TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209~218
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i1.648

Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine

Imam Much Ibnu Subroto*[1], Ali Selamat[2]
[1]Department of Informatics Engineering, Universitas Islam Sultan Agung, Indonesia
[2]Faculty of Computing, Universiti Teknologi Malaysia
*Corresponding author, e-mail: imam@unissula.ac.id, aselamat@utm.my

Abstract

Currently, most of the plagiarism detections are using similarity measurement techniques. Basically, a pair of similar sentences describes the same idea. However, not all like that, there are also sentences that are similar but have opposite meanings. This is one problem that is not easily solved by use of the technique similarity. Determination of dubious value similarity threshold on similarity method is another problem. The plagiarism threshold was adjustable, but it means uncertainty. Another problem, although the rules of plagiarism can be understood together but in practice, some people have a different opinion in determining a document, whether or not classified as plagiarism. Of the three problems, a statistical approach could possibly be the most appropriate solution. Machine learning methods like k-nearest neighbors (KNN), support vector machine (SVM), artificial neural networks (ANN) is a technique that is commonly used in solving the problem based on statistical data. This method of learning process based on statistical data to be smart resembling intelligence experts. In this case, plagiarism is data that has been validated by experts. This paper offers a hybrid approach of SVM method for detecting plagiarism. The data collection method in this work using an Internet search to ensure that a document is in the detection is up-to-date. The measurement results based on accuracy, precision and recall show that the hybrid machine learning does not always result in better performance. There is no better and vice versa. Overall testing of the four hybrid combinations concluded that the hybrid ANN-SVM method is the best performance in the case of plagiarism.

Keywords*: plagiarism detection, machine learning, k-nearest neighbors, artificial neural network, support vector machine*

1. Introduction

It is no doubt that the number of documents on the Internet is increasing every second. And with the tremendous increase in the size of the document is not difficult for anyone to seek and obtain the necessary documents quickly and accurately. In fact, it also applies to the uploaded documents on the internet which is relatively new. Actually, that is the contribution of search engine technology that is increasingly mature. Search engines make documents more easily and more quickly searched. This kind of situation also means that the opportunity for someone to cheat by way of plagiarism is also increasing. Some people use other people's ideas through his writings and claim as his idea and his worked. Of course, this kind of bad attitude, cannot be justified, and it is categorized as a crime of plagiarism. Plagiarism has always been considered as a serious problem so it becomes very important to prevent the recognition of copyright. That is why many of the techniques and technology offered for the prevention. Several tools have been commercialized as is quite famous, for example, is Turnitin [1] and ithenticate. There are also some tools that are free such as viper, plagiarismdetect, plagium and Plagiarism Checker with all its pros and cons. In general, almost all existing plagiarism detection technique is based on measuring similarity that is both local and global similarity [2].

Generally, cases of plagiarism can be solved by similarity approach, because a pair of similar sentences generally has the same idea. As known, the core problem rather than a crime of plagiarism is theft of ideas. However, there are also cases where things are not valid; for example, there is one word in the negation of the sentence that could mean the opposite idea. Another problem is related to the output of the similarity measurement. Similarity measurement produces a value between 0 and 1, where a value of 1 means 100% similarity value. Thus there must be a minimum value (threshold) to determine a document said to be plagiarism. The

Received September 13, 2013; Revised December 26, 2013; Accepted January 8, 2014

judgment of the plagiarism is difficult to do, although it could be done with the survey or the measurement precision and recall. Indeed, the advantages can be carried adjustment, but it also could be due to a deficiency point uncertainty. Another common problem is the definition of plagiarism that existed at the intersection. Inconsistency occurs expert in judging a case is considered plagiarism. Once a case is regarded as plagiarism, while in other very similar cases regarded as not plagiarism. In our opinion, all of three problems may be solved by a statistical approach. Much machine learning can be a solution to the statistical problem.

This paper offers a different approach that is machine learning. Machine learning is a classifier based on learning system using empirical data, which have been validated by experts. Based on the learning results will be obtained by a set of mathematical formulas with few constants, and variables are referred to as a model. Furthermore, this is considered a model of intelligence and will be used in general to classify cases other.

Consideration of machine learning approach is based on the assumption that although the theory of plagiarism has been understood by the experts, but in reality, they often differ in the same case. For example, a sentence that has a certain resemblance interpreted as a plagiarism by an expert, but it turns out; there are other experts disagree with regard not as plagiarism. For example, there are two very similar sentences in the text but one of them contains the word "not". The word "not" of course meaning is the opposite or no different than the "not" its. This can be confusing. This is just one example of the many examples of inequality perceptions of plagiarism. With accommodate these differences are expected to machine learning approaches could be the solution.

