Scientific Research Publishing

Journal of Software Engineering and Applications, 2018, 11, 167-179
http://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

Automatic Arabic Document Classification Based on the HRWiTD Algorithm

Ehsan Othman, Ayoub Al-Hamadi

Faculty of Electrical Engineering and Information Technology, NIT, Otto-von-Guericke-University, Magdeburg, Germany
Email: ehsan.othman@ovgu.de, Ayoub.Al-Hamadi@ovgu.de

How to cite this paper: Othman, E. and Al-Hamadi, A. (2018) Automatic Arabic Document Classification Based on the HRWiTD Algorithm. *Journal of Software Engineering and Applications*, **11**, 167-179.
https://doi.org/10.4236/jsea.2018.114011

Received: February 23, 2018
Accepted: April 25, 2018
Published: April 28, 2018

Copyright © 2018 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

The documents contain a large amount of valuable knowledge on various subjects and, more recently, documents on the Internet are available from various sources. Therefore, automatic, rapid and accurate classification of these documents with less human interaction has become necessary. In this paper, we introduce a new algorithm called the highest repetition of words in a text document (HRWiTD) to classify the automatic Arabic text. The corpus is divided into a train set and a test set to be applied to proposed classification technique. The train set is analyzed for learning and the learning data is stored in the Learning Dataset file. The category that contains the highest repetition for each word is assigned as a category for the word in Learning Dataset file. This file includes non-duplicate words with the value of higher repetition and categories and they get from all texts in the train set. For each text in the test set, the category of words is assigned to a specific category by using Learning Dataset file. The category that contains the largest number of words is assigned as the predicted category of the text. To evaluate the classification accuracy of the HRWiTD algorithm, the confusion matrix method is used. The HRWiTD algorithm has been applied to convergent samples from six categories of Arabic news at SPA (Saudi Press Agency). As a result, the accuracy of the HRWiTD algorithm is 86.84%. In addition, we used the same corpus with the most popular machine learning algorithms which are C5.0, KNN, SVM, NB and C4.5, and their results of classification accuracy are 52.86%, 52.38%, 51.90%, 51.90% and 30%, respectively. Thus, the HRWiTD algorithm gives better classification accuracy compared to the most popular machine learning algorithms on the selected domain.

Keywords

Automatic Text Classification, Confusion Matrix, SPA, Machine Learning Algorithms

1. Introduction

The internet is a very effective technique for obtaining a huge amount of information in different forms such as documents. Recently, there are millions of documents from various sources, most of which contain valuable information. Manual classification of documents consumes time and is very difficult, especially when people must estimate the category based on the information included. Therefore, the automatic text classification is used to discover the basic information of text documents automatically while saving human effort and time [1].

Automatic text categorization is assigning and categorizing texts by using a set of predetermined categories based on the contents of the text. Specifically, it is filtering and routing, clustering information in related texts, and then classifying the texts into specified topics [2]. The text classification process is divided into three main phases. First, compile training data. Second, select a set of features to represent the texts categories. Third, test testing data with selected machine learning algorithm [3]. The concept of machine learning (ML) refers to automatic methods of learning automatically without human intervention to make predictions accurate or behave intelligently. Text classification (TC) is one of the important areas in ML. TC is a method in data mining field; it is set categories of texts in a web page, book library, media articles, gallery etc. Predetermined categories are based on their content and then give valuable information from a large unstructured text resource such as email filtering (spam or legitimate) [4].

The classification of Arabic texts has received great attention in many recent researches based on the importance of the Arabic language and the huge population who speak Arabic. In this paper, we introduce the HRWiTD algorithm used to automatically analyze Arabic texts to estimate classifications (categories). The proposed algorithm abbreviation refers to highest repetition of words in a text document. The proposed algorithm abbreviation refers to highest repetition of words in a text document. The proposed technique for classifying text is built based on three main stages, pre-processing stage to remove noisy data. Feature extraction stage to learn dataset and build Learning Dataset file based on the extracted features from the train set. Learning Dataset file includes non-duplicate words with its highest repetition values and categories. Classification stage is estimating the classification of texts by using HRWiTD algorithm (the expected classification of the text is the category with the largest number of words). If the average of total repetition for all words in a text (that contains a predetermined classification (categories)) is less than 33.33%, the proposed classification of text sets is "General" category.

