RESEARCH ARTICLE

Improving Kurdish Web Mining through Tree Data Structure and Porter's Stemmer Algorithms

Ari M. Saeed[1], Tarik A. Rashid*[2,4], Arazo M. Mustafa[3], Polla Fattah[2] and Birzo Ismael[2]

[1]Department of Computer Science, College of Science, University of Halabja, Halabja, Iraq
[2]Department of Computer Science and Engineering, University of Kurdistan Hewler, Erbil, Iraq
[3]School of Computer Science, College of Science, University of Sulaymaniyah, Sulaymaniyah, Iraq
[4]Department of Software and Informatics Engineering, Salahaddin University-Erbil, Erbil, Iraq

***Corresponding author's email:** tarik.ahmed@ukh.edu.krd

Received: 27 April 2018 **Accepted:** 18 June 2018 **Available online:** 30 June 2018

ABSTRACT

Stemming is one of the main important preprocessing techniques that can be used to enhance the accuracy of text classification. The key purpose of using the stemming is combining the number of words that have the same stem to decrease high dimensionality of feature space. Reducing feature space causes to decline time to construct a model and minimize the memory space. In this paper, a new stemming approach is explored for enhancing Kurdish text classification performance. Tree data structure and Porter's stemmer algorithms are incorporated for building the proposed approach. The system is assessed through using support vector machine and decision tree (C4.5) to illustrate the performance of the suggested stemmer after and before applying it. Furthermore, the usefulness of using stop words is considered before and after implementing the suggested approach.

Keywords: Kurdish Text Classification; Porter's Stemmer Algorithm; Stemming; Tree Data Structure

1. INTRODUCTION

The number of visitors to internet has been expanded intensely and the quantity of content has grown exponentially. Thus, the challenge is now significantly higher as to how you may conduct searching and/or recognizing the data. Several data categories in web applications are available such as texts, audios, videocassettes, charts, and diagrams. The most common and important type of data media that have been used by users on the internet is text. Texts are mainly used for scientific purposes such as communication, conversation, and exchange ideas, facts, and opinions. Hence, it is used for classification, handling, and unifying huge amount of information of different applications such as clarifying bulletins, health coding, web surfing, and information retrieval. Information contents in text classification are characterized depending on each document meaning. Thus, automatic machine learning techniques as replacements for a rule-based approach are used for classification (Mohammed et al., 2018; Karthik et al., 2016).

A wide range of machine learning classifiers has been used to classify documents such as C4.5, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) are examples of machine learning classifiers that have been used to classify documents (Sharma et al., 2016). Nonetheless, some imperative preprocessing steps are needed to be conducted on the data sets before classification task to ensure the performance and accuracy of classifiers are achieved. These imperative steps are tokenizing, removing stop words, and stemming (Bahassine et al., 2014).

Stemming is one of the above steps that can be employed on the text document for obtaining the root of word or the Stem. Usually, it is used for word standardization through taking

Access this article online

DOI:10.25079/ukhjse.v2n1y2018.pp48-54 **E-ISSN:** 2520-7792

away the affixes of the words. It is supportive in numerous applications such as information retrieval, compression of texts, computational linguistics, and categorization of texts (Duwairi et al., 2007; Esmaili et al., 2013).

Several scientists have applied stemmers on various languages such as English, Dutch, Slavic, and others. The efficiency of using stemmers appears more on some languages that have extra complex morphology than others for instance, the usefulness of using stemmers in Slovene language is more visible than English or Dutch, and this is because Slovene is more complex than English and Dutch (Tanja Gaust, 2002). The Kurdish language also has complex morphology, and in addition, it has huge inflectional and derivational affixes. Hence, a Kurdish stemmer is very significant to stem Kurdish documents. Nonetheless, such challenges are still considered to be at their early days (Saeed et al., 2018).

