Journal of Theoretical and Applied Information Technology
31st July 2018. Vol.96. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FILTERING APPROACH AND SYSTEM COMBINATION FOR ARABIC NEWS CLASSIFICATION

[1] DR. ZAINAB A. KHALAF, [2] KHADEIJA A. HASSAN
[1,2] Basrah University, College of Science, Department of Computer Science, Basrah, Iraq
E-mail: [1]zainab_ali2004@yahoo.com, [2]khah1987@yahoo.com

ABSTRACT

Text classification is one of the important research in text mining applications because of the increasing growth of digital data from different resources. It became urgent need to find systems that help users to get information from this large number of texts easily and faster. One of the most significant challenges facing the text classification is how to reduce the number of features in high dimensional data spaces without reduced the performance. The objective of this research is to design an automatic text classification system and improve its performance with less number of features.

Two approaches are proposed in this paper: N-gram features filtering and system combination. The main aim of using different N-gram types is to use deep semantics and less number of features. Three classification systems based on SVM algorithm are applied with different N-gram filtering to reduce the number of features. These systems were compared with conventional classification approaches, cosine similarity and SVM.

The system combination fusion the hypotheses produced by the above classification systems in order to select the best class label via voting. These systems are applied on Arabic corpus.

The filtering approach reduced the number of features from (5,799) and (12,474) for cosine and SVM systems to roughly (3,400) by using N-gram types representation without reduce the performance. The performance of text classification is enhanced from (81.6) and (91.2) for cosine and SVM respectively to (94.2) for system combination.

Keywords: *classification , filtering, system combination, N-gram, Arabic news*

1. INTRODUCTION

One of the most important applications for text mining using machine learning tools is the text classification. Text classification is the task that assigns predefined classes to the text documents depending on their contents. It is a supervised learning because the output class label of examples known and predefined [1, 2, 3].

Manual text classification is too expensive and takes a long time. Hence, without an automatic classification system, access to text archives and searches within them would be restricted to the limited number of textual documents that have been manually classified by humans [4, 5].

In classification system, the document examples divide for two parts: training and testing dataset. The training stage involves training the classification system to establish the correct label for each document based on learning approach. The testing approach tries to assign the correct label to the unknown document based on the highest score of the similarity between the features of the predefined classes and unknown given document. This division of the training and testing datasets can be under different ways such as percentage or cross validation. Example of percentage division, which is used in this paper, is a 70:30% that means 70% of the datasets are used for training and 30% for testing [3, 6].

One of the most significant challenges facing the text classification is how to reduce the number of features in high dimensional data spaces without reduced the performance. The accuracy of the text classification algorithms can be enhanced by reducing the dimensionality of the effective features and data spaces. Researchers have addressed the problem in the text classification in many ways. One way is to find the deep semantics in order to reduce the number of the features. Another way is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to combine multiple text classification systems to improve the result by exploiting the different text classification outputs to take advantage of their potential complementarities to improve the performance [7].

The goals of this paper are: to design an automatic text classification system that reduce the number of features in high dimensional data spaces, and improve its performance with less number of features.

In this paper, two approaches are proposed. The first proposed approach is used to reduce the number of features by using filtering approach for the features (N-gram types). While the second proposed method combines multiple text classification systems that use different type of features improve the performance.

The rest of the paper is organized as follows: after showing the introduction in section 1, the related work is given in section 2. Section 3 shows the text classification system. Section 4 explains the proposed system. whilst Section 5 provides the experiments to evaluate the proposed approaches. Section 6 gives the conclusions. Finally, suggestion for future works is given in the section 7.

2. RELATED WORK

There are many researchers have been worked on the text classification in English and other European languages such as German, French and Spanish, and in Asian languages such as Japanese, Chinese and Indian. So, researches on the classification of text for Arabic language are limited.

According to previous studies in Arabic text [8], classification systems can be divided into three categories: applying new algorithms, comparing between different classification algorithms, and investigates the impact of the preprocessing stage.

The researchers [9, 10, 11] proposed news algorithms. Khreisat, 2009 [9] executed tri-gram for dimension reduction in Arabic text classification. The dataset was collected from Jordanian Arabic newspapers. Whereas, Belkebir et al. 2013 [10] used support vector machine (SVM), and neural network. The Bee Swarm Optimization (BSO) and CHI square techniques are combined with SVM. The light and root-based stemmer were used in pre-processing for Arabic dataset. While Hughes, 2017 [11] used neural networks for medical text classification.