The HRWiTD algorithm has been applied to convergent samples of six categories namely culture, economic, public, political, social, and sports to obtain the best classification accuracy. The selected corpus has got from SPA (Saudi Press Agency), it contains 1421 Arabic texts (Newswire), it was divided into two sets, 70% train set and 30% test set and this division is the best to get the best classifi-

cation accuracy based on [5]. The train set is analyzed to obtain predetermined categories for each word in all texts and then constructs the Learning Dataset file that will use to predict the categories of test set, then the classification of each text in the test set will be classified based on the learning process [6].

Based on recent research, various automated learning algorithms have been successfully applied to Arabic text. The most famous techniques to classify Arabic text from the best to the worst are C5.0 classifier, Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (C4.5), and K-nearest neighbor (KNN) [5]. These classification techniques recognize as simple and efficient methods for classifying texts [7] [8]. In this research, these techniques did not perform satisfactory performance in accuracy, and the best average accuracy from all categories is 52.86% using the C5.0 classifier. On the other hand, the HRWiTD algorithm achieved the best performance of the text classification and obtained the highest average accuracy (86.84%) compared to those techniques.

The second section presents some of the relevant work, the third section introduces the proposed work including the HRWiTD algorithm and the evaluation method used in the details, the fourth section presents the experimental results of the proposed algorithm and the most popular machine learning algorithms with their comparison, the latter part is the conclusion

2. Related Work

Text classification (TC) in data mining field is the process of extracting useful knowledge from text by analyzing complex and textual data [1]. The TC process is the automatic classification of a set of texts in categories based on content [9].

In many text mining algorithms, pre-processing is one of the main components of text classification. Typically, the TC framework begins with the pre-processing, then the extraction feature, and finally the classification steps [10]. In detail, the process of classifying texts is divided into nine steps. These steps in the order are 1) Data collection, 2) Word processing to remove noisy data, 3) Data segmentation into the train set and test set, 4) Extraction features to extract and generate the repetition list of data set features, 5) Feature selection based on 10 feature from selection methods [term frequency (TF), document frequency (DF), information gain (IG), CHI squared (CHI), NG, Goh and Low (NGL) coefficient, Darmstadt indexing approach (DIA) association factor, mutual information (MI), odds ratio (Odds), the Galavotti, Sebastiani, Simi (GSS) a coefficient and relevancy score (RS)] and seven weighting methods (Boolean, frequency, relative frequency, TFiDF, TFC, LTC and entropy), 6) Features representation, 7) Machine learning, 8) Applying a classification model, and 9) Performance evaluation [5].

The automatic text classification is used to classify texts in many languages such as Arabic. Arabic is the native language of more than 300 million people and is widely spread in the world [2].

Recently, many types of research have been published in machine learning al-

gorithms for the classification of Arabic text. Naïve Bayes is used to automatically classify Arabic documents in El-Kourdi *et al.* [2]. Sawaf *et al.* [11] used a statistical approach based on the Maximum Entropy to classify and cluster news articles. Sawaf *et al.* also described a method based on Association Rules to classify Arabic documents [3]. Al-Harbi *et al.* [12] compared the SVM algorithm and the Decision Tree algorithm. Al-Kabi, and Al-Sinjilawi [13] compared the classification of Arabic documents in Vector Space Model and Naïve Bayesian. Khreisat [14] compared KNN and SVM algorithms. Kanaan *et al.* [15] used Naïve Bayesian classifier to classify Arabic texts and distributed equally into many categories.

Different Machine learning algorithms that are applied to Arabic texts have produced the different classification accuracy that is presented in [5]. The most popular machine learning algorithms for classifying Arabic documents based on the most frequent selection methods (CHI, TF, DF, IG and None) are C5.0, SVM, NB, C4.5 and KNN, respectively [2] [16] [17].

