Recognition of Modern Arabic Poems

Abdulrahman Almuhareb, Waleed A. Almutairi*, Haya Altuwaijri, Abdulelah Almubarak, Marwa Khan.

King Abdulaziz City for Science and Technology (KACST), The National Center for Computation Technology and Applied Mathematics (CTAM), Riyadh, Kingdom of Saudia Arabia, Saudi.

* Corresponding author. Email: walmutairi@kacst.edu.sa.
Manuscript submitted February 5, 2015; accepted April 8, 2015.
doi: 10.17706/jsw.10.4.454-464

Abstract: We propose a machine learning method for recognizing modern Arabic poems based on the common poetic features of modern Arabic poetry. The poetic features include: rhyming, repetition, use of diacritics and punctuations, and text alignment. The method can classify text documents as poem or non-poem documents with a very high accuracy of 99.81%.

Key words: Meter, poem, poetry, rhyme, verse.

1. Introduction

This paper presents a method for modern Arabic poetry recognition. The method can classify text documents into poem and non-poem (prose) documents using supervised machine learning techniques. The classifier is applied on distinctive features of modern Arabic poems such as rhyming style, word repetition, use of diacritics and punctuation marks, and verse alignment (lineation). Arabic poetry can be classified into two categories: classical and modern poetry. The second category shares more poetic features with modern poetry of other languages.

Even though the proposed method targets the Arabic language, the basic model is language independent and can be applied to any language. The main goal of this work is to identify representative features of modern Arabic poetry that are computable and leads to correct classification. The direct application of this method is in the search engine domain. A search engine can apply this technique to construct a vertical search engine specialized in poetry . Vertical search engines, also called topical or specialty search engines, offer better support for domain specific search and tasks compared to typical web search engines [1].

The paper is organized as follows. Section 2 reviews related work. Section 3 presents the general features of modern Arabic poetry. In Section 4, we introduce the proposed method. Section 5 gives details of the experimental results and the training and testing dataset. Finally, the conclusions and future plans are described in Sections 6.

2. Related Work

Tizhoosh and Dara and Tizhoosh et al. proposed a method for poem recognition for poems written in English. Their proposed method classifies documents as poem and non-poem documents using classification techniques such as Naive Bayes, Decision Trees, and Neural Networks. The classifiers were applied on poetic features such as rhyme, shape, rhythm, meter, and meaning. Tizhoosh and his colleagues reported classification accuracy above 90.00%. Their results showed that all of the proposed features are useful for the poem recognition task and that features related to shape are the best [2], [3].

Almuhareb et al. proposed a rule based method for recognition of classical Arabic poems. The proposed method can recognize poem and non-poem segments when intermixed in the same text document. The method utilizes the common features of classical Arabic poems related to structure, writing style, and rhyme; and employs them in the recognition process. The authors reported a precision of 96.94% and a 92.24% recall. A prototype search engine for Arabic poetry was constructed using the proposed method [4].

Jansson analyzed the features of modern Arabic poetry based on three representative poems and compared them to classical Arabic poetry. The analysis showed that modern Arabic poetry lacks the distinctive features of classical Arabic poems which are verse symmetry, single meter and rhyme, and self-contained lines. Modern Arabic poems are recognized by using uneven lines, irregular rhyming, and partial meter. Other features of modern Arabic poetry that were analyzed by Jansson include punctuation marks usage, and word repetition [5].

3. Features of Modern Arabic Poems

Modern Arabic poetry differs significantly from classical Arabic poetry. In particular, verse symmetry and unified rhyme and meter are not maintained in modern Arabic poetry. In this section we review all of the common features of modern Arabic poems that can be helpful for the recognition process.

3.1. Presence

Instances of modern Arabic poems, as well as other types of poems, can be found in all sorts of printed and electronic documents including books, newspapers, magazines, and websites. An instance of modern Arabic poems can represent a complete poem or a poem portion. A single document can contain several modern Arabic poem instances. Poems can occur in designated documents by themselves or intermixed with normal text. In addition, poems can be found in non-textual media including audios, videos and images. On the web, Arabic poem instances can be found on designated websites. Only-poem websites normally organize poems in categories and adapt a unified style format that is preserved for the entire website.

3.2. Structure and Style

Typically, modern Arabic poems are written in consecutive short and uneven lines. Each line represents a verse. The lines can be arranged into sets of lines called stanzas. In some poems, all the stanzas contain the same number of lines. There is no limit on the number of lines in a poem. In many cases, the lines are arranged as separate paragraphs where a blank line is left between each two lines in the poem. It is also common that the entire poem is written in a single paragraph, or in several paragraphs (stanzas) separated by numbers, blank lines, or punctuation marks. The structure and style in modern Arabic poetry is loose compared to classical poetry. Fig. 1 and Fig. 2 show two short examples of a typical modern and classical Arabic poetry.

