A Robust Categorization System for Kurdish Sorani Text Documents

Article *in* Information Technology Journal · March 2017

DOI: 10.3923/itj.2017.

3 authors:

Tarik A. Rashid
University of Kurdistan Hewlêr (UKH)
463 PUBLICATIONS **10,764** CITATIONS
SEE PROFILE

Arazo M. Mustafa
University of Sulaimani
12 PUBLICATIONS **118** CITATIONS
SEE PROFILE

Ari M. Saeed
University of Halabja
13 PUBLICATIONS **136** CITATIONS
SEE PROFILE

http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

OPEN ACCESS Information Technology Journal

ISSN 1812-5638
DOI: 10.3923/itj.2017.27.34

Research Article
A Robust Categorization System for Kurdish Sorani Text Documents

[1]Tarik A. Rashid, [2]Arazo M. Mustafa and [3]Ari M. Saeed

[1]Department of Computer Science and Engineering, University of Kurdistan Hewler, Hewler, Kurdistan
[2]School of Computer Science, College of Science, University of Sulaymaniyah, Sulaymaniyah, Kurdistan
[3]Department of Computer Science, College of Science, University of Halabja, Halabja, Kurdistan

Abstract

Background: Text classification is a process of automatically assigning sets of documents into class labels depending on their data contents. It is also considered as an important element in the management of tasks and organizing information. Seemingly, the text classification process depends hugely on the quality of preprocessing steps. **Materials and Methods:** In this study, a novel pre-processing method (Normalizing, stemming, removing stopwords and removing non-Kurdish texts and symbols) was evaluated by means of comparing the performance of two text classification techniques, namely; decision tree (C4.5) classifier and Support Vector Machine (SVM) classifier. Two automatic learning algorithms for text categorization were compared using a set of Kurdish Sorani text documents that was collected from different Kurdish websites. The set of documents falls into 8 main categories namely: Sports, religions, arts, economics, educations, socials, styles and health. A set of preprocessing steps was performed on text documents such as normalizing some characters, stemming, removing stopwords and removing non-Kurdish texts and symbols, next, the documents were changed into a appropriate file format and finally the classification was conducted. **Results:** The findings of this study illustrated that the highest accuracy value 93.1% and the smallest time taken to building classifier was achieved with the SVM classifier after pre-processing and feature weighting steps were performed. **Conclusion:** The experimental results of this study can be utilized in future as a baseline to compare with other classifiers and Kurdish stemmers.

Key words: Documents classification, Kurdish stemming, machine learning algorithm, information retrieval

Received: September 01, 2016 Accepted: November 14, 2016 Published: December 15, 2016

Citation: Tarik A. Rashid, Arazo M. Mustafa and Ari M. Saeed, 2017. A robust categorization system for Kurdish Sorani text documents. Inform. Technol. J., 16: 27-34.

Corresponding Author: Tarik A. Rashid, Department of Computer Science and Engineering, University of Kurdistan Hewler, Hewler, Kurdistan

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Apparently, the amount of text documents available on the World Wide Web (WWW) grows rapidly in electronic forms; therefore, an automatic document classification is becoming an important field in computer science. Text categorization is a technique for organizing and managing these data text documents, at the same time is improving the precision of retrieval. Text categorization is also the process of classifying unstructured documents into one or more pre-defined categories such as art or sport, etc., based on linguistic features and content. Besides, it has been applied in many applications, examples of these applications are automatic web pages categorization[1], spam filtering[2], e-mail filtering[3], word sense disambiguation[4] and many others. Further, text categorization tasks can be divided into two types: Supervised and unsupervised document classification. In the supervised document classification, there are some external mechanisms that provide information on the correct classification for defining classes via the classifier and in the unsupervised document classification, the classification must be done without any external references, thus, the system does not have predefined classes[5,6].

In this study, supervised texts classification is adopted which is more accurate and is a sub-domain of data mining and machine learning techniques. At present, researchers presented many studies in text classification in various languages, but for the Kurdish language, there are very few studies. However, developing a text classification system for Kurdish documents requires various challenges such as the differences and the complex morphologies of the Kurdish language and the main factors behind these complexities are the large uses of inflectional and derivational affixes, also, there are challenges related to Kurdish Sorani dialect's writing system definiteness markers, possessive pronouns, enclitics and many of the widely-used postpositions are used as suffixes[7,8].