In general, the source used in the detection of plagiarism comes from two different sources. The first source is the primary repository that has built itself where all documents are collected, processed, and indexed so that it becomes easy and suitable for the purpose of detecting plagiarism. The second source is derived from secondary sources that the search engines are already proven mature like google, yahoo and others. The first source would be faster in the detection process but requires a resource that is large enough to hold the data that is very large and very fast growing. Turnitin is the one that uses this type of source. While the second source is a bit slow in the detection process for using secondary data and existence of additional processes to ensure cases of plagiarism. However, both types of sources are better able to detect new uploaded documents on the internet because it is generally owned by the search engine crawler is able to update the documents quickly. With these advantages in addition to the resources used in this experiment is the second kind of the Internet.

2. Machine Learning and Similarity Approach

Most of the existing techniques in detecting plagiarism using the similarity measurement approach. This technique is actually similar to the technique used in information retrieval (IR) is to determine the rank retrieval based on measuring the similarity to a query.

The similarity-based plagiarism detection can be divided into three groups, namely text-based similarity (Cosine, fingerprint, etc.)[3, 4], graph similarity (ontology, etc.)[5, 6], and line matching (bioinformatics, etc.)[7]. All techniques are based on similarity measurements that return a degree value of similarity from 0 to 1. A value equal to 1 is the greatest value of the mean level of similarity is 100%, the smaller value means getting away from a similar thing. The problem is the number of how many digits an appropriate degree of similarity can be regarded as plagiarism? Value of 90% may be considered as the minimum threshold of plagiarism but could also 95% if you want a higher level of similarity as a category of plagiarism. It means the measurement of the similarity-based plagiarism requires similarity threshold adjustment [8].

However, the threshold value is not required in the machine learning approach. The level of similarity in machine learning has been never visible because of plagiarism decision depends on the outcome of learning from the experts that have been presented in a numeric value. Experts have been asked to assess a number of comparisons couple sentences, and then he had to determine whether it is in the category of plagiarism or not. The decision data will be an intelligence of machine learning.

Several machine learning techniques that has proven its performance are k-Nearest Neighbors (KNN), Support Vector Machine Learning (SVM), and Artificial Neural Network (ANN). KNN is a simple theory but has proved very good accuracy. This technique will categorize a member of X based on its nearest neighbors. The number of nearest neighbors (n)

which determines the classification results are generally more than one but also not be too much. To optimize the accuracy of the variation in the value of n needs to be tried. If X has some neighbors are mostly in the category of plagiarism then X is a member of plagiarism, and vice versa. Although KNN has a high accuracy but this technique is slow because of high computing to calculate the distance to all neighbouring members. That is why this technique is also known as lazy classifier.

SVM is a classifier that also proven has good performance, especially in cases related to text. This technique is based on a statistical approach. Basically, the formula of this technique is to find the boundary between the two classes. The learning technique is to find an optimal threshold for each class with the goal furthest from the two boundaries.

The set of coordinates that determining boundary is exactly hereinafter called the support vectors. In this case the two classes plagiarism are plagiarism class and not plagiarism class.

ANN is a learning classifier based on the amount of data by modeling the brain works with mathematical models, in this case the data plagiarism and non-plagiarism. The workings of this machine refer to the relationship between neurons with other neurons in the other layer. Mathematically, a function neuron network $f(x)$ is defined as the composition of other functions $g_i(x)$. This in turn can be defined as a function of composition between interdependent. It really depends on how the network structure is designed that describes the relationship between the dependency. In general, the most widely used is the nonlinear weighted sum as in the formula below. Where K is the activation function, for example using the hyperbolic tangent, as simply a vector $g = (g_1, g_2, \ldots, g_n)$.

$$f(x) = K(\sum_i w_i g_i(x)) \quad (1)$$

A common use of the phrase ANN model really means the definition of a class of such functions (where members of the class are obtained by varying parameters, connection weights, or specifics of the architecture such as the number of neurons or their connectivity).

Some hybrid method proved successful in improving performer in some cases, such as KNN hybrid language for detection [9].

Work in this paper will prove that the only cases of plagiarism cannot be solved by similarity measurement approaches but also can be solved by machine learning approaches. This paper will show the experimental results of three single learning machines KNN, SVM and ANN like some hybrid of the single engine is. The purpose of the hybrid is to get a better performance than a single machine, especially on plagiarism cases.

3. Plagiarism Data Representation

A sentence generally contains an idea that will be delivered[10]. It is common in natural language. This reason is strong enough to construct machine learning algorithms based on a comparison of the level of the sentence.