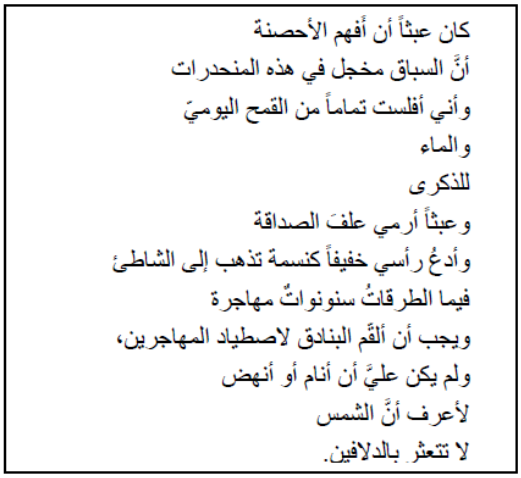
كان عبثاً أن أفهم الأحصنة
أنَّ السباق مخجل في هذه المنحدرات
وأني أفلست تماماً من القمح اليوميّ
والماء
للذكرى
وعبثاً أرمي علفَ الصداقة
وأدعُ رأسي خفيفاً كنسمة تذهب إلى الشاطئ
فيما الطرقاتُ سنونواتٌ مهاجرة
ويجب أن ألقّم البنادق لاصطياد المهاجرين،
ولم يكن عليَّ أن أنام أو أنهض
لأعرف أنَّ الشمس
لا تتعثر بالدلافين.

Fig. 1. Modern Arabic poetry example.

إذا ساءَ فِعْلُ المرْءِ ساءَتْ ظَنُونُهُ	وَصَدَقَ مَا يَعتَادُهُ من تَوَهُّمِ
وَعَادَى مُحِبّيهِ بقَوْلِ عُداتِهِ	وَأَصْبَحَ في لَيلٍ منَ الشّكّ مُظلِمِ
أُصَادِقُ نَفسَ المرْءِ من قبلِ جسمِهِ	وَأَعْرِفُهَا في فِعْلِهِ وَالتّكَلّمِ
وَأَحْلَمُ عَنْ خِلّي وَأَعْلَمُ أنّهُ	متى أجزِهِ حِلْماً على الجَهْلِ يَندَمِ
وَإنْ بَذَلَ الإنْسانُ لي جودَ عابِسٍ	جَزَيْتُ بجُودِ التّارِكِ المُتَبَسِّمِ
وَأَهْوَى مِنَ الفِتيانِ كلّ سَمَيذَعٍ	نَجيبٍ كصَدْرِ السّمْهَرِيّ المُقَوَّمِ

Fig. 2. Classical Arabic poetry example.

3.3. Rhymes, Meters and Diacritics

Poetic meters define the basic rhythm of the poem. Each meter is described by a set of ordered feet which can be represented as ordered sets of consonants and vowels. The use of meter and rhyme is not strict in modern Arabic poetry. Nazik Al-Malaika, one of the pioneers of modern Arabic poetry, has identified eight poetic meters in modern Arabic poems. These eight meters are a subset of the sixteen popular meters of the classical Arabic poems which were modeled by Al-Khalil bin Ahmed in the 8th century and his student Al-Akhfash who later added the 16th meter. Rhymes, on the other hand, are the repetition of the same sound at the end of the verse [6]. Unlike classical poems, modern poems can use multiple rhymes in the same poem without a firm order. Fig. 3 shows an example of a short modern poem with seven verses with two different rhymes. Arabic vowels in rhymes can be long or short. Short vowels are represented as diacritics above or below the letter while long vowels are represented as regular letters. The long and short versions of each basic vowel are considered equivalent for rhyme purposes. Usually, in modern Arabic writing, diacritics are ignored, however in poetry typing, poets tend to use diacritics more, especially with classical poems. Also, diacritics are used to resolve ambiguity.

خلصت مرساتي من الصخرِ،
ومشيت من بحر إلى بحرِ :
الريح رباني
وسفينتي انقاض انسانِ
فوجدتني، في آخر الدهرِ،
ارسو على الصخرِ،
فحضنته، وبكيت من ذعري !

Fig. 3. Example of short modern poem with two rhymes.

هل بحرٌ بعزّة موجه يخجَلْ ؟
أنا يدهُ
أنا المغزَلْ .
أأخجلُ منهُ ..
هل بحرٌ بعزّة موجه يخجَلْ ؟
أنا للخصب مصدرهُ
أنا يدهُ
أنا المغزَلْ .

Fig. 4. Example of phrase/line repetition in a poem.

3.4. Word Repetition and Punctuation Marks

Modern Arabic poems are also characterized by other features such as word repetition and punctuation mark

usage. Word repetitions do not comprise individual words only but also repetition of letters, phrases, a whole line/verse, similar sounds and sections. These repetitions help in building up the tempo, organize the relation of words, and emphasize the main theme of the poem. The other feature is the usage of punctuation marks such as full-stops, dashes, question marks, commas, and ellipses (three dots) inside the poem. They are added to divide clauses and phrases into rhythmic units and to increase the tension to a stanza. They are used at the end of lines as well as in the middle of them and several punctuation marks can be used in the poem. Fig. 4 and 5 show example of repetition and punctuation marks usage.

