Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 4, August 2020, pp. 1701~1710
ISSN: 2302-9285, DOI: 10.11591/eei.v9i4.1898

English poems categorization using text mining and rough set theory

Saif Ali Alsaidi, Ahmed T. Sadiq[2], Hasanen S. Abdullah[3]
[1,2,3]Department of Computer science, University of Technology, Iraq
[1]College of Edcation Pure Science ,Wasit University, Iraq

Article Info

Article history:

Received Nov 13, 2019
Revised Jan 20, 2020
Accepted Feb 3, 2020

Keywords:

Poem categorization
Poems
Rough set theory
Text mining

ABSTRACT

In recent years, text mining wasan important topic because of the growth of digital text data from many sources such as government document, email, social media, website, etc. The English poemsare one of the text data to categorization English Poems will use text categorization, text categorization is a method in which classify documents into one or more categories that were predefined the category based on the text content in a document. In this paper we will solve the problem of how to categorize the English poem into one of the English Poems categorizations by using text mining technique and machine learning algorithm, Our data set consist of seven categorizations for poems the data set is divided into two-part training (learning) and testing data. In the proposed model we apply the text preprocessing for the documents file to reduce the number of feature and reduce dimensionality the preprocessing process converts the text poem to features and remove the irrelevant feature by using text mining process (tokenize, remove stop word and stemming), to reduce the feature vector of the remaining feature we usetwo methods for feature selection and use rough set theory as machine learning algorithm to perform the categorization, and we get 88% success classification of the proposed model.

Corresponding Author:

Saif Ali Alsaidi,
Department of Computer Science,
The University of Technology,
Alsenaaa St, Baghdad, Iraq.
Email: salsaidi@uowasit.edu.iq

1. INTRODUCTION

Text mining aims to make users able to extract useful information from multi-text source, and text mining operation such as categorization or classification information retrieval and text summarization [1] Text mining Technique becomes very important especially with the growth of text data in many aspect Web site, social media, email, government document, health report, security etc. text categorization (TC) is the same meaning to text classification or topic spotting [2]. Text categorization is defined as how computer automatically can categorize unlabeled documents into one of the predefined categories [3]. According to the amount of content similarity. It has been deployed successfully in many applications such as topic detection [4], e-mail spam detection and filtering [5], author identification to the text documentation [6], web page classification [7], document classification [8], popular issues are facing the researcher in text categorization the first issue is challenging to build the classifier model with high dimensionality of text data because the feature vector will be large and that will decrease the performance, The second issue is feature selection for classifier model there are redundant or irrelevant feature that maybe affect the classifier result,The third issue

is when doing preprocessing data cleaning to remove noise and unnecessary feature from text data may be delete the keyword which is consider good feature to classification [9]. One of the digital text data available is English poems with many categorizations we use only seven categorizations of poems which is (sad poems, life poems, mother poems, friend poems, death poems, funny poems) there are two challenges thefirst one the used poems category is semi interrelated such as the love poems and mother poems the love word that using to express of love in love poems may be using to express mother love also.

The second challenge is how to categorization the poems into one of this category. By using text mining technique and machine learning algorithm rough set theory.Will build our modelin our proposed system uses one of machine learning algorithms is rough set theory was developed by Zdzislaw Pawlak, in the early 1980s. The rough set theory is one of the machine learning algorithms that been used for classification in many application with supervised data. The main aim of the rough set analysis is to synthesize an approximation of concepts from the collect data [10]. When used rough set for the classification of English poemswe get a good result and success in the classification of the English poem categorization. Rough set classifier performance is evaluated by computing its precision [11], recall [12], rough set theory is semi-similar with other approaches like the fuzzy set, genetic algorithm and statistic methods [13], The data set our corpus consists between 70-80 poems for each categorization as a training data and 50 poems with unbalance number of poemsfor each categorization as test data. The test data is 50 poems some poetry belong to the love, death, sad, mother, friend category and the other poem is don't belong to any class to measure the model performance all the English poetry collected from accessible web page [14]. Figure 1 shows classifier fellow work.

