Computers, Materials & Continua CMC, vol.60, no.2, pp.421-433, 2019

A Comparative Study of Machine Learning Methods for Genre Identification of Classical Arabic Text

Maha Al-Yahya[1, *]

Abstract: The purpose of this study is to evaluate the performance of five supervised machine learning methods for the task of automated genre identification of classical Arabic texts using text most frequent words as features. We design an experiment for comparing five machine-learning methods for the genre identification task for classical Arabic text. We set the data and the stylometric features and vary the classification method to evaluate the performance of each method. Of the five machine learning methods tested, we can conclude that Support Vector Machine (SVM) are generally the most effective. The contribution of this work lies in the evaluation of the five machine learning methods for the task of genre identification for classical Arabic text using stylometric features.

Keywords: Genre classification, Arabic text, supervised machine learning.

1 Introduction

Genre is defined as "a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" [Yates and Orlikowski (1992)]. Genre classification is valuable for writers, readers, and critics, it allows them to comprehend the accepted conventions for a specific class of literature. Another important aspect of document genre is that it helps the reader understand the content and reduces the cognitive effort [Crowston and Kwasnik (2003)]. Moreover, in information retrieval tasks in search engines and digital libraries, the genre classification of a document is as important as the document content since it provides significant information that would yield the results more relevant to the user.

In this study, we approach the problem of genre identification as an attribution task [Jockers (2013)]. Attribution studies have been mainly associated with author attribution studies, studies in which the author of a disputed text is identified among several candidate authors [Juola (2006)]. However, they have been used for other text analysis tasks such forensic linguistics [Rocha, Scheirer, Forstall et al. (2017); Afroz, Brennan, Greenstadt et al. (2012)], plagiarism detection [Ramnial, Panchoo, Pudaruth et al. (2016)], chronology studies -observing the developing voice of an author over a period of years [Juola (2007)], stylistic inconsistencies in collaborative writing [Glover and Hirst (1995)],

[1] Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

[*] Corresponding Author: Maha Al-Yahya. Email: malyahya@ksu.edu.sa.

literary influence [Jockers (2013)], and genre detection [Jockers (2013)] .

In general, attribution methods can be classified into unsupervised and supervised. Unsupervised methods are used when the data is unlabeled, and when the focus is on how the data is structured or grouped. However, supervised methods need labeled data, training data that has been annotated with class tags. These methods use machine learning algorithms to create a classification model (classifier) using the training data to classify new unseen data. A detailed explanation of machine learning methods for text classification can be found in Aggarwal et al. [Aggarwal and Zhai (2012); Juola (2006)].

For genre identification, the effectiveness of supervised machine learning methods have been evaluated for attribution studies for the English language [Jockers and Witten (2010)], however no comparative evaluation exists for supervised methods for the genre identification task of classical Arabic texts. In this study we aim to fill this gap by comparing a number of supervised methods for the genre identification task. Our objective in this study is to answer the following two questions:

1. How do supervised machine learning methods compare for the task of genre identification in classical Arabic text using the most frequent words of a corpus (MFW) as features?
2. For the best performing supervised machine learning method identified from the first question, what is the optimal value of MFW for the task of genre identification for classical Arabic text?

This article is organized as follows: Section 2 presents the related work for the genre identification task. Section 3 presents the research methodology. Section 4 presents the experiments. Section 5 presents the results obtained and Section 6 presents the analysis and discussion of the results. Finally, Section 8 presents the conclusions and future work.

2 Related work

The genre identification task can be described as a classification problem or as an attribution problem. Thus, our review of relevant work will include work in the fields of Arabic text classification and Arabic text attribution. A recent comprehensive review on Arabic text classification primary studies [Alabbas, Al-Khateeb, Mansour et al. (2016)] indicates that the majority of Arabic text classification studies apply the classification task on datasets which use the Modern Standard Arabic variation rather than the Classical Arabic variation, therefore there is shortage on studies for Classical Arabic. Modern Standard Arabic (MSA) is the standard for Arabic-speaking countries and is the formal language used in news and most media. Classical Arabic on the other hand, is the medieval dialects of Arabs, and is the language of the Quran (Holy book for Islam), it is still in wide use today in literature, language, and Islamic scholarship.

