ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse11952020.pdf

https://doi.org/10.30534/ijatcse/2020/11952020

Improved Term Frequency Inverse Document Frequency (TF-IDF) Method for Arabic Text Classification

Hamza Sulimansallam Alshuraiqi[1], Mohammad Said El-Bashir[2]
[1]Faculty of Computer Science and Information Technology
Al-al Bayt University - JordanAffiliation, Country, alshouriqi@yahoo.com
[2]Faculty of Computer Science and Information Technology
Al-al Bayt University - Jordan, mohdelb@aabu.edu.jo

ABSTRACT

In this paper, an Arabic text classification (ATC) system is proposed. The system is based on Support Vector Machine (SVM) and two improved feature selection methods that are modified versions of the traditional Term Frequency Inverse Document Frequency (TFIDF) method. The proposed ATC system performs three processes; pre-processing, feature selection, and classification. In the pre-processing phase, the input document string is tokenized into words and normalized. Then, the stop words are removed and the numbers and special symbols are deleted. Afterwards, light stemming is applied to remove the suffixes and prefixes of words. In the feature selection phase, the improved TFIDF (ITFIDF) method that was developed by Junkai et al. (2016) and the TFIDF-CF method that was developed by Mingyong et al. (2012) were customized to provide the inputs for the SVM for text classification. The suggested ATC system uses a distribution term to improve performance of the TFIDF. Performance of this proposed system was tested on a dataset of 20,230 Arabic documents. In addition, its performance was compared with performance of the traditional TFIDF method using the same dataset in terms of precision, recall, and the F-measure criteria. The values of these performance criteria obtained using the proposed ATC system and ten-fold cross validation were 87%, 85%, and 86%, respectively. A comparison in performance between the proposed system and previous systems revealed that the proposed system has better performance than other systems.

Key words : Feature Selection, SVM, Text Classification, TFIDF.

1. INTRODUCTION

Artificial Intelligence (AI) is a very important field of computer science. It has already established varying important applications such as text classification (TC), which is a process that allots a document to one predefined category or more [1][2]. This process is broadly applied for various purposes such as electronic mail (e-mail) classification, information retrieval, topic spotting, and junk e-mail filtration [3][4].

Importance of TC has grown in the recent decades because of the massive generation and accumulation of data, most of which (>80.0%) is categorized as text [5]. So far, numerous systems and algorithms have been suggested by researchers for categorization of texts. However, most of these systems and algorithms were designed for English texts and only few were intended for Arabic texts [6]. Indeed, information retrieval of Arabic language is a relatively recent practice and research in this area is limited relative to researches that have been performed on other languages such as the English, Chinese, Greek, and German languages.

The Arabic text classification (ATC) systems consist of several phases that start with pre-processing, followed by feature selection, and, then, classification [7]. Pre-processing is a crucial step in ATC that boosts performance of the sequent processes of feature extraction and text classification. It includes the processes of tokenization, normalization, cleaning, removal of stop words, and stemming. Feature selection, which is a process that is performed in the second phase of text classification, is the process of choosing the most relevant subset of features from the data that represent known categories [6]. Several methods are used for feature selection, including Term Frequency Inverse Document Frequency (TF-IDF), Chi-squared, and correlation methods [8]. The text classification phase employs different approaches for classification of texts, most common of which are the Support Vector Machine (SVM) algorithm, k-nearest neighbor (kNN) algorithm, Decision Trees (DTs), and the Naïve Bayes (NB) classifier [9]. In this research, the SVM classifier is used because evidence supports that it outperforms the NB classifier in ATC.

The present study suggests an ATC system based on the SVM and two improved feature selection methods: (i) the improved TF-IDF (ITFIDF) method, which was developed by Junkai et al. (2016) for the Chinese language; and (ii) the TF-IDF-CF method, which was developed by [10] for the English language.

2. LITERATURE REVIEW

The text classification system is a multi-stage system. In general, the classification systems differ in their mechanisms but they usually proceed in five steps, namely, document collection, pre-processing, feature selection, classification, and performance evaluation [5].

The first step in text classification is document collection. Documents can be collected from various sources [5]. Usually, those documents can be in any of a variety of file formats like txt, pdf, html, and doc.

Pre-processing of documents/texts is a highly-critical step in ATC, particularly since it influences accuracy of feature extraction and the eventual text classification. In ATC, pre-processing often includes four processes: tokenization, normalization, elimination of stop words, and stemming [11].