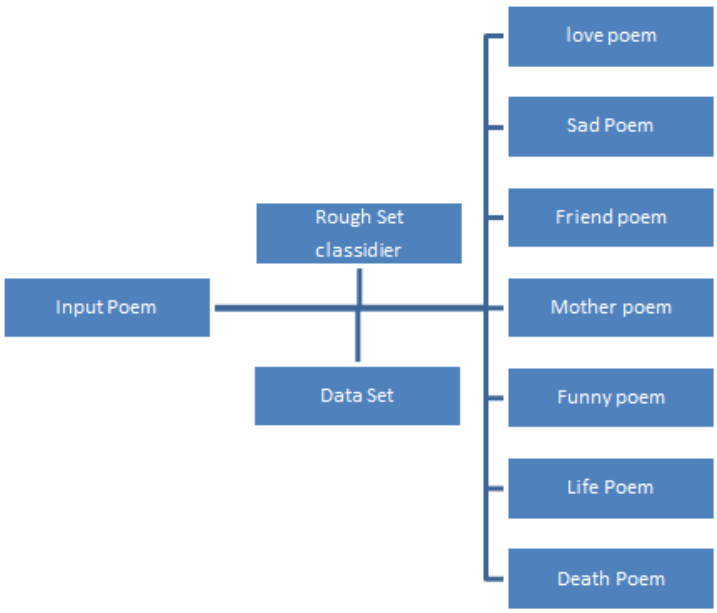

Figure 1. Classifier follow work

2. RELATED WORK

There are so many studded did on the text categorization in different type text data such as news, social media posts, political articles and in different text language English Arabic china we will mention the modern and related paper in text categorization. In [14], authors have proposed a system which categorizes the text automatically by use specific feature class which to build Bayesian classification. The proposed method selected apowerful feature for every class. The Naive Bayes rule designed by using the theorem of Baggenstoss PDF projection to classify the specific class feature. In [15], the authors proposed a new method for term weights calculating by partition the document into the segment. Tests with Dyna Part-FiLa (dynamic partitioning of text documents with first and last partitions) the document should be segmented before calculating term weight will be improving f-measure, The F-measure improve the classifier when work with DynaPart-FiLa by improving the F-measure at the start of document relevant term will appear. Increasing they're significant enables to improve the results of classification. Finally, they anticipated that the positional meaning of terms is a good sign of the document's context. In [16] researchers has

proposed a system which is categorized text automatically of Marathi files base on history browsing of a user profile. Vector space model provides a better result than probabilistic models.

The precision of the outcomes related to the system is way good compared with the Tamil language. The best clustering method is LINGO algorithm which provides high quality than another clustering method. In [17] researcher proposed system for text categorization by using hybrid intelligence technique for text categorization by using the text preprocessing for feature reeducation and term frequency for feature selection and used the rough set theory as a classifier to classifying document into three main categories of the labelled document which is computer science, mathematics and physics. In [18] is the most relevant paper researchers proposed a system which can categories the English poetry to their poet by using chi-square (CHI) technique are used for feature selection and apply five algorithm for classification. These algorithms are sequential minimal optimization (SMO), C4.5 decision tree, Naive Bayes (NB), random forest (RF) and k-nearest neighbours (KNN) the categorization result is different from each other, the best result for success classification is got by SMO technique. In [19] in this paper the case study is Arabic poetry and the researchers doing categorization to four classes of Arabic poetry the four types: love poems, Islamic poems, social poems, and political poems and use linear on nonlinear equation represent the similarity between the document classes and the features, in this paper the use three techniques for classification NB, support vector machines (SVM), and linear support vector classification (SVC) for the classification task.

3. THE PROPOSED CATEGORIZATION MODEL

In this section will discuss entire model steps start with dataset and the stages of proposed system and mechanism, there are seven main steps for this system the first step data collection which is digital English poems text collected from multi website source the text data is unstructured data so to deal with these type of data we need to convert the text data to feature also known as (keyword, term, token and attribute) to get feature text data preprocessing function used to convert text to split feature which each word in document is considered feature,then delete the useless feature to reduce the feature size by removing the stop words but still, there is big size of data by appalling stemming will reduce the data dimension and feature vector size, in the follow step is select the useful feature by applying two methods of feature filter to select the best feature that can help in categorization the input poem and the good feature by apply determine the number threshold to selected feature which is frequently larger than threshold (T >=4) these can increase the performance of system accuracy. In the first process, in the five-step, the model perform the learning and testing by using supervised classification techniques in this process build the classifier by feed the machine learning classifier algorithm to construct the sets of rough set from the feature selection in previous step for each categorization from the training data then test the system performance and accuracy to categorize the input poem to one of seven category and the last step is evolution Result system and performance measure of the model by count system efficiency in category and the error rate for the proposed system. The proposed system framework steps are showing in the Figure 2. and all the details and information of how to apply the main steps and subprocess in each main steps will discuss process follow work for the proposed category system are explained in the next section.

