بإضافة علامات الترقيم إلى أحرف *n*-grams فإن الأحرف ذات الأطوال 5-grams و 4-grams رفعت من أداء الاسناد بشكل أكبر من الأطوال ذات الأحجام الصغيرة كما في حالتي 2-grams و 3-grams. كما أثبتث نتائج هذه الدراسة أنه باستخدام علامات الترقيم ضمن أحرف *n*-grams مع أطوال متغيرة (n)، أظهرت فعالية كبيرة في تحسين دقة الاسناد بنسبة 7.5%، حيث ساعدت هذه الطريقة في الحصول على نسبة اسناد عالية بلغت 93% من معدل القياس F_1 (F_1- measure) للنصوص العربية. وبالتالي فإن علامات الترقيم توفر معلومات إضافية ومزيداً من المعرفة عن أسلوب كتابة المؤلف. وبذلك فإن هذه الطريقة التي تعتمد على تضمين علامات الترقيم مع أحرف *n*-grams ذات الأطوال المتغير تساهم بشكل كبير على تحسين أداء مهمة اسناد التأليف للنصوص العربية القصيرة.

الكلمات المفتاحية: إسناد التأليف، تصنيف النصوص، علامات الترقيم، أحرف *n*-grams، تعلّم الآلة.

Introduction

Authorship attribution (AA) is the process of identifying the authors of an anonymous document, usually from a pre-defined set of writers [1][2]. Automatic AA provides a valuable tool for a variety of applications including forensic examinations, academic plagiarism detection, and intelligence source identification. While text content is the only element employed in text classification (TC), writing style plays a significant role in AA. This makes AA a different kind of classification challenge. Writing style can be broadly understood as the underlying approaches of sentence constructions that are subject to analysis through the use of a range of components known as stylometric analysis. Different metrics of lexical repetition are among the stylometric traits that are essential in this context [3]. Because they may capture nuances at the lexical, syntactic, and structural levels, character n-gram characteristics are a very effective way to represent texts for stylistic reasons [4][5].

The field of AA has seen several studies on very short texts in multiple languages in recent years. Promising outcomes have been observed in certain studies of short texts using 500 character [6] or 500 words [7][8]. Short texts, ranging in length from 290 to 800 words, are used in this study to approximate the length of ancient Arabic texts. This enables to test the scalability of our methodology using very short text documents to show the impact of excluding and including punctuation marks within character n-grams on Arabic Authorship Attribution.

In the case of AA, machine learning techniques have produced results that are satisfactory. Among other classifiers, the estimates from the Naive Bayes (NB) and Support Vector Machine (SVM) classifiers have demonstrated high efficiency and accuracy in classification tasks [9][10]. As a result, these classifiers seem suitable for determining the author of unknown documents. Thus far, numerous studies in various languages have benefited from the remarkable outcomes of n-gram-based algorithms. These findings have led to the successful application of the n-grams method for text classification in AA in recent research projects [8][10][11][12]. In this study, three machine learning algorithms of SVM, NB, and KNN were trained using a variable length of character n-gram features including and excluding punctuation marks.

In the present study, several features of character n-gram with a variable length were used to generate vectors of features for Arabic short texts written on the same topic. Furthermore, a pre-processing step that leaves punctuation marks in texts intact for use as character features within n-grams is also included. This for improving the performance of classification for AA task in short Arabic texts. Using methods from Information Retrieval (IR), Machine Learning (ML), Natural Language Processing (NLP), and Data Mining, the comparative performance of each feature vector is examined. In this work, three distinct ML algorithms SVM, NB, and KNN—were employed because they enable taking into account a broad range of pertinent features without experiencing a discernible decline in accuracy if the majority of these features turn out to be irrelevant [13].

1. Material and methods

This study approached AA as a text classification task. Documents are automatically categorized by TC based on a predetermined set of authorship classes. The majority of TC systems use two stages. Figure 1 illustrates the TC approach concept used for the AA task. A pre-processed data set serves as the starting point, after which the data is divided into train and test sets. Based on predictive features that are extracted from data, training and test instances are generated. Using the training data, an ML model is constructed in the second phase. Finally, using previously unidentified documents (the "test data"), the generated model is tested. The term frequency of each selected feature is represented by numerical feature vectors in the training and test instances, which are then followed by the label of author.