After pre-processing of the text, the main problems to deal with are the high dimensionality of the data (large number of words) and identification of the words which best represent the categories. The feature selection process aims at selecting the important tokens or features from the document [5]. Selecting a subset of features will solve the high-dimensionality problem while selecting the best feature will improve the accuracy of text classification.

Many algorithms have been developed to calculate the weights of features, including the Document Frequency (DF) and Term Frequency (TF) algorithms, in addition to their integration, i.e., the Term Frequency Inverse Document Frequency (TF-IDF) algorithm. The TF-IDF is a feature selection algorithm which is widely used [12]. Some researchers (e.g., [10] and [12]) identified drawbacks for this algorithm and improved it. Two of the improved versions of this algorithm, that is, the ITFIDF and TF-IDF-CF algorithms, were applied on Chinese and English texts and they proved to be having better performance than the traditional TF-IDF algorithm. In the present study, these two algorithms were applied to Arabic text for classification purposes.

In the classification step, the documents are classified according to the nearest category. Various classification methods are already known, including the SVMs, the Artificial Neural Network (ANN), the Bayesian classifier, the DTs, and the kNN algorithm. The SVMs proved to be effective in dealing with high-dimensional data since classification is based on the number of support vectorsrather than by dimensionality of the data. Accordingly, a SVM with low number of support vectors can support good generalization, even if the data dimensionality is high [13],[14],[15],[16].

Review of the literature brings to light that, so far, various studies have been performed on Arabic text classification. For instance, El-Kourdi implemented automatic classification of Arabic texts by using the NB classifier. Experimentation disclosed that this classifier had a maximal classification accuracy of 92.8% and a mean classification accuracy of 68.78% [17].

Al-Shalabi et al. applied the kNN algorithm on Arabic text and employed the TF-IDF method as the weighting scheme. They compiled their text data from many newspapers. The results indicated recall and micro-average scores of 0.95 [18].

Saad examined influence of pre-processing and term weighting in a TC system for the Arabic language. The researcher used three common classification algorithms, namely, the DTs, SVMs, and kNN. The results disclosed that term pruning and light stemming performed better than other examined feature reduction approaches and that the NB and SVM performed better than the other investigated classification algorithms [11].

Alsaleem compared performance of the NB algorithm with that of the SVM in classification of Arabic texts. They employed dataset compiled from Saudi newspapers that consisted of 5,121 Arabic documents classified into seven classes. Their results indicated that the SVM outperforms the NB algorithm in ATC [19].

Duwairi compared levels of performance of four statistical feature selection methods by using the NB classifier. The four studied feature selection methods were the Chi-squared, correlation, uncertainty, and deviation methods. The study found that the correlation method gave the best accuracy and that it needed a short running time for selection of features, but the highest accuracy was associated with the case when no filter was employed. However, in this case the classifier required double the time needed by the correlation method [8].

Hawashin et al. developed an efficient feature selection approach that is based on the Chi-squared statistic and which outperformed various known approaches to feature selection on the basis of its influence on classification of Arabic texts. The proposed feature selection method outpaced the approaches of Information Gain, MeanTF-IDFand Chi-squaredwith SVM and Best Search for Feature Subset Selection [20].

Abu-Errub designed dual-stage approach to ATC; a stage of categorization that employs TF-IDF measurements and a stage of classification that is founded on the normal Chi-squared technique. This proposed approach was tested on 1,090 Arabic documents that were classified into 10 major categories and 50 sub-categories. The testing results gave evidence on that this algorithm could classify the documents into their most suitable sub-categories [21].

Fodil et al. suggested two methods in the feature extraction step of text classification: (i) the Semi-Automatic Classification Method (SACM) and (ii) the Automatic Classification Method (ACM). The results pointed out that the SACM using the TF-IDF had the best performance (score of 95.0%). The methods which used the relative frequency approach came in the last place, with no higher scores than 88.0% [4].

Odeh et al. developed a new method that identifies the principal words in documents by weighting each word using the TF-IDF method. Then, the two words of the highest weights are compared with key words in the testing categories. The study employed a set of Arabic documents that consists of 38,081 words. After deletion of the recurrent words, almost 8,112 words were left. The results indicated that the suggested method can categorize the Arabic text documents into suitable categories with a high precision [22].