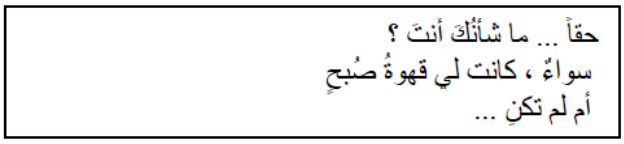

Fig. 5. Example of punctuation marks usage in a poem.

4. Method

The proposed method for modern Arabic poem recognition is to build a classifier based on the common features of modern Arabic poems as described in the previous section. The classifier can classify any text document as a poem or non-poem document based on the following features:

1) Average line length: line length is computed by counting all characters including the spaces in each line. The average is then calculated by dividing the sum of the line lengths over the number of lines in the document.
2) Standard deviation of line length.
3) Block average number of lines: we define a block (of text) as multiple text lines separated by one or moreblank lines, or single text line separated by multiple blank lines. Blocks of text can also be separated by lines that contains punctuations or numbers. For each block, we calculate the number of lines, then divides the sum over the number of blocks in the document.
4) Standard deviation of block number of lines.
5) Word repetition rate: ratio of the total number of repeated words to the total number of words in the document.
6) Line repetition rate: ratio of the total number of repeated lines to the total number of lines in the document.
7) Diacritic rate: the number of lines that have diacritics divided by the total number of lines.
8) Rhyme rate: the number of distinct rhymes divided by the total number of rhymes in the document. The rhyme of each line is the last letter in the line. If the last letter in the line is a vowel, then the rhyme is the last two letters of the line.
9) Punctuation rate: the total number of punctuation marks in the document divided by the total number of lines.

The first four features in the proposed model entail visual information about the document that can even be captured by the eye. The remaining five features provide surface linguistic information about the document. Even though the method is for modern Arabic poems, only two features are tied to the Arabic language which are the diacritic rate and the rhyme rate features. The method can also work for other languages once these two language dependent features have been customized to meet the requirements of the target language. Furthermore, these two features can be eliminated altogether to have a language independent basic method. Most of the poetic features in this method were also used by Tizhoosh et al. The main addition here is considering block of text, and customizing the rhyme feature for Arabic, and adding the diacritic feature.

5. Experiment

In this experiment, we used the features of the proposed method to represent documents and built several

Naive Bayes [7] and Decision Tree [8] classifiers using different feature combinations. The classifiers were trained using the dataset described in the next section and evaluated using 10-fold cross-validation. We run the experiment using Weka 3.6 [9]. The classification accuracy is measured using Equation 1 in addtion to the F-Measure as in Equations (2) to (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$Precision = \frac{TP}{TP + FP}, \quad (2)$$

$$Recall = \frac{TP}{TP + FN}, \quad (3)$$

$$F - Measure = 2.\frac{Precision.Recall}{Precision+Recall} \quad (4)$$

where :

TN : number of true negative cases.

FP : number of false positive cases.

FN : number of false negative cases.

TP : number of true positive cases.

5.1. Dataset

The dataset comprises 2067 plain text documents and is divided into two subsets. The first subset contains 513 modern Arabic poem documents collected randomly from the web. The collection includes poems written by about 150 different poets. The poems were collected using Google search results using poet names as search terms. For each poet, we select the top web pages that contain poems from the search results. For each selected webpage, we capture the poem portion of the page and store it in a text file. The second subset contains non-poem documents of articles and list items. We decided to enrich the non-poem set to include documents that contain list items because lists can visually resemble poems more than typical articles. This non-poem subset contains 1554 non-poem Arabic documents, including 1000 articles and 554 lists both of varying length, short and long. The articles were collected from the KACST Arabic corpus from different sources. The sources include newspapers, magazines, news agencies articles, and book segments [10]. The list documents were collected and generated randomly from a set of about 100 Arabic websites. List items were identified using the HTML tags < ul> and <li> in web pages. Fig. 6 shows an example of list items that looks like a short poem.

المنظومة الجامعية
بيانات أعضاء هيئة التدريس
البريد الإلكتروني للمنسوبين
البريد الإلكتروني للطلاب والطالبات
أدلة الاستخدام
نظام إجراءات سير العمل
نظام تسجيل المحاضرات
كيف تستفيد من الموقع
الموقع الجغرافي للجامعة
دليل هاتف الجامعة
نماذج الخدمات
المناقصات
نظام إدارة التعلم

Fig. 6. Example of a list from the dataset

5.2. Result

We built several classifier using the Decision Tree and Naive Bayes algorithms. The first classifier was a

decision tree using all of the features. This classifier achieved the best overall accuracy, 99.81%. The remaining classifiers were all Naive Bayes. Table I shows the results of these classifiers using different sets of features. The best accuracy was achieved using only the visual features at 99.71%. The all feature classifier scored a slightly lower performance of 99.61%. On the other hand, the linguistic features achieved only 87.13%. The baseline accuracy is 75.18% based on the majority class in the dataset. The results for the single feature classifiers varied widely. The best single feature classifier is the classifier that was built using the "Block Average Number of Lines" feature, 97.58%. This is a very excellent performance; about 2% less than the best achieved result. The best linguistic feature is "Diacritic Rate", 82.44%. Almost all of the remaining single feature classifiers performed below or equal to the baseline. Even though the "Block Average Number of Lines" feature achieved a very high accuracy, the effect of removing it from the all features set is negligible, at an accuracy of 98.94%.