There are several factors which play an important role in the text classification task and the text attribution task and can influence the performance. These factors include the choice of algorithm, the features selected, and the preprocessing of the features. Comparative work usually focuses on any of these aspects to see how it influences the performance. For example, for Arabic text classification, the work presented in Al-Thubaity et al. [Al-Thubaity, Alhoshan, Hazzaa et al. (2015)] evaluates the use of n-

grams for Arabic text classification. Features selected for the study include word n-grams: single words, 2-grams, 3-grams, and 4-grams. The data set used is a collection of Arabic news articles by the Saudi Press Agency, a Modern Standard Arabic (MSA) dataset which consists of 1,526 texts evenly divided into six news classes; cultural, sports, social, economic, political, and general. The classification methods are supervised methods including K-Nearest Neighbor (KNN), Naive Bayes (NB), and Support Vector Machines (SVM). The results indicate that single words achieve better results for the classification task, and the SVM classifier achieved the best accuracy at 72%. A similar comparative study [Khorsheed and Al-Thubaity (2013)] uses a corpus of seven datasets including news articles, websites, writers, forums, religious topics and Arabic poems. It compares five supervised algorithms for Arabic text classification: KNN, NB, C4.5, Artificial Neural Networks (ANN) and SVM. The results indicate that SVM achieved the best accuracy at 72%.

Another comparative study of Arabic text classifications algorithms using various Arabic stemmers is presented in Hmeidi et al. [Hmeidi, Al-Ayyoub, Abdulla et al. (2015)]. The authors compare four supervised classification algorithms: SVM, NB, KNN, Decision Tress, and Decision Tables. Using a dataset of news articles in MSA, and applying preprocessing that includes removing stop-words, punctuation and diacritics, the results of the comparative study shows that SVM obtained good results with light stemming, an accuracy of (98%).

Regarding text attribution studies for Arabic, the majority of work targets the task of authorship attribution and aims to identify the author of the text from a given set of candidate authors. Since our study is focused on genre identification for classical Arabic, we will limit our review to studies on classical Arabic text attribution, as they are the most relevant to our work.

The work in Howedi et al. [Howedi and Mohd (2014)] presents an experiment on authorship attribution for short historical Arabic texts written by 10 authors (3 texts per author). They use n-grams on the word and character level as features, and compare the NB supervised classification method against SVM. The best results were obtained by the NB classifier (96%) using word uni-grams (single words). The SVM classifier achieved an accuracy of 76.67%.

Another attribution study that focuses on comparing different variations of NB for the authorship attribution task for classical Arabic books is presented in Altheneyan et al. [Altheneyan and Menai (2014)]. The authors use a number of features including word level features, character level features, punctuation and other statistical features. Using a similar approach but on classical Arabic poetry, the work in Ahmed et al. [Ahmed, Mohamed, Mostafa et al. (2015)] explores authorship attribution by comparing three supervised machine learning methods: NB, SVM, and Sequential Minimal Optimization (SMO). In this comparative study, features such as characters, sentence length, word length are used. In addition, the authors exploit other poetry related features such as rhyme, and meter. They report the best performance at 98.15% by the SMO method using word length features.

The work in Ouamour et al. [Ouamour and Sayoud (2012)] presents an experiment on author attribution for short (209-800 words) classical Arabic historical texts using

character n-grams as features, and a supervised SMO-SVM classifier. The results indicate a classification accuracy of 80%. Using the same dataset (short historical texts), another study [Ouamour and Sayoud, (2013)] present a comparative study of seven supervised methods for attribution: distance based classifiers including Manhattan distance, Cosine distance, Stamatatos distance, and Canberra distance, a Multi Layer Perceptron (MLP) classifier, the SMO-SVM classifier and a linear regression classifier. The best results are obtained by SMO-SVM at 80% accuracy.