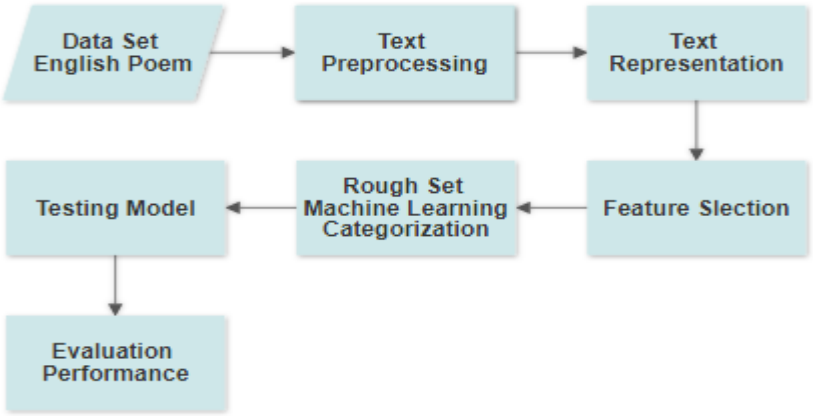

Figure 2. Framework categorization model

4. DATA SET

The dataset is of text data of English poem, which is collected from the internet web page https://www.poetrysoup.com. Text data of English poems the total number of poems we use is in this paper is 568 poems. The dataset is divided into two-part training and testing. The training data set consist of number of poems in English which is more than 518 poems, we use seven categorizations of English poems the information and number of poems that we use in each category shown in Table 1. Also the number of line in each poem is different some poem is long, sonnet, short poem, the test data content 20 poems. Each category poems put in one document where data set =(D1, D2, D3, D4, D5, D6, D7) where (D1=Love poem, D2=Sad Poems, D2=Life Poems, D3=Mother Poems, D4=Friend Poems, D5=Death poems, D6 Funny Poems) The data set has a large number of feature (words) in form as text data to reduce these number of feature we should doing some preprocessing method and reduction procedure. The advantage of preprocessing operation and reduction procedure is reducing the size of storage space for feature vector and the turning time and also increase the system performance by reducing the amount of feature, Figure 3 show the data set preprocessing and reduction.

Table 1. Data set information

Type of Poem	Traning poem No.	Test Poem No.
Love poem	78	8
Sad poem	75	7
Friend poem	75	7
Mother poem	75	6
Funny poem	73	4
Life poem	72	6
Death poem	70	4
Other poems	0	8

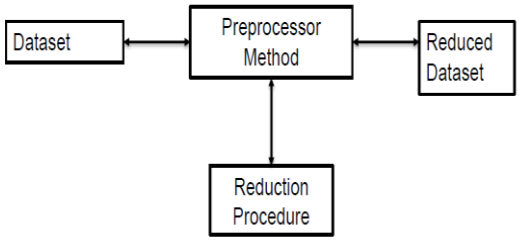

Figure 3. Dataset pre-processor

5. PREPROCESSING OPERATIONS

One of the necessary and essential steps in the text mining and text analysis is text preprocessing because the text data is unstructured and it's challenging to work with it in the original form [20] we apply three steps of text data preprocessing as showing in Figure 4.

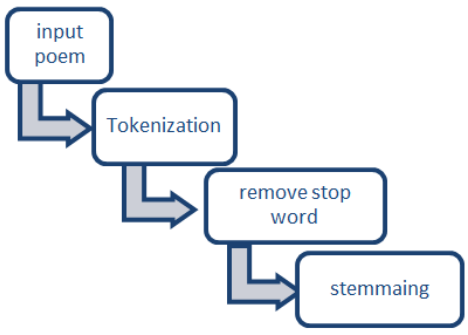

Figure 4. Preprocessing step

5.1. Tokenization

It is the first step in the text preprocessing; there are two types of tokenization which is word tokenization and sentence tokenization, in this paper we use word tokenization because the model will category according to the word wight frequent. Word tokenization is segmenting the paragraph or document of text content into split words called token based on the white space.

5.2. Remove stop word

The next step in text data preprocessing is Stop word, it's one of dimension reduction process in text mining to delete the words that are useless or not effect to system categorizationperformance or the most common words such as punctuation, the conjunction that does not affect the model accuracy performance.