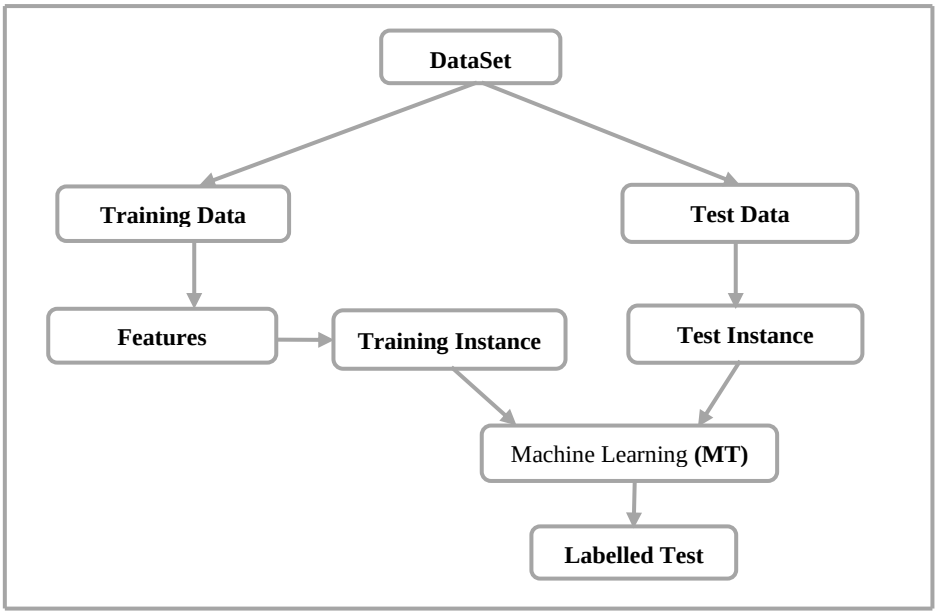

Figure 1: Conception approach of AA as a TC task.

The following section explains the process of the AA task which consists of features extraction that was computed in numerical information, and selecting features, and then the classification task was applied using ML algorithms of NB, SVM, and KNN which rely on extracted features, finally, evaluation of the method are presented.

1.1 Features Extraction

The frequency of punctuation is probably a reliable indicator of authorship; some authors, understandably, avoid using punctuation to save time by substituting commas and question marks for periods and periods [15][16]. By including punctuation marks within character n-grams and excluding them in certain experiments, this study presents particular AA experiments. The procedure of splitting the input texts into character n-gram here was preceded by replacing all punctuation marks with spaces before feature extraction in some experiments, while in other experiments the punctuation marks were split within character n-gram using a variable length of n. The language used in the text affects the features and how to extract them [18]. In this study, different lengths (n) of characters were extracted, including character 2-gram, character 3-gram, character 4-gram, and character 5-gram. Table 1 shows a description of each extracted feature of character n-grams using the following sentence example:

".“متى ذهب إلى مكة؟”

Table 1. Description of each extracted character n-gram feature (note: the underscore symbol ” _” represents spaces)

Characte n-grams	Length of grams (n)	n-grams name	The features extraction of the sentence example: “متى ذهب إلى مكة؟”
C2	2	Bi-grams	مت تى ى_ _ذ ذه هب ب_ _إ إل لى ى_ _م مك كة ة؟
C3	3	Tri-grams	متى تى_ ى_ذ _ذه ذهب هب_ ب_إ _إل إلى لى_ ى_م _مك مكة كة؟
C4	4	Tetra-gram	متى_ تى_ذ ى_ذه _ذهب ذهب_ هب_إ ب_إل _إلى إلى_ لى_م ى_مك _مكة مكة؟
C5	5	Penta-gram	متى_ذ تى_ذه ى_ذهب _ذهب_ ذهب_إ هب_إل ب_إلى _إلى_ إلى_م لى_مك ى_مك _مكة؟

1.2 Selecting Features

The present study used the technique of *Chi*-squared (Chi-x^2) as the method of features selection (FS). All features selection methods have as their common denominator the ranking of the features based on their independently calculated scores, followed by the use of an evaluation function to choose the highest scoring features [10]. By eliminating features that are unnecessary for the classification task, FS methods seek to reduce the dimensionality of a data set [14]. Reducing the curse of dimensionality to produce better classification accuracy is another

advantage of FS [14]. For the best features with the highest *Chi-x*^{2} value to be included in the feature vectors fed into the NB, KNN, and SVM classifiers.