Table 1. Naive Bayes Classifiers Performance Using Different Sets of Features

Features	Accuracy (%)	Precision(Weighted Avg.)	Recall(WeightedAvg.)	F-Measure (Weighted Avg.)
(-) Block average number of lines	98.94	0.989	0.989	0.989
All Features	99.61	0.996	0.996	0.996
Average Line length	66.67	0.848	0.667	0.687
Diacritic rate	82.44	0.565	0.824	0.645
Line repetition rate	7518	0.565	0.752	0.645
Linguistic Features	87.13	0.908	0.871	0.878
Block average number of lines	97.58	0.978	0.976	0.976
Punctuation rate	49.49	0.803	0.495	0.5
Rhyme rate	77.70	0.874	0.777	0.76
SD of Line length	68.41	0.857	0.684	0.704
SD of block number of lines	53.12	0.776	0.531	0.548
Visual features	99.71	0.997	0.997	0.997
Word repetition rate	75.18	0.565	0.752	0.645

The F-Measure scores were similar. The "Block Average Number of Lines" feature achieved the best F score for a single feature (97.6%). The excellent performance of the "Block Average Number of Lines" visual feature is attributed to that poems are usually arranged in multiple blocks of text (stanzas) with a bunch of short lines (verses). While plain unformatted text articles are normally organized as one or more blocks of less but longer lines of text representing paragraphs. In the dataset, the average value of this feature is 10.7 lines for poems, and 10.2 lines for article documents. For item list documents, this value is 13.4. Fig. 7 shows the class distribution using the "Block Average Number of Lines" and "Average Line Length" features.

Four other features achieved good precision rates, but with reduced recall: Rhyme Rate, Punctuation Rate, Average and SD of Line Length. This shows that rhymes and punctuations are good indicators for poems, however, in some poems, they may not be emphasized enough. Fig. 8 and 9 shows the poem and non-poem distribution in the rhyme and punctuation feature space. The same applies to the line length features. Poem lines tend to be short and with low standard deviation in most cases but not always. The mean of the average and SD of line length of poems in the dataset is 23.6 and 8.5 characters, respectively, compared to 390.4 and 375.7 characters for non-poems. Fig.10 shows the distribution in the line length space.

The precisions of the remaining four features were low, namely, Diacritic Rate, SD of Block Number of Lines, Line and Word Repetition Rates. These features, found to be not very indicatives for poems. On the other hand, the Diacritic Rate feature achieved a good recall score (82.4%) but the low precision score (56.5%) is due to having a similar usage rate of diacritics in poem and non-poem samples in the dataset. Fig. 11 illustrates the class

distribution in the diacritic feature space.

The results of this experiment confirm the findings of Tizhoosh et al. That common poetic features are useful in poem recognition and that visual (shape) features in particular are very adequate for this purpose.

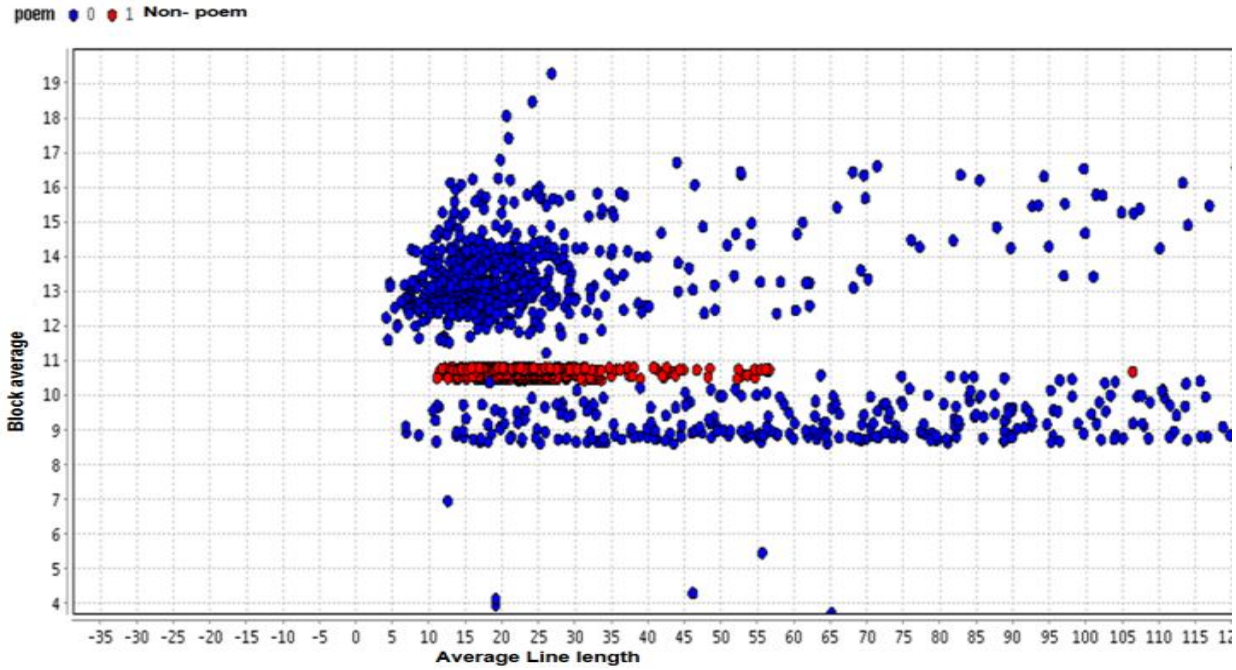

Fig. 7. Block Average Number of Lines and Average Line Length features.