In our earlier work [Al-Yahya (2018)], we explored and evaluated the use of stylometric analysis of classical Arabic text to support the task of automated genre detection. In the study, unsupervised clustering and supervised classification were applied on the King Saud University Corpus of Classical Arabic texts (KSUCCA) using the most frequent words in the corpus (MFWs) as stylometric features. Four popular distance measures established in stylometric research, namely, classic delta, Eder's delta, Argamon's linear delta and the Canberra distance are evaluated for the genre detection task using unsupervised methods. The results of the experiments show that stylometry-based genre clustering and classification align well with human-defined genre. However, in that study only one measure was explored for supervised classification, the Delta classifier, which is a distance-based classifier. In this study, we attempt to compare other supervised methods to investigate which methods perform well for the genre classification task for Arabic text.

From the review of relevant work, we can see that no comparative studies exist that compare supervised methods for the genre identification task using stylometric features of the text and the field require studies which aim to fill the gap in the research on classical Arabic genre identification.

3 Methodology

A framework enables us to concentrate on the important factors which might have influence on the task of genre identification. Juola [Juola (2006)] identifies a theoretical framework to compare different methods for authorship attribution. This method is comprised of three phases: standardization/normalization, determination of the event set, and the analysis method. Normalization is the process of converting the events to a standard normal canonical form. The event set refers to the features we wish to include as indicators of style. The analysis method is the process by which the features are processed to produce the results; they range from distributional methods to machine learning methods. We will adopt this framework for our comparative analysis.

Regarding normalization, we apply tokenization to the text, converting it into a set of tokens. For the event set, There are a number of features which have been used as style markers in stylometric studies, these include [Juola (2006)]: Lexical features such as vocabulary, vocabulary properties, function words; Syntactic features such as part of speech; Orthographic features such as stemming or character n-gram. For any of the chosen features, the collection of documents is then converted to a set of feature vectors which are analyzed to detect genre. The vectors that are most similar probably belong to the same genre. We use single words as they have shown good results for Arabic text classification [Al-Thubaity, Alhoshan, Hazzaa et al. (2015)]. A popular and very

successful feature that has been reported in author attribution literature using single words is the most frequent words (MFWs) [Juola (2006)]. These words are usually non-content bearing words such as determiners, pronouns, conjunctions, prepositions, etc. For the analysis method, since our dataset is already labeled with genre labels, we will use supervised methods to evaluate the effectiveness for the genre identification task. Five supervised methods common in stylometric literature are compared; Delta, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), and nearest shrunken centroids (NSC). The methods are described in the following sections. Using this framework, we conduct our experiments to evaluate and compare the effectiveness of supervised methods for the genre identification task.

3.1 The delta classifier

The delta classifier is a distance based classifier based on the Burrows Delta [Burrows (2002)] which is used in stylometric analysis for the task of authorship attribution. The classifier calculates the distances between documents in the training set, and the test document is then classified with the same class as the document with the smallest distance [Argamon (2008)]. The implementation we use for this classifier is the implementation provided by Mike et al. [Mike, Rybicki and Eder (2016)]. In this study, we use four different kernels for the distance computation; classical delta, Eder's delta, Argamon's delta, and Canberra distance.

3.2 The k-nearest neighbor (KNN)

The K-Nearest Neighbor (KNN) is an instance based learning algorithm. In the KNN classifier, the similarity between a new document and documents in the training set is computed, and the k most similar documents are selected. The target document is assigned the most common class of its k nearest documents. There are a number of similarity measures, usually the cosine similarity or the Euclidean distance is used [Hotho, Nürnberger and Paaß (2005)]. The choice of k and the choice of distance metric can be critical and can affect the results [Han, Kamber and Pei (2011)]. Its main advantages include its simplicity, and ease of implementation and good accuracy results if the parameters are selected sensibly [Hmeidi, Al-Ayyoub, Abdulla et al. (2015)]. The major drawbacks of this method are the high computation and memory requirements.