5.3. Stemming

It is one of useful step in text preprocessing, the main goal of stemming is back the words to his root such as (connected, connection, connecting) is the same word which is the root (connect), so it crucial step to reduce the feature vector size by decrease the number of terms and reduce the storage space [21].

5.4. Text representation

By using machine learning methods for text categorization and text mining, we need to represent the text document in one forms that make learning algorithm easy to work and applies text categorization using machine learning methods, in this paper, we use word weight frequency count method to represent the words of document by the numerical value for the term present in the document [22]. And there is another method for text representation by using binary value to represent the feature if the term X present in document D then true, else falsebut in this paper, we use the term weight to represent the text data.

6. FEATURE SELECTION

In this step, we will select the most important feature with a high correlation to the class, the feature which can help to increase the categorization model performance. The one of the most famous problem in machine learning application is feature selection [23]. To select the relevant term or attribute to build the information system to apply the machine learning algorithm rough set theory we used feature selection to choose the best feature from the large vector space feature. There are two methods used for feature reduction to delete the not good feature and minimize the dimensionality of feature space. Feature selection is the first method and feature transformation/extraction is the second one. In this paper used the first method which is feature selection, we use the selected features to build the sets of the rough set theory algorithm for the classifier [24]. There are many method for feature slection in [25] suggested two new feature slection metrics the feature selected according to the term weighted one of feature selection method is:

6.1. Filtered-based feature selection

In this module, we count the weight of each term which is the output from tokenization, stemming and removing the stop word. To create a feature vector which is sets of the rough set theory for each category poem.

6.2. Term weighted concept

To determine import feature from features vector to feed the classifier which is machine learning algorithm roughest theory and build the sets, Term weighting is to give weight which is value for each term (feature) in each document file of category poems to determines the importance term format. To give weight for the different terms, we used term. Frequency for each word, in this paper the word (feature) is weighted according to the occurrence in the total of poems in file category and how many poems contain this word and the count of the word in each poem, using term frequency method to select the best feature is not suitable for this type of data because the poem is different in length and the number of poems is also changed so we will use two filter method to select the best feature. The first filter is to choose the best feature inside the document file which content number of poems belong to one of the seven categories we take in consideration the number of frequent terms and the number of frequent poems which contain the term.

$$Fw = \frac{T}{Nw} \quad (1)$$

Where **Fw**=weight of feature **W, T=**word frequent in whole poems in file, **Nw**=Poem frequency the number of poem contain the term **W** and use two threshold one for term frequency and one for number of poems contain the term. For example, the word or feature (love)was 25 frequent in love poems document

file which it contains 24 poems these features may be all the24 frequent in one poem of 24 total poems, these is drawback to use the term frequent So we select the Term with high frequency and number of poems which contain the word is also high. **The second filter** is to select the feature between the different files of poem category by using TF-IDF where the TF is the frequent term and inverse document frequent, **TF**=the number of terms frequent in each document file, and

$$DF = \log_2(N/F) \quad (2)$$

Where **N** represents the total number of the document in the dataset, **F r**epresen**ts** the number of the document containing the term. For example, the word (Heart) was selected as a good feature for two or more categories, so this is a problem to give weight to each term we use.

$$TF.IDF = \text{TF} * \text{IDF} \quad (3)$$

To select the discriminative feature for each category that increases the model performance.term frequency is obtained for each term to produce the training file. In the categorization process for text files the weights assigned to a different term.

7. SUPERVISED CLASSIFICATION TECHNIQUES ROUGH SET THEORY

Rough set theory (RST) it's one of machine learning algorithm developed by Pawlak [26] used for text classification the RST is used to work with the not certain information, the rough set basic concept is indistinguishable, reduction, and core that can be used for data classification and knowledge reduction. The reduction concept is the smallest subset of feature that be used for describe all classes of in decision Information system while keeping the indistinguishable relation obtained by using full set of feature [27], the first point of using rough set theory which is depend on analysis datato build sets of data which is formatted as a information table, where all row is considered as object (poem). Every column consider as attribute (terms) which is selected feature, this table is called information system, the information table form as pair of S=(U, A), where **U** is finite and nonempty set of objects called Universe and **A** is finite and nonempty set of the features such that a: U →Va for every a ∈ A. The set of value of (a) is Va, then for each subset of feature A like B where B⊆ A there is an equivalence relation IND A(B)

$$\text{INDA(B)}=\{(x, y) \in U2 \mid \forall\, a \in B\ a(x) = a(y)\} \quad (4)$$