In this paper, a new stemming approach for Kurdish Sorani texts is suggested for obtaining the root of the words. Approximately 40–45 million people speak the Kurdish language in four states: Turkey, Iran, Iraq, and Syria. In general, the Kurdish language is part of Indo-European family, and specifically, it is part of the Iranian group. The Kurdish language has several dialects, but the most common types are Sorani and Kurmanji. The Sorani dialect is used only in Iraq and Iran, whereas the Kurmanji dialect used in all the four parts of Kurdistan. The Kurdish language has two different official scripts. The Sorani dialect uses an Arabic script in which the writing style is from right-to-left, whereas the Kurmanji dialect uses Latin script and writing style is from left-to-right. The dataset in this research was collected from different online websites that were written in Arabic script with Kurdish Unicode Fonts (Rashid et al., 2017; Salavati et al., 2013). The dataset can be obtained from the following link https://archive.ics.uci.edu/ml/datasets/ KDC-4007+dataset+Collection. With the availability of this data, a problem of not recognizing Kurdish texts with several local fonts which are available on the internet will be avoided by users or readers when uploading data or downloading data.

In this research work, SVM and C4.5 are implemented on this stemmer through utilizing (Rashid et al., 2017).

2. RELATED WORK

There are numerous techniques that have been developed to produce stemmers for different languages such as English, Arabic, and Persian for enhancing the accuracy in text classification; nonetheless, the number of stemmers in Kurdish are few and the research works on this area are few.

The KNN, Naïve Bayesian (NB), J48, and sequential minimal optimization (SMO) classifiers were applied for evaluating both the Light and Khoja stemmers on Arabic texts. Features are reduced through using normalization process. The digits, formatted tags, special marks, punctuation marks, and Latin words are removed. Their experimental results demonstrated that the light stemmer outperformed the Khoja stemmer when 10-fold cross validation technique was used. The evaluation process is performed on the classifiers through using the weighted average of F-measure, precision, and recall. In SMO, all instances are classified correctly when the light stemmer was applied (Khalid et al., 2016; Mamoun and Mahmoud, 2016).

In an evaluation research work, a novel method was applied for evaluating a classifier using Information Gain (IG) as a feature selection technique. Both Bayesian networks (BN) and multinomial NB (MNB) were applied as classifiers. Their experiential results were assessed through three datasets for training and testing the proposed method. The datasets were Reuters-REO, Review Polarity, and Reuters-REI. The proposed method outperformed on six classifiers (BN, KNN, MNB, SVM, NV, and decision tree) (Rahman and Usman, 2016).

A new research work for assessing information retrieval on Kurdish Sorani texts through Pewans dataset was proposed. The Pewan is the first typical test collection for Sorani language. It was collected from both Peyamner and VOA as two online news agencies. The dataset size was 1 KB and collected between 2003 and 2012 (18420 articles from VOA and 96920 articles from Peyamner). In this research study, a list of stop words was created that involved 282 extremely regular words and the N-gram was applied for accumulating 30 lists of affixes. The light stemmer was constructed for assessing the effectiveness of information retrieval on the documents. The experiment indicated that the quality of information retrieval system can be improved through applying the stemmer and the performance of information retrieval is enhanced in Sorani texts (Esmaili et al., 2013).

A stemmer for Kurdish Sorani texts was developed for reducing discrepancies of words to roots. Their experimental results showed that the effectiveness of information retrieval was increased and the dimensionality of feature vectors in documents was decreased when the stemmer was used. It was concluded that the processes that were applied to

Kurdish Sorani texts could be revised and applied in Kurdish Kurmanji too for greater efficiency (Saeed et al. 2018; Salavati et al., 2013).

3. PROPOSED APPROACH

In this section, an innovative approach is used for stemming Kurdish text classification. There are some comprehensive steps in classification before conducting it. These steps are raw data collection, pre-processing on data, data representation, and finally, classification (Esmaili et al., 2013) as shown in Figure 1.

