BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Electrical Science & Engineering
https://ojs.bilpublishing.com/index.php/ese

ARTICLE

A New Model for Automatic Text Classification

Hekmatullah Mumivand[1*] Rasool Seidi Piri[1] Fatemeh Kheiraei[2]

1. Software Engineering Department, Lorestan University, Aleshtar Higher Education Center, KhorramAbad, Lorestan, IR Iran
2. Engineering Department, Lorestan University, KhorramAbad, Lorestan, IR Iran

ARTICLE INFO

Article history
Received: 30 April 2021
Accepted: 1 June 2021
Published Online: 3 June 2021

Keywords:
Text classification
Machine learning
W-SMO
N-gram

ABSTRACT

In this paper, a new method for automatic classification of texts is presented. This system includes two phases; text processing and text categorization. In the first phase, various indexing criteria such as bigram, trigram and quad-gram are presented to extract the properties. Then, in the second phase, the W-SMO machine learning algorithm is used to train the system. In order to evaluate and compare the results of the two criteria of accuracy and readability, Macro-F1 and Micro-F1 have been calculated for different indexing methods. The results of experiments performed on 7676 standard text documents of Reuters showed that the best performance is related to w-smo bigram criteria with accuracy of 95.17 micro and 79.86 macro. Also, the results indicated that our proposed method has the best performance compared to the W-j48, Naïve Bayes, K-NN and Decision Tree algorithms.

1. Introduction

We live in a world that the information has much value for us. With increasing the amount of information available on Internet, the tools are needed very much to help searching, filtering and managing the resources.

Text classification is referred to the thematic indexing practice of natural language texts based on a predetermined set. Now, text classification is applied to many of the fields from text indexing based on a controlled dictionary to text filtering, automatic production of the meta-data, word sense disambiguation, production of hierarchical catalogues of web resources and generally in any application requiring documentation organization or special selective and comparative distribution of documentation [1]. The other applications of text classification may be included the automatic systems of responding to the questions, information filtering, identifying the data themes, worthlessness of electronic mails, identifying the title and the other related fields [2]. The main challenge of documents classification is the bigness of the features space in this type of matters. In many of the present algorithms such this one, a large space causes that the classifier becomes much slow and inefficient. Moreover, there are the features that not only cause no better documents classification but also lower the precision of classification [3].

A text may not be interpreted directly by a classifier or a classifier algorithm but using an indexing process it is mapped to an array that its contents are stated by the dimensions. This practice helps to provide the necessary consistency and homogeneity for the texts of training and trial set sand validation [1].

In this paper, for automatic text classification three in-

**Corresponding Author:*
Hekmatullah Mumivand,
Software Engineering Department, Lorestan University, Aleshtar Higher Education Center, KhorramAbad, Lorestan, IR Iran;
Email: mhekmat.m@gmail.com

DOI: https: //doi.org/10.30564/ese.v3i1.3170

dexing methods including "bigram, trigram and qudgram" are used and machine learning algorithm W – SMO also utilized. The results show that the best method of text indexing is bigram. To classify the text, the 7676 - paper dataset of Reuters news agency has been used. This dataset has been collected titled as Reuters – 21578 that the constructing method and statistical information of this dataset explained in [4].

Of the selected documents regarding their contents classified in 8 groups, 70% have been considered as learning sets and 30% of documents placed in the test set on which the automatic classification is performed.

The rest of the paper is organized as the following. In the second section a review of text classification is given. Then in the third section the details of proposed method studied completely and finally in the fourth section the results and evaluation of proposed method given.

2. Review of Text Classification

With respect to the extension of electronic text information size being available through Internet and the other resources if no suitable indexing and classification is present, the practice of retaining and processing unclassified text information will be exposed to many problems. Text classification has lots of applications including documents pursuit, document management, document extension and lowered information size. A lot of machine learning methods have been used about text classification during the recent years including neural networks [5], K – nearest neighbor (K – NN) [6], Naïve Bayes networks [7], and decision tree [8] that each of these methods has different calculations and precision.