The data is taken from a comparison of the query sentence with another sentence, which is the result of an Internet search. From the data of the search results are then validated by experts who understand the definition of plagiarism category well. With this method of data collection as we expect the experimental results are not much different from that resulting at the time of real application of Internet-based plagiarism detection. Figure 1 summarizes the validation process starts with the collection of data with the sample input document in a standard format PDF, DOC and TXT. After going through the process of pre-processing will produce a bunch of sentences that will be validated by experts. Experts mark next section for suspected plagiarism and what is not suspected further.

Including unnecessary suspicion is the bibliography, the phrase in quotation marks, numeric data, images, and special characters that do not form a sentence. Data validation result is beneficial for the filter engine to throw in a process that does not detect plagiarism as mentioned above. Output of this process is a collection of sentences that will be used as a query in a search engine to find some sentences potential plagiarism.

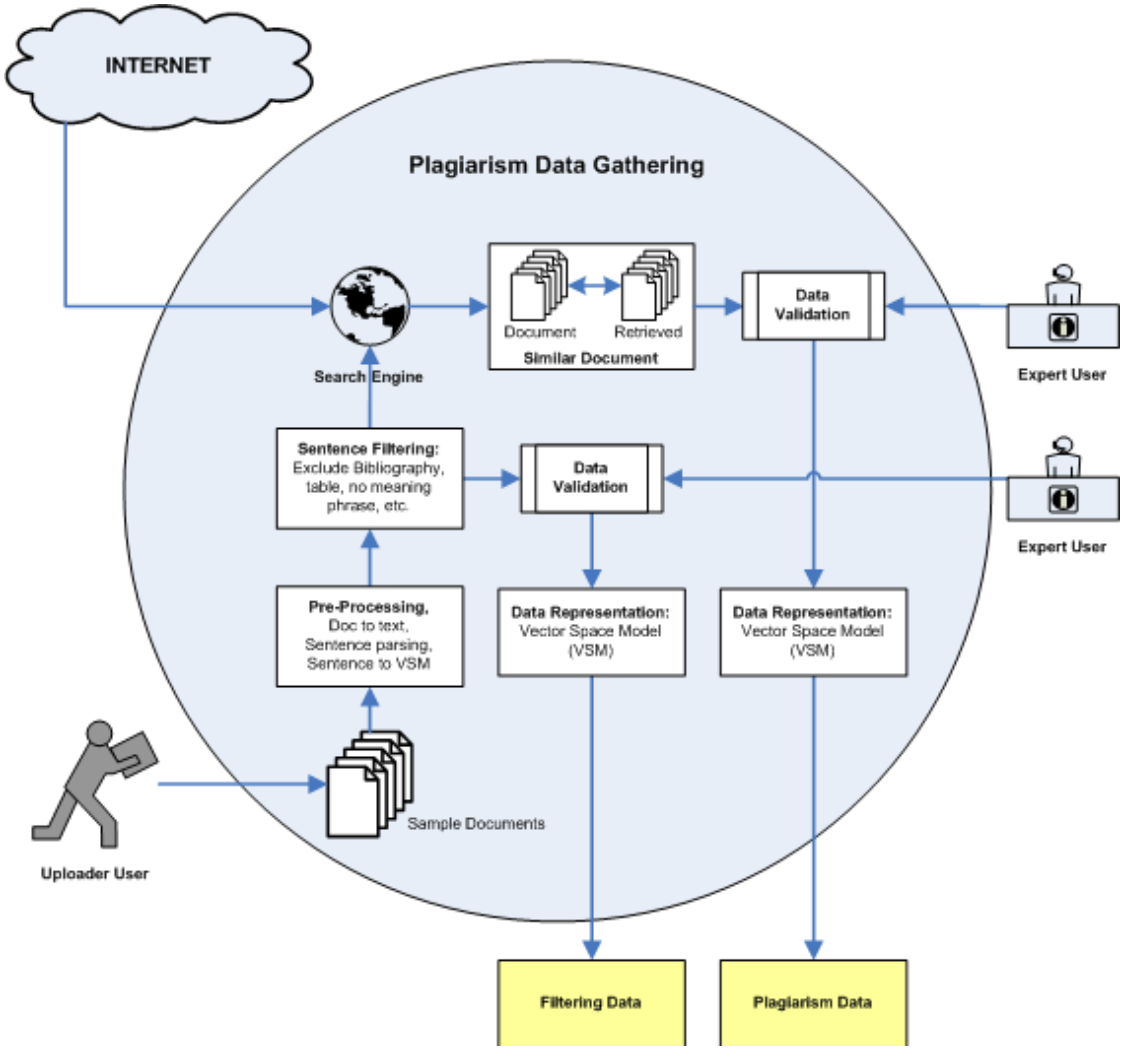

Figure 1. Plagiarism Validation

Pre-processing

Data sentence validated by experts further processed with the tokenization. Generally, the processing of a sentence using the standard text in natural language programming (NLP) is a word stemming and stopping removal. Synonym search process is also applied in this work.