On other hand many researchers are try to improve the preprocessing stage, such as [8, 12]. Saad, 2010 [8] used four algorithms: K-nearest neighbor (K-NN), Decision Tree (DT), support vector machine (SVM) and Naïve Bayes (NB). This study focused on the effect of preprocessing (stemming techniques) and term weighting methods on the classifiers. The dataset collected from CNN, BBC and other Arabic open websites. Elhassan et al. 2015 [12] used SVM, NB, K-NN and DT. They studied the effect of pre-processing. They used Basic Stop Words File (BSWF) that consists of 216 words, and they added Extended Stop Words File (ESWF) which contain the words that most appearances in training dataset. ESWF consists of 240 words like stop words. The dataset contained (750) Arabic documents.

Many researchers have tried different text classification system with various methods as comparative study, such as [13, 14, 6, 15].

Al-Shargai et al. 2011 [13] executed a comparative study between SVM, NB and DT on Arabic dataset. Alsaleem, 2011 [6] produced a comparative study between Naïve Bayes (NB) and support vector machine (SVM) on Arabic documents.

Hmeidi et al. 2015 [14] executed a comparative study between NB, SVM, K-NN, and DT. The research studied the effect of applying different types of Arabic stemmers (root-based and light stemmer). Trivedi et al. 2015 [15] executed a comparative study between NB, SVM and DT. They used a percentage splitting dataset for training and testing. They used two datasets: the first contained (768) instances and (9) features. The second contain (40) food items with (4) features.

3. TEXT CLASSIFICATION SYSTEM

Text classification (TC) task consists of six main steps: documents collection, pre-processing, indexing, dimension reduction, classification algorithm, and performance evaluation. As shown in figure (1) [16].

Journal of Theoretical and Applied Information Technology
31st July 2018. Vol.96. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

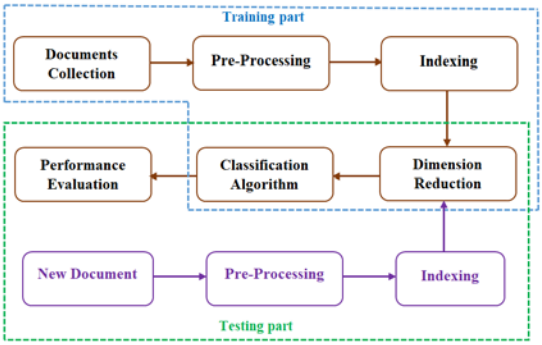

Figure (1): Typical Text Classification System

3.1 Documents Collection

The first step of text classification is collecting the documents which may be collect from different resources. The documents can be in different format (such as html, doc and pdf) and in different topics such as news, reports …. etc. [16].

3.2 Pre-Processing

The preprocessing stage is a preparatory step which aims to prepare and provide documents for text classification. Multi-subsequent steps are applied in preprocessing stage as: tokenization, normalization, removing stop words, and stemming words [16, 6, 12, 17].

(a) Tokenization

The document is a string of characters, therefore, it need to partition into sentences. Then each sentence tokenizing into a list of words. Token is a sequence of characters grouped together for getting useful semantic unit [16, 18].

(b) Normalization

Each document in dataset processes to remove the punctuation marks, digits, non-letters and the words that not belong to the used language. In English language, the capital letters converts to small letters. While in Arabic language, the diacritics must deleted, and normalize some Arabic letters such as "hamza (ا) or (أ) " and "mada (آ)" in all its forms to (alef (ا)) [6, 12, 9].

(c) Removing stop words

Stop words are the words that frequently occurring in most of documents, but not useful in getting significant meaning of description the document content. In general, stop words contain pronouns, preposition, question tools, connectivity tools, negative, adverbs and some other words [6, 12, 19].

In Arabic language, there are many stop words in addition to different suffixes and prefixes made the stop words more than other languages. The advantages of removing stop words are to reduce the size of documents (corpus size) and noise reduction. Therefore, it can increase the performance of classification system [20, 19].