3.1. Dataset Collection

The dataset in this research work is collected from different websites. The data contain 4007 text files that are categorized manually for eight different classes; they are economy, education, style, sport, art, health, religion, and social. Each class is equally distributed to have 500 class documents (Rashid et al., 2017).

3.2. Data Pre-processing

The preprocessing steps are applied for reducing the noise of data and number of the features. These preprocessing steps are as follows (Alajmi et al., 2012; Alami et al., 2016):

a. Tokenization, this is applied to break up the sentences into words through removing spaces.
b. Eliminate all non-Kurdish characters, digits, punctuation marks, and numbers.
c. Eliminate stop words and useless words. This research works prepares a list of stop words. This list contains 240 stop words (Rashid et al., 2017).
d. Replacing characters such as:
“ه”+ space with “ە”

- ك with ک
- ة with ە
- يwith ی
- ؤ with و

e. Stemming, this is the most significant part of this research work as it has greater impact on the classification of Kurdish text. The stemming approach can be divided into two common approaches. These are called a rule-based approach and a statistical approach. In the rule-based approach usually, a set of rules are applied for removing suffixes of each word that terminated with a set of suffixes. The algorithm of Lovins is the initial implementation, which contains 35 transformation rules, plus 294 endings with 29 circumstances. In addition, there is Porter's stemmer that contains 60 rules collected in five steps as base rules for obtaining the root of words. On the other hand, there are some statistical methods, which can use a dictionary for determining the root, and for instance, there is Yass stemmer through which distance measure is used for identifying the clustering-based approach for distinguishing the class. On the other hand, the Yass stemmer is an example of statistical method via which a dictionary is used for determining the root. The Yass stemmer is applied on the French and Bengali for enhancing the information retrieval performance. Yass stemmer is evaluated against Lovin and Porter stemmers; nonetheless, it requests an inclusive statistical linguistic resources and tools for implementation (Salavati et al., 2013). In this research work, a list of prefixes and suffixes

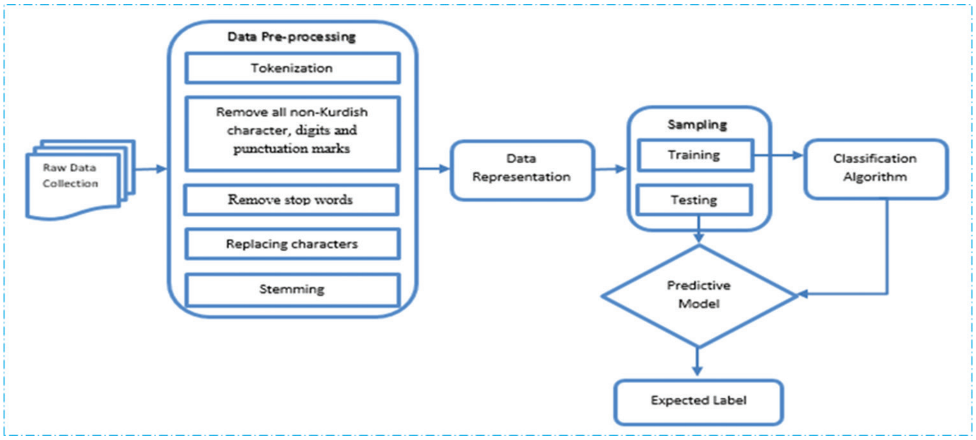

Figure 1. Proposed method

are used for determining the stem of the word as shown in Table 1. The following points are the main features of the proposed approached:-

1. The approach consists of nine levels that start with a feature or token and ends with a leaf or a stem.
2. The token can hold three prefixes and six suffixes. Hence, three levels are assigned for removing three prefixes and six levels for removing six suffixes.
3. If the token starts with one of prefixes and the length of token is >3, then the prefixes are removed and the process is repeated 3 times, for instance, in token (جێبەجێکردن), in the first process, (جێ) is removed in the first level, then (بە) is removed in the second level, after that another (جێ) is removed in the third level, and (کردن) is remained.
4. The rest of the levels (4 till 9) are used for removing the suffixes of the token, and if the token is ended with one of the suffixes and the length of the token is >3, then the suffixes are removed as the first level and the process is repeated till the last level.
5. The last level is leaves (stems), the best stem is selected by evaluating all stems to get the smallest one, and the length of stem must be >2 as shown in Figure 2.