As to the machine learning data are represented in the vector space model (VSM). It is standard technique in Information Retrieval[11]. Slightly different from other learning machines, which features a numeric VSM is calculated based on the features of a data query, but in the case of plagiarism, the data represented by the formula of a combined two sentences. This combined show of relationships between two sentences of suspected plagiarism. The process of how the merger of the two climates has been illustrated in Figure 2. The function F (x) = {s, d} is a relation between sentences comparator (s) of the suspected sources of plagiarism sentence (d).

The more similar words in a sentence with the sentence comparison is then possible to plagiarism will be even greater. In this work this feature is used with the range between 0 and 1. Maximum value 1 if a sentence is all token are the same/similar to the comparator, and will be zero if none of with the sentences the same token comparison.

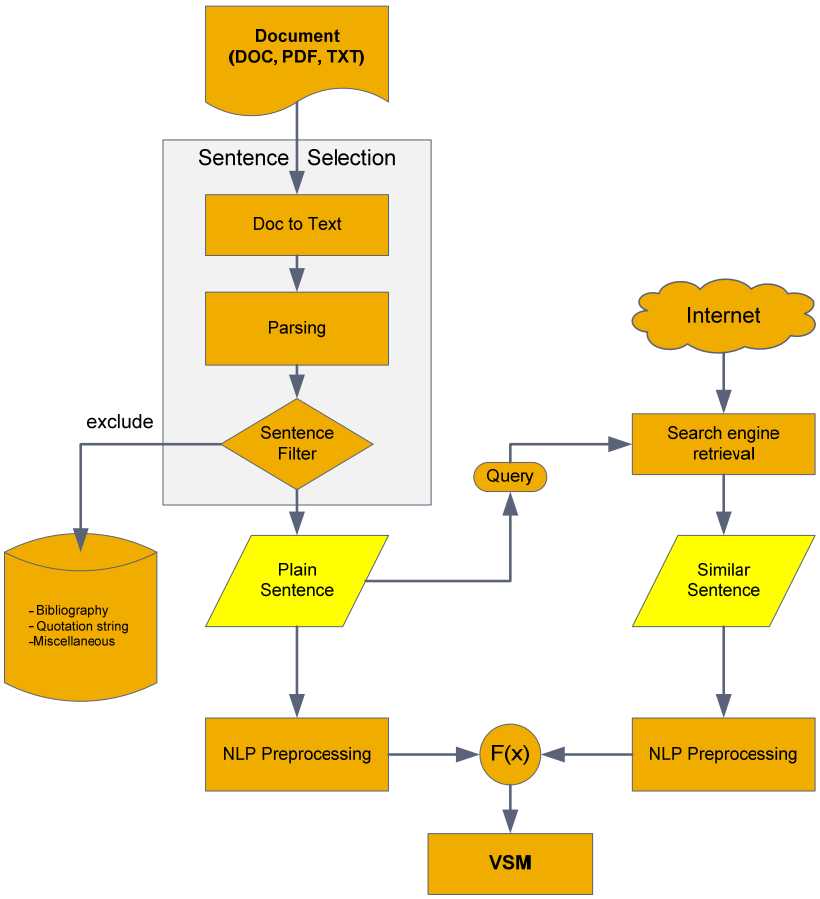

Figure 2. Plagiarism Data Preprocessing

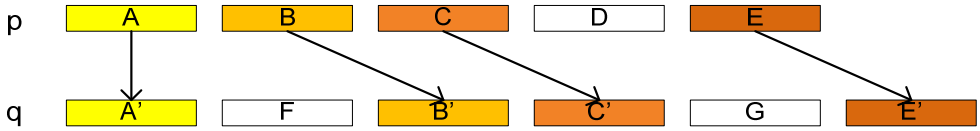

Figure 3. Density Map

The graph in Figure 3 tells the mapping of our plagiarism data collection to the frequency of same / similar token and token density. Dark blue indicates plagiarism while a light yellow color indicates no plagiarism. From the graph appears that in general the more words / tokens are similar then chances are detected as plagiarism will be greater but there is no guarantee that it is definitely plagiarism. There are even cases where all the tokens in a sentence there are similarities with the tokens in other words (frequency = 1) but is not considered as plagiarism. This is particularly likely to occur if the value of the tokens density comparison is small. Density calculations illustrated with the picture and formulas in Figure 4 and equation 2.