Kurdish language is a branch of Indo-Iranian languages and it is the official language of Kurdish people who live in four countries, namely, Iraq, Turkey, Iran and Syria. The Kurdish language consists of 33 letters and written from right to left like Urdu, Persian and Arabic languages and it has two main dialects: Sorani and Kermanji. Kurdish Sorani is the official dialect of Kurdish people in Iraq and Iran. Generally, the documents in this study are purposed and prepared in Kurdish Sorani dialect. The main goals of this study were using a new pre-processing method and then classifying Kurdish documents through utilizing different methods and finally evaluating them.

The problem of text classification in other languages (such as English, Arabic, etc.) has been handled by presenting many different studies and using different techniques of classifying documents with acceptable performance.

Mohammed *et al.*[9] used the N-gram frequency statistics for classifying Kurdish text. An algorithm called Dice's measure of similarity was employed to classify the documents. A corpus of Kurdish text documents was build using Kurdish Sorani news articles collected from the online websites of several Kurdish newspapers. It consisted of 4094 text files divided into 4 categories: Art, economy, politics and sport. Each category was divided equally per their sizes (50% as a training set and 50% as a testing set). The documents in all categories go through the pre-processing steps which involve replacing some characters and removing Kurdish stopword. For the training and the test documents, the N-gram word level 1 g and character level (2-8) frequency profile were generated for each document and saved in text files. The recall, precision and F1 measure were used to compare the performance. The results showed that N-gram level 5 outperformed the classification with the other aforementioned N-gram levels.

Al-Harbi *et al.*[10] proposed two popular classification algorithms SVM and C5.0 for Arabic text classification. In general, documents used in this study consisted of 17,658 text documents, which were collected from different sources. Arabic text classification was implemented to accomplish both feature extraction and selection tasks, then the chi-square technique was used to calculate features which were important. The text documents were divided into training and testing sets (70% for training and 30% for testing in each corpus). The results showed that C5.0 classifier produced the average accuracy rate of 78.42% which was better than SVM that produced the average accuracy rate of 68.65%.

Zhang *et al.*[11] evaluated the performance of three document representation methods (TF.IDF, Latent Semantic Indexing (LSI) and multi-word) in text classification. In the study, they used two different languages namely; Chinese and english. The dataset for the Chinese corpus was TanCorpV1.0, which consisted of 14,150 documents with 20 categories. Then, 4 categories were selected randomly from original corpus, thus, totally 1200 documents were used. There was reuters-21578 distribution 1.0 for the english corpus, which contained 21,578 documents with 135 categories, again 4 categories assigned, thus, totally 2042 english documents

were used. In addition, stopwords were eliminated from the english documents. The SVM was used to estimate the performances of the above methods. Besides, information gain feature selection method was applied. Experimental results demonstrated that Latent Semantic Indexing (LSI) outperformed others methods in both document collections in text categorization. In this study, the proposed method is explained.

MATERIALS AND METHODS

In this study, the methodology to categorize Kurdish Sorani documents was introduced. Figure 1 describes the specifics of the proposed methodology which were elaborated in the following.

Dataset collection: A set of Kurdish text documents collected from different websites. The documents consist of 8 categories, i.e., sports, religions, arts, economics, educations, socials, styles and health, each of which consists of 500 text documents, where the total size of the corpus is 4,000 text files. Table 1 shows the distribution of the documents among these 8 categories.

Table 1: Numbers of classes and documents in selected sources

Class	No. of documents	No. of classes	Sources
Sport Health Economy	1057	3	http://www.rudaw.net/sorani
Sport Economy Education	753	3	http://www.nrttv.com
Style Socials Education	1705	4	http://www.snnc.co
Art Religion	504	1	Randomly compiled from the internet

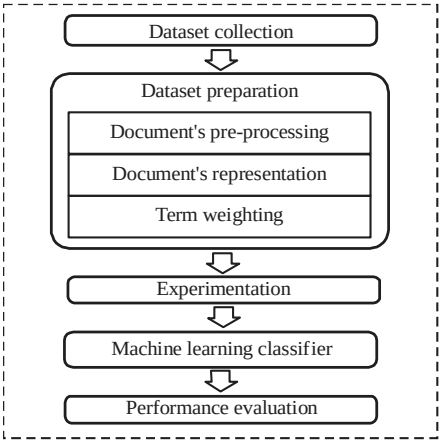

Fig. 1: Methodology of building Kurdish Sorani classifier

Dataset preparation: In this study, Kurdish text documents have been prepared per the following steps.

