THE COMPARISON OF THE TEXT CLASSIFICATION METHODS TO BE USED FOR THE ANALYSIS OF MOTION DATA IN DLP ARCHITECT

Murat TOPALOĞLU

Kesan Yusuf Çapraz School of Applied Sciences, Trakya University, Kesan, 22880, Turkey

ABSTRACT

Text classification is used for the purpose of preventing the leakage of the data which is highly important within the institution through unallowed ways. The results obtained from the text classification process should be integrated into the DLP architecture immediately. The data flowing through the net requires instant control and the flow of the sensitive data should be prevented. The use of the machinery learning methods is required to perform the text classification which will be integrated into the DLP architecture. The experimental results of the comparison of text classification methods to be used in the interface written on the ICAP protocol have been prepared in the networked architecture developed for the DLP system. Also, the choice of the text classification method to be used in the instant control of the sensitive data has been carried out. The DLP text classification architecture developed helps decide the classification method through the examination of the data in motion. The method to be chosen for the text classification is applied to the ICAP protocol, and the analysis of the sensitive data and confidentiality are provided.

KEYWORDS

Decision support systems, Data Leak Prevention, Data in Motion, Security, ICAP

1. Introduction

This study involves the experimental results of the comparative text classification methods to be used for the interface which will be written in the ICAP (Internet Content Adaption Protocol) in the networked architecture which has been developed for DLP (Data Loss Prevention). In addition, this study looks into the methods aiming at the classification of the data in motion in data loss prevention. The networked architecture developed for DLP controls the data flow using ICAP. Our purpose is to choose the most suitable text classification method for the interface which will be written in the ICAP with C programming language. In this study, the classification method which will be chosen and programmed will be determined.

Text classification is the name given to the process in which the written documents are divided into certain classes depending on their contents [1]. The aim in the text classification is to determine which preset category the data will be included in by taking the features of the data into

DOI : 10.5121/ijcsit.2013.5507

consideration. It is highly required that the data be used under the conditions the institute allows and the probable damage to the data arrangement of the institution be minimized. The aim followed in the text classification is to determine which category, sensitive, confidential, or normal, the data will be included in.

DLP text classification allows for the classification of the data shared on the institutional network and prevents the data of high importance for the institution automatically. The DLP architecture developed involves the comparison of performance values of the classification algorithms through feature extraction techniques and weighting methods used in the text classification.

2. Materials and Methods

The software "text2arfff" which was developed by Amasyalı et al. performs the feature extraction of the texts with various methods, digitize the texts with the help of the weighting methods, and converts the texts to ARFF format which is the input file of WEKA [2] program, and it has been used to form arrf files in the system developed [3].

16 different feature vectors of 2-gram and words which were obtained through the use of parsing method have been extracted using grammatical and statistical features like K-Means algorithm, one of the clustering methods, and classification has been managed through these feature vectors. Whether the document contains sensitive data or not will be determined with the use of Naive Bayes, which is one of the machinery learning methods, Support Vector Machine (SVM), k-nearest neighbor algorithm (IBK) and decision trees (J48). 10-time cross validity and education choice have been used for DLP architecture and f-measure value has been used to determine the performance of classifiers.

2.1. Data Set

The DLP architecture determines whether the data flow will be allowed according to the outcomes obtained from the text classification categories. The sizes of the documents used consist of 800 documents in English ranging from 1 kb to 36 kb. The sensitive data are composed of 400 documents labeled as secret, confidential, and sensitive which involves warfare correspondence belonging to the USA [4]. The other normal data have been taken from 400 pieces of news concerning economics, sports, social, and other subjects. The classification begins with the 10 % of these data. 40 documents involving sensitive data (class1) and 40 documents involving normal data (class2) have been put in these two classes, and the number has been increased by five other documents each time. For instance, while the number of the documents was 40 for the first try, it was 45 for the second try and 50 for the third try. Finally, the number of the documents for the last try reached 400 for each classes, adding up to 800 in total. The classification operation has been achieved through using the two feature vector (2-gram and words) separately for each data set. During the preparation of the feature vectors, maximum frequency value for 2-gram and words has been changed as 10 and 50 and it has been decided that the frequency value for the k value has will be taken as 50, which is the constant value.