3. THE PROPOSED METHOD

This research aimed at developing an ATC system based on the SVM and two improved TF-IDF feature selection methods. This section illustrates the procedure followed to develop this proposed system. It gives details on the dataset used in the study, the various steps of document pre-processing, and the measures adopted for evaluation of classification performance. Results of this research are given and discussed in the next section.

3.1 The Arabic Document DatasetFinal Stage

In this study, the researchers used the Arabic document dataset presented in Al Watan 2004 [23]. It consists of 20,230 Arabic documents that belong to six categories: culture, economy, international news, local news, religion, and sports. The number of documents in every one of these categories is listed in Table 1.

Table 1: Number of documents in each of the six categories of the dataset used in the current study.

Category	# of Documents	Average number of words in each category
Sports	4540	323
Religion	3860	838
Local news	3560	441
International news	2030	431
Economy	3460	433
Culture	2780	518

3.2 The Proposed Arabic Text Classification System

In this study, the researchers propose an ATC system that is based on the SVM classifier and two improved TF-IDF feature selection methods. This system is a multi-phase classification system that performs classification by means of three major processes: document pre-processing, feature selection, and text classification (Figure 1). These processes are illustrated in the following sections.

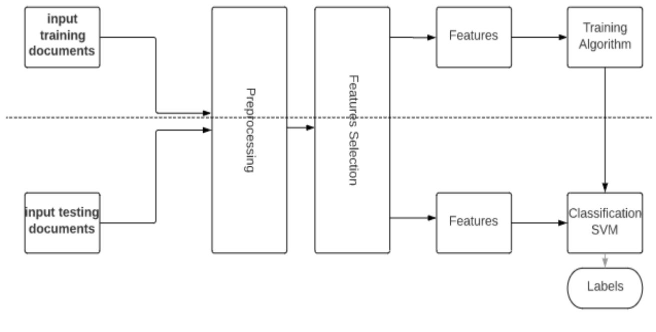

Figure 1: The text classification process

3.3 Pre-processing

In comparison with texts of other languages like the Chinese and the English languages, the Arabic texts are characterized by very sophisticated morphology [24]. They are often written cursively and they include diacritics that give different meanings to the same word, depending on their nature and positions on it.

Text pre-processing is a critical step in the ATC process, mainly because it has direct implications on performance of the later processes of feature extraction and text classification. In the current study, five steps define text pre-processing (Figure 2): stemming, tokenization, normalization, stop word removal, and cleaning.

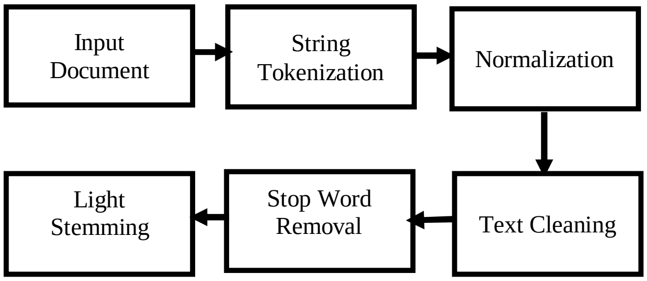

Figure 2: The text pre-processing steps in the proposed ATC system

In text pre-processing, strings of the document under processing are tokenized into words and normalized. Thereafter, all stop words are eliminated and any unnoticed numbers, words, and special symbols are deleted. Then, light stemming is applied so as to remove prefixes and suffixes from letters.

3.4 Arabic Text Feature Selection

Another important process in text classification is feature selection. Its main goal is selection of a sub-set of features from the original, input text documents [25]. In this respect, the present study employed two feature selection methods (the TF-IDF-CF and ITFIDF methods) in the proposed ATC system in order to lower dimensionality of the feature space of the data and enhance the efficiency and accuracy of the classifiers.

3.4.1 The Traditional TF-IDF and TF-IDF-CF Methods

The traditional TF-IDF method is frequently used to reduce dimensionality of the input space of the text document by selecting a subset of the text features that leads to an improved classification. The traditional TF-IDF neglects the distribution of the features and gives the same weight to both terms, which is considered as a shortcoming in this algorithm.

The first method developed to improve the traditional TF-IDF in terms of dispersion of the features in the same category was the TF-IDF-CF method [10], which improved the traditional TF-IDF by adding new parameter to represent the in-class characteristics.

3.4.2 The Improved TF-IDF (ITFIDF) Method

The second improvement to the TF-IDF method was suggested by Junkai et al. [12]. It calculates the weights according to the distribution in the same category and the distribution between categories.