3. Proposed Work

In this paper, there are three main phases to classify Arabic texts, pre-processing, feature extraction and classification. In the pre-processing stage, the selection feature is used to remove noisy data such as numbers, punctuations, kashida, stop words and diacritics [18]; in the feature extraction stage, features are then identified when learning the train set, and then building a Learning Dataset file. This file includes unduplicated words with the highest repetition values and categories, and these words are not repeated (just keep the word and category of the category of the highest repetition). In the classification stage, the classification of each text in the test set by using HRWiTD algorithm is based on matching the words of each text with the words in the Learning Dataset file to obtain a prediction classification (category) for each word. Typically, when more than two thirds (66.67%) of words with undefined categories are found in the text, the classification for this text is ambiguous and it is difficult to determine a particular classification. In fact, the "General" category includes all type of texts, some of which may belong to a specific category and some may belong to an unspecified category. Therefore, the best-predetermined classification of ambiguous text is "General" classification. In the suggested approach, if the average of the total of the repetition for all words in a text containing a predetermined classification (category) is greater than third (33.33%), the expected classification of the text is the category with the largest number of words. Otherwise, the proposed classification will be "General".

The accuracy of using the HRWiTD algorithm for classifying is evaluated through the confusion matrix. This method evaluates the predicted classification of the texts with the actual classification (from six categories) in the Arabic news (SPA).

This section describes the main stages of classification of Arabic texts in details. **Figure 1** shows the stages which include data collection, documents

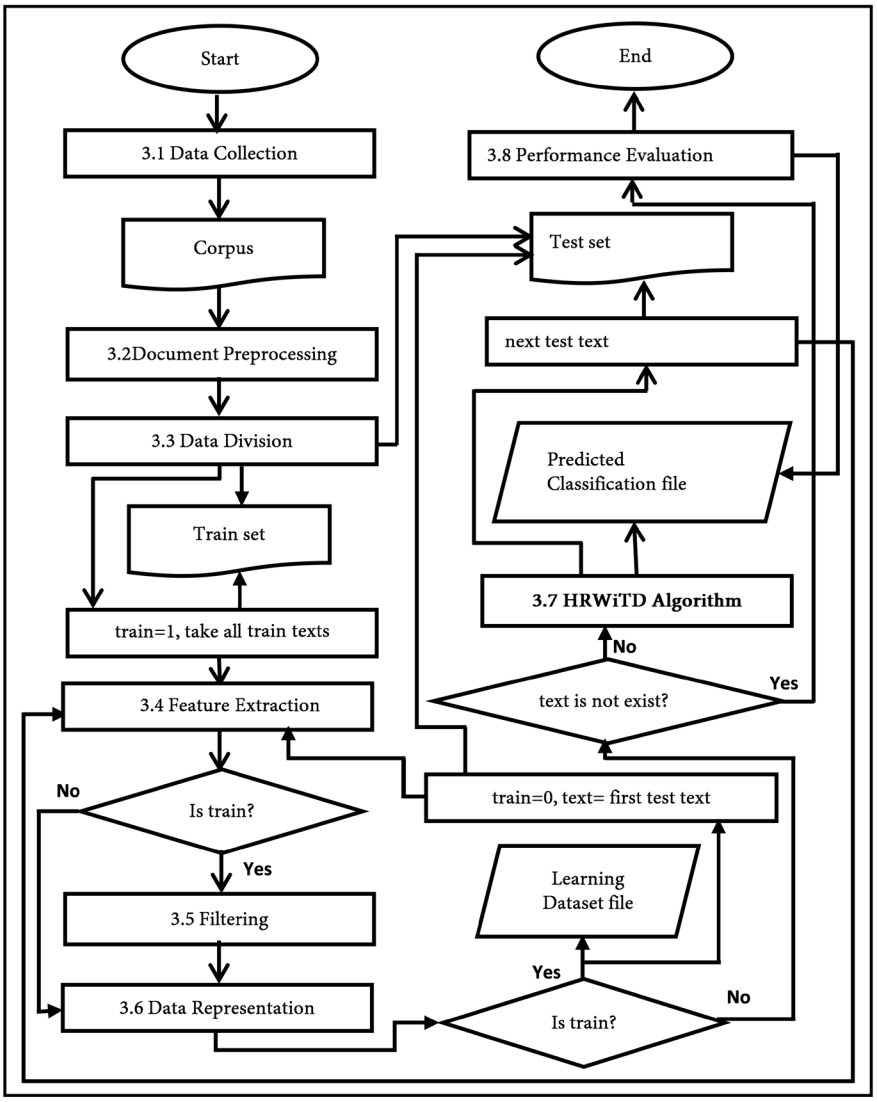

Figure 1. Arabic text classification stages.

processing, data division, feature extraction from train set, filtering, feature extraction, data representation, applying a HRWiTD algorithm, and performance evaluation.