Text pre-processing: Kurdish text documents have been pre-processed per the following steps:

- **UTF-8 encoding:** The first step in data preprocessing is transforming text documents to UTF-8 encoding
- **Tokenization:** Converting from a string of characters into a sequence of tokens (features) delimited by white space, punctuations, tab, new line, etc
- **Removal of number and punctuation:** A simple tool is created using Java programming language to remove hyphens, punctuation marks and any occurrences of digits, such as (1), (0.2), (100), (!), (?), (,), (:), etc
- **Normalization:** This step is aiming to unify words typed differently that are caused by multi-unicode character:
  - Replacing Arabic letter YAA (Y in english) to Arabic letter-KURDISH YEH (Y in english)
  - Replacing Arabic letter ALEF MAKSURA to Arabic letter-KURDISH YEH (Y in english)
  - Replacing Arabic letter KAF (K in english) to Arabic letter-Kurdish letter KEHEH (K in english)
  - Replacing ea (e in english) which is consisted of "ZWNJ1+HEH to Kurdish letter AE
- **Stemming:** It is a technique for reducing an inflected or derived word to its stem (roots). The major advantage of this step is to reduce the total number of terms (features) in the text documents and therefore, the size and complexity of the data storage requirements of text classification algorithms are reduced. In this study, a Kurdish Stemming algorithm (K Stemming algorithm) is presented. It is a module which contains a set of simple rules depending on conditions to strip affixes from the given word to find possible roots. Stemming is used to enhance a high dimensional feature space problem to improve the accuracy of text classification systems
- **Removing stopwords:** A stopword is a word like a pronoun, conjunctions and prepositions. The stopwords have little semantic content and occur frequently in a document. These stopwords will increase the noise of the results because they are so common and do not help in discrimination among categories. In this study, a list of Kurdish stopwords is prepared (nearly 240 stopwords)

Documents representation: Before starting the classification task, documents must be transformed into a format that is

recognized by a computer, Vector Space Model (VSM) is the most commonly used method[12]. This model represents the text documents as a vector of words. The text documents are converted from the full text into a document vector.

Term weighting: In the text classification problem, terms that appear in documents are represented to machine learning classifiers as real-numbered vectors of weights. Term weighting is being determined by several ways: Such as Boolean weighting, word frequency weighting, TF-IDF, entropy etc. In this study, the TF-IDF weighting function is used. It is based on the distribution of the terms within the document and within the collection, where the higher value indicates that the word occurs in the document and does not occur in many other[13,14]. This can be expressed as following Eq. 1:

$$TF.IDF(t_i, d_j) = TF(t_i, d_j) \times \log(N/DF(t_i)) \quad (1)$$

where, TF is the frequency of term t_i in document d_j and DF (t_i) is the number of documents that contain term t_i, after stopword removing and word stemming and N is the total number of documents.

Experimentation of methodology: This study explained the methodology and classification processes for two classification algorithms. It was proposed to experiment the same data, but differently, this means that the testing was done for more than once for two algorithms for evaluating them. Table 2 shows the description of 5 tests (Test 1-5) and their methodology experimentations (Including normalization, stopword removal, stemming and term weighting).

Machine learning classifier: Support Vector Machine (SVM) is a supervised machine learning algorithm. It has been proposed for text classification by Cortes and Vapnik[15]. Researchers have used it widely in text categorization. The SVM builds a hyperplane that perfectly separates a set of positive patterns from a set of negative patterns with a maximum margin in case of linear. Likewise, Decision Tree (DT) algorithm is widely used in machine learning and data mining fields. The DT is simple and can be easily understandable and converted into a set of humanly readable if-then rules[16]. The ID3 algorithm is one of the most well-known decision tree algorithms. The C4.5 is an extension of ID3. In this study, the C4.5 algorithm applied.

Table 2: Characteristics of five tests and their methodology experimentations

Tests	Experimentation of methodology
1	Without pre-processing (original dataset)
2	Normalization and Kurdish stemming
3	Normalization and stopword removal
4	Normalization, Kurdish stemming and stopword removal
5	Normalization, Kurdish stemming and stopword removal+TF-IDF weighting

Table 3: Confusion matrix for a binary classifier

Class	Positive	Negative
Positive	True positive	False negative
Negative	False positive	True negative

Evaluation of the classification: In general, confusion matrix is adopted for effective evaluation in classification problems to estimate correctly and incorrectly classified instances for each class. The confusion matrix has only two classes, positive and negative for a binary classification problem[17]. Confusion matrix is shown in Table 3. Where, TP, FP, TN and FN are the true positive rate, the false-positive rate, the true negative rate and the false-negative rate, respectively.

