Journal of College of Education for pure sciences(JCEPS)
Web Site: http://jceps.utq.edu.iq/ Email: jceps@eps.utq.edu.iq
Volume 8, Number 1, March 2018

Simple Approach for documents classification

Assist.prof.Dr.yossra Hussein , Safa abdel jalil Ibrahime

yossra_1@yahoo.com , Safa.abdel.jalil@gmail.com

Computer science-university of technology

Abstract

There are a lot of classification methods that have been developed over the last years to work on database, whether numbers, images or documents for controlling and analyzing data. This paper offers a proposed classification approach to classify documents and determine the correct category for each document.The experiment results showed the success rate of the proposed classification approachis 91% by using F-measure, micro-average and macro average.

Keywords: classification, text categorization, document, F-measure.

1. Introduction

The increase in the amount of data in recent years has led to presence a huge numbers of electronic documents, such as electronic libraries, e-mails messages, social networking data, Internet web pages, books, electronic articles and many others, there should be an available, important and appropriate techniques to organize these documents [1], the classification of the document is the task to classify the document under a specific category according to the characteristics of this document and may be classified under one or more classes, if a document is assigned to only one class, it is called "single-label" and if the document is assigned to more than one class, it is called "multi-label" [2].

2. Related works

Several researchers were preformed to cover some of the related works and to provide an overview of the previous important works in document classification

C. Zanchetitin .et al (2012) [3]developed hybrid algorithm between support vector model (SVM) and K-Nearest Neighbor (KNN), this hybrid reduce KNN confusing in distinguishing characters, SVM help KNN with ambiguous characters and separate them, so SVM work as classifiers when KNN ran into confusion, the results of hybrid algorithm was better and more accuracy then KNN results.

vishwanath bijalwan et al. (2014) [4]compared between three algorithms on text classification those algorithms are KNN naïve bayes and term-graph, KNN has best accuracy in spite of its problems in time where it has a high time complexity compared with other two algorithms

Sayali D. Jadhav et al. (2016) [5] compared decision tree, KNN and Bayesian algorithms, the analysis on those three techniques showed that decision tree has best results from accuracy and error rate as well as decision tree is easier in implementation, where Bayesian algorithms has similar accuracy as decision tree, KNN has less quality results compared to other algorithms techniques, this comparative study shows that each algorithm as its environment for presenting good results and no algorithm can satisfy all criteria and measures.

Natalia Labuda et al.(2017) [6] modify KNN and improve it by reducing KNN sensitivity in selecting number of nearest neighbors parameter (K), original KNN has varying K in each data point depending on test point lying region where the improvement has been done by fixed value of K in all data point.

3. Proposed Approach

This work introduced a new method to classify documents based on their terms to obtain good results, the proposed approach has several phases to get the final results which is classified tested documents, Figure (1) describe the proposal approach.