3.1. Data Collection

Data collection is the first and very important stage for the classification of Arabic texts. We chose an Arab source (Newswire) from the Saudi Press Agency (Saudi Press Agency), which includes convergent samples of six categories. We choose a SPA source for two reasons: availability of actual classification (category) for each text in corpus and availability of SPA texts on the Web. SPA statistics are shown in Table 1.

3.2. Documents Preprocessing

The process of pre-processing is actually a process of improving the classification

of text documents by removing the data that is worthless. The data may include worthless numbers, punctuations, kashida, Hamza ",," diacritics, and stop words. Some words do not belong to any classification such as prepositions, pronouns, etc., so we append them to a stop word list see Table 2. Preprocessing also normalize text documents by changing TaaMarboutah "ة" to "ا". ATC Tool is used to remove worthless data from the selective corpus.

3.3. Data Division

At this stage, ATC Tool is used to dividing corpus into two partitions, the train set, and the test set. The train set contains 70% of a selected corpus and a test set

Table 1. SPA statistic of selected corpus.

Source	Classes	No. of Texts	No. of Words	No. of Unique Words
Saudi Press Agency	Cultural	251	47,499	9993
	Economic	248	40,065	7780
	General	171	32,395	8592
	Political	250	35,350	7430
	Social	251	49,615	10,124
	Sports	250	41,657	7332
Total	**6**	**1421**	**246,581**	**51,251**

Table 2. Removable stop words.

Classes	No. of Texts
Demonstrative pronouns	هذا, هذه, ذلك, تلك, هذان, هذين, هتان, هتين, هؤلاء, أولائك, ...
Relative pronouns	الذي, التي, اللذا, اللذين, اللتان, اللتين, الذين, اللاتي, اللواتي, ...
Subject pronouns	انا, انتَ, انتِ, هو, هي, نحن, أنتما, هما, نحن, أنتم, أنتن, هم, هن, ...
possessive pronouns	عند, مع, ل, لها, ...
Numbers	واحد, اثنين ,ثلاثه , ...
Special converters to accusative	كان وأخواتها, إنّ وأخواتها, ظنّ وأخواتها
Prepositions of time	صباح, ظُهر, ساعة, سنة , أمس, ... حين,
Prepositions of place	تحت, أمام, وراء, حيث , دون, ... فوق,
Prepositions	الواو, الفاء, ثم, حتى, أَو, أَم, بل, لا, لكن, ...
Conjunction	من, عن, على, في, الباء, إلى, اللام, الكاف, حتى, رُبَّ, مذ, منذ, التاء, الواو, ...
Countries & Cities	اليمن, امريكا, المانيا,..... & صنعاء, نيورك , برلين, ...
Proper Noun	احمد, علي , خالد, محمد, عمر, عبدالله , ...
Nationalities	يمني, امريكي, الماني,....
Others	بنت ,بن , ابن, ام, اب, اخ' اخت, جد, جده, حفيد, حفيدة, عم, عمه, خال, خالة, اليوم, غدا

(Suffix or prefix of singular/dual/plural/feminine/masculine with any Stop Words mentioned above in this table) or (Article with any Stop Words mentioned also above in this table).

contains 30%, and this division is best for the best performance of the classification based on [5]. The user can manually select the percentage of the train set and the test set.

3.4. Feature Extraction

In this stage, we use data from train set and test set from internal or external source. Features extract and the repetition list of words generates by using the ATC tool. The ATC tool lists and saves the repetitions of each word in all texts of the train set in a train list file. It also lists and saves the repetitions of each word in all texts of the test set in a test list file. In addition, add a field to train the list file and the test list file to label the category of each word. The category of words in the train list is the actual category. On the other hands, the word categories in the test list are set from the Dataset Learning file of the same words.