3.4.2.1 Inter-Category Dispersion

The inter-category dispersion refers to distribution of the feature term among several categories. Inter-category dispersion has large values when terms appear in a single category or few categories. Based on the standard deviation formula [12], inter-category dispersion is definedin Equation (1) as follows [12]:

$$D(t)\begin{cases}\dfrac{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(f_i(t)-\overline{f(t)}\right)^2}}{\overline{f(t)}}, & \overline{f(t)}\neq 0\\ 0, & \overline{f(t)}=0\end{cases}\quad (1)$$

Where

f$_i$(t): number of documents which include the term *t* in the category C$_i$.

$\overline{f\ (t)}$: mean frequency of occurrence of a feature term in all categories.

m: total number of categories.

3.4.2.2 Intra-Category Information Entropy

The intra-category information entropy (ICIE) expresses extent of importance of a term in every category. It is commonly employed to describe distribution of a feature term in a given category based on definition of the information entropy. The terms with the highest spread are the terms most suitable for representing the category [12]. The ICIE may be calculated by using Equation(2):

$$E(t,C_i)=-\sum_{j=1}^{n}e_i \quad (2)$$

The term e$_j$is definedin Equation (3) as follows:

$$e_j = \begin{cases}\dfrac{Nd_j}{NC_i}\log_2\dfrac{Nd_j}{NC_i}, & NC_i\neq 0 \ and\ Nd_j\neq 0\\ 0, & NC_i=0\ or\ Nd_j=0\end{cases}\quad (3)$$

Where

n: number of documents in the category*C$_i$*.

N$_{dj}$: number of feature term*t* in the *j*th document of the category *C$_i$*.

NC$_i$: total number of feature term *t* in all documents of the category *C$_i$*.

3.4.2.3 Category Discrimination

Analysis indicates that the higher the inter-category dispersion or intra-category dispersion information entropy, the higher the power for category distinction in Equation (4):

$$CD(t,C_i)=D(t)*E(t,C_i) \quad (4)$$

In conclusion, Equation (5) improves the TF-IDF:

$$w_{ij}(t) = tf_{ij}*CD(t_j, C_i) \quad (5)$$

Where

tf$_{ij}$: frequency of the term *t*$_j$.

CD(*t$_j$*, *C$_i$*): Category discrimination of the feature term *t* to category *C*$_i$ (computed by using Equation (4)).

In the improved TF-IDF algorithm, the inverse document frequency (IDF) method is substituted with a category discrimination CD(t$_j$, C$_i$) term. This overcomes the shortcoming of the traditional TF-IDF algorithm, which does not take into consideration the inter-category and intra-category distributions of the feature term.

Through comparison, it was found that the weight results of the improved TF-IDF algorithm reflect the actual distribution of the feature term and are highly close to reality. Hence, we conclude that the category discrimination term can help in vanquishing the shortcoming of the classical TF-IDF algorithm.

3.5 Arabic Text Classification using Support Vector Machines

In the ATC system proposed here, which is based on the SVM, the Arabic text document is first prepared (pre-processed) by removing from it unnoticed words, stop words, and any special symbols. Thereafter, light stemming is implemented so as to remove any prefixes and suffixes. Afterwards, features are carefully selected and extracted using the ITFIDF and TF-IDF-CF methods. Then, the extracted feature vectors are classified by supervised SVM.

The researcher chose the SVM owing to that it is one of the best supervised machine learning algorithms. It is widely employed in text classification in general and in ATC in particular. It was applied to TC [26] and it gave better results in terms of the classification accuracy than other machine learning methods like the NB classifier, the kNN algorithm, and the DTs [19]. As well, it is of wide use by virtue of its inherent advantages, which mainly include [24]:

1. It is powerful in high-dimensional spaces. Over-fitting has limited influence on the calculations of the eventual decision margin.
2. Each and every feature is quite important. Even the features which may be counted as inappropriate have been found to be suitable when computing the decision margin.
3. It is robust against small sample sizes.

In the ATC system suggested in this study, training was performed by using the SVM and a dataset of 20,230 Arabic documents that are classified into six categories: international news, culture, local news, sports, religion, and economy. The results of classification of these documents that were obtained by using the proposed ATC system are outlined and discussed in the following section.