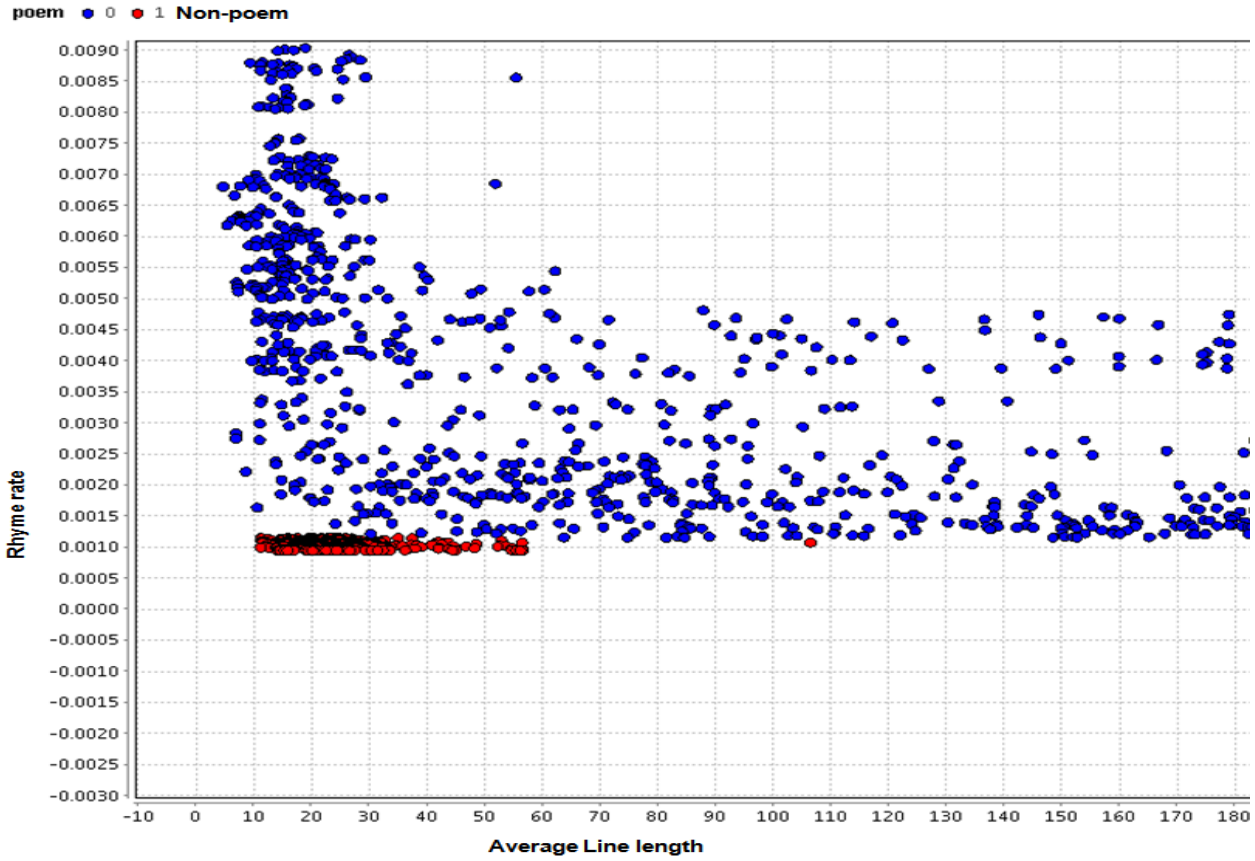

Fig. 8. Rhyme rate and Average Line Length features.