3.5. Filtering

At this stage, train file will filter by remove the duplication words with their classifications. The word that has the highest repetition will remain with its relative data (repetitive number and category) and delete the same words and its relative data with less repetition.

3.6. Data Representation (Train Set/Test Set)

At this stage, the train list file that is produced from the filter stage will format into Learning Dataset file. The test list file that is produced from the extract feature stage will be used for classifying text with HRWiTD algorithm. The data will be represented as an array with n rows and m columns where rows correspond to words in text and columns that correspond to repetition and category.

3.7. Classification Algorithm (HRWiTD)

In this step, the Learning Dataset file is produced from the data representation stage and the test list file will be used in the classification algorithm (HRWiTD). The test list file is used to store the predicted classification (which gets from Learning Dataset file) for all words in each text. Predicated classification file is used to store the predicted classification of all test texts. Details of the HRWiTD algorithm process are given in **Figure 2**.

3.8. Performance Evaluation

The performance of using the HRWiTD algorithm for classifying texts has been evaluated using the confusion matrix [19]. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed

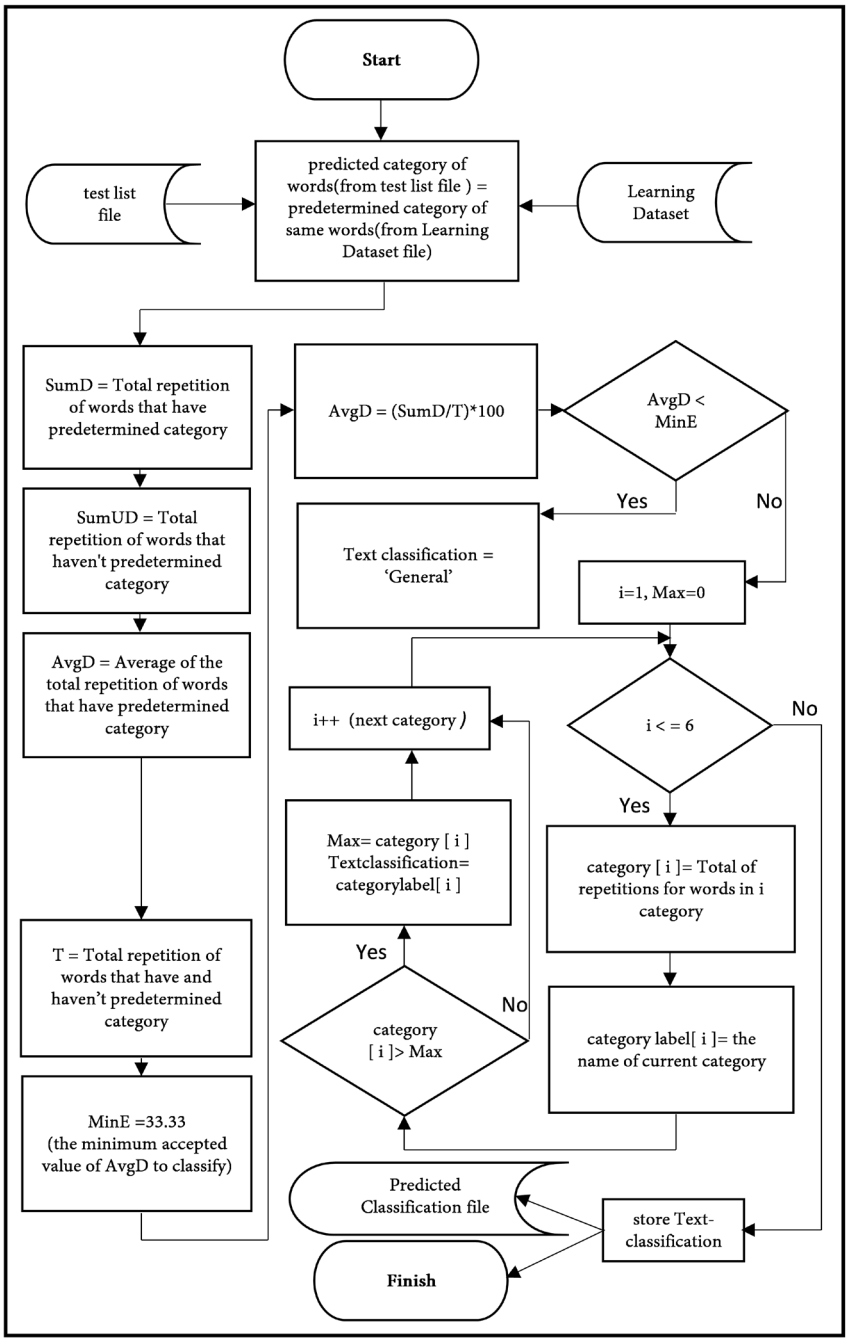

Figure 2. HRWiTD algorithm to classify Arabic texts.

from the confusion matrix. The actual and predicted information (classification) will be assigned by using HRWiTD algorithm. The confusion matrix should evaluate the performance using the actual and predicted information in the matrix, see Table 3.

Entries in the confusion matrix have the following meaning in the context of our study:

- **True negative (TN)** is the number of correct predictions that an instance is negative.

Table 3. Confusion matrix.

n = 420 (30% of Data set)		Predicted		
		Negative	Positive	
Actual	Negative	TN	FP	TN + FP
	Positive	FN	TP	FN + TP
		TN + FN	FP + TP	**Total**

- **False positive (FP)** is the number of incorrect predictions that an instance is positive.
- **False negative (FN)** is the number of incorrect of predictions that an instance negative.
- **True positive (TP)** is the number of correct predictions that an instance is positive.
- **Total** is the summation of all above variables. See Equation (1).