3.3 The support vector machine (SVM)

The Support Vector Machine (SVM) classifier is a supervised machine learning algorithm that transforms the training data into a higher dimension and builds an N dimensional hyperplane which splits the data into two classes using training tuples called "support vectors" [Han, Kamber, and Pei (2011)]. The aim of a SVM classifier is to determine the separators that best separate classes [Aggarwal and Zhai (2012)]. SVM is considered an excellent classifier for text documents, and is widely used in text classification where high dimensionality is the norm. However, its disadvantages include complexity and high memory requirements [Hmeidi, Al-Ayyoub, Abdulla et al. (2015)].

3.4 The naive bayes (NB) classifier

The Naive Bayes (NB) classifier is a probabilistic classifier based on the assumption that the class of a document has some relation to the words that appear in it, which can be described as the conditional distribution [Hotho, Nürnberger and Paaß (2005)]. Using the Bayesian formula, the probability of a class given the words of the document is computed. The NB classifier is known to perform well in terms of accuracy and speed with high dimensional data and is generally considered a good classifier for the text categorization task [Han, Kamber and Pei (2011)].

3.5 The nearest shrunken centroid (NSC)

The nearest shrunken centroids (NSC) classifier [Tibshirani, Hastie, Narasimhan et al. (2003); Hastie, Tibshirani and Friedman (2009); Jockers and Witten (2010)] uses the training documents to compute a standardized centroid for each class (profile). It then compares the centroid of the new document to each of the class centroids. The class whose centroid is the closest to the target centroid is considered the predicated class. The NSC classifier is considered suitable for high dimensional data such as text.

3.6 The dataset

We used the King Saud University Corpus of Classical Arabic (KSUCCA) as our dataset [Alrabiah, Al-Salman and Atwell (2013); Alrabia, Al-Salman, Atwell et al. (2014)]. Since the dataset is already labeled with genre it will be suitable for use in supervised machine learning classifiers. The division of training and test data set varies in the literature, 60/40, 70/30, or 80/20. Since our dataset is relatively small, we started with a (80/20) percentage split. The dataset was divided between a training set including 80% of the texts selected randomly from each genre, and a test set comprising the remaining 20% of texts. The corpora statistics are presented in Tab. 1.

Table 1: KSUCCA corpus statistics

Genre	No. of texts in corpus	No. of words	No. of training texts	No. of testing texts
Religion	150	23645087	120	30
Linguistics	56	7093966	45	11
Literature	104	7224504	83	21
Science	42	6429133	34	8
Sociology	32	2709774	26	6
Biography	26	3499948	21	5
Total	**410**	***50602412***	**328**	**82**

4 Experiments

Each classifier has different parameters which need to be set, for the Delta classifier, the number of features used need to be identified. We use a MFW range from 50-500 with an increment of 10 MFW for each distance measure, generating 46 models for each distance

measure giving a total of 184 distinct classification models. Since we are interested in finding the best performance by evaluating how the distance measure behaves when the number of MFW changes, a graph was plotted taking into consideration quantities of MFW taken for experimentation. We then compare the accuracy of genre identification.

For the other four machine learning methods, we set the parameters according to the implementation used. For the KNN classifier, the tuning parameter is the number of nearest neighbors (the value of k). We set the value of k to 5 (5 nearest neighbors). We experimented with various values of K on our data, starting from K=1, up to K=9. We found that the best value of k is equal to 5, which gives us the minimum error rate. The KNN uses all features present in the data (MFWs). For SVM, all features present in the data are used to generate the classification model (MFWs). For the NB classifier, the default delta distance measure is used. For the NSC classifier, the default classic delta measure is used. We ran the four classifiers: NSC, NB, SVM, and kNN with an MFW ranging from 50-500 in increments of 10 MFW, thus 46 models were generated for each classifier, giving a total of (46*4=184) classification models (classifiers).