$$\delta(q,p) = \sqrt{\frac{\sum_{i=1}^{N}(q_i - p_i)^2}{N-1}} \quad (2)$$

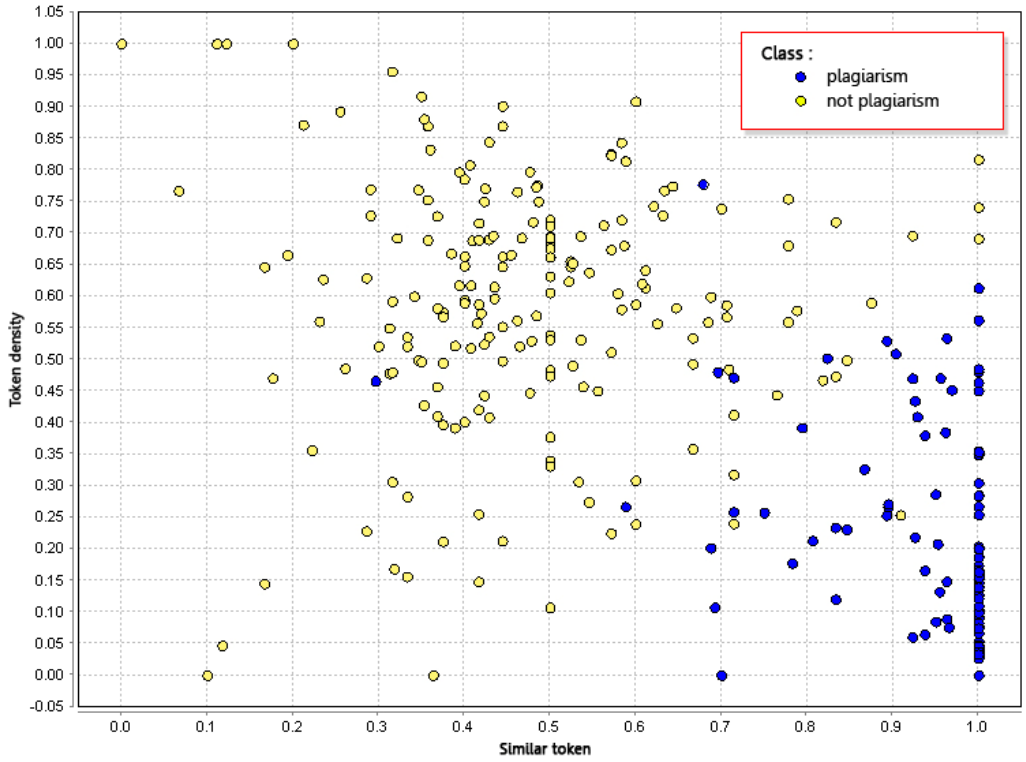

Figure 4. Plagiarism Data Map

Figure 4 describes the relationship between the two sentences p and q. There are 4 the same token, namely A, B, C, and E show the frequency same/similar token. Similar token means synonymous. The frequency normalization produces a value between zero and one. In this experiment the frequency used as the first feature. While the second feature is called with the density. The density describes the relationship between the densities of the two groups of tokens. The relationship of this density can also be calculated based on the two token groups. Formula eucledian distance is commonly used, as in equation 2. Density value of zero means the two tokens groups are identic, while the density is close to 1 means that the two groups are very different density. Two features are enough to detect the presence of elements in a document plagiarism.

4. Hybrid Machine Learning

Machine learning KNN, SVM, ANN has two inputs in the form of training dataset and testing dataset, with the same data format that is VSM (vector space model). The machine output is a classifier which is usually called a model. The KNN model is a reduction of the training set, the SVM model is a collection of vectors, and ANN models is the structure of neurons and their weights. The same input data format of them make relatively easy to combine between single machine learning. This work will experiment four combinations that are KNN-SVM, KNN-ANN, SVM-ANN and ANN-SVM.

Figure 5 shows the design of hybrid A-B of two learning machines A and B. Machine A is a machine that performs the training set to get the optimal value in the form of prediction sets. Based on this prediction set that will be found some results are not the same prediction with input from the testing dataset that called error. Error has two possibilities, namely real error and misclassification of data. Real error means the results of the validation expert is right, but the machine is not able to predict accurately the data testing. While misclassification of data is very likely to occur in cases of plagiarism due to two things, namely due to lack or inconsistency carefully situations. Inaccuracy due to lack of care-giving an expert in plagiarism signed during data validation. Inconsistency occurs if it finds dubious cases, so when encountering other

similar cases sometimes do a different decision. In short misclassification is human error. Assuming that the error containing a mixture of misclassification, the predicted result that was reduced will be used as a training dataset again on machine B. This is the most important part of a hybrid system that will improve the performance of the classifier.