Where INDA(B)=indiscernibility relation of feature column B. If (x, y) ∈INDA(B), then objects x and y are indiscernible and X, Y are similar to each other concerning attribute B, the equivalence classes of the B-indiscernibility relation are denoted [x]B. To assign every subset of the objects X ⊆ U into two sets (LB(x) lower approximation, UB(X) upper approximation) of X, respectively and define as follow:

$$\text{UB (X)}=\{x \mid [x]B \cap X \neq \varphi\} \quad (5)$$

Where the lower approximation is the set of the objects which is exactly dealing with a certain class, and the upper approximation is the set of the objects which can be lead to the certain class or not. The test model where each documents features was selected is represented as an array containing the features and class of its feature. We build information system each row of the information system is the object which represents. One poem of the predefine English poems category, and many columns to each attribute of the keyword. Where the tested poems after we apply the same preprocess steps for the training data and reduce the amount of feature then selected the feature with high weight to build the list of the critical term for the test poem. Based on the lower and upper approximation mesures between the features of the tested poem which consider the new objects and the features of the trainings objects, if all test poem features can be considered as lower approximation Set with the specific class so this poem exactly belong to this class, and if there is some feature is distributed with other class this mean rough. So use upper approximation equation for rough set to decide to which class the poem belong. The following pseudocode is showing the main process of the proposed model:

Step 1*: INPUT: D1,D2,….,D7 number of the document where Each Di contain Number of Poems. C1, C2,….C7 the number of categories*
Step 2:*For each category Ci Do, For each Document Dj for Ci Do*
Step 3 *:Tokenize the Dj into features Dj→Fj*

***Step 4**: Delete Stop word, number, the special character from Fj →F1j*
***Step 5**: Stemming the F1j →F2j*
***Step 6**: Count the frequency for all F2j*
***Step 7**: Set the T1 threshold for term frequency and T2 for Poem frequency*
***Step 8**: Select the F2j which frequent >= T1*
***Step 9**: Count the Poem frequent >= T2*
***Step 10**: Term Weight =Term frequent/ No. poem frequent*
***Step 11** apply TF-IDF between the Di*
***Step 12**: Build the Decision information system F3j*
***Step 13**: Compute Upper Approximation for each Ci using the*

$$UB(X) = \{X|[Xi] \cap X \neq \emptyset\}$$

***Step 14**: Compute the lower approximation for each Ci using*

$$LB(X) = \{X|[Xi] \subseteq X\}$$

***Step 15**: classify a poem*
***Step 16**: End*

8. RESULTS PERFORMANCE MEASURE

The model performance accuracy is measured by using the confusion matrix to measure the decision of categorizing model poem to the appropriate class, as shown in Table 2. True positive (TP) is the number of poems which is successful categorize to his class true category, true negative (TN) is the number of the poem which does not belong to anyone of seven category's and the model also give successful result true negative, false positive (FP) if the poem actually don't belong to any seven categories and the model label as one of these categories, false negative (FN) if the poem belong to one of the seven categories of the English poem and the model failed in label.

Table 2. Presents the confusion matrix of the model

N=50	Actually Positive (true)	Actually Negative (False)
Predicted Positive (true)	38	2
Predicted Negative (False)	4	6

Accuracy (Ac): Is the ratio between the number of success document category and document correctly not a category to the total number of the test document

$$Ac = \frac{TPi + TNi}{total\ test\ document}$$

$$Ac = \frac{38 + 6}{50} = 0.88$$

Error rate (E): Is the ratio between false document categorized and the total a number of documents.

$$Ei = \frac{FPi + FNi}{total\ test\ document}$$

$$Ei = \frac{4 + 2}{50} = 0.12$$

Precision (P): Is the percentage of success document category to all document belong to that category

$$P = \frac{TPi}{TPi + FPi}$$

$$P = \frac{38}{38 + 4} = 0.90\%$$

The model result is considered good according to the accuracy result which is 88% while the error rate is only 12% and the precision show good result 90% to categorization, Table 3 show the information about test poem type and number of correct categorization and number of error categorization and Figure 5 is bright show the result of categorization evaluation, Figure 6 show the performance evaluation.