In machine learning and statistics, the accuracy rate (Acc) and error rate (Err) are used to examine the performance of classification algorithms. The accuracy rate is the percentage of correctly predictions of the classifier and the error rate is the percentage of incorrectly predictions of the classifier, this is also called misclassification[18]. These are expressed as following Eq. 2 and 3:

$$\text{Accuracy rate} = \frac{TN + TP}{TP + FP + TN + FN} \quad (2)$$

$$\text{Error rate} = \frac{FN + FP}{TP + FP + TN + FN} \quad (3)$$

Moreover, recall, precision and F-measure are three widely used algorithms to evaluate the effectiveness of text categorization[19]. The F1-measure, introduced[20] is the harmonic average of both precision and recall. Recall, precision and F-measure are expressed respectively as following Eq. 4-6:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{Recall+precision}} \quad (6)$$

RESULTS AND DISCUSSION

After conducting a comprehensive study, some insightful thoughts and conclusions can be discussed. Table 4 shows the evaluation metrics for applying an SVM classifier on five tests respectively. As expected for the weighted averages of the precision, recall and F-measure values are shown in Table 4. Table 4 shows how close they are in almost all the tests considered. In addition, as can be seen from the Table 4 that the precision and recall of the 8 categories for both test 2 and test 4 were better than test 1 (using the original dataset). This is because the processing included the K Stemming via which the size of features was reduced and ultimately the final performance of Kurdish text classification is improved. On the other hand, it can be noticed that precision, recall and F-measure for the normalization and stopword removal processes did not affect or slightly affected in test 3. Per test 5 results, the SVM with TF-IDF term weighting yields better performance than the C4.5 with TF-IDF term weighting. The other related observation is noticed in the experimentations while using two popular processes which are stemming and removal of stopwords that reduced the building times for the classifiers compared with the original dataset. KStemming in general can considerably increase accuracy and reduce the learning times for the SVM classifier.

Table 4 shows that the C4.5 results, the weighted averages for the precision, recall and F1-measure in test 1 (the original dataset is involved) are very small, whereas the performances for the same dataset and the same classifier used in test 2 and 4 have increased very significantly compared to the original dataset. The reason for this is that the two tests in preprocessing step contained the KStemming technique, whereas, the performance for test 3 have increased very slightly which contained the normalization and stopword removal in the preprocessing stage. On the other hand, the performance of test 5 which includes feature weighting TF×IDF, produces same levels of precision, recall and F1-measure compared to test 4. In fact, if one carefully looks at the two Table 4 and 5 as the evaluation metrics for the SVM experiments and C4.5 experiments are shown, it is obvious that C4.5 more responsive for certain preprocessing steps. Table 4 and 5 show the weighted average for each test.

The dataset was experimented using two methods for measuring accuracies which are the percentage split method, where 70% of the dataset used as a training set and the remaining 30% used as testing set and the k-fold cross validation method via which the data was divided into 10 folds, a fold is used as testing and the remaining folds are used as training. All documents for training and testing passed through the pre-processing steps, which include tasks of

Table 4: Experimental results of recall, precision and F1-measure by classes for SVM classifier on 5 tests

	Test 1			Test 2			Test 3			Test 4			Test 5		
Classes	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Religion	0.91	0.78	0.84	0.93	0.84	0.88	0.93	0.75	0.83	0.95	0.78	0.86	0.96	0.78	0.86
Sport	0.89	0.91	0.90	0.93	0.97	0.95	0.90	0.95	0.92	0.93	0.99	0.96	0.92	0.99	0.95
Health	0.87	0.89	0.88	0.91	0.89	0.90	0.90	0.90	0.90	0.90	0.93	0.92	0.91	0.92	0.92
Education	0.87	0.92	0.89	0.91	0.93	0.92	0.88	0.93	0.91	0.91	0.92	0.92	0.93	0.94	0.94
Art	0.92	0.90	0.91	0.94	0.93	0.94	0.92	0.91	0.91	0.92	0.95	0.94	0.92	0.95	0.93
Social	0.90	0.92	0.91	0.89	0.91	0.90	0.88	0.91	0.90	0.91	0.93	0.92	0.90	0.92	0.91
Style	0.89	0.92	0.90	0.91	0.93	0.92	0.89	0.93	0.91	0.91	0.97	0.94	0.90	0.97	0.93
Economy	0.92	0.97	0.95	0.95	0.98	0.96	0.92	0.98	0.94	0.96	0.97	0.97	0.97	0.97	0.97
Weighted average	0.90	0.90	0.90	0.92	0.92	0.92	0.90	0.91	0.90	0.92	0.93	0.93	0.93	0.93	0.93