(d) Stemming words

The number of word forms is often very big and increasing depending on the document length. For example (organize "ينظم" , organizing "تنظيم" , organized "نظم" , organization "منظمة") . Therefore, we need to reduce the number of word forms and removing redundant forms. There are some techniques of filtering can be applied for words reduction such as stemming technique [12, 18].

Stemming is the task of grouping words that derived from a common stem and replacing each of them with the original stem [20, 16].

Arabic language morphology is more complicated than other languages, and stemming techniques is more important [20]. Stemming can be root based stemming which return the word to its original stem or light stemming which removing the suffixes and prefixes [12].

There are many stemmers of Arabic language such as Khoja, Motaz, Light10 stemmer [21,22].

3.3 Document representation

Some terms are more important to the meaning of a document than others. Term weights are recommended to describe a document rather than the frequency statistics from the inverted index file. The most effective and widely used approach for calculating term weights is a term frequency-inverse document frequency (TF/IDF) weighting scheme [23, 24].

TF-IDF method for term weighting is used in this paper.

Term frequency (TF) is the number of times a given term appears in that particular document and it is given in equation (1) [23, 2].

$$TF_i = tf_{ij} \qquad (1)$$

Where tf_{ij} represented as the frequency of term i in document j.

The Inverse document frequency (IDF) is defined as the logarithm of the number of all documents

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

divided by the number of documents containing that term. the IDF equation is [25, 23]:

$$IDF_{ij} = log_2 (N/n_j) \quad (2)$$

Where (N) the total number of documents in the collection

(n_j): the number of documents that contain at least one occurrence of the term i [23, 25]

Then TF-IDF can be calculated weighting by the following equation [2]:

$$TF\text{-}IDF_{ij} = TF_i * IDF_i \quad (3)$$

3.4 Dimensionality Reduction

Dimensionality Reduction is defined as the task that reducing the size of a document representation by selects or extracts the suitable group of features. The aim of dimensionality reduction is to save the storage space, make the problem more manageable from the learning method, make the classification more effectiveness and efficiency of computational terms, and reduce over fitting [25, 26, 2, 24].

3.5 Classification Algorithms

The automatic text classification techniques are supervised machine learning. There is no singular algorithm can be considered the best for classification in text mining problem. This can be because of the data type, data size, number of classes, and documents number of each class [2]. Text classification consists of two major fields: linear fields and hierarchical fields. Linear text classification is the approach that addresses classification by means of lexical chains and mainly uses the repetitions of the words to text classified. Cosine similarity is examples of linear text classification. Whereas the hierarchical text classification is based on deep similarity. Similarities can be analyzed relative to the relationships between related words as well as the word repetitions. Decision trees (DT), latent semantic analysis (LSA) and support vector machine (SVM) are examples of hierarchy text classification [5, 8, 27]. Some of these algorithms are explained in following subsections.

3.5.1 Cosine Similarity

Cosine similarity is a measure of similarity between two vectors that can be used with high dimensional spaces [28, 29].

The cosine similarity computed by the following equation [30]:

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x.y}{\sqrt{\sum x^2}.\sqrt{\sum y^2}} \quad (4)$$

Where x and y are two vectors of features weight.

The work of cosine similarity in text classification can be illustrated as the following: after organize the training and testing documents as vectors, the cosine similarity computed of each testing vector by comparing it against all training vectors. Then select the class of training vector that have the highest similarity [31, 32].

3.5.2 Support Vector Machine (SVM)

SVM is a supervised learning algorithm that can be used for text classification. It was developed and used by Vapnik and the work team. They constructed a separating hyperplane with the objective of separating the data into different classes. The separating hyperplane should have maximum distance to the nearest data point in each class [33, 14].

Advantages of SVM are high accuracy, good theoretic assurances about overfitting and, with an appropriate kernel, and used with high dimensionality dataset. The disadvantages of SVM are poor interpretability, complexity and high memory requirements [14].

In the linearly separable data, the arrangement of the data points as it shown in Figure (2). Suppose there are two classes (x and o), the figure (2) illustrating that the x's and o's data points can be linearly separated [33].