Table 1. List of prefixes and affixes

List of prefixes	List of suffixes
”دە“,”کۆ“,”پڕ“,”جێ“,”لە“,”بە“,”ڕا“,”هەڵ“ دەر“,”سەر“,”یان“,”ی“,”تان“,”ت“,”مان“,”م”, ب“,“”	ایە"," دا","ش", "ی", "گا", "یان","تان " " "," مان"," ەکان ", " و ", " ێک ", " ەکە ", "ەک" ,بە" ","وە"," ەوە","تن", "ین", "ان", "دن", "ە" ,"ن","من","ێت", هات","کار","بوون", "تر"”"
List of Translated Kurdish Prefixes into English “ten”,”plural”,”full”,”filler”,”than”,”be”,”filler”,”filler” ,”out”,”head”,”them”,”him/her/its”,”your”,”your”,”us”,”me”,”filler”	List of Translated Kurdish Suffixes into English “ filler”,” filler”,”filler”, “him/her/its”, “ox”, “them”, your “, “us”,” filler “, “ and “, “ filler “, “ the “, “a” , “be”,”filler”, “filler”,”filler”, “filler”, “filler”, “filler”, “filler” ,”filler”,”i”,”filler”, “filler”, “been”,”done”, come””

Table 2. Different illustration of stem

Kurdish word (features)	English meaning	Stem
ڕاکردن	Running	کرد
هەڵمانکردایە	We turned on	کرد
دەرتانکردینەوە	Did you get them out	کرد
لەبەسەرکردنەوەمان	We have visited	کرد
بەکۆکردنەوەی	By collecting	کرد
جێبەجێکردن	executing	کرد
کردەوەی	Opened	کرد
نەتکردووە	Did not do it	کرد

The proposed approach guarantees to choose the best path for obtaining the stem. The above graph explains that in level four (کردنەوەمان) is ended with three different suffixes. The suffixes of (مان) and (ان) are selected as the best suffix to arrive at the stem that is (کرد) while (ن) is not selected since the length of character of (کرد) is greater than two are and smaller than (ن کردنەوەما). However, after removing the suffixes of (مان) and (ان), still there are suffixes, which have to be removed such as (وە) and (نە), there are suffixes, which have to be removed such as (هوە) ,(وە) ,(ە) ,(م), and (ن). Selecting the stem in this approach depending on suffixes, furthermore, high dimensionality of feature space has a greater impact on the performance and efficiency of the text classification algorithm. Thus, usefulness of this approach is to reduce feature space as presented in Table 2.

Table 2 illustrates the post implementing of this approach on eight features, only one stem is produced. Accordingly, high dimensionality of features spaces is decreased in the proposed approach.

4. DATA REPRESENTATION

One of the important steps in text classification is to convert unstructured text documents to a form that can be ready to be used by machine learning algorithms. Vector space model (VSM), N-gram, and Bag-of-Words are algorithms that can be implemented for this purpose. VSM is the common method that can be used to represent the text as a vector $\vec{d} \cdot \vec{d} = \{w_1 \ldots w_n\}$ Where where w_n is the weight term in text documents (Danisman and Adil, 2008). Moreover, vectors have certain weights that are utilized to increase the accuracy of classification when term weighting is executed.