To answer our second question: "what is the optimal value of MFW for the task of genre identification for classical Arabic text?", we conducted additional tests to examine if any improvements on the results will appear if the size of the MFWs list is changed for SVM. We performed additional testing on the SVM classifier for varying sizes of MFWs ranging from 500-5000, with a step size of 100.

5 Results

5.1 Delta classifier distance measure effect on genre identification accuracy

Tab. 2 shows the minimum accuracy, maximum accuracy, and general attributive success (average accuracy) for Delta classifiers using the four distance measures. Classic delta achieved the maximum accuracy of 80% at 130 MFWs, and Eder's delta achieved the highest overall attributive success among the distance measures with a value of 75%. The chart in Fig. 1 depicts the classification performance of each distance for the the range of MFWs from 50 until 500.

Table 2: Details of classification performance for delta

Measurement	Distance Measures			
	Classic Delta	Argamon's Delta	Eder's Delta	Canberra
Minimum accuracy	69%	66%	66%	65%
Maximum accuracy	80%	75%	78%	77%
General Attributive success	74%	70%	75%	72%

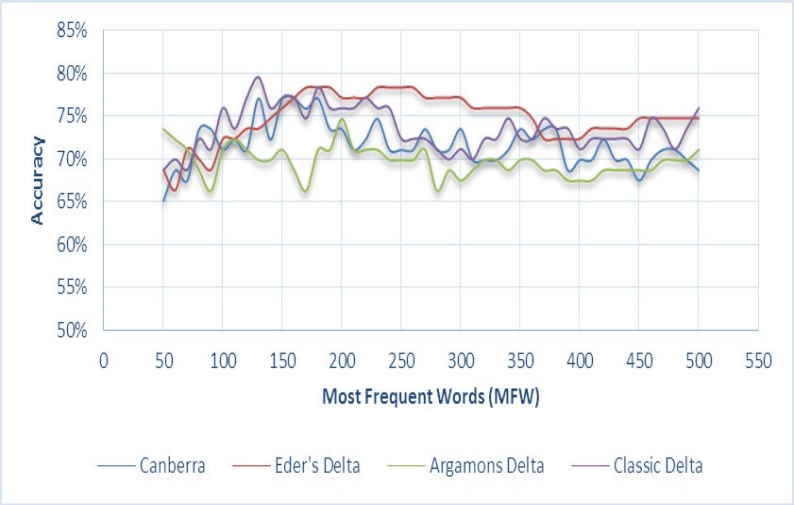

Figure 1: Classification performance relative to number of MFWs

5.2 Classification method accuracy for genre identification

We also compared the classification accuracy for genre identification among all eight methods. Tab. 3 shows the minimum, maximum, and general attributive success of genre identification for all eight models, and Fig. 2 show how each classifier performs for varying number of MFWs.

Table 3: Accuracy of supervised methods for genre identification

Classification Model	Accuracy		
	Minimum accuracy	Maximum accuracy	General accuracy
Delta with Classic	69%	80%	74%
Delta with Argamon's	66%	75%	70%
Delta with Eder's Delta	66%	78%	75%
Delta with Eder's Canberra	65%	77%	72%
NSC	53%	59%	57%
NB	49%	68%	60%
SVM	61%	86%	80%
KNN	60%	68%	65%