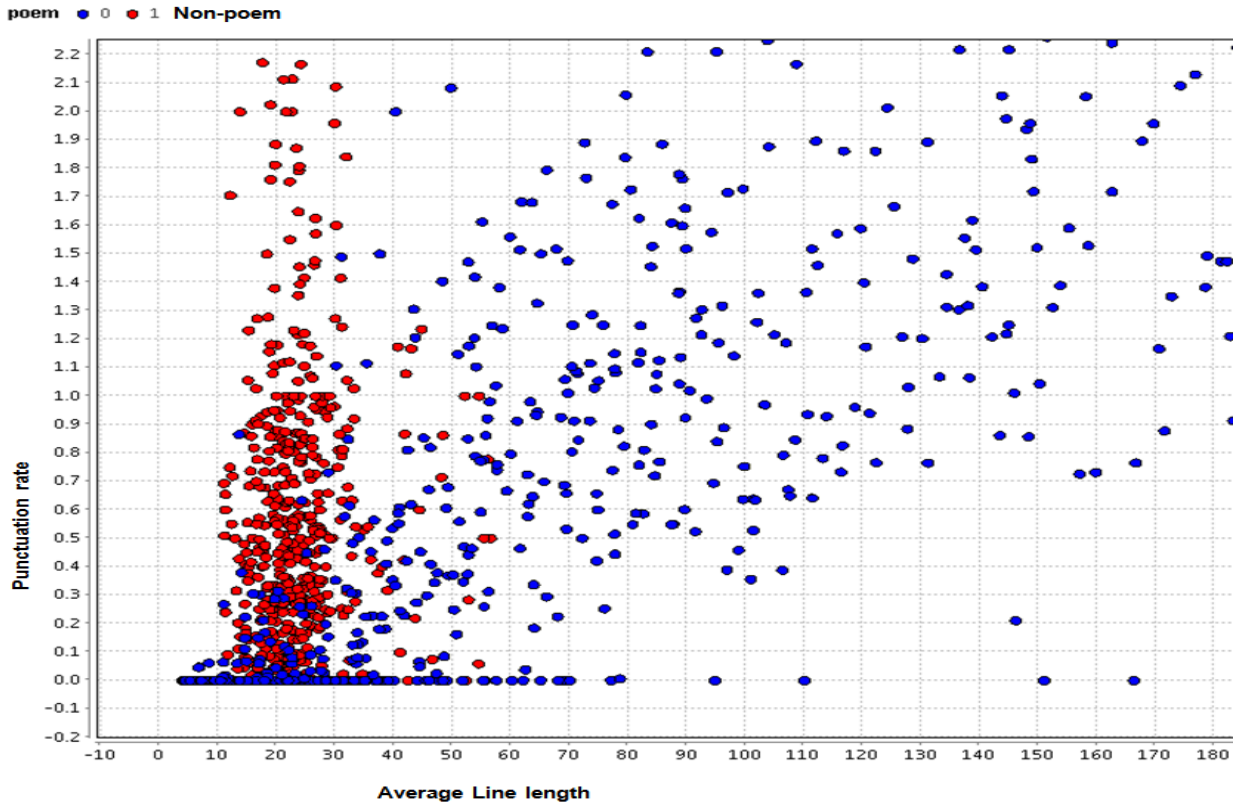

Fig. 9. Punctuation rate and Average Line Length features.