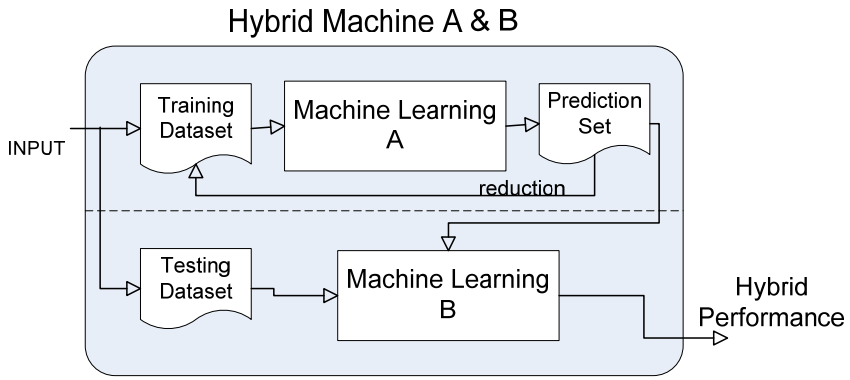

Figure 5. Hybrid Machine Learning [A-B]

With a training dataset that is relatively cleaner from misclassification, the results of training on machine B is expected to produce a better performance of classifier model. Hybrid Machine A-B will determine the overall accuracy from the classifier system while machine B is the part that determines the speed of classification of each other similar case X. That is why in this experiment did not include a hybrid ANN-KNN and SVM-KNN caused KNN is a lazy classifier. Explanation of Figure 4 will be described by the following algorithm model.

```
01      D = load dataset();
02      split_dataset(D): [Training_dataset, Testing_dataset]
03      while(error > 0){
04          learn(machine_learning_A,Training_dataset) : Prediction_dataset, Model_A;
05          New_traning_dataset = remove_error(Prediction_dataset);
06          error = calculate_error(Prediction_datase);
07          Training_dataset = New_training_dataset;
08          Model_A = save_machine_learning_model(model_A);
09      }
10      Learn(machine_learning_B,Training_dataset,Model_A): model_B,Prediction_B
11      Predict(Testing_dataset,model_B): Prediction_B;
12      Validate(Prediction_B): Accuracy, Precision, Recall;
```

Table 1 shows the performance of several experiments conducted. All the experiments show that the hybrid method is better than the performance of single learning machines premises. In some experiments we did not hybridise it ANN-KNN and SVM-KNN, because KNN slow in doing so it does not fit that classification is placed at the tip when the tip of a hybrid is a hybrid that determines the speed of the classification process. Hybrid ANN-SVM shows the best performance with an accuracy of 97.6%, a precision of 98.92% and recall of 97.37%.

Table 1. Machine Learning Performance

Method	Accuracy	Precision	Recall
KNN	94.67	95.46	96.32
SVM	95.01	96.84	96.32
ANN	96.01	96.53	97.37
KNN-SVM	95.67	95.43	95.26
KNN-ANN	97	96.29	97.25
SVM-ANN	95.67	95.43	95.26
ANN-SVM	97.6	98.92	97.37

Precision denominates what percentage of all instances that a detection method reports as suspicious are plagiarism. Recall denominates what percentage of all plagiarized instances in the collection a detection method reports.

Figure 6 shows a comparison of the performance of accuracy, precision, and recall of machine learning SVM with hybrid KNN-SVM and ANN-SVM. In the hybrid architecture seen that SVM always in the back position, this means that the SVM is the main engine of a hybrid, while the front machine position is supporting. Pure SVM actually be quite good performance when looking at that the accuracy of 95%. Even in terms of precision and recall turned out better than hybrid KNN-SVM. Although the hybrid has better accuracy. This shows that the hybrid is not necessarily better than pure machine learning. Hybrid architecture largely determines the overall performance. Seeing the overall variation of the hybrid SVM can be concluded that ANN is a very significant support to improve the performance of SVM, especially in cases of plagiarism. Hybrid ANN-SVM has the overall best performance of the three. With accuracy, precision, recall all of the above 97% then the hybrid is said to be the best of hybrid SVM.