In [9], text classification in Turkish language using n – gram has been studied. In this research using unigram, bigram, trigram and qudgram, the text has been classified.

The tests in this paper have been conducted on 600 documents allocated to six sets and the efficiency of this research reported 95/83%. In [10], text classification has been given using an integrated algorithm that is composed of SVM and K – NN algorithms. The results of this research performed on dataset of Reuters news agency show that at the best state the efficiency of this integrated method is 81/48% and at the worst state 54/55%.

3. Phases of Proposed Method

In Figure 1, the general phases of the proposed method have been shown that in the following each of the phases described and the obtained results explained.

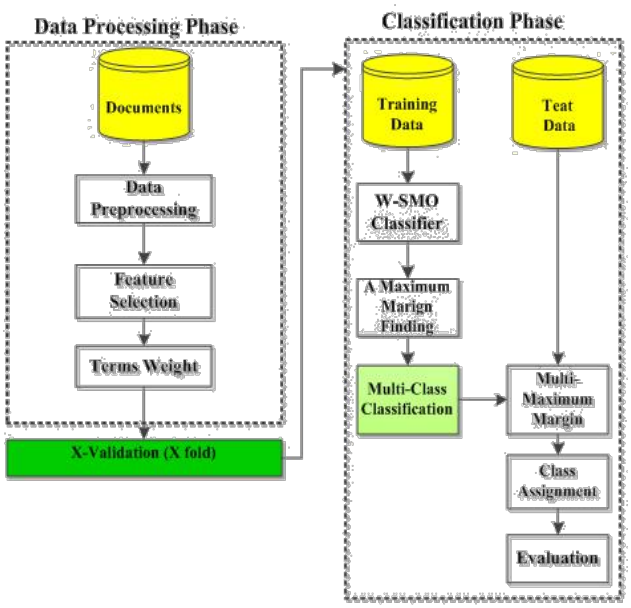

Figure 1. Proposed method for text classification

3.1 Data Processing Phase

A set of operations resulted in producing a set of refined data in order to achieve suitable features of text is namely called text processing. This operation includes the phases of text preparation, documentation indexing and indices weighting described in the following.

The phase of text preparation

In the phase of text preparation, the text including sequential characters is changed to a display suitable for learning algorithms and classification.

This phase in our proposed method is usually included the following cases:

- Obtaining the word root
- Omitting prefixes and suffixes
- Omitting Stop Words and writing symbols

Indexing phase

In this phase, a dj text is shown by a vector of its phrases weight. In other words, $dj = <w1_j, w2_j,..., w|T|_j>$ where T is the set of phrases brought at least once over the throughout training set (sometimes also called features) and $0 \leq wk_j \leq 1$. The difference of approach is usually due to one of the following reasons in this case:

- Different definitions of a thing named " phrase ".
- Different ways of calculating the weight of terms.

In this paper, text indexing method has been used as simple words and N – gram (qudgram, trigram, bigram) method that N- gram method explained in the following.

 DOI: https: //doi.org/10.30564/ese.v3i1.3170

N – gram method

In this method, indexing is as sequential from N letters successively. A word of text as a set of N – grams overlapping each other has been shown [11]. For example, word " TEXT " is composed of the following N – grams:

Bigram :_T, TE, EX, XT, T_

Trigram :_TE, TEX, EXT, XT_, T_ _

Quadgram :_TEX, TEXT, EXT_, XT_ _, T_ _ _

Where " – " is an indicator of distance. The advantage of N – gram is its nature. Since any strain is composed of a few words, the errors are not dispersed and effect on a few numbers s of strains.

Weighting phase

To weight the features, different approaches may be utilized. The simplest of this weighting state may be done as binary.