Journal of Theoretical and Applied Information Technology
31st July 2018. Vol.96. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

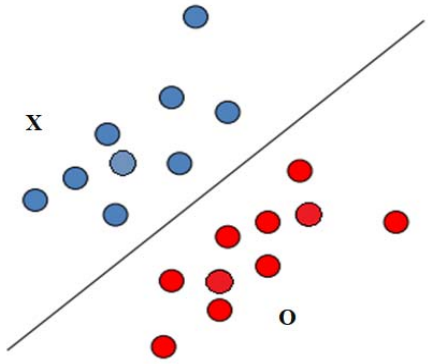

Figure (2): A Separating Data Points (Hyperplane Separable Line)

In this case, a set of labeled training points $\{x_i,y_i\}$, were i=1,..,L , and $x_i \in R^d$, $y_i \in \{-1,1\}$. Since the data points are linearly separable in this case a linear hyperplane or decision boundary can be defined by the equation (5) [34]:

$$f(x) = w.x + b \qquad (5)$$

It is very important to notice that the data points that lie on the hyperplane satisfy the following formula that given in the equation (6) [34]:

$$f(x) = \begin{cases} w.x + b > 0, & Postive\ data\ points \\ w.x + b < 0, & Negative\ data\ points \end{cases} \qquad (6)$$

Linear classification hyperplane calculation for the both cases "positive" and "negative" is illustrated in Figure (3) [34].

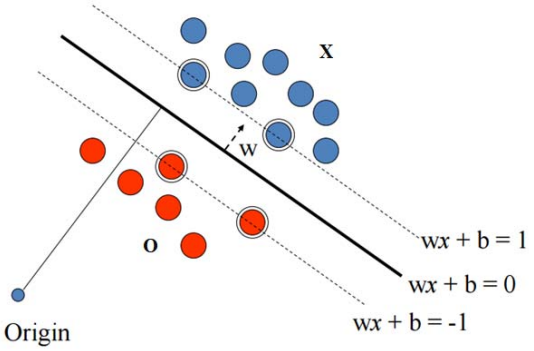

Figure (3): Linear Classification Hyperplane

3.6 Performance Evaluation

The last stage of text classification is the performance evaluation. The evaluation is made experimentally. An evaluation metrics can be used to measure the effectiveness and performance of the text classifier [16, 35].

There are many measures can be used such as precision (P), recall (R), and F1-measure (F1) [35, 3, 16].

$$P = \frac{TP}{TP + FP} \qquad (7)$$

Where, TP is true positive which is referred to the number of labels that are correctly recognized. FP is false positive which is referred to the number of labels that are wrongly recognized.

$$R = \frac{TP}{TP + FN} \qquad (8)$$

FN is false negative which is referred to the number of labels that not identified as labels but should have been recognized as a label.

$$F_1 = \frac{2PR}{P + R} \qquad (9)$$

The F-measure is a compromise of both precision and recall for measuring the overall performance of text classification.

4. PROPOSED SYSTEM

The proposed classifier framework for Arabic documents consists of two main parts: filtering features of N-gram types and system combination. Figure (4) illustrate the proposed framework for text classification.

Journal of Theoretical and Applied Information Technology
31st July 2018. Vol.96. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

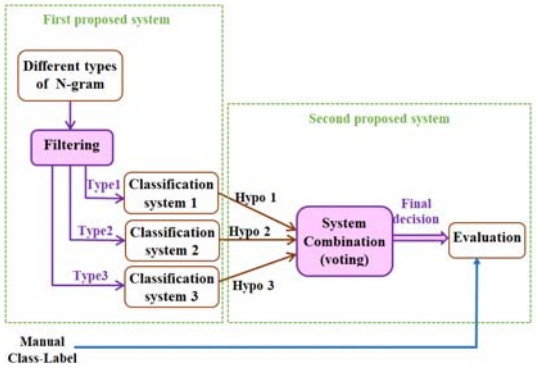

Figure (4): The proposed text classification framework

The proposed system consists of two parts: filtering approach and system combination. In the following subsections, the proposed system will explain in more details.

4.1 Filtering approach

The features are the most important part that effect on the classification system performance. To enhance the features, some researchers are relied on word repetition between words, while others are based on deep semantics rather than term repetition principles. The word repetition doesn't concern of semantic and relations between words. There are some words give the same meaning (synonymy) but with different graphemes such as (جاء) and (أتى). Moreover, there are some words come together as one word in most times such as (دار السلام) and (كرة القدم).