$$\text{Total} = \text{TN} + \text{FP} + \text{FN} + \text{TP} \tag{1}$$

Overall, the accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined by using the Equation (2):

$$\text{AC} = \left(\text{TN} + \text{TP}\right) / \text{Total} \tag{2}$$

- There are two possible predicted classifications: "Positive" and "Negative". If we were predicting the target classification (ex. "Sport") of text, for example, "Positive" would mean it belongs to that target classification, and "Negative" would mean it doesn't belong to that target classification.
- The classifier (HRWiTD algorithm) has a total of 420 (test data) out of 1421 predictions for each of six categories, including 70 text per category.
- Out of those 420 cases, the classifier predicted "Positive" FP + TP times, and "Negative" TN + FN times.
- In reality, FN + TP classification in the table is belong to target classification, and TN + FP classification do not.

4. Experimental Results and Discussion

The HRWiTD algorithm is used to classify Arabic texts. The confusion matrix method was used to determine the classification accuracy of the HRWiTD algorithm, which is 86.84% in this experiment, see Table 5 for details. On the other hand, the same data set has been applied in various famous classifier techniques, models have been developed based on using C5.0, decision tree C4.5, NB, KNN and SVM classifiers (models create by using Rapid Mine Software 5.0) [13]. We test the performance of the models on the test set and evaluate the accuracy based on the use of a Cross-validation technique and set the number of validations to X-Validation operators. The previous classifiers were evaluated based on two advanced methods for term selection: CHI square (CHI) and Information gain (IG), and different weight methods (Boolean, Entropy, Frequency, LTC,

Relative Frequency, TFC and TFiDF). Moreover, two sample methods for term selection: TF (term frequency) and DF (document frequency) were selected. The top 10, 15, 20, 25, and 30 terms for each classification in the dataset were selected as the representative terms, based on their related to TF and DF. The classification accuracy results for the classifiers are shown in Table 4.

The data in Table 4 showed the best classification accuracy is C5.0 classifier and then were KNN after that NB then SVM, C4.5 is the worse. Based on the operation of different weight methods on the data set, Boolean, Frequency, and TFiDF have shown the best weighting methods for the different classifiers that used to obtain the best classification accuracy.