1.3 Classification

Features extraction and classification are the two main steps in the AA task. In the current study, the step of feature extraction is focused on extracting different lengths of character n-gram features for each candidate author. The second step was to train ML algorithms by implementing three different classification algorithms SVM, NB, and KNN using the extracted features of the first stage and then calculating and comparing the results with the features of anonymous texts in order to identify the writer of anonymous texts.

In the experiments here, authorship was classified using three classifiers SVM, NB, and KNN. In addition, a 3-fold cross-validation technique was used to divide the AAAT dataset into test data and training data, which were used for building a probabilistic classification model of each classifier per author. Figure 2 shows the feature extraction and classification processes of AA that are applied in the present study.

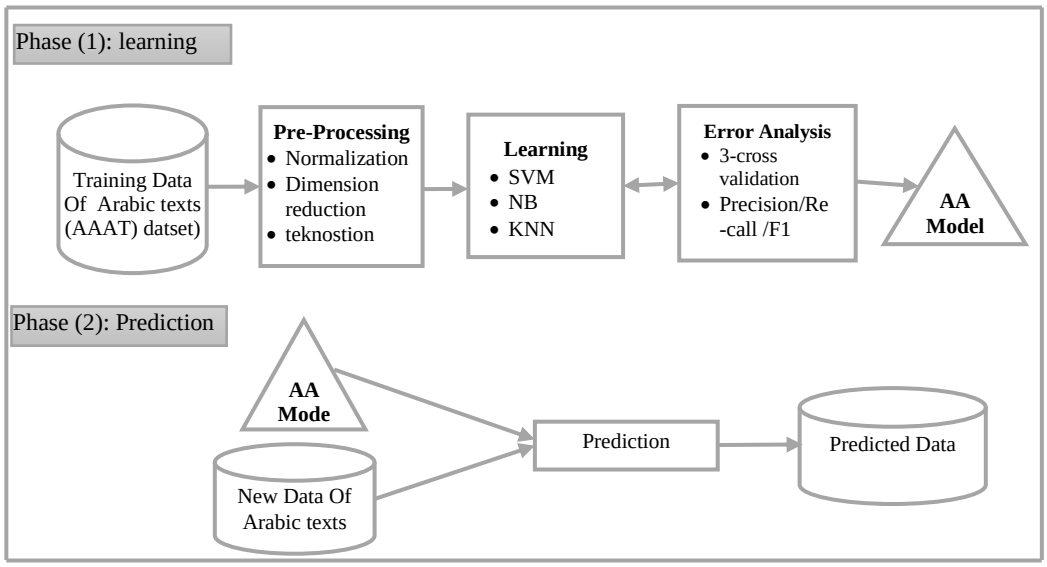

Figure 2: Classification task stages of AA process in the present study.

1.4 Evaluation

For evaluating the classification performance, the dataset was divided into 3 subsets according to the technique of *K*-fold cross-validation, using 3-folds. One subset becomes a test set and the remaining becomes training set for each iteration. With this experimental configuration, we ensure that all the text data are part of both training and test. Moreover, the measures of Macro averaged recall "$\boldsymbol{Rr}^{(M)}$", Macro averaged precision "$\boldsymbol{Pr}^{(M)}$", and Macro $\boldsymbol{F_1}$ were used to justify the experiment results.