N-gram is a method for representing the features. It extracts sequential characters or words from the document string. N-gram method is illustrated in figure (5).

To extract the semantic relationships between words, filtering approach is used to find the best features that gives good result with minimum number of features. The main characteristic to select filtering approach with different N-gram types is to achieve the purpose of information filtering to select the best features and removing noise.

The idea of filtering is that if the sequence of words has a weight larger or equal to the weight of individual words that mean the sequence of words is more confidence than individual words. The assumption here is that if the confidence score (N-gram probability) of the words sequence is larger than low N-gram type, so that means it the reliable set of word sequence.

To explain the filtering algorithm, suppose that F is the initial feature set in classification system that generated by the traditional way.

Let $F=\{f_1, f_2, f_3,..., f_n\}$, where f_i is an uni-gram word (individual word), and suppose W_1, W_2, W_3 are the weight of uni-gram, bi-gram, and tri-gram (f_1, f_2, f_3) word sequence respectively. Then the following formula is used to find the high confidence score:

$$F_{new} = \begin{cases} f_1 f_2 f_3 \ldots\ldots\ldots \text{ if } W_3 \geq (W_2 \text{ and } W_1) \\ f_1 f_2 \ldots\ldots\ldots\ldots \text{ if } W_2 \geq W_1 \\ f_1 \ldots\ldots\ldots\ldots \text{ otherwise} \end{cases} \quad (10)$$

F_{new} is the list of the affected features after applied filtering approach. The assumption here is that if $W_3>=W_2$, and $W_3>=W_1$ that means the word sequence $f_1\ f_2\ f_3$ can be estimated as good feature more than $f_1\ f_2$ and f_1.

Figure (5) is illustrated a simple example of filtering approach with N-gram.

Suppose there are 25 documents, and want to extract features for the phrase " بغداد عاصمة العراق " , if the frequency f of N-gram types in a given document as follow:
Uni-gram: f (بغداد)=5, f (عاصمة)=3, and f (العراق)=6
Bi-gram: f (بغداد عاصمة)=3 and f (عاصمة العراق)=3
Tri-gram: f (بغداد عاصمة العراق)=3
And suppose the number of documents that contain the term as the following:
Uni-gram: n (بغداد)=6, n (عاصمة)=4, and n (العراق)=8
Bi-gram: n (بغداد عاصمة)=4 and n (عاصمة العراق)=4
Tri-gram: n (بغداد عاصمة العراق)=2
Then TF-IDF for each term (according to equation 3) as the following:
Uni-gram: *TF-IDF*(بغداد)=10.29, *TF-IDF*(عاصمة)=7.929, and *TF-IDF*(العراق)=9.858
Bi-gram: *TF-IDF*(بغداد عاصمة)=7.929 and *TF-IDF*(عاصمة العراق)=7.929
Tri-gram: *TF-IDF*(بغداد عاصمة العراق)=10.929

Figure (5): An example of filtering approach

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this example, the weight of tri-gram (بغداد عاصمة العراق) is larger than the weight of bi-gram ("بغداد عاصمة" , "عاصمة العراق") and uni-gram ("العراق" , "عاصمة" , "بغداد"). According to our assumption equation (10), the system firstly checks the tri-gram weight, if the weight of tri-gram is the highest one; then, the tri-gram sequence (بغداد عاصمة العراق) is used as feature.

To recap, the purpose of using the filtering approach is to reduce the features number without affecting on the classifier performance. To apply filtering method in this study, SVM system is used with filtering approach. The N-gram types are extracted and computed their weights by using TF-IDF. Three types of N-gram are used in this study individually. They are uni-gram with bi-gram filtering type, uni-gram with tri-gram filtering type, and uni-gram, bi-gram with tri-gram filtering type. Then, the filtering approach is used to extract only the affected features. Next, the selected features (F_{new}) are represented by vector space model. Final, the vector is used in the SVM classifier in order to predicting the class labels of testing documents.

4.2 System combination

System combination is the system that combines the results of different classifiers and then used voting to get better results. The idea of voting is selecting the most frequency hypothesis result. That means if two or more classifiers are given the same class to the test document, the result is that class. A system combination is selected because it provides many advantages such as multiple parallel classification systems can be used at the same time; different knowledge can be adopted in each classification system.