See Table 4, machine learning settings for the best classification accuracy of C5.0 classifier when representation = Frequency, training size = 70% with DF, term selection = CHI square, and terms = top 30 terms of each category. The best classification accuracy of KNN classifier when representation = Frequency, training size = 70% with DF, term selection = CHI square, and terms = top 30 terms of each category. The best classification accuracy of NB classifier when representation = TFiDF, training size = 70% with DF, term selection = IG, and terms = top 30 terms of each category. The best classification accuracy of SVM classifier when representation = LTC, training size = 70% with TF, term selection = CHI square, and terms = top 30 terms of each category. The best classification accuracy of C4.5 classifier when representation = Boolean, Frequency, TFiDF, training size = 70% with TF and DF, term selection = CHI square and IG, and terms = All top terms of each category.

Table 5 shows the details of results of the best two classification techniques based on the performance, namely C5.0 classifier, and HRWiTD algorithm. The average of the accuracy for the six categories texts is calculate by use Equation (3), the accuracy for each categories are namely AC (Culture), AC (Economic), AC (General), AC (Political), AC (Social) and AC (Sport). Moreover, it shows the accuracy of using C5.0 for those six categories, the best result when using the frequency weight method, CHI method to evaluate weight selection and DF with 30 terms per category. The train set was 70% for those two classification techniques. HRWiTD algorithm has got 86.84% as total accuracy, it is better than C5.0 and other classification techniques. However, C5.0 was better than HRWiTD algorithm to classify "General" category.

Table 4. Results of the best classification accuracy for different classifier techniques.

Classifier	Accuracy (%)	Dataset
C5.0	52.86	Frequency, 70, DF, CHI, 30
KNN	52.38	Frequency, 70, DF, CHI, 30
NB	51.90	TFiDF, 70, DF, IG, 30
SVM	51.90	LTC, 70, TF, CHI, 30
C4.5	30	Boolean, Frequency, TFiDF, 70

Table 5. The best results of classification accuracy C5.0 classifier and HRWiTD algorithm.

Category	C5.0 Classifier	HRWiTD Algorithm (Train 70%)
	Frequency, 70%, CHI, DF, 30	
Culture	40.00%	85.24%
Economic	47.14%	91.43%
General	80.00%	64.58%
Political	32.86%	94.29%
Social	35.71%	86.19%
Sport	81.43%	99.29%
Total	**52.86%**	**86.84%**

$$\text{Total} = \left[\left(\text{AC}(\text{Culture}) + \text{AC}(\text{Economic}) + \text{AC}(\text{General}) + \text{AC}(\text{Political}) + \text{AC}(\text{Social}) + \text{AC}(\text{Sport})\right) * 100\right] / 6 \quad (3)$$

5. Conclusion

In summary, this paper was carried out to classify Arabic texts automatically using the HRWiTD algorithm. We have applied it to 1421 Arabic Newswire from the Saudi Press Agency (SPA). The corpus includes convergent samples of six categories (culture, economic, public, political, social, and sports). In this paper, the average of the overall classification accuracy for six categories is 86.84 %; confusion matrix method is used to evaluate the classification accuracy. The classification technique in this paper is constructed based on three main phases which are preprocessing, features extraction and classification by using HRWiTD algorithm. The repetition for a predetermined category of each word in the text is calculated. If the average of the total of those words is less than 33.33%, the expected classification of text is "General" category; otherwise, the expected classification of text is the category with the largest number of words. We compared the accuracy of the proposed algorithm (HRWiTD) with the accuracy of the most popular techniques and the accuracy of C5.0, KNN, SVM, NB and C4.5 classifies are 52.86%, 52.38%, 51.90%, 51.90% and 30%, respectively. The best classification performance was when techniques used advanced methods for term selection (CHI, IG, None), different weight methods (Boolean, Entropy, Frequency, LTC, Relative Frequency, TFC and TFiDF), and two sample methods for term selection (TF and DF). Thus, we conclude that the best technique to classify Arabic texts in the selected domain is obtained from the HRWiTD algorithm. In addition, the HRWiTD algorithm gives the best classification accuracy for each individual classification except the "General" category. In future work, first, the HRWiTD algorithm needs to be improved to get better results to classify all text categories; here we cover only six categories and other categories were assigned general category as general. Second, it needs to extend the expe-

rimental corpus from different resources to demonstrate efficiency. In this research, we applied the proposed algorithm on 1421 texts, and there are a number of words in the texts that their categories are unknown and which can lead to a poor classification of texts. Therefore, the corpus must be much larger to get the best learning.