3. Body

While weighting was being carried out, the features of the arrf files in which text2arrf software was used were determined according to the choices given in Table – 1.

Table – 1.Parameter choices of the experiments performed

Experiments	**Method**	**Tf / Tfxidf**	**Frequency**	**k**
Experiment 1	2-Gram	Tfxidf	10	50
Experiment 2	2-Gram	Tfxidf	50	50
Experiment 3	Words	Tf	10	50
Experiment 4	Words	Tf	50	50

The choices given above have been applied in four machinery learning methods, which are Naïve Bayes, SVM, IBK and J48. K values of the K nearest neighbor algorithm, being changed from 1 to 30, has been calculated separately. The k value which gives the best outcome has been calculated separately for all the experiments. nu-SVC classification has been used as SVM type in Support Vector Machines.

In the first experiment, the education clusters obtained through Text2arrf software are learnt by the classification algorithms in the experiments carried out with the use of WEKA software. The average f-measure outcomes obtained as a result of this learning process are shown in Figure – 1. According to these outcomes, the comparison of the classifiers and the evaluation of the performances are conducted.

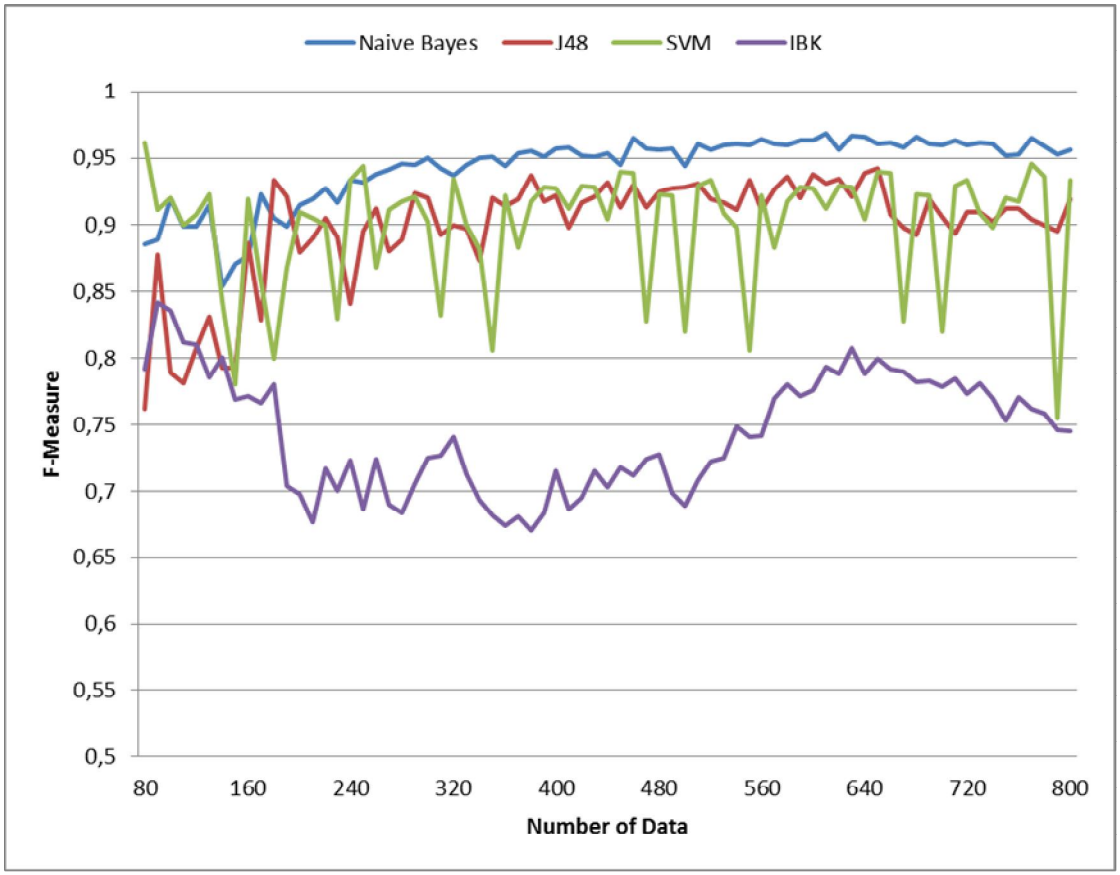

Figure – 1 Experiment 1 Result

In Table – 2, the highest and the lowest accuracy rates of the classifiers used in the experiment 1 at the end of the learning process are shown. According to this table, the Naive Bayes classifier's rate of classifying the samples accurately is higher than other classifiers.

Table – 2.The highest and the lowest accuracy rates of the classifiers according to the experiment 1

The percentage of the accurately classified samples	Naive Bayes	J48	SVM	IBK
Maximum (Max)	96,8852	94,1538	96,25	84,4444
Minimum (Min)	85,7143	76,25	75,5932	69,2105

In the algorithms above, all parameters have been used with their assumed values. K value has been chosen 4 only for the k nearest neighbor algorithm value.

Of the algorithms, the highest success belongs to Naive Bayes. The existence of a waving structure has been observed in Support Vector Machines and they have ranked as the second algorithm with the learning rate. Decision trees are more stable than the support vector trees and

they have ranked third with their learning rate. The worst learning structure is found in k nearest neighbor algorithm.

The average f-measure outcomes obtained at the end of the learning process in the second experiment are shown in Figure – 2. According to these outcomes, the comparison of the classifiers and the evaluation of the performances are conducted.