Table 3. Test poem information

Type of Poem	Test Poem No.	Correct categorization	Erro categorization
Love Poem	8	8	0
Sad poem	7	7	0
Friend poem	7	6	1
Mother poem	6	5	1
Funny poem	4	4	0
Life poem	6	4	2
Death poem	4	4	0
Other poem	8	6	2

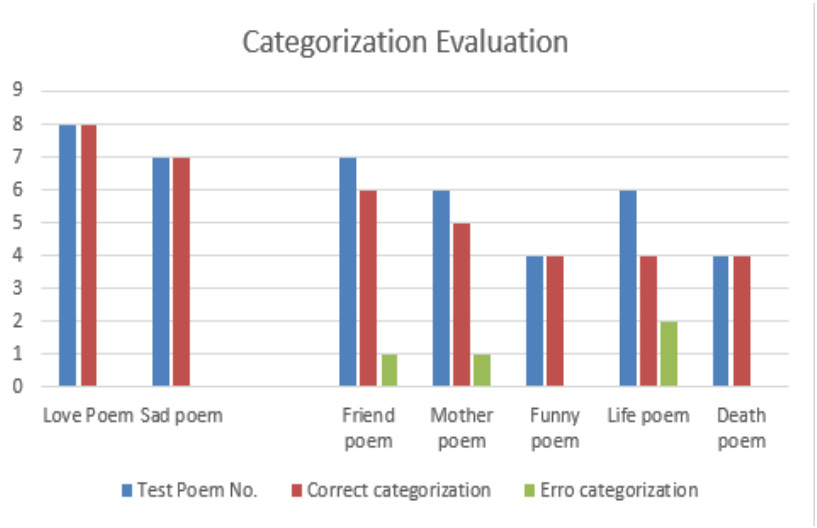

Figure 5. Categorazaton evaluation

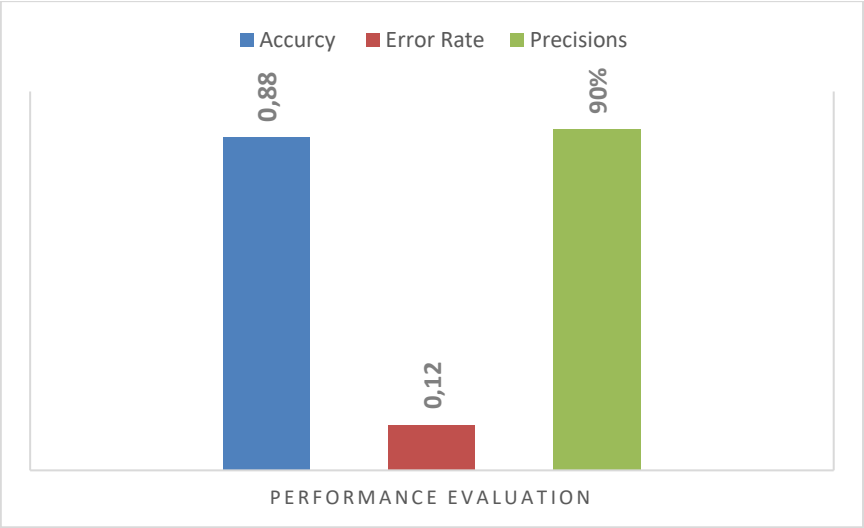

Figure 6. Performance evaluation

9. CONCLUSION

The categorization English poems is hard to do without perform some of the preprocessing steps to reduce the number of the features, by appling remove the stop word will reduce the number of feature by delete the unimportant feature but still have large number of feature to reduce the feature by apply stemming to back many words to the same root word then apply the filter for feature selection to select the most essential feature which help to reduce the storage space and process time also increase the performance of the proposed model, the model result shows that using the roughest theory algorithm as supervised machine learning can minimize the error rate to perform English poems categorization and give good accuracy result 85% to classify English poems.

FEATURE WORK

In the feature work, we try to improve the results of accuracy by using other feature selection method, and we strive to categorize the poems according to the poet.

REFERENCES

[1] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *International Journal of Artificial Intelligence & Applications,* vol. 3, no. 2, p. 85, 2012.

[2] M. E. R. Ruiz and P. Srinivasan, "Combining machine learning and hierarchical structures for text categorization," *Doctoral Dissertation, University of Iowa*, 2001.

[3] J. J. G. Adeva and J. M. P. Atxa, "Intrusion detection in web applications using text mining," *Engineering Applications of Artificial Intelligence,* vol. 20, no. 4, pp. 555-566, 2007.