Acknowledgements

This research is supported by German Academic Exchange Service (DAAD).

References

[1] Al-Diabat, M. (2012) Arabic Text Categorization Using Classification Rule Mining. *Applied Mathematical Sciences*, **6**, 4033-4046.

[2] Kourdi, M.E., Bensaid, A. and Rachidi, T.-E. (2004) Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, 28 August 2004, 51-58. https://doi.org/10.3115/1621804.1621819

[3] Al-Thubaity, A., Almuhareb, A., Al-Harbi, S., Al-Rajeh, A. and Khorsheed, M. (2008) KACST Arabic Text Classification Project: Overview and Preliminary Results. *Proceedings of The 9th IBIMA Conference on Information Management in Modern Organizations*, Morocco, 1 January 2008, 1239-1244.

[4] Mohammad, A.H., Alwada'n, T. and Al-Moman, O. (2016) Arabic Text Categorization Using Support Vector Machine, Naïve Bayes and Neural Network. *GSTF Journal on Computing*, **5**, 108-115, 2016.

[5] Khorsheed, M.S. and Al-Thubaity, A.O. (2013) Comparative Evaluation of Text Classification Techniques Using a Large Diverse Arabic Dataset. *Language Resources and Evaluation*, **47**, 513-538. https://doi.org/10.1007/s10579-013-9221-8

[6] Menon, A.K. (2009) Large-Scale Support Vector Machines: Algorithms and Theory. UCSD, San Diego, 1-17.

[7] Aliwy, A.H. and Ameer, E.H.A. (2017) Comparative Study of Five Text Classification Algorithms with Their Improvements. *International Journal of Applied Engineering Research*, **12**, 4309-4319.

[8] Sharef, B., Omar, N. and Sharef, Z. (2014) An Automated Arabic Text Categorization Based on the Frequency Ratio Accumulation. *The International Arab Journal of Information Technology*, **11**, 213-221.

[9] Saad, M.K. (2010) The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification. Master Thesis, Islamic University, Gaza.

[10] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B. and Kochutet, K.A (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. https://arxiv.org/pdf/1707.02919.pdf

[11] Sawaf, H., Zaplo, J. and Ney, H. (2001) Statistical Classification Methods for Arabic News Articles. *Arabic Natural Language Processing*, *Workshop on the ACL* 2001, Toulouse, 6 July 2001.

[12] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. and Al-Rajeh, A.A. (2008) Automatic Arabic Text Classification, *The 9th International Conference on the Statistical Analysis of Textual Data*, Lyon, 12-14 March 2008, 77-83.

[13] Al-Kabi, M.N. and Sinjilawi, S. (2007) A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text. *The University of Sharjah Journal of*

Pure and Applied Sciences, **4**, 13-26.

[14] Khreisat, L. (2006) Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study. *International Conference on Data Mining*, Las Vegas, 26-29 June 2006, 78-82.

[15] Kanaan, G., Al-Shalabi, R. and Al-Azzam, O. (2005) Automatic Text Classification Using Naïve Bayesian Algorithm on Arabic language. *IBIMA* 2005 *Conference on the Internet & Information Technology in Modern Organization*, Cairo, 13-15 December 2005.

[16] Galathiya, A.S., Ganatra, A.P. and Bhensdadia, C.K. (2012) Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. *International Journal of Computer Science and Information Technologies*, **3**, 3427-3431.

[17] Wu, X.D., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D. and Steinberg, D. (2008) Top 10 Algorithms in Data Mining. *Knowledge and Information Systems*, **14**, 1-37. https://doi.org/10.1007/s10115-007-0114-2

[18] Veeraswamy, A., Alias, S. and Kannan, E. (2013) An Implementation of Efficient Datamining Classification Algorithm using Nbtree. *International Journal of Computer Applications*, **67**, 26-29. https://doi.org/10.5120/11448-7043

[19] Kohavi, R. and Provost, F. (1998) Glossary of terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning*, **30**, 271-274. https://doi.org/10.1023/A:1017181826899