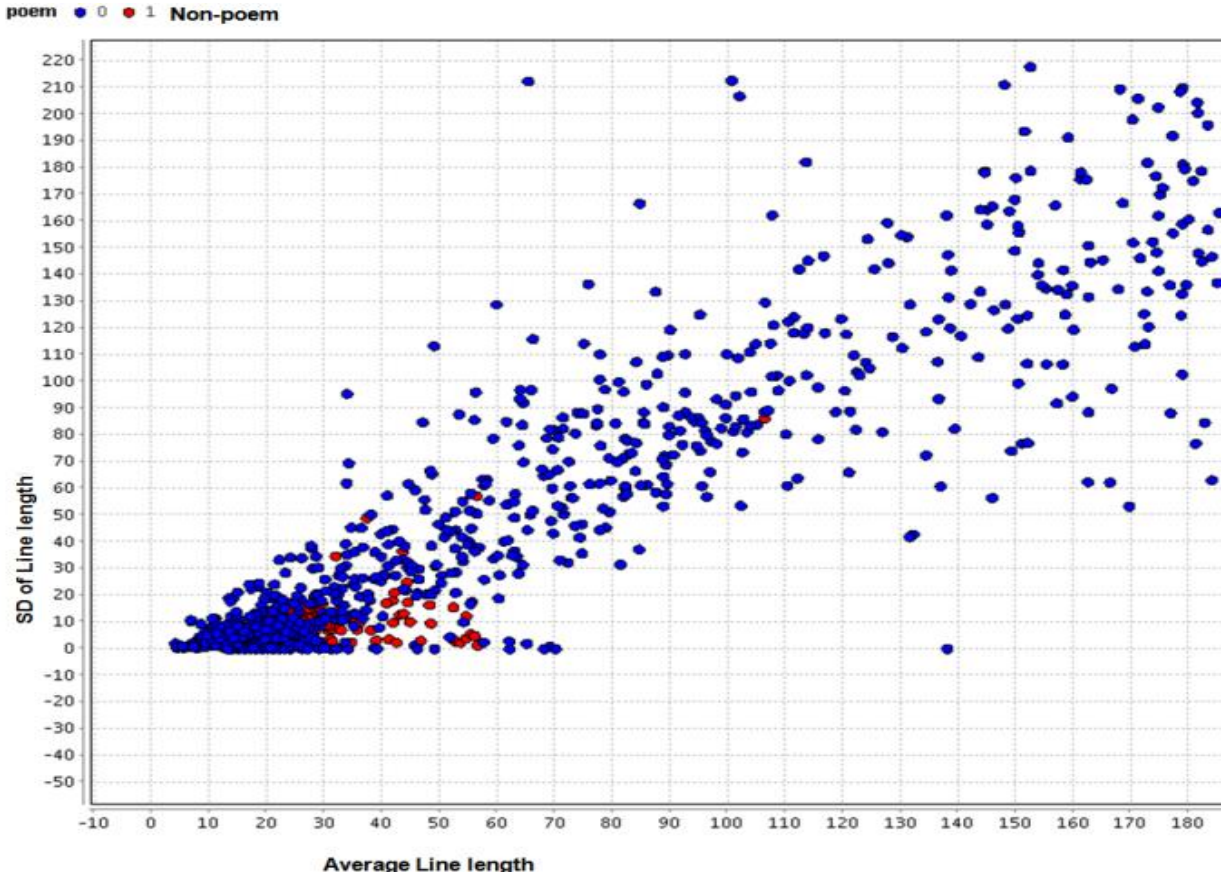

Fig. 10. SD of line length and average line length features.

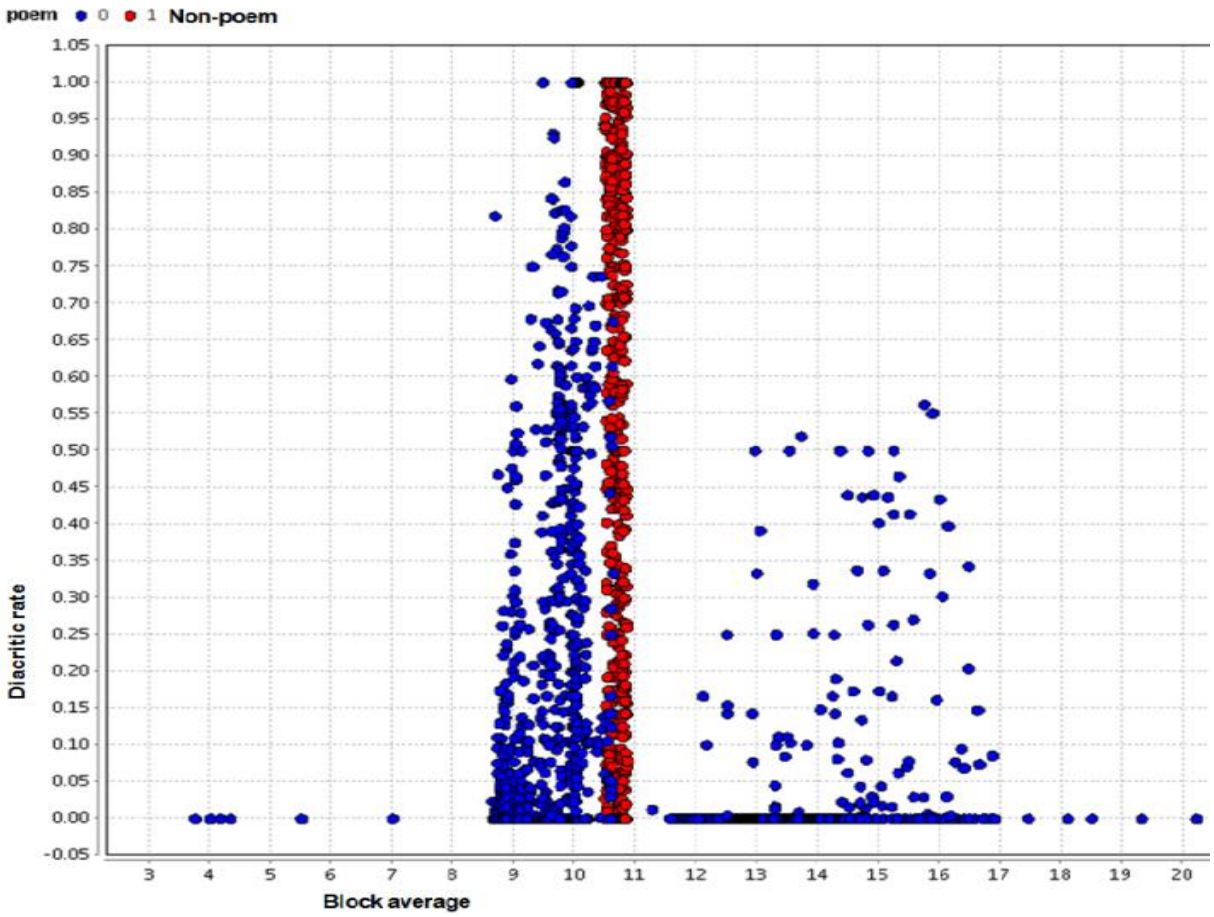

Fig. 11. Diacritic rate and Block average number of lines features.

6. Conclusions and Future Work

In this paper, we proposed a method for modern Arabic poem recognition. The method can classify any text document as a poem or non-poem document with a very high accuracy of 99.81%. The classifier exploit the common visual and linguistic features of modern Arabic poems such as line length, rhyme, punctuations, and diacritics. In the future, we would like to apply this method on other languages. We would like also to extend this work to cover the cases where poem and non-poem text are intermixed in a single document.