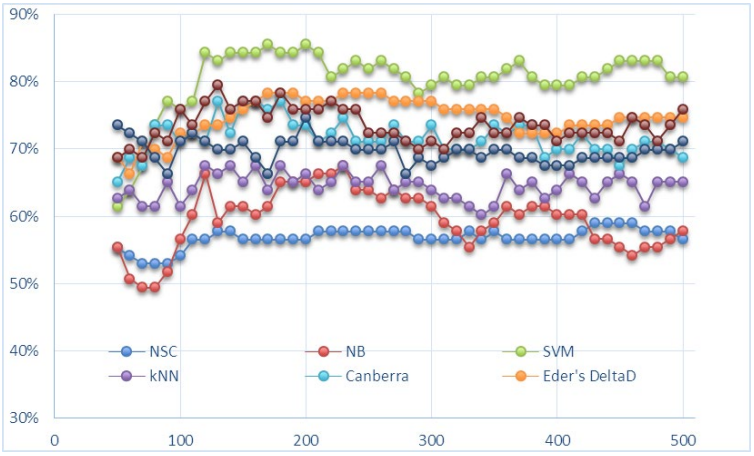

Figure 2: Classifiers performance using 50-500 MFW

Among the eight classifiers, the Support Vector Machine (SVM) classifier yielded the highest overall percentage of correct genre attribution among all the classification models used. From the chart we can see that accuracy is low at small number of MFW (50 MFW) and increases with the increase in MFW, it peaks to 86% around 170 MFW, and 200 MFW, and then drops down slightly to 78% around 290 MFW, after that performance keeps in the 80s% range until the maximum MFW is reached at 500 MFW.

5.3 Optimal MFW value for SVM

Results from additional testing conducted to examine if any improvement in SVM accuracy is achieved for MFW ranges 500-5000 are shown in Fig. 3. A slight improvement (86.7%) is achieved at 2100-3000 MFW range. The overall accuracy for the SVM classifier using (500-5000 MFW) is 84.8%.

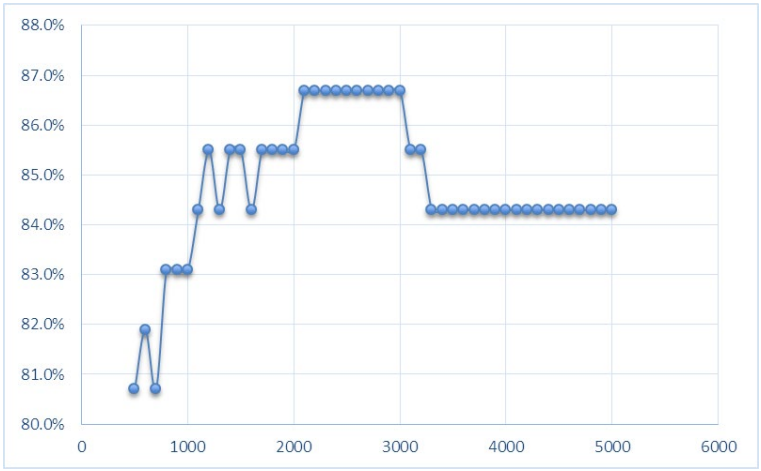

Figure 3: SVM Performance for MFW range 500-5000

6 Discussion

Comparing the effect of Delta distance measure on genre identification accuracy for the Delta classifier, we can observe that classic delta has achieved the best results at 80% accuracy using 130 MFWs. Eder's delta distance has shown the best general attributive success among the four tested distance measures at 75%. Comparing to results reported in the literature, our findings are excellent, as the maximum for distance based classifiers is 60% as reported in Ouamour et al. [Ouamour and Sayoud (2013)].

From the comparison of genre identification accuracy for all eight models, the Support Vector Machines (SVM) yielded the highest overall percentage of correct genre attribution (80% correct attributions) among varying MFWs setting from 50 to 500. It also recorded the highest accuracy of 86%.

The SVM classifier recorded the minimum accuracy of 61% at 50 MFW and the maximum accuracy at 86% using 170 MFWs. An accuracy of 86% is considered acceptable for the genre detection task, as the results reported in the literature are 69% for SVM on German novels using stylometric features of text [Hettinger, Becker, Reger et al. (2015)], and 64% for SVM on genre identification for English using the the Brown Corpus [Wu, Markert and Sharoff (2010)]. For Arabic genre identification, there are no previous studies to compare to, however, comparing to Arabic text classification literature our results are also reasonable, as the accuracy of Arabic text classification ranges between 61% and 98% [Alabbas, Al-Khateeb, Mansour et al. (2016)]. Comparing our results to attribution studies, our results are comparable to those reported in the literature, Howedi et al. [Howedi and Mohd (2014)] report 76.67% accuracy for SVM, and [Ouamour and Sayoud (2012); Ouamour and Sayoud (2013)] indicate performance of 80% by an SVM variant.