4. Results and Discussion

In this section, we discuss effectiveness of the proposed ATC system, which is based on TF-IDF and the ITFIDF and TF-IDF-CF methods. A comparison of performance was made between the proposed feature selection methods and the normal TF-IDF method.

Performance of the herein developed ATC system was assessed and compared with performance of the classical TF-IDF method using the same dataset, text processing processes, and SVM classifier. The adopted performance evaluation criteria were the recall, precision, and F1-measure. Definitions of these measures are provided next.

Consider documents in the testing sub-set that belong to category A. The classifier predicts category for every document. The various predictions group into any of four categories:True positive (TP), True negative (TN), False positive (FP)and False negative (FN) which are used to calculate the following.

Precision

Precision is the proportion of documents that were correctly predicted to be belonging to the right category. It is calculated using Equation (6) [27]:

$$\text{Precision} = \frac{TP}{TP+FP} (6)$$

Recall

Recall is the proportion of actual category A documents that were correctly predicted [27]. It can be estimated by using Equation (7):

$$\text{Recall} = \frac{TP}{TP+FN} (7)$$

The F1-measure

The F1 measure is the harmonic mean of the recall and precision (p) [27]. It is calculated according to Equation (8):

$$\text{Precision} = 2 * \frac{recall*precision}{recall+precision} (8)$$

4.1 Results of the Proposed Classification System

The classification results of the proposed ATC system that were obtained using the improved feature selection methods were compared with the results produced by the traditional TF-IDF method for the same dataset under similar processing conditions. The results uncover the all three methods have comparable precision values, even though the ITFIDF method produced slightly better results than the other two methods. The resultant precision values are 0.87 for the ITFIDF, 0.86 for the TF-IDF-CF, and 0.86 for the traditional TF-IDF (Figure 3). Thus, the ITFIDF method has nearly 0.01 higher precision than the TF-IDF-CF and the traditional TF-IDF methods. This suggests that the ITFIDF method reduces the number of incorrectly classified documents.

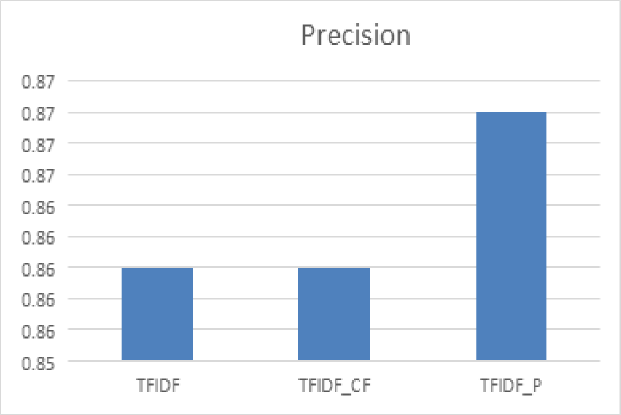

Figure 3: Values of precision of the ITFIDF, TF-IDF-CF, and traditional TF-IDF methods.

The values of recall for the ITFIDF, the TF-IDF-CF, and the traditional TF-IDF are 0.85, 0.82, and 0.78, respectively (Figure 4). These results confirm that the improved feature selection methods led to an enhancement in the classification performance of the proposed ATC system and that the ITFIDF performs slightly better than the TF-IDF-CF. As Figure 4 shows, the ITFIDF method has approximately 7.0% higher recall values than the traditional TF-IDF method. This means that the ITFIDF method boosts the number of documents that are allotted to their correct classes.

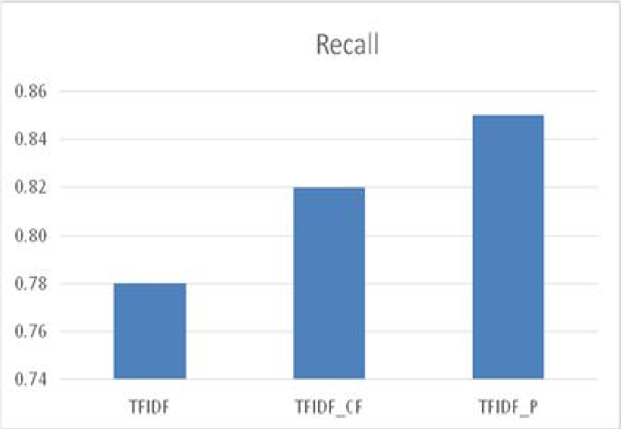

Figure 4: Values of recall of the ITFIDF, TF-IDF-CF, and traditional TF-IDF methods

The values of the F1-measure for the ITFIDF, TF-IDF-CF, and traditional TF-IDF are 0.85, 0.82, and 0.78, respectively (Figure 5). These outcomes disclose that the improved methods are superior to the traditional TF-IDF method. In general, using the distribution term (CD(tj, Ci)) in the same class by the TF-IDF-CF method led to enhancement in the classification accuracy over that of the traditional TF-IDF method. However, using the distribution in the same class and between classes brought about better results than those obtained from the ITFIDF method.