Acknowledgment

This work was funded by King Abdulaziz City for Science and Technology (KACST) under Grant 33-823.

References

[1] Clifford, L. (1997). Searching the internet. *Scientific American*, *276(3)*, 52–56.

[2] Hamid, R. T., & Rozita, A. D. (2006). Pattern analysis and applications. *On Poem Recognition*, *9(4)*, 325–338.

[3] Hamid, R. T., Farhang. S., & Rozita, D. (2008). Poetic features for poem recognition: A comparative study. *Journal of Pattern Recognition Research, 3(1)*, 24–39.

[4] Abdulrahman, A., Ibrahim, A., Lama, A. S., & Haya, A., (2013). Recognition of classical arabic poems. *Proceedings of the Workshop on Computational Linguistics for Literature* (pp. 9–16), Atlanta, Georgia. Association for Computational Linguistics.

[5] Anna, J. (2010). A Prosodic Analysis of Three Modern Arabic Poems. *Modern Arabic Poetry.* SPL magisteruppsats i arabiska SPL 2010-012.

[6] Shmuel, M. (1976). The Development of Its Forms and Themes Under the the Influence of Western Literature. *Modern Arabic Poetry*.

[7] McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. *Proceedings of the workshop on learning for text categorization*, (pp. 41– 48), Citeseer.
[8] Ross, Q. J. (1986). Induction of decision trees. *Machine learning, 1(1)*, 81–106.
[9] Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., & Ian, H. W. (2009). The weka data mining software: An update. *ACM SIGKDD explorations newsletter*, *11(1)*,10–18.
[10] Thubaity, A. A., Khan, M., Mazrua, M. A., & Mousa, M. A. (2013). New language resources for arabic: Corpus containing more than two million words and a corps processing tool. *Proceedings of the 2013 International Conference on in Asian Language Processing* (pp. 67–70).

Abdulrahman Almuhareb is an assistant professor of computer science at the National Center for Computation Technology and Applied Mathematics at KACST, Saudi Arabia. He received his B.Sc. degree in information systemsfrom King Saud University in 1993 and He received his M.Sc. degree in computer science from Ball State University in 1998, and He received his Ph.D. degree in computer science from the University of Essex in 2006. Almuhareb's research focuses on Arabic language processing, search engines, and machine translation. Other research interests include information extraction, data mining, and artificial intelligence.

Waleed A. Almutairi was born in Riyadh, Saudi Arabia in 1980. He received his B.Sc. in computer science from IMAM University Riyadh, Saudi Arabia. And he completed his M.Sc. degree in computer science from NYU-POLY, New York, United States of America in 2009.

In summer of 2004 he worked in STC as TRAINEE for about 3 months. After graduation, he worked for about one year in GDMS as a SYSTEMS ENGINEER. Since 2009 he worked for King Abdulaziz City for Science and Technology (KACST) in the National Center of Computation Technology and Applied Mathematics as a RESEARCHER Riyadh, Saudi Arabia. During that time he worked on two projects on Arabic search engine and Arabic poem search engine. Machine learning, especially Text mining interested him.

Mr. Almutairi has co-authored the paper “Search tools using named entity recognition" in 2013. And published a paper “Opinion ranking based on lists in search engines” in 2014.

Haya altuwaijri is an information technology specialist. She received her bachelor degree in information technology field from Computer & information Science College, King Saud University, Riyadh, Saudi Arabia in 2010.

She is a researcher in the National Center for Computation Technology and Applied Mathematics in King abdulaziz city for Science and technology, she Attend "Summer Training Program in The Saudi British Bank in Information Technology Department", held on July 2008 in Riyadh, Saudi Arabia.

Ms. Altuwaijri participate in enriching the Arabic content of Information Security in Excellence Centre and (Google Knol) with two articles titled “Classification of electronic crime according to the use of computer", and “Terms in the world of network”, 2009. Also she participates in publishing a research paper titles “Recognition of Classical Arabic Poems”, 2013.

Abdulelal M. Almubarak was born in Taif, Saudi Arabi, in 1986. He completed his BSc in computer science from King Saud University, Riyadh, Saudi Arabi in 2010.

Since 2010, Abdulelah has worked as a researcher at King Abdul-Aziz City for Science and Technology in the National Center of Computation Technology and Applied Mathematics, Riyadh. He has been involved in three successful research projects involving an Arabic search engine, an Arabic poem search engine and a high-performance computing (HPC) project. He was also a member of the Saudi Computer Society for more than two years.

Mr. Almubarak's research interests include search engines, high-performance computing (HPC), data mining and artificial intelligence. And He co-authored the paper "An energy-efficient multi-GPU supercomputer," which was published in 2013.

Marwa M. Khan was born in Jeddah, Saudi Arabia, in 1987. She received her B.A in information technology (IT) from King Saud University (KSU), Riyadh, Saudi Arabia, in 2010.

She started working as a researcher at King Abdulaziz for Science and Technology (KACST) in Riyadh in March 2011 until now. She worked as a senior programmer in a cooperated project between the KSU and King Abdulaziz University Hospital for profoundly deaf children from June 2010 to March 2011.