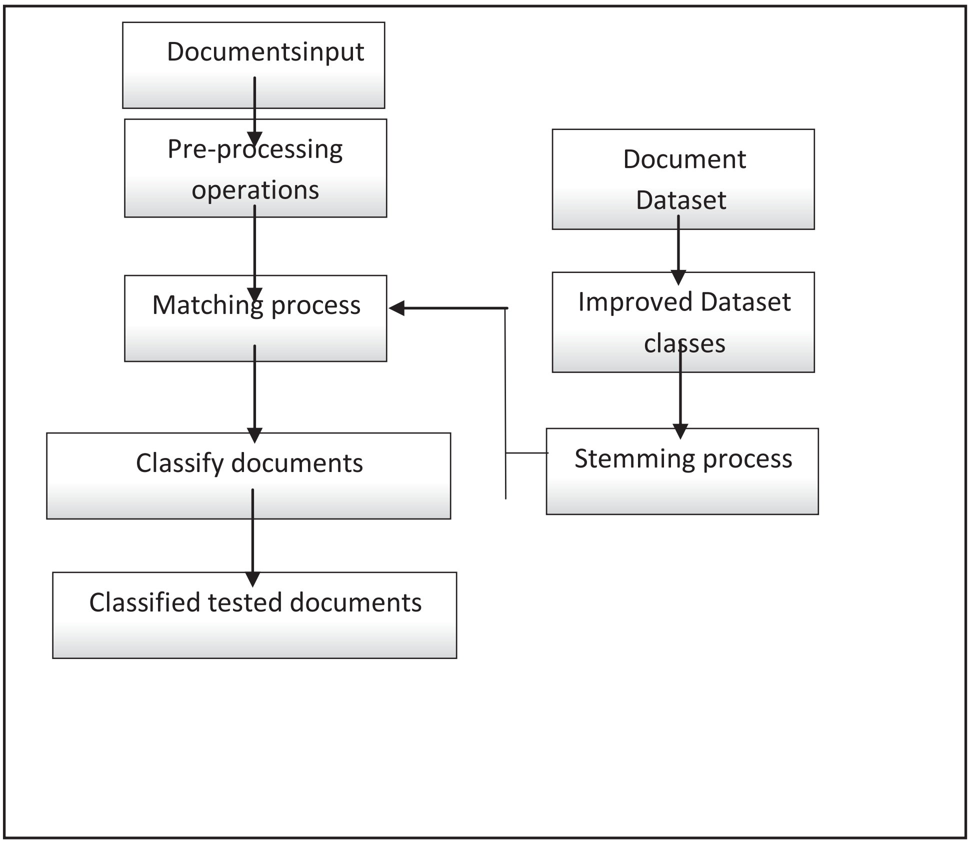

Figure (1): Block Diagram of Classification Proposal Approach

The proposed approachworks on three phases: first phase applies preprocessing operations to the database which are three parts tokenizing, stop words removal and stemming, tokenizing performs on every word in the document and put it in a token to be handled easier, stop word removal removes the unwanted words that are unimportant for clustering the dataset these words like is, are, an, un... etc.

The stemming process is done by reducing words to their root form by removing the changes between several forms for the same word as example computer, computing they back to their root compute, stemming process in addition remove the difference between lowercase and uppercase of the words.

The second phase work on prepare a mini-database derived from the original database, this paper adopted six categories each class presented by file contains terms and concepts that represents class`s subject, the process of choosing concepts and terminology was mainly based on two main condition:

1- The terminology is extracted from the database exclusively.
2- The terms are defined by one category only and do not exceed the rest of the categories

The process of extracting concepts was done manually, which reduces the percentage of noise terms and reduces the rate of error effectively.Developed database are also passed by the stemming process which is the last operation from pre-processing operations to be unified all the terms,

The third phase is classify each document to their class, this work use six classes which are (Windows, hardware, graphics, space , electrons and cryptography) the documents that are classified may belong to the those classes or may not belong at all, matching processis the most important step in the proposal approach, this process depends mainly on the calculate the words that are matched between the words of the entered document and the developed database words for each category, Matching process produce number of words matching between the tested Document and the words of the developed database, it will be easy to know the type of the tested document, it is assigned into the category with the highest value matching between all the classes.

There is two anomalies cases, first if all values are equal to zero, the class of this document is undefined, and second in the case of the highest two values are equal, this document will assign to classes of these two values, algorithm (1) describe the steps of proposal classification approach.

```
Algorithm (1): proposal approach
Input:
Developed database DD
Tested document dataset TD
Counter=0
Max = 0
Output:A classified TD
Begin:
Step 1: for each class in DD
            For each term in class
                Stemming_process (term)
            Next
        Next
Step 2: for each document in TD
```

```
                For each word in document
                    Stemming_process (word)
                Next
            Next
Step 3: For each document in TD
                For each word in document
                    For each class in DD
                        For each term in class
                            If word = term then
                                counter(class) = counter(class)+1
                            end if
                        next
                    next
                next
                for each class in DD
                    if max <= counter (class) then
                        max = class of counter (class)
                    end if
                next
            document_class = max // document classified according to  the
                                        class of highest counter
        next
End
```

4. Experimental results

The approach was tested on a sample of Newsgroup 20, which is 100 -document contain a rate of 21000 word which led to a longer execution time, the proposed approachwas written in vb.net language version 2015 on widows10.

The success rate of the proposed approachexplained by using F-measure and it`s two different types of averages, **micro-average** and **macro-average**. The results displayed in Table 1 and 2

Evaluation Criteria		Proposed approach
Micro Average	P	0.91
	Π	0.91
	F1-Measure	0.91

Table 1. Micro average criteria results

Approach Classes	Π	P	F1-Measure	macro-average
Class windows	1	0.85	0.91	0.91
Class hardware	0.87	1	0.93	
Class graphics	0.86	0.92	0.88	
Class Space	1	1	1	
Class electrons	1	0.72	0.83	
Class cryptography	0.88	1	0.93	

Table 2. macro-average criteria results

The ratio of Classification Error are calculated by using Error Rate table (3).

Number of document	Number of Misclassified Documents	Error Rate $E=\frac{number\ of\ misclass\ Doc}{Number\ of\ doc}*100$
100	9	9%

Table 3. Rate of Error

From above values it`s obviously that proposed approach has a success ratio equal to 91% in both micro and macro average when its work on Newsgroup 20derived database, this a high success ratio but still do not show the full effectiveness of the approach, although there are fault ratios during the classification, but it should be noted that most of the documents that were classified by mistake are already belonging to that wrong class, and to clarify more, for example, there is a document under the category of Windows, but its mention The windows graphics thus it uses the terms of the graphics science more than others and the nature of the proposed approach, it works to classify this document under the graphics category, scientific that this classification is wrong but in practically it is a true classification, this situation repeated in the proposed approach, but for the accuracy of the approachits calculated as a wrong classification.

The figure (2), (3) shows the percentages of Table (1) and (2) respectively.