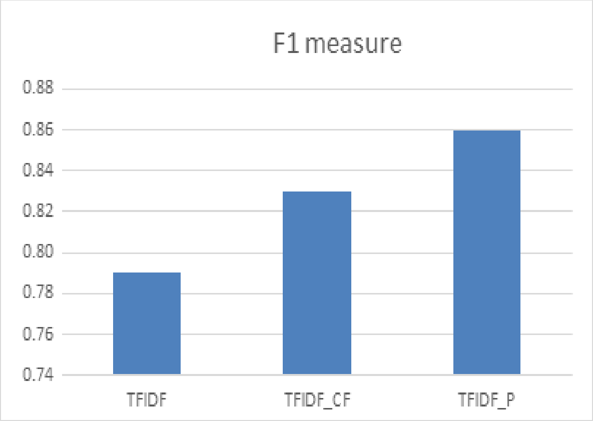

Figure 5: Values of the F1-measure of ITFIDF, TF-IDF-CF, and traditional TF-IDF methods

A potential reason why the proposed, improved ATC system did not give better than expected classification results is that the dataset employed in this research includes a 'local news' category. This category includes numerous popular words that cause vagueness. Consequently, this method had the lowest values of recall, precision, and the F1-measure.

4.2 Comparison with other Classification Systems

To assess the comparative performance of the proposed text classification system, its classification results were compared with results of previous studies conducted under the same conditions. Hawashin et al. used Alwatan 2004 dataset, but without the culture category. The number of all documents used in their study was 5,250. These documents were split into a training sub-set of 250 documents (50 documents for every single category) and a testing sub-set of 5,000 documents (1,000 documents for every one category). The value of the F1-measure associated with the traditional TF-IDF method and the SVM classifier is 0.77 whilst the value obtained by Hawashin et al. is 0.6 (Figure 6). The difference between these values may be related to difference between these two studies in pre-processing of the input documents [20].

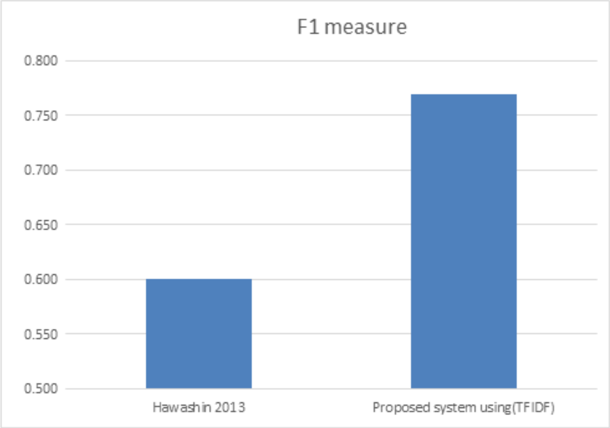

Figure 6: Values of the F1-measure produced by the proposed ATC system using the TF-IDF method and that produced by the system of Hawashin (2013)

Elsehemy et al. applied the traditional TF-IDF method and the SVM classifier on Alwatan 2004 dataset (20,291 documents) with 10-fold cross-validation. The precision values obtained are presented in Figure 7. It is noticed that the herein proposed ATC system has higher precision than the system developed by Elsehemy et al. [23],[28].

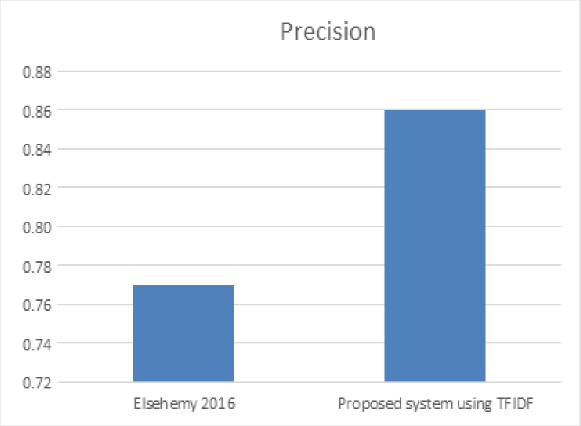

Figure 7: Values of precision produced by the proposed ATC system using the TF-IDF method and that produced by the system of Elsehemy et al

The recall values summarized by Figure 8 illustrate that both the herein proposed ATC system and that of Elsehemy et al. have similar recall values [28].