In general, it's difficult to select a single classifier as the optimal one for the all problems and datasets. The best classifier that select for limited dataset may be the worst one for another dataset. Therefore, system combination can take advantages of the different classifiers to improve the performance of the system by producing the most likely result through voting approach.

The idea of voting is selecting the most frequency hypothesis result. That means the classifiers that given the high number of the same class to the test document is chosen as the best result.

A system combination is chosen to meet the objectives of the study, namely to improve classification system performance.

In this study, five classifiers are combined. Two traditional classifiers with three proposed systems are used. The two traditional classifiers are cosine similarity and SVM classifier.

For the proposed system, filtering approach is used to extract three types of N-gram in order to create three different vector space models that used with SVM algorithm. These types are: uni-gram with bi-gram, uni-gram with tri-gram, and uni-gram with bi-gram as well as tri-gram.

next, the above five classification systems are combine to select the best class label via voting. The voting is used to select the best class label with the highest number of votes. The combination can be extended for any number of text classifiers (>=3).

5. EXPERIMENTAL RESULTS

This section presents the experiments to evaluate the proposed approaches. Dataset that we used is collected manually and contain (2750) Arabic documents. Also we percentage for division the dataset (50:50, 65:35, and 75:25) for trained: tested. In the validation phase, each hypothesis that produced from the text classification is compared against a gold-standard (reference) that manually classified by humans using F-measure for evaluation. Section 5.1 will explain the baseline results. Experimental results of the proposed system will discuss in section 5.2.

5.1 Baseline systems

Two methods are used as baseline systems without any modifications:

5.1.1 Cosine Similarity Results

The cosine distance is a common method for measuring similarities by calculating the distance between any two vectors (texts). The highest cosine score means more similarity between these vectors.

Different percentage to select the number of features is utilized. The best result is obtained with 25% for features and (65:35) for trained/tested news. This system achieves 81.6% F-measure score.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.1.2 SVM Results

SVM is a good classifier by finding a hyperplane in order to separating the data into different classes. SVM is the most common approach of text classification that works well with unstructured and semi-structured texts.

To determine the threshold value, experiments with several threshold values for text similarity are done (0.3, 0.4, 0.5, and 0.6), and thereafter the best threshold is chosen. The best threshold is 0.4 that achieved by experiments with SVM classifier. The best result is obtained with percentage division (75:25) for trained/tested news and 50% for the number of features. This method achieves a 91.2% F-measure.

The results show that the SVM baseline classifier gives good results compared to cosine system.

5.2 Proposed systems

The proposed system is consisting of two parts: SVM system with filtering approach and systems combination. The following subsections explain the details of the results.

5.2.1 SVM system with filtering approach

To reduce the number of features without negative affected on the performance of SVM system, different N-gram filtering types are applied. The N-gram filtering contains three types as follows:

(A) Uni-gram and bi-gram filtering type

The best result of SVM with this type of filtering is obtained when the percentage division (65:35), and 15% for trained/tested news and number of features respectively. This system achieves an F-measure of 90.8.

SVM algorithm with this type of filtering reduced the number of the used features from 12,474 in the baseline SVM system without filtering approach to 3,404 without slightly negative impact of the classification performance.

(B) Uni-gram and tri-gram filtering type

The best result of SVM with this type of filtering is obtained when the percentage division (65:35), and 15% for trained/tested news and number of features respectively. This system achieves an F-measure of 91.

SVM algorithm with this type of filtering reduced the number of the used features from 12,474 in the baseline SVM system without filtering approach to 3,466 without negative impact of the classification performance.

(C) Uni-gram, bi-gram and tri-gram filtering type

The best result of SVM with this type of filtering is obtained when the percentage division (65:35), and 15% for trained/tested news and number of features respectively. This system achieves an F-measure of 90.8.

SVM algorithm with this type of filtering reduced the number of the used features from 12,474 in the baseline SVM system without filtering approach to 3,400 without negative impact of the classification performance.

5.2.2 System combination

This system combines multiple classification hypotheses to select the best result via a voting approach. Five classification systems are implemented to find the best label. Each classification system uses the same tools for pre-processing but a different number and type of features. The systems that explained in this study are: cosine similarity system, SVM system, SVM with uni-gram and bi-gram filtering, SVM with uni-gram and tri-gram filtering, and SVM with uni-gram, bi-gram and tri-gram filtering. The generated hypotheses from the above systems are then combined via voting approach to find the best text class label.