Table 5: Experimental results of recall, precision and F1-measure by classes for C4.5 classifier on 5 tests

	Test 1			Test 2			Test 3			Test 4			Test 5		
Classes	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Religion	0.66	0.62	0.64	0.83	0.63	0.72	0.69	0.49	0.57	0.84	0.62	0.71	0.83	0.62	0.71
Sport	0.66	0.65	0.65	0.84	0.89	0.86	0.67	0.61	0.64	0.83	0.92	0.87	0.83	0.92	0.87
Health	0.56	0.61	0.59	0.77	0.80	0.79	0.57	0.57	0.57	0.79	0.82	0.80	0.78	0.82	0.80
Education	0.62	0.68	0.65	0.82	0.83	0.82	0.70	0.80	0.74	0.85	0.83	0.84	0.85	0.83	0.84
Art	0.71	0.59	0.65	0.76	0.84	0.80	0.65	0.74	0.69	0.78	0.87	0.82	0.78	0.88	0.83
Social	0.70	0.68	0.69	0.74	0.80	0.77	0.68	0.70	0.69	0.76	0.80	0.78	0.76	0.80	0.78
Style	0.65	0.70	0.68	0.81	0.84	0.82	0.62	0.71	0.66	0.78	0.87	0.83	0.78	0.87	0.82
Economy	0.72	0.80	0.76	0.87	0.87	0.87	0.73	0.83	0.78	0.90	0.89	0.90	0.90	0.89	0.90
Weighted average	0.66	0.67	0.66	0.80	0.81	0.81	0.66	0.68	0.67	0.82	0.83	0.82	0.82	0.83	0.82

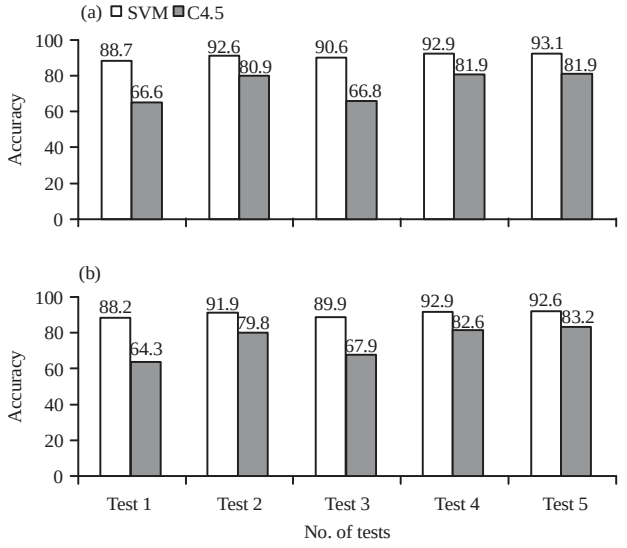

Fig. 2(a-b): Effect of pre-processing on the experimental results for SVM and C4.5 using (a) 10 folds cross validation and (b) Percentage split 70%

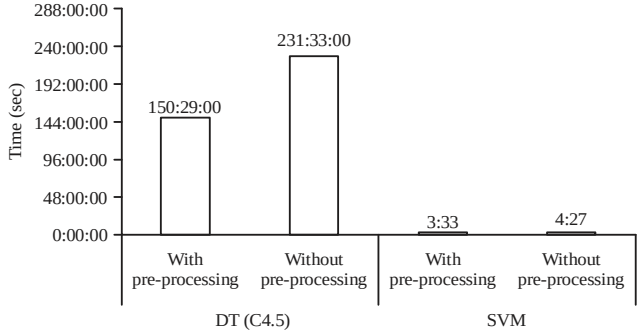

Fig. 3: Time taken to build classifiers of SVM and C4.5 with and without pre-processing

implementing, stopword removal, word stemming and feature weighting TF-IDF as explained in Table 2.