Another selection of weighting on each word is with respect to the number of repeating each word. But, one of the suitable and considerable techniques is to use *tf-idf* [15]. is the frequency product of each word at inverse of document frequency usually defined as the following:

$$tf - idf(t_k, d_j) = tf(t_k, d_j) \times \log \frac{|N|}{N(t_k)} \quad (1)$$

Where: N is the candidate of the total number of documents, t_k is the number of documents from training set in which the word t_k has occurred at least once $tf(t_k,d_j)$ indicates to the number of replications of kth word at j^{th} document. Therefore, more occurrence of a word is effective in its increased weight if it has not been replicated in all of the other texts. Regarding that this method is reflexive and its other dimensions have been used in many practices and its good efficiency approved on different datasets, this reflective method selected in this paper.

3.2 Classification Phase

This phase includes two learning and test phases that each of them are described in the following.

Learning phase: Automatic text classification algorithm

In this paper, W – SMO algorithm [12,13] has been applied. Using support vector machine algorithms in text classification problems is a new approach attracted much attention within the recent years. In learning phase, W – SMO approach tries to select the margin of decision such a way that it maximizes its least distance with any of desired sets. This type of selection causes that our decision tolerates noise conditions in practice well and also responds appropriately. This type of margin selection is done based on the points named support vectors.

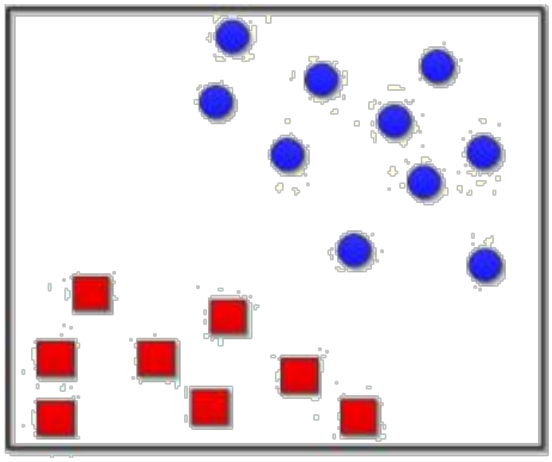

Figure 2. The set of points related to two sets

Figure 2 shows the set of records related to two sets in a problem of binary classification. Respecting this figure, it is identified that there is the possibility of separating these two sets using various linear classification. Now, suppose that we have two lines b1 and b2 drawn in Figure 3. The aim is to find the best line among these two drawn lines. The best line is simply recognizable using the algorithm of support vector machines.

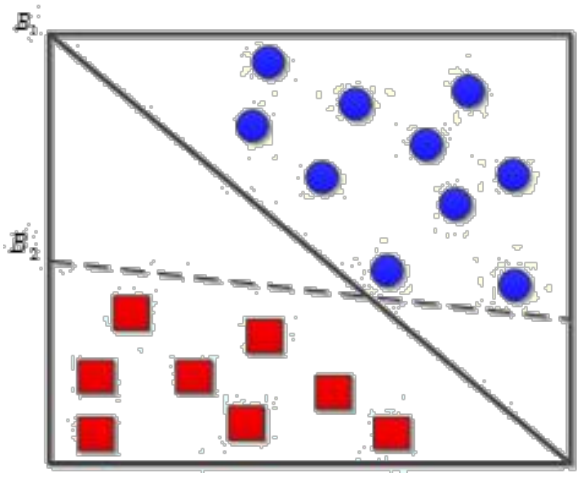

Figure 3. Classification lines of the sample

The algorithms based on support vector machines are algorithms trying to maximize a margin. These algorithms, to find the line of separating the sets, begin from two parallel lines and move these lines in opposite direction of each other until each of the lines reaches a sample of a special set at its side. After doing this phase, a band or margin is formed between two parallel lines. Whatever the width of this band is more, it means that the algorithm could maximize the margin and the aim is also at maximizing this margin. In fact, our aim is at selecting the

DOI: https: //doi.org/10.30564/ese.v3i1.3170

most possible value for this margin. In the center of image margin, the line separating the sets, that is, central line is placed. Now, among the lines drawn the algorithm selects a line with the maximum side margin as the separator of the sets. The margin related to two lines b1 and b2 is shown in Figure 4.