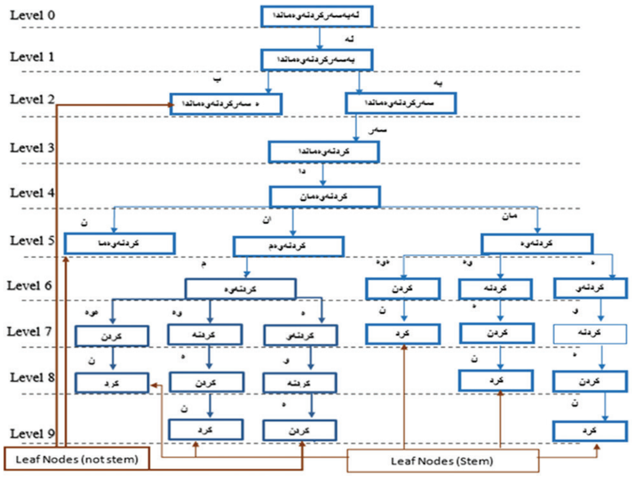

Figure 2. The level steps of removing prefixes and affixes from the Kurdish word

5. CLASSIFICATION

In this research work, KDC-4007 dataset was used to evaluate the proposed approach. SVM and decision tree (C4.5) (Mustafa and Rashid, 2017; Rashid et al., 2017; Saeed et al., 2018) are applied as two common machine learning algorithms for comparing the results. The dataset is partitioned for two inequality portions. The first portions are training that is 70% while the rest is testing that is 30%.

6. RESULT AND DISCUSSION

Various experimental tests are used in this paper. Table 3 shows the results of five different experiments and the impact of stemming is exhibited:

1. The first experiment applies only natural sentences (NS).
2. In the second experiment stop words are removed (RSW).
3. In the third experiment, the RSW and then the stemmer are applied (RSW then SK).
4. In the fourth experiment the stemmer is applied, and then the RSW (SK then RSW).
5. The fifth experiment uses the stemmer but without removing stop words (SK not RSW).

Table 3 exhibits the results of J48 experiments.

Tables 3 and 4 indicate the performance of the proposed stemmer on text classification. The values obtained for F1 measure increased dramatically and the success of this stemmer is depicted when RSW then SK is applied for each class (education, art, social, and style) in decision tree (C4.5) and the classes (health, education, art, social, and economy) in SVM as well while the rest of classes decreased slightly.

However, F1 measure for the stemmer is implemented before and after removing stop words and compared with removing stop word without stemming. According to Tables 3 and 4, the F1 measures for each SK then RSW has different values for each SVM and decision tree (C4.5), for classes of health,

Table 3. F1 Performance results for the decision tree (C4.5)

F1 measure					
Classes	NS	RSW	RSW then SK	SK then RSW	SK not RSW
Religion	0.61	0.496	0.684	0.688	0.707
Sport	0.666	0.667	0.829	0.831	0.824
Health	0.573	0.496	0.746	0.745	0.751
Education	0.688	0.655	0.775	0.759	0.759
Art	0.675	0.696	0.784	0.783	0.773
Social	0.7	0.692	0.83	0.796	0.775
Style	0.681	0.673	0.863	0.829	0.824
Economy	0.759	0.734	0.782	0.857	0.87

Table 4. F1 performance results for the SVM

F1 measure					
Classes	NS	RSW	RSW then SK	SK then RSW	SK not RSW
religion	0.84	0.833	0.864	0.862	0.868
Sport	0.922	0.919	0.938	0.936	0.942
Health	0.877	0.888	0.913	0.902	0.9
Education	0.897	0.906	0.921	0.918	0.921
Art	0.905	0.915	0.957	0.948	0.949
Social	0.9	0.912	0.928	0.924	0.926
Style	0.901	0.918	0.927	0.927	0.933
Economy	0.947	0.95	0.967	0.965	0.964