Figure 2a and b shows the best result obtained was through the SVM compared to the C4.5 classifier for each of the five different tests. Figure 3 shows that as expected, the learning times for a classifier like the SVM was generally low, whereas, the decision trees took considerably longer time to build, such longer building times do not necessarily correspond to higher accuracies. As indicated by the data introduced in Fig. 2a and b, the highest accuracy value of 93.1% was achieved by using the SVM when all the pre-processing steps and TF×IDF weighting were used with test 5 applying the cross-validation method. Whereas, the highest accuracy value of 83.2% was achieved by DT (C4.5) when all the pre-processing steps and TF×IDF weighting were used with test 5 applying the percentage split method compared to the values obtained from test 1 (i.e., original datasets).

In the case of the percentage split method, the accuracy values for test 4 and 5 were identical, literally, 92.6%, this means that the SVM classifier obtained the same accuracy results compared to the values obtained from C4.5 classifier which were 82.6 and 83.2%, respectively, which increased slightly after performing feature weighting TF×IDF. However, in the case of the cross validation method, the accuracy values for test 4 and 5 were identical, literally, 81.9%. This indicated that the C4.5 classifier obtained the same accuracy results compared to the values obtained from SVM classifier which were 92.9 and 93.1%, respectively, which increased slightly after performing feature weighting TF×IDF.

Figure 2a shows that the accuracy performance value of the two classifiers on datasets is significantly increased after applying Kurdish stemming-step module. For example, before this process, the accuracy values on the test 1 (original data set) and test 3 (Normalization and stopword removal) for the SVM and the DT (C4.5) classifiers were 88.7, 90.6, 66.6 and 66.8%, respectively; nevertheless, they became 92.6, 92.9, 80.9 and 81.9% in test 2 and 4 after performing the Kurdish stemming-step module, respectively. The experimental results in Fig. 2a and b showed the accuracy results for the percentage split and the cross-validation methods obtained over the two selected classifiers for 5 tests.

From the experimental results, it is obvious to observe that the Kurdish stemming method influenced the performance of the C4.5 classifier on 5 tests significantly. Thus, the range of accuracy in the C4.5 was higher than the SVM. In other words, the range of accuracy in the SVM is less influenced by the dimension of the data set than the C4.5. Thus, in the latter, the Kurdish stemming reduces the dimensionality of the dataset drastically by grouping words of the same origin together. For example, the word "Feerga" which in english means "School", the word "Feerbun" which in english means "Learning" and the word "Fearkar" which in english means "Teacher" are all grouped under the same root/stem "Feer" which in english means "Teach"). While this reduction makes it easier to build a classification model (especially for classifiers suffering from the 'Curse of dimensionality'). It can be noticed that among the 5 tests for the two classifiers, still the SVM produces higher accuracies, better speed of learning, better classification and higher tolerance to irrelevant features and noisy data. Also, one exceptional property of the SVM is that its capability to learn can be independent of the dimensionality of the feature space. Figure 3 shows the time taken to build classifiers (SVM and DT).

CONCLUSION AND FUTURE RECOMMENDATIONS

This study aimed at comparing 5 tests, test 1 (without pre-processing (Original dataset)), test 2 (Normalization and KStemming), test 3 (Normalization and stopword removal), test 4 (Normalization, KStemming and stopword removal) and test 5 (Normalization, KStemming and stopword removal+TF-IDF weighting based) on two well-known classification techniques (Decision tree (C4.5) and Support Vector Machine (SVM)). Both the SVM and C4.5 were used to classify Kurdish Sorani text documents into 8 class labels. The concept started by non-implementing or semi-implementing or implementing pre-processing steps on Kurdish Sorani text documents. The results showed that the SVM classifier achieved the highest accuracy. These 5 tests showed that there were differences between the two classifiers in performance in terms of accuracy and time.

The future study may include other well-known classifiers techniques such as using deep learning neural networks or using some advanced nature inspired algorithms for optimizing the SVM classifier. Finally, we plan to apply nature inspired algorithms for selecting the best features on the dataset before using the text classification techniques.