The best result of this system is obtained when the percentage division was (65:35), and 15% for trained/tested news and the number of features respectively. This system achieves an F-measure of 94.2.

Finally, the performances of the baseline and the proposed system for the same dataset are compared and discussed, as shown in table (1).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: The performance details comparison

Algorithm	Number of features	Train/test	F-measure
Cosine	25%	65:35	81.6
SVM	50%	75:25	91.2
SVM with uni-gram and bi-gram	15%	65:35	90.8
SVM with uni-gram and tri-gram	15%	65:35	91
SVM with uni-gram, bi-gram and tri-gram	15%	65:35	90.8
System combination	15%	65:35	94.2

The result of the comparative analysis indicates that using the proposed systems reduce the number of used features and also improves the classification performance. However, some documents were not classified by SVM and SVM with N-gram filtering. This problem resolved by system combination in addition to enhanced the classifier performance.

6. CONCLUSIONS

The study proposes two approaches: one to reduce the number of the features and others to improve the text classification. Filtering algorithm is used instead of the common uni-gram for dimension reduction via select the best features to build the space vector in the text classification. Different N-gram types are used to find the deep semantics rather than term repetition principles. Besides, to improve the performance of the text classification system combination is used. Text classification system combination uses more than one text classification systems to take advantage of different techniques and knowledge in different systems. In general, it's difficult to select a single classifier as the optimal one for the all problems and datasets. The best classifier that select for limited dataset may be the worst one for another dataset. Therefore, system combination can take advantages of the different classifier to improve the performance of the system by producing the most likely result through voting approach. The proposed system achieved the objectives of the study, namely to improve classification system performance with minimum number of used features.

7. LIMITATION OF THE STUDY

This study has several limitations as listed below:

1- The systems were proposed and tested for only Arabic documents and using the pre-processing that suitable for Arabic documents.

2- The proposed classifier tested only news broadcasted in 2016.

3- The study used only the SVM algorithm in testing N-gram types.

8. FUTURE WORKS

Several suggestions are proposed for future works based on this study:

1-Test the proposed algorithms in other languages, such as English and Malay languages.

2-Apply the filtering approach with different text classification algorithms, such as latent semantic analysis, principle component analysis, and decision tree.

3-Test the proposed algorithms in high-processing modules such as information retrieval

REFERENCES

[1] Swadia, J., "A Study of Text Mining Framework for Automated Classification of Software Requirements in Enterprise Systems. " 2016: Arizona State University.

[2] Garcia Constantino, M., "On the use of text classification methods for text summarisation. " 2013, University of Liverpool.

[3] Chua, S.H.L., "An investigation into the use of negation in Inductive Rule Learning for text classification. " 2012, University of Liverpool.

[4] Joachims, T., "Text categorization with support vector machines: Learning with many relevant features. " Machine learning: ECML-98, 1998: p. 137-142.

[5] Zrigui, M., et al., "Arabic text classification framework based on latent dirichlet allocation. " Journal of computing and information technology, 2012. 20(2): p. 125-140.

[6] Alsaleem, S., "Automated Arabic Text Categorization Using SVM and NB. " Int. Arab J. e-Technol., 2011. 2(2): p. 124-128.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] Khalaf, Zainab A., et al. "ASystem COMBINATION FOR MALAY BROADCAST NEWS TRANSCRIPTION." JURNAL TEKNOLOGI 77.19 (2015): 35-44.

[8] Saad, M.K., "The impact of text preprocessing and term weighting on arabic text classification." in Gaza: Computer Engineering, the Islamic University. 2010.

[9] Khreisat, L., "A machine learning approach for Arabic text classification using N-gram frequency statistics. " Journal of Informetrics, 2009. 3(1): p. 72-77.

[10] Belkebir, R. and A. Guessoum. "A hybrid BSO-Chi2-SVM approach to Arabic text categorization. " in Computer Systems and Applications (AICCSA), 2013 ACS International Conference on. 2013. IEEE.

[11] Hughes, M., et al., "Medical Text Classification using Convolutional Neural Networks. " arXiv preprint arXiv:1704.06841, 2017.