2. Corpora and Experimental Settings

The experiments were done on an Arabic data set called AAAT, which was used in previous studies [8][10]. The AAAT dataset consists of three ancient Arabic documents per author. There are ten different authors of these texts. To make processing easier, the data was converted to UTF-8 format. Next, a pre-processing algorithm that we designed it for tokenization, punctuation mark removal, and normalisation processing received the data. The effectiveness of the feature extraction and classification phases depended heavily on the results of this text pre-processing step.

3. Results and discussion

This work investigates a pre-processing method to insert textual Punctuation Marks (PMs), so that they can be used as character features within n-grams. Moreover, *Chi*-squared scores were ranked in the experiments individually for each character n-grams condition that appeared in each document per author. Following that, the most frequent feature of each character n-grams including PMs and excluding them was selected using different frequency threshold values. The results of using different lengths of character n-grams with various sizes of the

features set by applying selection methods of *Chi*-squared are shown in Tables (2 -4) in terms of recall $Re^{(M)}$ and precision $pr^{(M)}$. The best results of each length of n-grams when including PMs and excluding them are in bold.

Table 2. The results of applying different lengths of character n-gram Including PMs within n-gram vs. Excluding PMs, using SVM.

Character n-grams	Including PMs within n-gram		Excluding PM from n-grams	
	$(Pr^{(M)})$	$(Re^{(M)})$	$(Pr^{(M)})$	$(Re^{(M)})$
C2	69.17%	**70.00%**	68.33%	66.67%
C3	**93.75%**	83.33%	92.50%	86.67%
C4	**90.00%**	83.33%	83.33%	80.00%
C5	**91.79%**	83.33%	70.83%	70.00%
Average of the results of all n-grams	86.17%	79.99%	78.74%	75.84%

Table 3. The results of applying different lengths of character n-gram Including PMs within n-gram vs. Excluding PMs, using NB.

Character n-grams	Including PMs within n-gram		Excluding PM from n-grams	
	$(Pr^{(M)})$	$(Re^{(M)})$	$(Pr^{(M)})$	$(Re^{(M)})$
C2	44.00%	**46.67%**	36.67%	40.42%
C3	77.67%	76.67%	**79.83%**	73.34%
C4	**90.17%**	86.67%	82.92%	83.33%
C5	**94.17%**	93.33%	77.50%	76.67%
Average of the results of all n-grams	76.50%	75.83%	69.23%	68.44%

Table 4. The results of applying different lengths of character n-gram Including PMs within n-gram vs. Excluding PMs, using KNN.

Character n-grams	Including PMs within n-gram		Excluding PM from n-grams	
	$(Pr^{(M)})$	$(Re^{(M)})$	$(Pr^{(M)})$	$(Re^{(M)})$
C2	45.58%	56.67%	**69.56%**	53.33%
C3	79.79%	**88.00%**	69.72%	63.33%
C4	61.58%	53.33%	**79.95%**	66.67%
C5	**75.58%**	53.33%	61.00%	53.33%
Average of the results of all n-grams	65.63%	62.83%	70.05%	59.16%

The summary in Tables (2 - 4) indicates that in most cases (especially when SVM and NB classifiers were used) character features are better when they include punctuation marks. This is because the feature set size of the AAAT data set increased when punctuation was included as a feature within the 2-, 3-, 4-, and 5-gram conditions. This indicates that this increase in the feature set size of short texts proved more details about the author. Punctuation therefore revealed something about the author. However, characters of 2-grams and 4-grams conditions using KNN seemed to have the best percentage of average precision ($Pr^{(M)}$) and recall when they excluded punctuation marks. Moreover. It is noteworthy that in all cases when using the classifiers of SVM, NB, and KNN, the most significant performance improvement occurred when punctuation marks were added to the character 5-grams condition.

From Table 1, it can be elicited that the results of the best classification improved from roughly 70.83% to 91.79% in terms of precision ($Pr^{(M)}$), and from 70.00% to 83.33% in terms of recall $Re^{(M)}$ that was when punctuation marks were added to 5-grams.