SIGNIFICANT STATEMENT

In general, text mining is interested in applying some technique of extracting interesting and useful knowledge from unstructured text collections. Text classification is one of the most important text mining tasks. With the vast number of various sets of documents uploaded every day to the world wide web and the internet. Thus, there is an expanding need to consequently compose these reports into their categories or classes to encourage the area and recovery of important documents. Consequently, text classification is a way towards characterizing an unstructured text document in its required category(s) depending upon its substance. In the context of text classification problem, applying the text classification problem on Kurdish Sorani text documents is one of the most challenging problems.

ACKNOWLEDGMENT

This study was supported by the University of Kurdistan Hawler, Hawler, Kurdistan.

REFERENCES

1. Bhalla, V.K. and N. Kumar, 2016. An efficient scheme for automatic web pages categorization using the support vector machine. New Rev. Hypermedia Multimedia, 22: 223-242.
2. Hu, W., J.L. Du and Y.K. Xing, 2016. Spam filtering by semantics-based text classification. Proceedings of the 18th International Conference on Advanced Computational Intelligence, February 14-16, 2016, IEEE., pp: 89-94.
3. Bahgat, E.M., S. Rady and W. Gad, 2016. An e-mail filtering approach using classification techniques. Proceedings of the 1st International Conference on Advanced Intelligent System and Informatics, November 28-30, 2015, Beni Suef, Egypt, pp: 321-331.
4. Hidalgo, J.M.G., M.D.B. Rodriguez and J.C.C. Perez, 2005. The role of word sense disambiguation in automated text categorization. Proceedings of the International Conference on Application of Natural Language to Information Systems, June 2005, Berlin, Heidelberg, pp: 298-309.
5. Nidhi and V. Gupta, 2011. Recent trends in text classification techniques. Int. J. Comput. Applic., 35: 45-51.
6. Nalini, K. and L.J. Sheela, 2014. Survey on text classification. Int. J. Innov. Res. Adv. Eng., 1: 412-417.
7. Walther, G., 2011. Fitting into morphological structure: Accounting for Sorani Kurdish endoclitics. Proceedings of the 8th Mediterranean Morphology Meeting, September 14-17, 2011, Cagliari, Italy, pp: 299-321.
8. Samvelian, P., 2007. A lexical account of Sorani Kurdish prepositions. Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, July 20-22, 2007, Stanford, CA., pp: 235-249.
9. Mohammed, F.S., L. Zakaria, N. Omar and M.Y. Albared, 2012. Automatic Kurdish Sorani text categorization using N-gram based model. Proceedings of the International Conference on Computer and Information Science, Volume 1, June 12-14, 2012, IEEE, pp: 392-395.
10. Al-Harbi, S., A. Almuhareb, A. Al-Thubaity, M.S. Khorsheed and A. Al-Rajeh, 2008. Automatic Arabic text classification. Proceedings of the 9th International Conference on Textual Data statistical Analysis, March 12-14, 2008, Lyon, pp: 77-83.
11. Zhang, W., T. Yoshida and X. Tang, 2011. A comparative study of TF*IDF, LSI and multi-words for text classification Expert Syst. Applic., 38: 2758-2765.
12. Odeh, A., A. Abu-Errub, Q. Shambour and N. Turab, 2014. Arabic text categorization algorithm using vector evaluation method. Int. J. Comput. Sci. Inform. Technol., 6: 83-92.
13. Robertson, S., 2004. Understanding inverse document frequency: On theoretical arguments for IDF. J. Documentation, 60: 503-520.

14. Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.
15. Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20: 273-297.
16. Last, M., A. Markov and A. Kandel, 2008. Multi-Lingual Detection of Web Terrorist Content. In: Intelligence and Security Informatics, Chen, H. and C.C. Yang (Eds.). Springer, New York, pp: 79-96.
17. Costa, E.P., A.C. Lorena, A.C.P.L.F. Carvalho and A.A. Freitas, 2007. A review of performance evaluation measures for hierarchical classifiers. Evaluation Methods for Machine Learning II: Papers from the AAAI-2007 Workshop, AAAI Technical Report WS-07-05, 2007.
18. Verbraken, T., W. Verbeke and B. Baesens, 2013. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Trans. Knowl. Data Eng., 25: 961-973.
19. Manicka, R. and C. Kanakalakshmi, 2015. Performance evaluation of machine learning techniques for text classification. Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, March 27, 2015, IJANA, pp: 53-57.
20. Van Rijsbergen, 1979. Information Retrieval. Butter-Worths, London, UK.