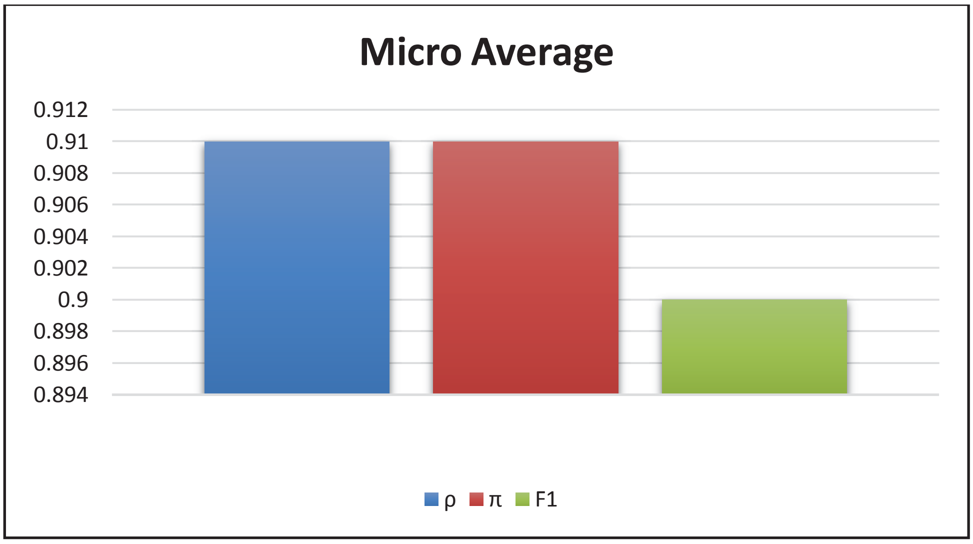

Figure 2. Micro Average

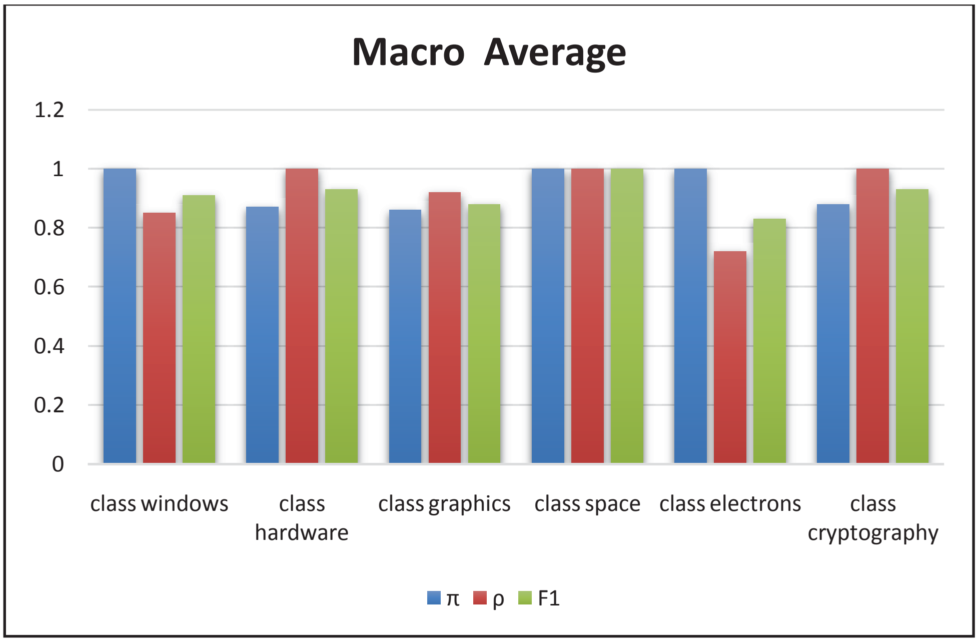

Figure 3. Macro Average

5. Conclusions

This paper introduce a new proposed technique to classify Documents with fast and simple execution, the proposed technique work on matching the tested documents terms with terms of developed dataset that contain terms refer to specific subject, this matching processcan be considered as a similarity measure with high accuracyThe proposal approachfor classification has a high value of accuracy, by using F-measure and its macro and micro average as shown in table (1) and (2) as well as figures (2) and (3).

6. Reference

[1] Irwan Bastian, Rozaliyana, Metty Mustikasari.(2016).Web Document Clustering System Using K – Means Algorithm. J. of Advanced Research in Computer Science and Software Engineering, 6(8), 181-186

[2] Rajni Jindal, Ruchika Malhotra, Abha Jain.(2015).Techniques for text classification: Literature review and current trends. Webology, 12(2),

[3] C. Zanchetitin¸ BL. Bezerra¸ W. Azevedo.(2012).A Knn-Svm Hybrid Model For Cursive Hand writing Recognition¸ The International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia.

[4] Vishwanath Bijalwan, Pinki Kumari , Jordan Pascual, Vijay Bhaskar Semwal, 2014, Machine learning approach for text and document mining.

[5] Sayali D. Jadhav, H. P. Channe. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. J. of Science and Research (IJSR), 5(1).1842-1845.

[6] Natalia Labuda, Julia Seeliger, Tomasz Gedrande, Karol Kozak.(2017).Selecting Adaptive Number of Nearest Neighbors in *k*-Nearest Neighbor Classifier Apply Diabetes Data, Journal of Mathematics and Statistical Science, 17(1), 1-13.