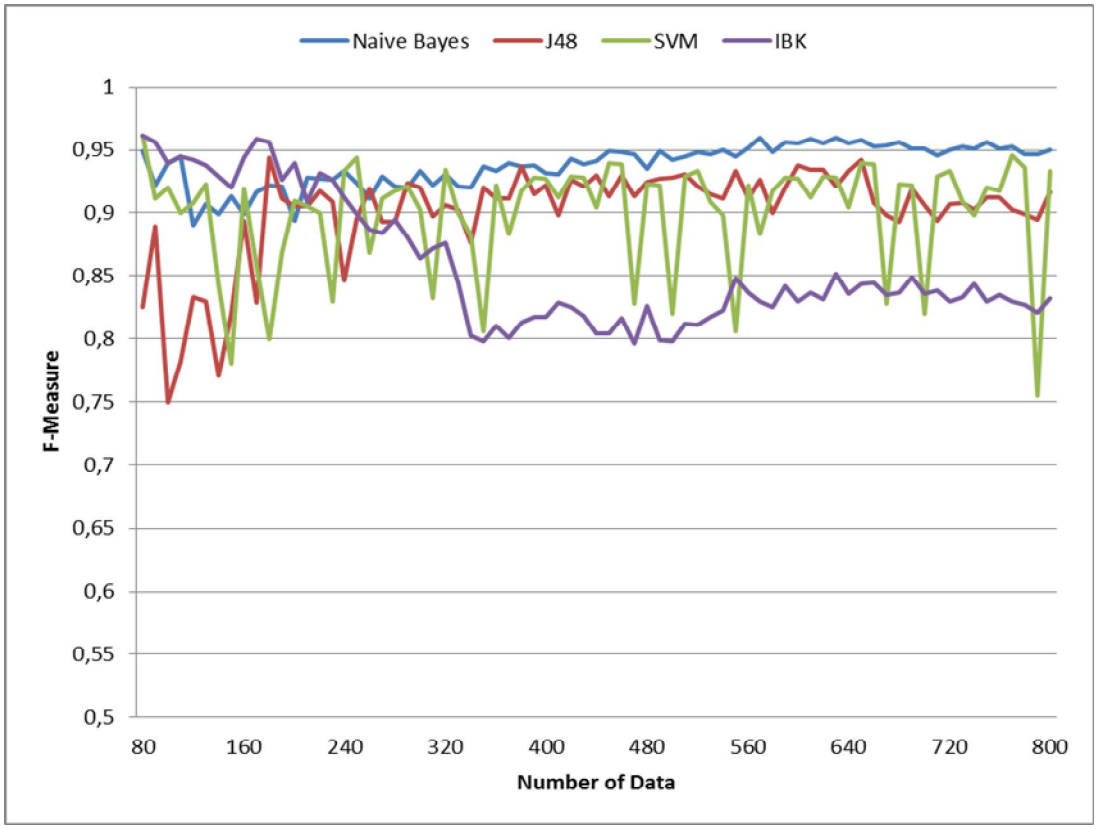

Figure – 2 Experiment 2 Result

In Table – 3, the highest and the lowest accuracy rates of the classifiers used in the experiment 2 at the end of the learning process are shown. According to this table, the IBK classifiers rate of classifying the samples accurately is higher than other classifiers.

Table – 3.The highest and the lowest accuracy rates of the classifiers according to the experiment 2

The percentage of the accurately classified samples	Naive Bayes	J48	SVM	IBK
Maximum (Max)	95,9649	94,4444	96,25	96,25
Minimum (Min)	89,1667	75	75,5932	80,4255

Support Vector Machines have produced the same outcomes as in the experiment 1. The decision trees have also produced a successful outcome as its frequency value of the experiment 2 of Table