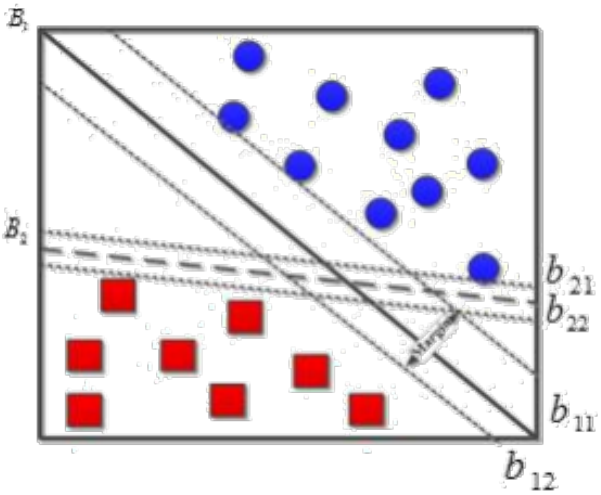

Figure 4. Classification lines margin of the sample

The relation of calculating the margin is as relation (2):

$$M \arg in = \frac{2}{\left\|\vec{w}\right\|^2} \quad (2)$$

Consequently, after calculating the margin, the algorithm selects B1 line as a separator line because the side margin of this line is more than that B2. After selecting the separator line, the algorithm calculates a function for calculating the classification of new records based on the equations set of separating line and equations set of parallel line. The equations set of separating line B1 and equations set of parallel line are shown in Figure 5. In this figure $\vec{w} \times \vec{x}_1 + b = 1$ is the equation of line b11. Consequently, $\vec{w} \times \vec{x}_1 + b \geq 1$ refers to the right side of this line and in fact refers to the zones in which the records from the type of circle set are located on it. When the algorithm reaches the relation $\vec{w} \times \vec{x}_1 + b \geq 1$ after placing the values of the new record features in function, it will return the value of 1 meaning that the new record is related to the circle set. In Figure 5, $\vec{w} \times \vec{x}_1 + b = -1$ is the line equation of b12. Consequently, $\vec{w} \times \vec{x}_1 + b \leq -1$ refers to the left side of this line and in fact refers to the zones on which the records from the type of square set are located on it. When the algorithm reaches the relation $\vec{w} \times \vec{x}_1 + b \leq -1$ after placing the values of new record features in function, it will return the value of -1 meaning that the new records belong to the square set.

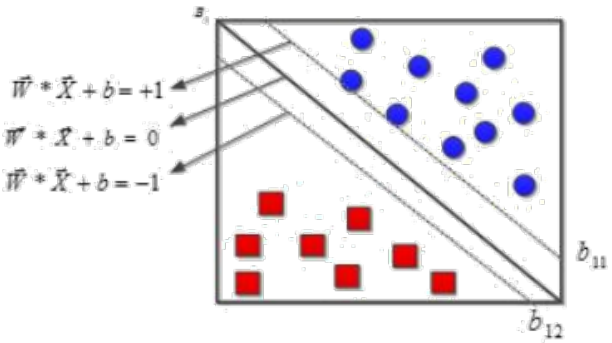

Figure 5. Minimizing margin of classifier line in support vector machine

In this section, at first the data set applied to test the classifier and learning are defined and then the implementation details are given.

3.3 Data Set

The documents used as the dataset have been collected from Reuters news agency. The dataset includes 7676 text documents with different sizes classified into 8 classes. Each document is labeled based on the contents and zones in which it found and placed in an individual file describing a set or a set collection marked by a label. Table 1 shows a list of the dataset from Reuters news agency for test phase.

To separate the test and training collections, X – validation has been used. The number of subsets for this practice is 5 and number of documents has been divided into 5 equal subsets. Each time, one subset has been regarded as the test set and the other four subsets as learning set. Finally, the mean of obtained results has been calculated.