Likewise, Table 2 indicates that the best classification results improved in terms of precision $(Pr^{(M)})$ from 77.50% to 94.17% and in terms of recall ($Re^{(M)}$) from 76.67% to 93.33%, this is when punctuation was added to the 5-

gram condition as opposed to the 2-gram,3-gram, and 4-gram conditions. The reason for this was the way of the combination of punctuation within 5-grams, where one punctuation mark could appear with character 5-grams in five different ways; Thus, the frequency of appearing each punctuation mark occurs in five different n-grams which leads to an increase the number of features of the texts which can help classification algorithms to classify the author of short texts correctly. While, when punctuation marks become a part of the 2-grams, indicated that one 2-grams might occur with one character. Thus, the number of features of these 2-grams does not necessarily increase, as the case of 5-gram. In this respect, punctuation marks have less of an impact on short texts when they are added to small lengths of n-grams, such as the constrictions of 2- and 3-grams, than when they are added to 5-grams, which increases the feature size of the data set and provides more information about the author of the short text. The punctuation therefore reveals the author's style.

Additionally, based on all of the data, it was found that adding punctuation marks to character n-gram constructions improved classification performance more when the n-gram length was five than when it was two, this also held when punctuation were combined with the 3-gram and 4-gram conditions.

Generally, in both cases when punctuation marks are included or excluded from the AAAT data, the best classification score was produced with 5-grams level using NB classifier, which produced the largest improvement in average precision ($Pr^{(M)}$). On the other hand, the best average of the results of all n-grams conditions was using SVM classifier in term recall $(Re^{(M)})$ when punctuation was included.

3.1 The effect of punctuation to potential robustness to algorithms of machine learning

The classifiers comparison experiments are shown in this section. In order to prevent situations where one authorship class was not more well-represented than another, the classifiers were presented with an equal number of instances for each author, a circumstance that can lead to incorrect classifications [17]. These experiments demonstrated the robustness of three distinct classifiers in performing AA using various features of character n-grams, including and excluding punctuation marks, using a small amount of data from short Arabic texts. Furthermore, the classifiers were employed using the RapidMiner toolkit. It was recalled that 3 text documents for each an author were collected and stored in AAAT dataset. Thus, a three-fold cross validation was performed. Two texts were used for the training phase and a third text was used for testing for each fold. The process was then carried out once more. As a result, every fold was saved for testing, necessitating three attempts at the classification task. The average results for the AAAT data set were then determined by combining the outcomes of these three classification tasks. Table 5 indicates the micro F_1- measure of the best attributions using SVM, KNN and NB classifiers with different lengths of character n-grams including punctuation and excluding them. While, table 6 shows Micro F_1- measure averaged of overall attribution.

Table 5. The results of Micro F_1- measure Averaged of overall attribution obtained by applying different lengths of character n-gram Including PMs within n-gram vs. Excluding PMs, using NB, SVM, and KNN

classifiers	Including PMs within n-grams				Excluding PMs from n-grams			
	2-gram	3-gram	4-gram	5-gram	2-gram	3-gram	4-gram	5-gram
SVM	0.67738	0.84454	0.82809	0.82571	0.63499	0.86238	0.72904	0.67047
NB	0.42539	0.74834	0.85738	**0.93238**	0.37079	0.73620	0.78952	0.76571
KNN	0.47970	0.69589	0.51166	0.54284	0.53776	0.61904	0.65809	0.51277

According to Table 5, the best classification result was obtained by applying the NB classifier on 5-gram when punctuation marks were included, yielding a score of 0.93238 of the *F1*- measure. Conversely, the application of the NB classifier on the 2-gram produced the lowest classification result when punctuation was excluded at 0.37079 of the *F1*-measure. This classification result of 0.37079 could not be used to ascertain whether a text was written by the author or not. This means that the NB classifier provided more information about the author of the short texts when punctuation marks were added to all n-grams conditions: 5-gram, 4-gram, 3-gram, 2-gram which respectively provided at the levels of 0.42539, 0.74834, 0.85738, and 0.93238 compared to the same experiments when punctuation marks were excluded, which reached less percentage at the levels of 0.37079, 0.73620, 0.78952, 0.76571 respectively.