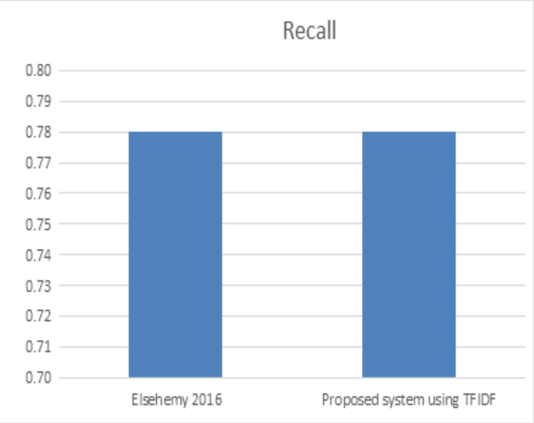

Figure 8: Values of recall produced by the proposed ATC system and the system of Elsehemy et al.

As Figure 9 shows, the value of the F1-measure of the herein proposed ATC system is larger than that associated with the system proposed by Elsehemy et al. [28].

5. CONCLUSIONS

This study proposed a multi-stage ATC system that is based on the SVM and a couple of improved TF-IDF feature selection methods. The system performs three processes in sequence: pre-processing, feature selection, and classification. In this system, the Arabic text document is first prepared by removing the stop words and any special symbols. In addition, this system uses light stemming for removing the suffixes and prefixes of words. Then, the features are selected and extracted by using the ITFIDF and TF-IDF-CF feature selection methods. Afterwards, the extracted features are classified using the SVM. These two improved feature selection methods utilize distribution of features in the same category and between the categories to compute the feature weight.

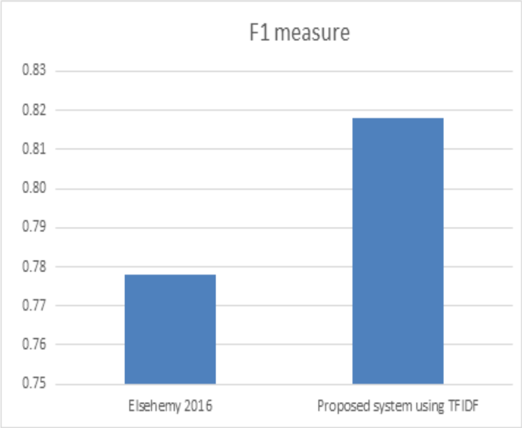

Figure 9: Values of the F1-measure produced by the proposed ATC system and the system of Elsehemy et al.

In the ATC system proposed by this study, the SVM was trained on a dataset of 20,230 Arabic documents categorized into six categories: local news, international news, economy, culture, religion, and sports.

This study reaches to the conclusion that the ATC system proposed here produces better classification results with the ITFIDF than with the traditional TF-IDF-CF and TF-IDF feature selection methods under the same document processing conditions in terms of the F-measure, recall, and precision performance measures. The values of these performance criteria that were obtained using the ITFIDF method were 87%, 85%, and 86%, respectively. Meanwhile, the values of these three measures produced by the normalTF-IDF method were 86%, 78%, and 79%, respectively. The corresponding values associated with the TF-IDF-CF method were 86%, 82%, and 83%, respectively.