Table 1. Document's list of test phase for data set Reuters – 21578

Sets	Trial phase	Test set	Sum
Acq	1596	696	2292
Trade	251	75	326
Ship	108	36	144
Interest	190	81	271
Grain	41	10	51
Crude	253	121	374
Earn	2841	1083	3924
Money-fx	206	87	293

3.4 Evaluation Criteria

In problems of text classification, recall, precision and F1 criterion are usually used as following formulas:

 DOI: https: //doi.org/10.30564/ese.v3i1.3170

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{F1} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5}$$

TP: Number of texts truly attributed to a class.

FN: Number of texts wrongly attributed to a class.

FP: Number of texts wrongly rejected from a class.

At last, to evaluate the efficiency on the total classes, mean taking method has been used. In macro mean- taking, the precision and recall rates of total classes are calculated. In this method, the total classes are given equal weight.

After obtaining the precision, recall and F1 for each set, two methods are applied to calculate the mean of these criteria [14]. In formulas (6) and (7), macro-precision is shown by π^M and micro-precision π^μ.

$$\pi^M = \frac{\sum_{i=1}^{|c|} \pi_i}{|c|} \tag{6}$$

$$\pi^\mu = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} (TP_i + FP_i)} \tag{7}$$

At each of two above formulas, |c| means the number of sets that is 8 in our trial.

4. Results and Evaluation

To evaluate the proposed method, Rapid Miner simulator software and a system with Intel processor 2,3GHZ, memory 4GB and operation system 7 have been used. In this paper, to automatic classification of the texts three indexing methods bigram, trigram and qudgram and also machine learning algorithm W – SMO have been applied. To classify the texts, 7676 text documents from Reuters news agency have been used. This dataset has been collected titled as Reuters – 21578. Of the selected documents regarding their contents classified in 8 sets, 70% have been considered as learning sets and 30% of documents placed in the test set on which the automatic classification is performed.

The results of evaluating classifier using indexing methods are shown in Figure 6 in which bigram indexing method from two criteria of Micro – F1 & Macro – F1 has a better efficiency to the methods of trigram and qudgram.

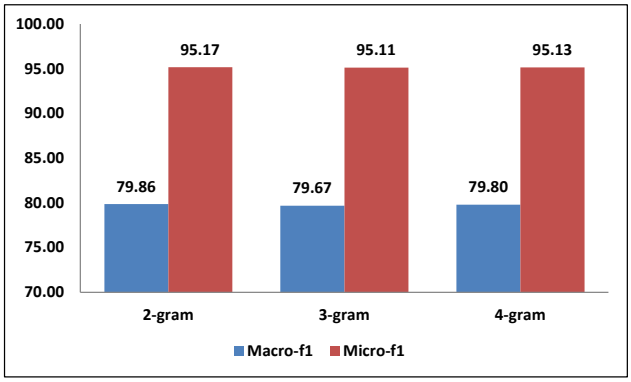

Figure 6. Obtained results for indexing methods

In addition, for each one of the news set the precision and recall criteria have been evaluated and obtained results in Table 2 show that the best efficiency of proposed algorithm is related to Earn set with precision 98.73% and recall 99.26%.

Table 2. Details of proposed algorithm results for each set

Set	Precision	Recall
Acq	94.06%	96.26%
Trade	92.31%	90.67%
Ship	70.00%	61.11%
Interest	87.43%	82.72%
Grain	56.52%	40.00%
Crude	93.42%	88.43%
Earn	98.73%	99.26%
Money-fx	83.94%	80.64%

Also, we have evaluated our proposed method with machine learning algorithm like W – j48, Naïve Bayes, K – NN, Decision Tree that the results of this evaluation shown in Figure 7. The results show that classification efficiency using W – SMO and bigram combination is more than that the other combinations. The precision of classification reaches 95.17% of Micro – F1 precision at the best state and 79.86% of Macro – F1 precision.