[4] J. Zeng and S. Zhang, "Variable space hidden Markov model for topic detection and analysis," *Knowledge-Based Systems,* vol. 20, no. 7, pp. 607-613, 2007.

[5] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications,* vol. 36, no. 7, pp. 10206-10222, 2009.

[6] C. Zhang, X. Wu, Z. Niu, and W. Ding, "Authorship identification from unstructured texts," *Knowledge-Based Systems,* vol. 66, pp. 99-111, 2014.

[7] S. A. Özel, "A web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Systems with Applications,* vol. 38, no. 4, pp. 3407-3415, 2011.

[8] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Engineering Applications of Artificial Intelligence,* vol. 52, pp. 26-39, 2016.

[9] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR),* vol. 34, no. 1, pp. 1-47, 2002.

[10] S. K. Pal and A. Skowron, "Rough-fuzzy hybridization: A new trend in decision making" *Springer-Verlag,* 1999.

[11] F. Debole and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets," *Journal of the American Society for Information Science and technology,* vol. 56, no. 6, pp. 584-596, 2005.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing,* Association for Computational Linguistics, vol. 10, pp. 79-86, 2002.

[13] Suraj Z., "An Introduction to Rough set Theory and Its Applications: A Tutoria," *proceeding of the 1st International Computer Engineering Conference (ICENCO') New Technologies for the Information Society,* Cairo, Egypt, pp. 1-39, December 2004.

[14] "*Data Collection Source*," [Online]. Available: https://www.poetrysoup.com/

[15] B. Tang, H. He, P. M. Baggenstoss, and S. Kay, "A Bayesian classification approach using class-specific features for text categorization," *IEEE Transactions on Knowledge and Data Engineering,* vol. 28, no. 6, pp. 1602-1606, 2016.

[16] A. Kulkarni, V. Tokekar, and P. Kulkarni, "Term weighting using contextual information for categorization of unstructured text documents," in *2015 Annual IEEE India Conference (INDICON)*, IEEE, pp. 1-4, 2015.

[17] J. J. Patil and N. Bogiri, "Automatic text categorization: Marathi documents," in *2015 International Conference on Energy Systems and Applications*, IEEE, pp. 689-694, 2015.

[18] A. T. Sadiq and S. M. Abdullah, "Hybrid intelligent technique for text categorization," in *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, IEEE, pp. 238-245, 2012.

[19] D. O. Sahin, O. E. Kural, E. Kilic, and A. Karabina, "A Text Classification Application: Poet Detection from Poetry," *arXiv preprint arXiv:1810.11414,* 2018.

[20] M. A. Ahmed, R. A. Hasan, A. H. Ali, and M. A. Mohammed, "The classification of the modern arabic poetry using machine learning," *Telkomnika,* vol. 17, no. 5, 2019.

[21] M. Anandarajan, C. Hill, and T. Nolan, "Introduction to Text Analytics," in *Practical Text Analytics*: Springer, pp. 1-11, 2019.

[22] M. J. Denny and A. J. P. A. Spirling, "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it," vol. 26, no. 2, pp. 168-189, 2018.

[23] K. Aas and L. Eikvil, "Text categorisation: A survey," ed: *Technical report, Norwegian computing center*, 1999.

[24] X. Meng, Q. Chen, and X. Wang, "Semantic feature reduction in chinese document clustering," in *2008 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 3721-3726, 2008.

[25] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70-79, 2018.
[26] D. Ö. Şahin and E. J. A. Kılıç, "Two new feature selection metrics for text classification," vol. 60, no. 2, pp. 162-171, 2019.
[27] Z. J. I. j. o. c. Pawlak and i. sciences, "Rough sets," vol. 11, no. 5, pp. 341-356, 1982.
[28] H. Rasiowa and A. Skowron, "Rough concepts logic," in *Symposium on Computation Theory*, Springer, pp. 288-297, 1984.

BIOGRAPHIES OF AUTHORS

Saif Ali Alsaidi
PhD. Candidate Computer science
University of Technology - Iraq
Employ at Wasit University - Iraq

Dr. Ahmed Tariq Sadeq
Prof.Doctor in Artificial Intelligence
Computer science department
University of Technology - Iraq

Dr. Hasanen S. Abdullah
Asst.Prof.Doctor in Artificial Intelligence
Computer science department
University of Technology - Iraq