REFERENCES

1. Bhumika, Prof Sukhjit Singh Sehra, and Prof Anand Nayyar. **A review paper on algorithms used for text classification.** International Journal of Application or Innovation in Engineering & Management 3.2 (2013): 90-99.
2. Maganti Syamala , N.J.Nalini. **A Deep Analysis on Aspect based Sentiment Text Classification Approaches.** International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.5, September - October 2019
3. Omaia Al-Omari , Nazlia Omari. **Enhanced Document Classification Using Noun Verb (NV) Terms Extraction Approach.** International Journal of Advanced Trends in Computer Science and Engineering,,Volume 8, No.1, January – February 2019.
4. Fodil, Leila, Halim Sayoud, and Siham Ouamour. **Theme classification of Arabic text: A statistical approach.** Terminology and Knowledge Engineering 2014.
5. Korde, Vandana, and C. Namrata Mahender. **Text classification and classifiers: A survey.** International Journal of Artificial Intelligence & Applications 3.2 (2012): 85.
6. Duwairi, Rehab M. **Statistical Feature Selection Techniques for Arabic Text Categorization.** 2013.
7. Ayadi, Rami, Mohsen Maraoui, and Mounir Zrigui. **A Survey of Arabic Text Representation and Classification Methods.** Research in Computing Science117 (2016): 51-62.
8. Duwairi, Rehab M. **Statistical Feature Selection Techniques for Arabic Text Categorization.** 2013.
9. Azara, Mohammed, Tamer Fatayer, and Alaa El-Halees. **Arabie text classification using Learning Vector Quantization.** Informatics and Systems (INFOS), 2012 8th International Conference on. IEEE, 2012.
10. Liu, Mingyoug, and Jiangang Yang. **An improvement of TFIDF weighting in text categorization.** International Proceedings of Computer Science and Information Technology (2012): 44-47.
11. Saad, Motaz K. **The impact of text preprocessing and term weighting on arabic text classification.** Diss. The Islamic University-Gaza, 2010.
12. Yi, Junkai, Guang Yang, and Jing Wan. **Category Discrimination Based Feature Selection Algorithm in Chinese Text Classification.** Journal of Information Science and Engineering 32.5 (2016): 1145-1159.
13. IH Witten, E Frank, MA Hall, CJ Pal. **Data Mining: Practical machine learning tools and techniques.** Second Edition, Morgan Kaufmann by Elsevier, 2005.
14. Duda, R. O., Hart, P. E., & Stork, D. G. & Stork, D. G. (2001). **Pattern classification.** A Wiley-Interscience Publication.
15. Han, J., (2006). **Data Mining: Concepts and Techniques**. Morgan Kaufinann.
16. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J (2008). **LIBLINEAR: A library for large linear classification.** Journal of machine learning research, 9(Aug), 1871-1874.
17. El Kourdi, Mohamed, Amine Bensaid, and Tajje-eddine Rachidi. **Automatic Arabic document categorization based on the Naïve Bayes algorithm.** Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Association for Computational Linguistics, 2004.
18. Al-Shalabi, Riyad, Ghassan Kanaan, and M. Gharaibeh. **Arabic text categorization using KNN algorithm.** The 4th International Multiconference on Computer and Information Technology, CSIT. 2006.
19. Alsaleem, Saleh. **Automated Arabic Text Categorization Using SVM and NB.** Int. Arab J. e-Technol. 2.2 (2011): 124-128.
20. Hawashin, Bilal, Ayman Mansour, and Shadi Aljawarneh. **An efficient feature selection method for Arabic text classification.** International journal of computer applications 83.17 (2013).
21. Abu-Errub A., **Arabic Text Classification Algorithm using TF-IDF and Chi Square**

Measurements. International Journal of Computer Applications 93 (6), 40-45, 2014.

22. Odeh, Ashraf, et al. **Arabic text categorization algorithm using vector evaluation method.** arXiv preprint arXiv: 1501.01318 (2015).
23. M. Abbas, K. Smaili, D. Berkani. **Comparing TR-Classifier and kNN by using Reduced Sizes of Vocabularies.** The 3rd International Conference on Arabic Language Processing, CITALA 2009, 4-5 May 2009, Mohammadia School of Engineers, Rabat, Morroco.
24. Gharib, Tarek F., Mena B. Habib, and Zaki T. Fayed. **Arabic text classification using support vector machines.** (2009).
25. Khan, Aurangzeb, Baharum B. Bahurdin, and Khairullah Khan. **An Overview of E-Documents Classification.** International Conference on Machine Learning and Computing, Singapore. 2009.
26. Joachims, Thorsten. **Transductive inference for text classification using support vector machines.** ICML. Vol. 99. 1999.
27. Chantar, H. K., & Corne, D. W. (2011, October). **Feature subset selection for Arabic document categorization using BPSO-KNN.** In Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on (pp. 546-551). IEEE.
28. Elsehemy, A., M. Abdeen, and T. Nazmy. **Enhanced Arabic Semantic Information Retrieval System Based on Arabic Text Classification.** World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering 3.1 (2016).