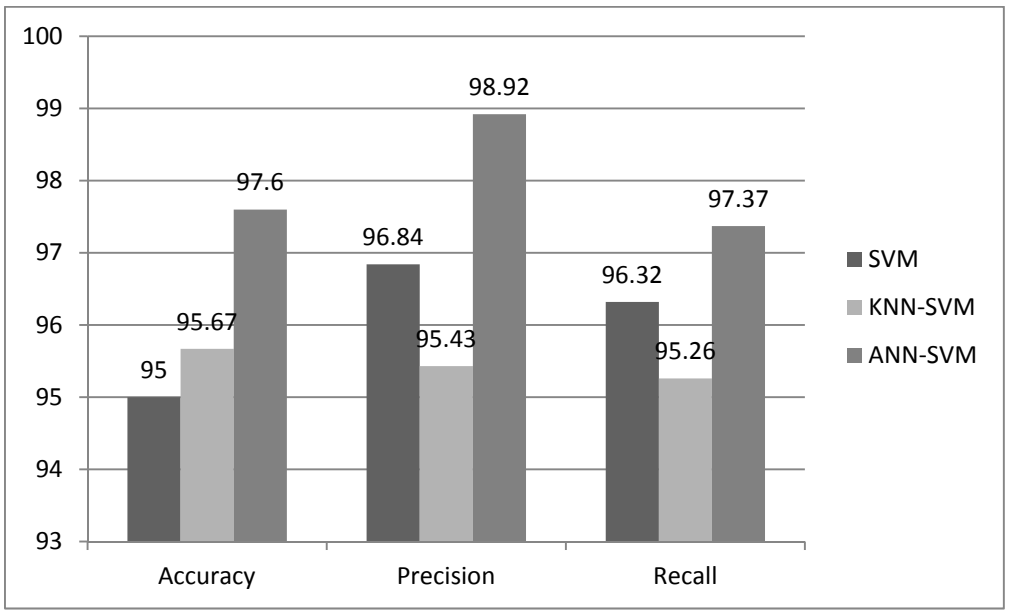

Figure 6. Accuracy, Precision and Recall of Hybrid SVM Method

Figure 7 illustrates the comparison of the performance of the hybrid SVM method. Here, accuracy, precision, and recall pure ANN compared with the hybrid of two, KNN-ANN and SVM-ANN. Like the earlier discussion that the hybrid method is not necessarily improve the performance proved here. Hybrid SVM-ANN produces no better performance than a pure ANN, be it accuracy, precision, and recall. The graph shows that the hybrid KNN-SVM method successfully improves the performance of ANN. KNN-ANN accuracy was the best of the other while the value of precision and recall almost the same as pure ANN. This graph concludes that the best performance of hybrid ANN method is a hybrid method of KNN-ANN.

Figure 7 shows a comparison of the best hybrids ANN with the best hybrid SVM. As a result, the hybrid ANN-SVM method is superior in every way. With an accuracy of 97.6%, the precision 98.92% and recall of 97.37%, we conclude that the hybrid ANN-SVM is the best in the experiments that have been carried out. ANN-SVM is able to improve the performance of pure SVM and pure ANN.