– 1 has been increased to 50 and it has worked in a more stable manner. K nearest neighbor value has been chosen as k = 7. In comparison to experiment 1, K nearest algorithm has increased its success and yielded a better outcome.

The average f-measure outcomes obtained at the end of the learning process in the third experiment are shown in Figure – 3. According to these outcomes, the comparison of the classifiers and the evaluation of the performances are conducted.

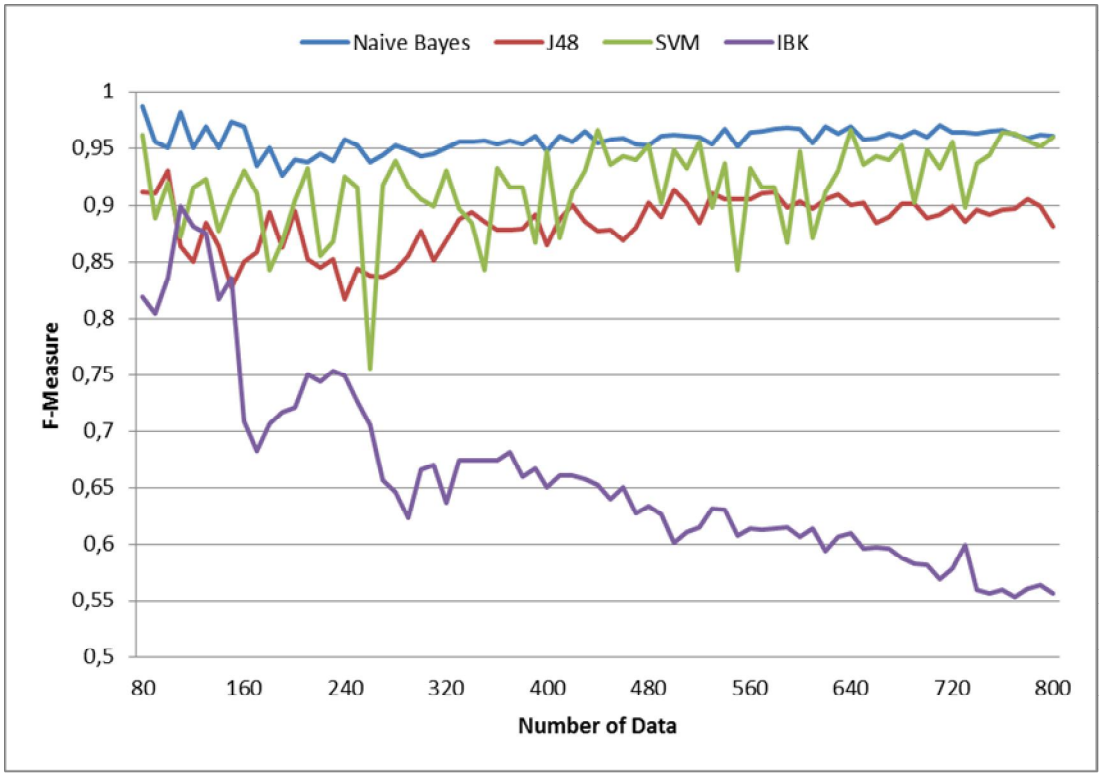

Figure – 3 Experiment 3 Result

In Table – 4, the highest and the lowest accuracy rates of the classifiers used at the end of the learning process are shown. According to this table, the Naive Bayes classifier's rate of classifying the samples accurately is higher than other classifiers.