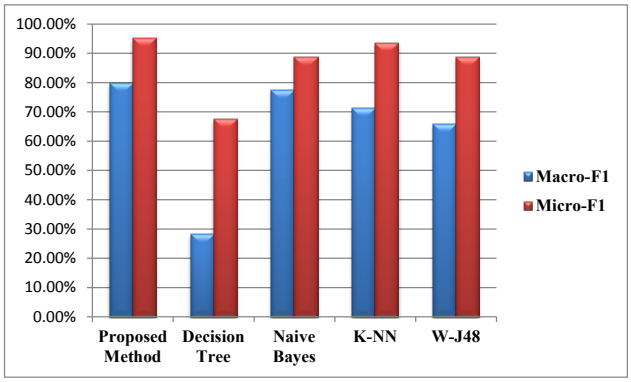

Figure 7. Obtained results based on two criteria Micro – F1 & Macro – F1 of our proposed method with W – j48 , Naïve Bayes , K – NN , Decision Tree methods

DOI: https: //doi.org/10.30564/ese.v3i1.3170

5. Conclusions

In this paper, a method has been presented for automatic text classification. This method has been evaluated with the standard dataset of Reuters news agency including 7676 documents classified within 8 different sets. Using different tests performed on indexing methods, combination of bigram method and learning algorithm W – SMO had the best efficiency. In addition, our proposed method was evaluated with machine learning algorithm W – j48, Naïve Bayes, K – NN, Decision Tree. The results of evaluation showed that our proposed method for this dataset has the best efficiency to these algorithms. Finally, our proposed method reached Micro – F1 & Macro – F1 criteria to 79.86%, 95.17% for this dataset, respectively.

References

[1] Weiyu Zhang; Can Xu, " Microblog Text Classification System Based on Text CNN and LSA Model", 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT), 2020.

[2] XiaoyuLuo, " Efficient English text classification using selected Machine Learning Techniques", Alexandria Engineering Journal, Volume 60, Issue 3, Pages 3401-3409, June 2021.

[3] Y. Lin,Y. Qu, Z. Wang, "A Novel Feature Selection Algorithm for Text Categorization", Expert Systems with Applications, Vol. 33, pp(1-5), 2007.

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[5] http://www.rapidi.com.

[6] C. H. Wan, L. H. Lee , R. Rajkumar , D. Isa," A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-nearest neighbor and Support Vector Machine", Elsevir 2012.

[7] J. Sreemathy, P. S. Balamurugan," An Efficient Text Classification Using KNN and Naïve Bayesian", International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 03, March 2012.

[8] Li Y. H. and Jain A. K. , "Classification of text documents".The Computer Journal 41(8), pp.537-546, 1998.

[9] A. Guran, S. Akyokus, N. G. Bayazit, M. Zahidbgurbuz, "Turkish Text Categorization Using n-gram word", International Symposium on Innovations in Intelligent Systems and Applicaitons, June 29 – July 1, 2009.

[10] Wan, C. H., et al. "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine". Expert Systems with Applications (2012). DOI: 10.1016/j.eswa.2012.02.068. Elsevir 2012.

[11] Cavnar, William B., "N-Gram-Based Text Filtering For TREC-2," to appear in the proceedings of The Second Text Retrieval Conference (TREC-2), ed. by, Harman, D.K., NIST, Gaithersburg, Maryland,1993.

[12] C. H. Wan, L. H. Lee , R. Rajkumar , D. Isa," A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-nearest neighbor and Support Vector Machine", Elsevir 2012.

[13] Y.Huang, "Support Vector Machines for Text Categorization Based on Latent Semanticindexing", Technical report, Electrical and Computer Engineering Department, Johns Hopkins University.

[14] Sebastiani, F "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No.1, pp. 107-131, 2002.

[15] M.H. Aghdam,N. Ghasem-Aghaee,M.E. Basiri." Text feature selection using ant colony optimization", Expert Systems with Applications,PP(6843–6853),2009.

 DOI: https: //doi.org/10.30564/ese.v3i1.3170