The least recorded performance was exhibited by the NB classifier at 49% using 70, and 80 MFWs. The lowest general attributive accuracy was shown by the NSC classifier at 57%. Looking at the SVM classifier performance using 50-500 MFWs, the chart in Fig. 1 shows that accuracy is low at small number of MFW (50 MFW) and increases with the increase in MFW, it peaks to 86% around 170 MFW, and 200 MFW, and then drops down slightly to 78% around 290 MFW, after that performance keeps in the 80s% range until the maximum MFW is reached at 500 MFW.

Looking at the SVM classifier's performance using 500-5000 MFWs, the chart in Fig. 3 shows the optimal value of MFW that provides the best performance for SVM at 86.7% using the range 2100-3000 MFWs. The overall accuracy for the SVM model with 500-5000 MFW is 84.8%, which is better than the overall accuracy for the SVM model using 50-500 MFW (80%). This result indicates that SVN performs better at larger values for MFW.

7 Conclusion

In this study we set out to investigate how supervised machine learning methods compare for the task of genre identification in classical Arabic text using the most frequent words of a corpus (MFW) as features. We compared five supervised methods using the KSUCCA corpus of classical Arabic text as the data set. The methods compared are: Delta, KNN, SVM, NB, and NSC. The results show that SVM outperformed the other

classifiers for the genre identification task in classical Arabic texts with a general accuracy of 80% and a highest accuracy of 86%. This result supports the results in the Arabic text classification literature, which indicates that SVM outperforms other supervised classification algorithms [Alabbas, Al-Khateeb, Mansour et al. (2016)].

The study also aimed to identify, for the best performing classifier, the SVM classifier, the optimal value of MFW for the task of genre identification for classical Arabic text. The results show that the accuracy improves for MFWs in the range of 2100-3000 MFWs, thus SVM performs better with larger number of MFWs for Arabic text genre identification.

Although this study provided important results, it is important to consider other factors that might have influence on the performance of the genre identification accuracy other than the number of MFWs. For example, preprocessing of MFWs such as stemming, or punctuation removal, or sampling of the corpus text. Future work could focus on evaluating the classification performance when preprocessing is applied to MFWs. Another limitation is that we did not do any sampling, the number of texts and the length of the texts were not comparable for all genres. This might have had an impact on the results we obtained, therefore it is important to address this issue in future studies. Moreover, new and emerging methods for natural language processing using deep learning and Convolutional Neural Networks have reported high accuracy and good performance for known natural language tasks [Xiong, Shen, Wang et al. (2018)], these methods could be examined for the task of genre detection for Arabic text in future work.

References

Afroz, S.; Brennan, M.; Greenstadt, R. (2012): Detecting hoaxes, frauds, and deception in writing style online. *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 461-475.

Aggarwal, C.; Zhai, C. (2012): A survey of text classification algorithms. *Mining Text Data*. Springer, Boston, MA, pp. 163-222.

Ahmed, A.; Mohamed, R.; Mostafa, B.; Mohammed, A. (2015): Authorship attribution in Arabic poetry. *Proceedings of the 10th International Conference on Intelligent Systems: Theories and Applications*, pp. 1-6.

Alabbas, W.; Al-Khateeb, H.; Mansour, A. (2016): Arabic text classification methods: systematic literature review of primary studies. *Proceedings of the 4th IEEE International Colloquium on Information Science and Technology*, pp. 361-367.

Alrabia, M.; Al-Salman, A.; Atwell, E.; Alhelewh, N. (2014): KSUCCA: a key to exploring Arabic historical linguistics. *International Journal of Computational Linguistics*, vol. 5, pp. 27-36.

Alrabiah, M.; Al-Salman, A.; Atwell, E. (2013): The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. *Proceedings of the 2nd Workshop on Arabic Corpus Linguistics*.