[12] Elhassan, R. and M. Ahmed, "Arabic Text Classification on Full Word. " International Journal of Computer Science and Software Engineering (IJCSSE), 2015. 4(5): p. 114-120.

[13] Al-Shargabi, B., W. Al-Romimah, and F. Olayah. "A comparative study for Arabic text classification algorithms based on stop words elimination. " in Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications. 2011. ACM.

[14] Hmeidi, I., et al., Automatic Arabic text categorization: "A comprehensive comparative study. " Journal of Information Science, 2015. 41(1): p. 114-124.

[15] Trivedi, M., et al., "Comparison of Text Classification Algorithms. " International Journal of Engineering Research & Technology (IJERT), 2015. 4(02).

[16] Korde, V. and C.N. Mahender, "Text classification and classifiers: A survey. " International Journal of Artificial Intelligence & Applications, 2012. 3(2): p. 85.

[17] Khalaf, Zainab Ali, and Tan Tien Ping. "Novel Noun Pronunciation Unification Approach to Improve Story Boundary Identification in the Transcription of Malay News Broadcasts." IJCSA 11.1 (2014): 37-55.

[18] Manning, C., P. Raghavan, and H. Schütze, "An Introduction to information retrieval. " 2009: Cambridge University Press, Cambridge, England.

[19] Alajmi, A., E. Saad, and R. Darwish, "Toward an ARABIC stop-words list generation. " International Journal of Computer Applications, 2012. 46(8): p. 8-13.

[20] Croft, W.B., D. Metzler, and T. Strohman, "Search engines: Information retrieval in practice. " second ed. Vol. 283. 2010: Addison-Wesley Reading.

[21] SAFAR:Software Architecture For Aranic Language Processing , 2013 (http://arabic.emi.ac.ma/safar/?q=examples).

[22] Al-Kabi, M.N., et al., "A novel root based Arabic stemmer. " Journal of King Saud University-Computer and Information Sciences, 2015. 27(2): p. 94-103.

[23] Xia, T. and Y. Chai, "An improvement to TF-IDF: Term Distribution based Term Weight Algorithm. " JSW, 2011. 6(3): p. 413-420.

[24] Dadgar, S.M.H., M.S. Araghi, and M.M. Farahani. "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. " in Engineering and Technology (ICETECH), 2016 IEEE International Conference on. 2016. IEEE.

[25] Khalaf, Zainab Ali. "Broadcast News Segmentation Using Automatic Speech Recognition System Combination with Rescoring and Noun Unification." Diss. Universiti Sains Malaysia, 2015.

[26] Addis, A., "Study and Development of Novel Techniques for Hierarchical Text Categorization. " University of Cagliari, 2010.

[27] Khan, A., et al., "A review of machine learning algorithms for text-documents classification. " Journal of advances in information technology, 2010. 1(1): p. 4-20.

[28] Nalawade, R., A. Samal, and K. Avhad, "Improved Similarity Measure For Text Classification And Clustering. " International Research Journal of Engineering and Technology (IRJET), 2016. 3(05): p. 214-219.

[29] Khalaf, Zainab Ali, and Tan Tien Ping. "Automatic identification of broadcast news story boundaries using the unification method for popular nouns." Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on. IEEE, 2013.

[30] Mohammad, A.H., O. Al-Momani, and T. Alwada'n, "Arabic Text Categorization using k-nearest neighbour Decision Trees (C4. 5) and Rocchio Classifier: A Comparative Study. " International Journal of Current Engineering and Technology, 2016. 6(2): p. 477-482.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[31] Khreisat, L., "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. " DMIN, 2006. 2006: p. 78-82.

[32] Masuma, M.R. and V. Losarwar, "A Similarity Measure for Text Processing. " International Journal for Research in Engineering Application & Management (IJREAM) 2016. 02(06).

[33] Guduru, N., "Text mining with support vector machines and non-negative matrix factorization algorithms. " 2006, University of Rhode Island.

[34] Parrella, F., "Online support vector regression. " Master's Thesis, Department of Information Science, University of Genoa, Italy, 2007.

[35] Jindal, R., R. Malhotra, and A. Jain, "Techniques for text classification: Literature review and current trends. " Webology, 2015. 12(2): p. 1-28.