The results of the SVM classifier on the 5-gram, 4-gram, and 2-gram conditions showed that excluded punctuation marks also produced the lowest classification results. These conditions yielded *F1 measures* of 0.63499, 0.72904,

and 0.67047, respectively. Included punctuation marks produced higher classification results at 0.67738, 0.82809, and 0.82571, respectively. The exception to this was the 3-gram condition, where excluded punctuation marks produced a high *F1*- measure of 0.86238 before falling to 0.84454 when punctuation marks were added, as shown in Table 5.

Regarding the KNN classifier, the classification results of including punctuation within character n-grams and excluding them provided different results of F_1- measure. As indicated in Table 5, some conditions of n-grams (3-gram and 5-gram) improved the classification when including punctuation, while some conditions of n-grams (2-gram and 4-gram) improved the classification when punctuation was excluded.

According to Table 6, both the NB and SVM classifiers performed better at classifying when punctuation was included within n-grams. Specifically, the classification performance of the average of all results was enhanced by 7.5% and 6.25%, respectively, when NB and SVM were used. This means that the procedure of adding punctuation within n-grams enhanced the performance of the proposed method. Therefore, It can be stated that punctuation frequency is probably a reliable indicator of AA.

Based on the improvement values shown in Table 6, the classification performance of the KNN classifier same to not be good when punctuation marks were included within character n-grams. This was probably caused by the fact that the KNN classifier was built to handle even larger amounts of data than were utilised in the experiments of this study.

In conclusion, table 6 illustrates that, among the three ML algorithms employed in this investigation, the NB classifier yielded the best-improved value of including punctuation marks within character n-grams. On the other hand, as Figure 3 illustrates, the SVM classifier produced the highest average of overall attribution. Additionally, the KNN classifier displayed the lowest classification based on figure 3.

Table 6. Micro F_1- measure Averaged of overall attribution obtained by using NB, SVM, and KNN, and the improvement of Including PMs within n-gram vs. Excluding PMs from n-grams.

classifiers	Including PMs within n-grams	Excluding PMs from n-grams	Improvement rate
SVM	**78.75%**	72.5%	6.25%
NB	**73.5%**	66.0%	**7.5%**
KNN	55.25%	**57.5%**	- 2.25%

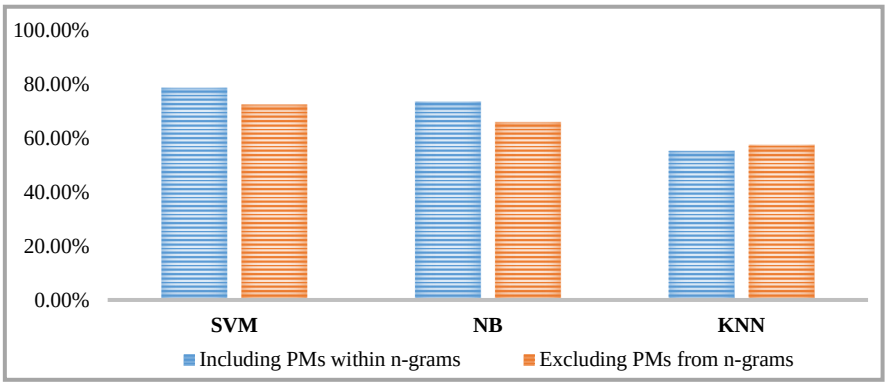

Figure 3: Micro F_1- measure the percentage of the best attribution obtained by using NB, SVM, and KNN, Including PMs within n-gram vs. Excluding PMs from n-grams.

4. Conclusion

This paper examines the impact of using punctuation marks within character n-grams as a feature representation by applying three different Machine Learning (ML) for Authorship Attribution (AA) of short Arabic texts, where each ML (SVM, NB, and KNN) was trained against two text documents per author. The selection method of *Chi*-squared was used to carry out the experiments independently for each feature condition, utilising different feature set sizes. The experimental results provided that the frequency of punctuation marks with character n-grams is likely to be a good indicator of authorship. This procedure improved the performance of the method at 7.5% of

macro F_1- measure as average of overall features. It can be concluded that the classification performance was improved by the use of punctuation marks within character n-gram features. Punctuation appears to reveal more details about the writing style of the author. This is conceivable because while some writers utilise punctuation frequently, others do not.