Altheneyan, A.; Menai, M. (2014): Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 473-484.

Al-Thubaity, A.; Alhoshan, M.; Hazzaa, I. (2015): Using word n-grams as features in Arabic text classification. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Studies in Computational Intelligence*, vol. 569, pp. 35-43.

Al-Yahya, M. (2018): Stylometric analysis of classical Arabic texts for genre detection. *The Electronic Library*, vol. 36, no. 5, pp. 842-855.

Argamon, S. (2008): Interpreting Burrows' Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, vol. 23, no. 1, pp. 131-147.

Burrows, J. (2002): Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguist Computing*, vol. 17, no. 3, pp. 267-287.

Crowston, K.; Kwasnik, B. (2003): Can document-genre metadata improve information access to large digital collections? *Library Trends*, vol. 52, no. 2, pp. 345-361.

Glover, A.; Hirst, G. (1995): Detecting stylistic inconsistencies in collaborative writing. *The New Writing Environment*, pp. 147-168.

Han, J.; Kamber, M.; Pei, J. (2011): *Data Mining: Concepts and Techniques*, Third Edition. Morgan Kaufmann, Burlington, MA.

Hastie, T.; Tibshirani, R.; Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Second Edition. Springer, New York.

Hettinger, L.; Becker, M.; Reger, I.; Jannidis, F.; Hotho, A. (2015): Genre classification on German novels. *Proceedings of the 26th International Workshop on Database and Expert Systems Applications*, pp. 249-253.

Hmeidi, I.; Al-Ayyoub, M.; Abdulla, N.; Almodawar, A.; Abooraig, R. et al. (2015): Automatic Arabic text categorization: a comprehensive comparative study. *Journal of Information Science*, vol. 41, no. 1, pp. 114-124.

Hotho, A.; Nürnberger, A.; Paaß, G. (2005): A brief survey of text mining. *LDV Forum-GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, no. 1, pp. 19-62.

Howedi, F.; Mohd, M. (2014): 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.

Jockers, M.; Witten, D. (2010): A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, vol. 25, no. 2, pp. 215-223.

Jockers, M. (2013): *Macroanalysis: Digital Methods and Literary History*, First Edition. University of Illinois Press, Urbana.

Juola, P. (2007): Becoming Jack London. *Journal of Quantitative Linguistics,* vol. 14, no. 2-3, pp. 145-147.

Juola, P. (2006): Authorship attribution. *Foundation and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233-334.

Khorsheed, M.; Al-Thubaity, A. (2013): Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation*, vol. 47, no. 2, pp. 513-538.

Mike, K.; Rybicki, J.; Eder, M. (2016): Stylometry with R: a package for computational text analysis. *The R Journal*, vol. 8, no. 1, pp. 107-121.

Ouamour, S.; Sayoud, H. (2013): Authorship attribution of short historical Arabic texts based on lexical features. *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 144-147.

Ouamour, S.; Sayoud, H. (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*, pp. 44-47.

Ramnial, H.; Panchoo, S.; Pudaruth, S. (2016): Authorship attribution using stylometry and machine learning techniques. *Intelligent Systems Technologies and Applications, Advances in Intelligent Systems and Computing*, vol. 384, pp. 113-125.

Rocha, A.; Scheirer, W.; Forstall, C. W.; Cavalcante, T., Theophilo, A. et al. (2017): Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5-33.

Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. (2003): Class prediction by nearest shrunken centroids with applications to DNA microarrays. *Statistical Science*, vol. 18, no. 1, pp. 104-117.

Wu, Z.; Markert, K.; Sharoff, S. (2010): Fine-grained genre classification using structural learning algorithms. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 749-759.

Xiong, Z.; Shen, Q.; Wang, Y.; Zhu, C. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 2, pp. 213-227.

Yates, J.; Orlikowski, W. (1992): Genres of organizational communication: a structurational approach to studying communication and media. *Academy of Management Review*, vol. 17, no. 2, pp. 299-326.