SVM: Support vector machine

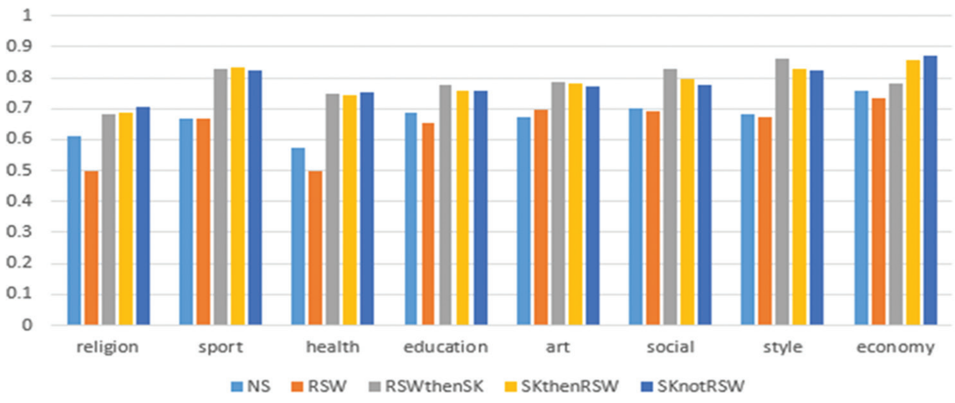

Figure 3. The performance of F1 measure for five experiments using decision tree *(C4.5)*

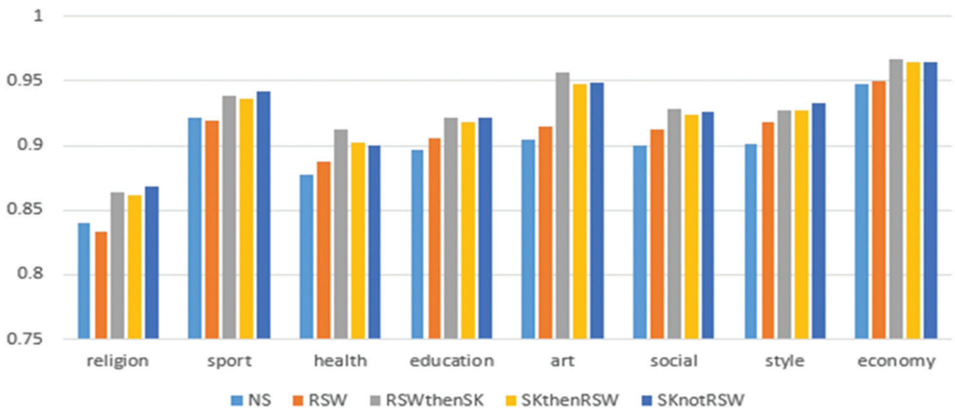

Figure 4. The performance *of F1* measure for five experiments using support vector machine

education, art, social, and style. In Table 3, classes of religion, sport, and economy have higher F1, while in Table 4, all the class go down gently without style remains in its state. Therefore, it can be said that the success of stemmer before and after removing stop words depended on the type of class in dataset. Figures 3 and 4 represent the performance results of F1 measure on five experiments for each class label via classifiers C4.5 and SVM respectively

In the decision tree or C4.5 classifier, unlike the experiment SK not RSW, the F1 measure result on the experiments RSW then SK and SK then RSW is decreased for all classes, except for style, which is increased steadily (Figure 3). In the same way, in SVM, the F1 measure result on the mentioned experiments for the classes of religion, sport, education, and style is gradually increased, but classes of health, art, social, and economy is slightly decreased (Figure 4).

7. CONCLUSIONS AND FUTURE WORK

In this research work, a new stemmer approach is examined on KDC-4007 dataset for improving text classification in Kurdish language. The proposed algorithm used tree data structure and Portes stemmer's techniques. It can be concluded that the F1 measure obtained the best result when the stemmer was implemented and stop words were not removed. Hence, the influence of stop words showed when NS experiment compared with removing stop word experiment.

Another important investigation in this experiment effect this stemmer on stop words.

There are various proposals that can be investigated to improve the efficiency of stemmer in Kurdish text such as mixing new techniques and showed the effects of affix in Kurdish stemmer.

REFERENCES

Alajmi, A., Saad, E. M., & Darwish, R. (2012). Toward an ARABIC stop-words list generation. *International Journal of Computer Applications (0975 – 8887)*, 46(8), 8-13.