According to the experimental results, the SVM and NB classifiers outperformed the KNN classifier using all conditions of n-grams. Furthermore, most of the conditions of n-grams obtained the best percentage of macro F_1-measure when they added to punctuation marks. In addition, the length of 5-grams increased the classification performance highly compared to the lengths of 2-grams, 3-grams and 4-grams conditions.

This study mainly contributes to the field of AA in short texts using a variable length of n-grams, and is specific for the Arabic language. This study paid a lot of attention to improving the classification performance of the issue of text size by using the specific procedure of using punctuation marks within character n-grams in the experiments of authorship attribution.

References

[1] A. B. Lopez-Monroy, M. Montes-y-Gomez, L. Villasenor-Pineda, J. A. Carrasco-Ochoa, and J. F. Martinez-Trinidad, "A new document author representation for authorship attribution", Conf. paper, 2012, "doi:10.1007/978-3-642-31149-9_29".

[2] D. Estival, Tanja Gaustad, Son Bao Pham, and Will Radford 2007. TAT: An Author Profiling Tool with Application to Arabic emails.*Proceeding of the Australasian language Technology Workshop*. pp 21-30.

[3] W. Oliveira Jr.a, E. Justino a, L.S. Oliveira b 2013. Comparing Compression Models for Authorship Attribution. *Forensic Science International Journal.* PP 100- 104.

[4] J. Houvardas &E. Stamatatos 2006. N-gram Feature Selection for Authorship Identification. *J. Euzennat and J. Domingue (Eds): Proceeding of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA).* PP 77-86.

[5] Grieve, "Quantitative authorship attribution: an evaluation of techniques", *Literary and Linguistic Computing*", PP 251-270, 2007.

[6] C. Sanderson & S. Guenter 2006. Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. *Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp 482-491.

[7] M. Koppel, J. Schler and S. Argamon 2009. Computational Methods in Authorship Attribution.*Journal of the American Society for Information Science and Technology.* PP 4-8.

[8] F. Howedi and M. Mohd, Z, Aborawi, and S. A.Jowan, "Authorship attribution of short historical arabic texts using stylometric features and a knn classifier with limited training data", *Journal of Computer Science 16(10):1334 – 1345, 2020*, "doi:10.3844/jcssp.2020.1334.1345".

[9] R. Hong, R. Tan and F. S. Tsai 2010. Authorship Identification for Online Text.*International Conference on Cyberworlds.* PP 155-162

[10] 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, 2014, pp: 48- 56.

[11] E. Stamatatos, "On the robustness of Authorship Attribution based on character n-gram features", *Jornal of Low & Policy*, 2013, pp: 427-439.

[12] H. J. Escalante, T. Solorio, and M. Montes-y-G'ozmez, "Local histograms of character n-grams for authorship attribution", *in Proc, of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, 2011, pp: 288-289.

[13] M. Koppel, J. Schler, S. Argamon and Yaron 2012. Fundamental Problem of Authorship Attribution.

[14] M. Ikonomakis, S. Kotsiantis and V. Tampakas 2005. Text Classification Using Machine Learning Techniques.*Wseas Transactions on Computers, Vol 4 (8). pp 966-974.*

[15] C. Rodriguez, D. A. P. Alvarez, C. A. M. Sifuentes, G. Sidorov, L. Batystian, and A. Gelbukh, "Authorship attribution through punctuation n-grams and averaged combination of SVM Notebook for PAN at CLEF 2019", 2019.

[16] U. Stanczyk, and K. A. Cyran, "Can punctuation marks be used as writer invariants? rough set-based approach to authorship attribution", *2nd European Computing Conf. (ECC'08)*, 2008, pp: 228-233.

[17] K. Luyckx 2010. Scalability Issues in Authorship Attribution.*PhD thesis, Department of Linguistics, Faculty of Arts and Philosophy, Dutch UPA University.*

[18] M. Eder, "Style-Markers in authorship attribution a cross-language study of the authorial fingerprint", *Polish Academy of Science, Institute of Polish Language, 2011*.