Table – 4.The highest and the lowest accuracy rates of the classifiers according to the experiment 3

The percentage of the accurately classified samples	Naive Bayes	J48	SVM	IBK
Maximum (Max)	98,75	93	96,5909	90
Minimum (Min)	92,6316	81,6667	76,1538	61,8182

As for the method, words have been used instead of n-gram and of all the algorithms; the highest success belongs to Naive Bayes. According to the words method, except for the k nearest neighbor algorithm, the values of the all other algorithms have increased. The value of the k nearest neighbor algorithm has been determined as k = 1. As to k nearest neighbor algorithm, as the number of texts has become more, the success of classification has decreased.

The average f-measure outcomes obtained at the end of the learning process in the fourth experiment are shown in Figure – 4. According to these outcomes, the comparison of the classifiers and the evaluation of the performances are conducted.

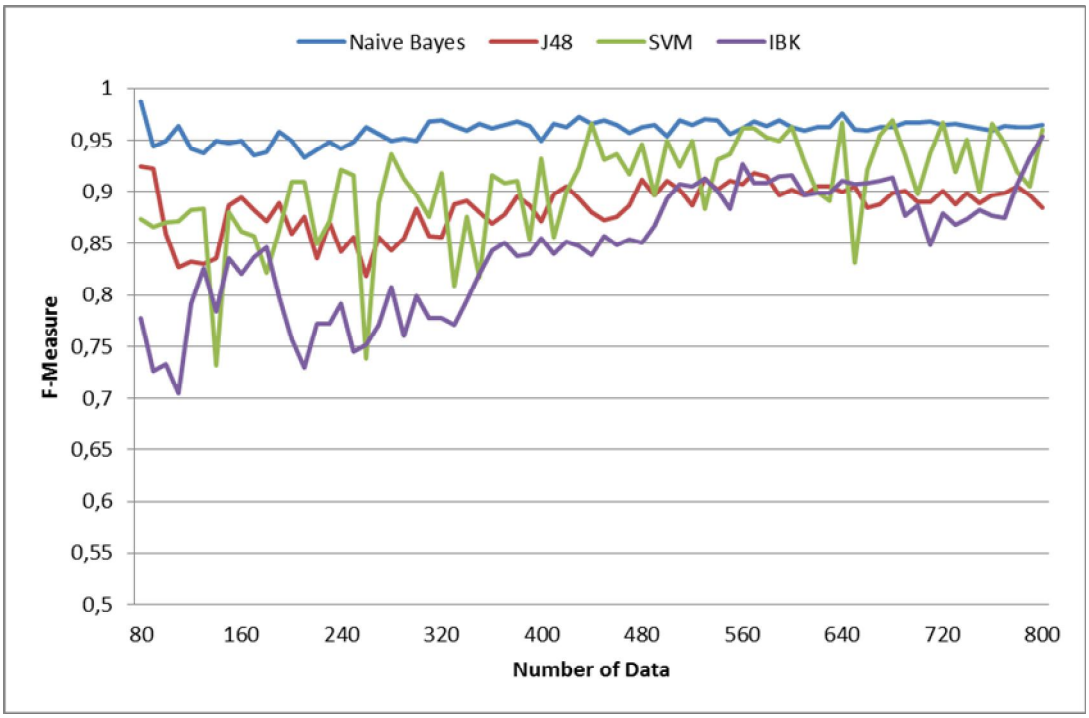

Figure – 4 Experiment 4 Result

In Table – 5, the highest and the lowest accuracy rates of the classifiers used in the experiment 4 at the end of the learning process are shown. According to this table, the Naive Bayes classifier's rate of classifying the samples accurately is higher than other classifiers.