Alami, N., Meknassi, M., & Ouatik, S. A. (2016). Impact of stemming on Arabic text summarization. *International Colloquium on Information Science and Technology (CiSt).* Tangier, Morocco: IEEE.

Bahassine, S., Mohamed, K., & Abdellah, M. (2014). New stemming for Arabic text classification using feature selection and decision trees. *5th International Conference on Arabic Language*. Oujda, Morocco: IEEE. p. 200-205.

Danisman, T., & Adil, A. (2008). Feeler: Emotion classification of text using vector space model. In*: AISB 2008 Convention Communication Interaction and Social Intelligence.* Vol. 2. Aberdeen, UK: AISB.

Duwairi, R., Al-Refai, M., & Khasawneh, N. (2007). Stemming versus light stemming as feature selection techniques for arabic text categorization. *Innovations in Information Technologies (IIT).* Dubai, Dubai: IEEE.

Esmaili, K. S., Donya, E., & Shahin, S. (2013). Building a Test collection for Sorani Kurdish. *International Conference on Computer Systems and Applications (AICCSA).* Ifrane, Morocco: IEEE.

Khalid, A., Zakir, H., & Baig, M. A. (2016). Arabic stemmer for search engines information retrieval. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 7(1), 407-411.

Mamoun, R., & Mahmoud, A. (2016). Arabic text stemming: Comparative analysis. *Conference of Basic Sciences and Engineering Studies (SGCAC).* Khartoum, Sudan: IEEE.

Mohammed, F. S., Zakaria, L., & Omar, N. (2012). Automatic Kurdish SORANi text categorization using N-gram based model. *International Conference on Computer and Information Science (ICCIS).* Kuala Lumpeu, Malaysia: IEEE.

Mustafa, A. M., & Rashid, T. A. (2017). Kurdish stemmer pre-processing steps for improving information retrieval. *Journal of Information Science*, 44(1), 15-27.

Karthik, P., Saurabh, M., & Chandrasekhar, U. (2016). Classification of text documents using association rule mining with critical relative support based pruning. *International Conference on Advances in Computing, Communications and Informatics (ICACCI).* Jaipur, India: IEEE.

Rahman, A., & Usman, Q. (2016). A Bayesian classifiers based combination model for automatic text classification. *International Conference on Software Engineering and Service Science (ICSESS).* Beijing, China: IEEE.

Rashid T.A., Mustafa A.M., & Saeed A.M. (2018). Automatic Kurdish text classification using KDC 4007 dataset. In: Barolli, L., Zhang, M., & Wang X., editors. *Advances in Internetworking, Data and Web Technologies. EIDWT 2017. Lecture Notes on Data Engineering and Communications Technologies*. Vol. 6. Cham: Springer.

Saeed, A. M., Rashid, T. A., Mustafa, A. M., Al-Rashid Agha, R. A., Shamsaldin, A. S., & Al-Salihi, N. K. (2018). An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification. *Iran Journal of Computer Science*, 1(2), 99-107.

Salavati, S., Sheykh, E.K., & Akhlaghian, F. (2013). Stemming for Kurdish information retrieval. In: Banchs, R.E., Silvestri, F., Liu, T.Y., Zhang, M., Gao S., & Lang, J., editors. *Information Retrieval Technology. AIRS 2013. Lecture Notes in Computer Science.* Vol. 8281. Berlin, Heidelberg: Springer.

Sharma, N., A. S., & V. T. (2016). Text classification using combined sparse representation classifiers and support vector machines. *4th International Symposium on Computational and Business Intelligence (ISCBI).* Olten, Switzerland: IEEE.

Tanja Gaust, G. B. (2002). Accurate stemming of Dutch for text classification (language and computers: Studies in practical linguistics). In: Theune, M., Nijholt, A., & Hondorp, H., editors. *Computational Linguistics in the Netherlands.* Amsterdam: Rodopi. pp. 104-117, 14.