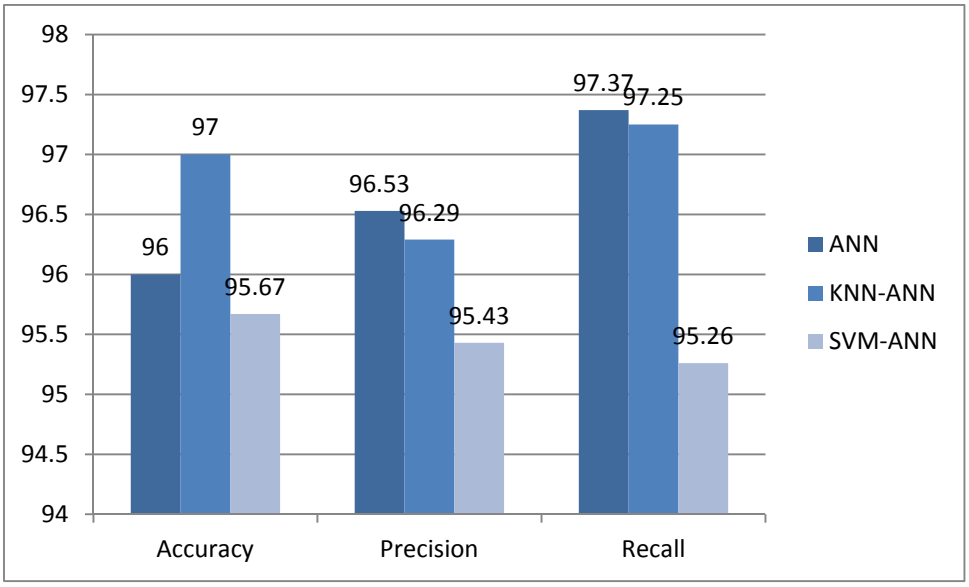

Figure 7. Accuracy, Precision and Recall of Hybrid SVM Method

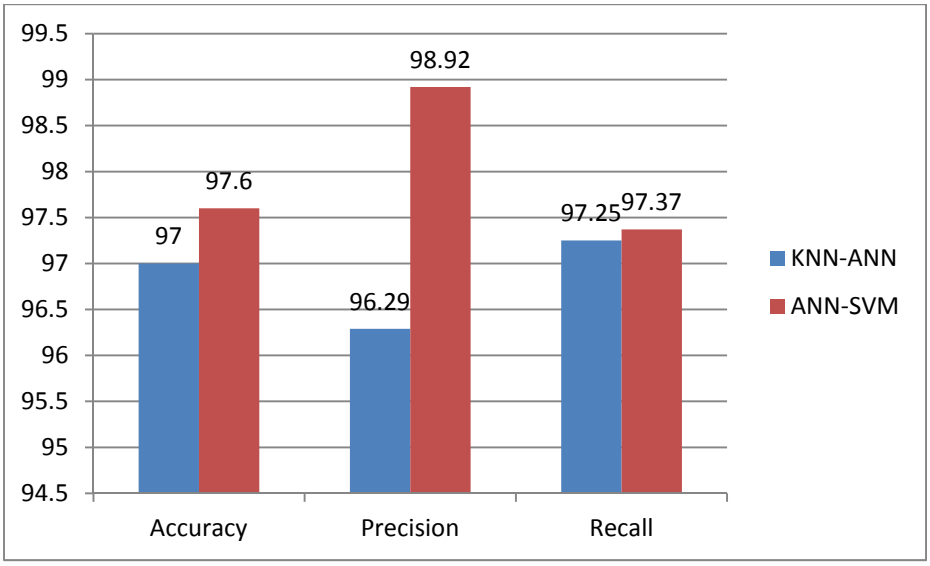

Figure 7. Hybrid ANN vs. Hybrid SVM Method

5. Conclusion

The general conclusion from the results is that the machine learning eksmerimen suited to solving detection of plagiarism. This is shown by the results of all methods of learning machines that do produce an average value above 90%.

The experimental results show that in general there is improvement performance in the use of hybrid machine learning methods in the case of plagiarism. In terms of accuracy, precision and recall are better than pure KNN, SVM, ANN. However, the hybrid method does not always produce better performance, can even reduce performance. As happened in the performance of SVM-ANN which is worse than a pure ANN. Comparison of all methods in this work, can be concluded that the hybrid ANN-SVM method with the best performance techniques.

This paper shows that the search engines can be optimized functions for detecting plagiarism, ie. by adding computing machine learning in it. It has been proved with the results of this experiment in which the use of machine learning has an average high accuracy.

References

[1] M Bill. Turnitin.com and the scriptural enterprise of plagiarism detection. *Computers and Composition.* 2004; 21: 427-438.
[2] N Meuschke, B Gipp. State of the Art in Detecting Academic Plagiarism. *International Journal for Educational Integrity*. 2013; 9(1): 50-71.
[3] MS Pera, Y.-K. Ng. SimPaD: A word-similarity sentence-based plagiarism detection tool on Web documents. *Web Intelli. and Agent Sys.* 2011; 9: 27-41.
[4] N Gustafson*, et al. Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity*. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '08)*.* 2008; 690-696.
[5] P Foudeh, N. Salim. A Holistic Approach to Duplicate Publication and Plagiarism Detection Using Probabilistic Ontologies. Advanced Machine Learning Technologies and Applications (A. Hassanien, et al., Eds.). Springer Berlin Heidelberg. 2012; 322: 566-574.
[6] C Liu *et al. GPLAG: detection of software plagiarism by program dependence graph analysis*. *12th* ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA. 2006.
[7] C Xin *et al.* Shared information and program plagiarism detection. *IEEE Transactions on Information Theory.* 2004; 50: 1545-1551.
[8] C.-Y. Chen *et al.* Plagiarism Detection using ROUGE and WordNet. *Journal of* Computing. 2010; 2(3): 34-44.
[9] A Selamat*,* IMI Subroto, Choon-Ching Ng. Arabic Script Web Page Language Identification Using Hybrid-KNN Method. *International Journal of Computational Intelligence and Applications. 2009;* 8(3): 315-343.
[10] DR White, MS Joy. Sentence-based natural language plagiarism detection. *J. Educ. Resour. Comput.* 2004; 4(4): 1-20.
[11] A Ekbal *et al.* Plagiarism detection in text using Vector Space Model. *12th International Conference on Hybrid Intelligent Systems (HIS)*. 2012: 366-371.