Table – 5.The highest and the lowest accuracy rates of the classifiers according to the experiment 4

The percentage of the accurately classified samples	Naive Bayes	J48	SVM	IBK
Maximum (Max)	98,75	92,5	96,9118	95,375
Minimum (Min)	93,3333	81,9231	73,5714	72,7273

Words have been used instead of N-gram again and frequency value has been changed to 50 during the formation of arrf files. Naive Bayes has had the highest success and stable learning value among the algorithms. In the support vector machines, the waving lessens and the success

rate reaches the highest point. In the decision trees, the lowest success rate is seen at the experiment 4; however, learning continues in an increasing manner after a certain point. K nearest neighbor algorithm's success is as high as it was in the experiment 2. This algorithm has shown a successful increase starting from the lowest learning level. The value of K nearest neighbor point has been chosen as K = 12. As stated in Table – 1, the success of the k nearest neighbor increased when the k value was increased to 50 in the formation arrf files of the experiment.

4. Discussion

There exist the following problems and discussions for the algorithms to be used in the DLP text classification architecture.

The differences among the algorithms may be due to the following reasons;

- Carrying out functions affecting the model extraction in the operations like the qualification choice made at the level of data preprocessing and data completion, which can influence analysis outcomes.
- The data formed in different ways of preprocessing with different analysis results.

The factors affecting the classification can be as followings;

- Differences in algorithms,
- Features special for the data set,
- Incompatibility between the method and the problems.

The features special for the data set can be as followings;

- Class ambiguity,
- Insufficient number of samples

Class ambiguity indicates the situations in which no distinction can be made with the features given within the classification problem using any classification algorithm. Another factor which makes classification more difficult is the scarcity of data. Classifying the situations which are not exemplified with enough examples to limit the generalization mechanism of the classifiers is most likely to be randomly done. Naïve Bayes, which is a linear classifier with normal distribution assumption [5], can turn into a nonlinear classifier when a kernel density estimator is used [6].

5. Conclusion

Naive Bayes has given the highest learning of DLP text classification architect using the words weighting. However, the high learning rate has not changed for the words weighting which was prepared with an altered k value. The happy graph of the k nearest neighbor classifier has been achieved at the intended level on the education data prepared with the experiment 4 in the DLP text classification architecture. Yet, its success level does not prove as successful as that of Naive Bayes.

When all these outcomes have been analyzed, it has been determined that experiment 4 environment and Naive Bayes classifier will be based in the DLP to be developed due to the fact that Naive Bayes classifier has achieved 98.75 accuracy over the education data prepared with words feature extraction method and the f-measure/data graph, or the happy graph, has a less wavy structure. Therefore, during the preparation of education data, words will be chosen as the feature extraction method, idf will be chosen as the weighting method, 50 will be the frequency value, and finally 50 will be for k, which is the repetition number. Following that, Naive Bayes has been employed as a classifier in learning education data process and assumed parameters of this classifier has been using in the WEKA environment. Besides these, happy graph provides us with important information about the performance of the classifier. In parallel to this, with 550 education data instead of 800, the previous accuracy level of the classified has been approximately caught and there has been a considerable increase in its performance.

References

[1] Jackson P. & Moulinier I. (2002). "Natural language processing for online applications: text retrieval, extraction, and categorization". Amsterdam.

[2] Weka Project – http://sourceforge.net/projects/weka/

[3] Amasyalı, M. F., Davletov, F., Torayew, A., & Çiftçi, Ü. (2010). "text2arff: Türkçe Metinler İçin Özellik Çıkarım Yazılımı". SİU 2010. Diyarbakır.

[4] Torture Archive – http://www.aladin0.wrlc.org/gsdl/cgi-bin/library?c=torture&a=q

[5] Manning, C. D., Raghavan, P., & Schütze, H. (2008). "Introduction to Information Retrieval". Cambridge University Press. New York.

[6] John, G. H., & Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345.