CLASSIFICATION OF NEWS ARTICLES USING K-NEAREST NEIGHBOR BASED ON PARTICLE SWARM OPTIMIZATION

Afdhalul Ihsan[1], Ednawati Rainarli[2]
Teknik Informatika-Universitas Komputer Indonesia
Jalan Dipatiukur No.112-116 Bandung, 40312
Email : afdhalulihsan@email.unikom.ac.id,ednawati.rainarli@email.unikom.ac.id

ABSTRACT

This study will classify the news by implementing the K-Nearest Neighbor algorithm based on Particle swarm optimization. Particle Swarm Optimization is used as a method for selecting features. This study uses a dataset of 250 news documents, 25 documents as test data and 225 documents as training data. Stages in this study are divided into 2 training and testing process. Based on the results of the study showed that the use of Particle Swarm Optimization as a selection feature can provide better accuracy results compared to using only the K-Nearest Neighbor method. Accuracy results obtained after selecting features using particle swarm optimization by 80% with K value = 9, many particles are raised as many as 50, the value of c1 and c2 = 1 and iteration of 100. In this study the particle swarm optimization method has not met the stop conditions determined so that the results of the selection feature process are not optimal. Even though the process in particle swarm optimization has not met the stopping condition, the selection feature can increase the accuracy of the classification process by 20% compared to without using selection feature which only gets the highest accuracy by 60%.

Keyword *: K-Nearest Neighbor, Feature Selection, Particle Swarm Optimization, Dataset, Classification*

1. INTRODUCTION

News is information about an incident that contains facts in it. Nowadays, it is easier to gain access to news, for example through an online news portal. The numerous amount of news articles as well as the topics discussed in the news sometimes makes it difficult to find an article with the desired information or topic discussuion of interest. Therefore, we need a system that can identify the news article. In order to identify news articles, the classification method can be used.

Research that has discussed the classification of news documents, such as the research of Andi Ahmad Irfa and friends [1] who use the K-Nearest Neighbor method in classifying news article documents [1]. In research done by Andi Ahmad Irfa and his friends, the result obtained from the research conducted was an f-measure value of 69.9% [1]. Research on the classification of documents using the K-Nearest Neighbor method was also carried out by Claudio Fresta Suharno and his friends [2]. In the research conducted by Claudio, it explained the comparison of document classification using K-Nearest Neighbor with feature selection and without feature selection [2]. In Claudio's research, he explained that the use of feature selection in the classification stage gives better system accuracy results compared to document classification without feature selection.

Feature selection is the stage to choose the most important features in a data or document [2]. The purpose of feature selection is to improve the performance of document classification by removing features deemed irrelevant in the classification to reduce the dimensions of the feature set [2]. Feature selection is a very important step in optimizing the performance of the classification method [3].

There are currently many and varied methods for selecting features. Mehdi Hosseinzadeh Aghdam and Setareh Heidari conducted research on the comparison of 4 feature selection method, which in this study compared the methods of Particle Swarm Optimization, Information Gain, Chi-Square, and Genetic Algorithm. In Mehdi's reserach, K-Nearest Neighbor's research is used as a classification method. In the research, Mehdi found that the use of Particle Swarm Optimization as a feature selection can give better results compared to the use of another method. [4]. Based on that, in this study Particle Swarm Optimization and K-Nearest Neighbor will be used to classify the article topic from a news.

2. RESEARCH CONTENT

2.1. System Description

In this research, the constructed system is designed to be able to classify the topic of an article. The built system implemented the K-Nearest Neighbor and Particle Swarm Optimization methods. In order to build the system, there are two stages being involved, namely the training phase and the testing phase. The research phase is a feature search *phase by using Particle Swarm Optimization where the result of this training phase are a list of selected words based on Particle Swarm Optimization search. The result of the training phase will be used at the testing stage where the testing phase is the classification stage of news articles using Particle Swarm Optimization and K-Nearest Neighbor and in the testing stage will also measure the accuracy generated by the system.*

2.2. Document Classification

Document classification is one area of research in the acquisition of information that develops methods to determine or categorize a document into groups that have been previously known automatically [5]. Document classification aims to categorize unstructured documents into groups that describe the contents of the document [5].

2.3. K-Nearest Neighbor

K-Nearest Neighbor is one method that can be used to do clasification. K-Nearest Neighbor is a supervised learning method which means there are training data in the classification process, and the purpose of this method is to predict or map the data based on pre-existing training data [6]. K-Nearest Neighbor is a classification technique that performs the classification process by calculating the distance from the data to be tested against training data [6]. The value of the calculated distance will be used as the value of closeness or similarity between the test data and the training data. Measurement of proximity or similarity in the K-Nearest Neighbor method is calculated using the euclidean distance equation [6]. The formula for the Euclidean distance equation as follows :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)}^{2} \quad (2.1)$$

Where x and y are points in the n dimensional vector space, while xi and yi are scalar quantities for the i dimension in the n dimensional vector space.

2.4. Particle Swarm Optimization

PSO is a population-based solution search algorithm based on the behavior patterns of birds, bees or fishes that are moving in group [7]. Particle Swarm Optimization has the advantage of easiness to be implemented,and convergence that is fast and simple [8]. The word particle in Particle Swarm Optimization shows a bird in a flock, when one of the bird in that flock finds a short path to the food source, the rest of the other birds in the group will also immediately follow the path that the bird found even though the distance is far apart [8]. In particle swarm optimization each individual is represented by a vector in a multidimensional solution search space [7] .

2.5. Feature Selection

Feature selection is the process of reducing features from m features to as many as n features by removing non-informative features or by selecting the most informative features [9]. Some methods in feature selection are filter based, wrapper based, and hybrid.

2.6. Term weighting Tf-Idf

Tf-Idf Term Weighting is one of the feature extraction methods that is widely used in the document classification process [10]. Feature extraction is a step to convert text or documents into a more repressive form, for example vector [10]. The Tf-Idf method combines 2 concepts in calculating the weight of a term (word), which are calculating the frequency of occurrence of a word in one document, and inverse of the document containing the term (word).

To calculate the weight of a term (words) the following formula could be used.

:

$$W_t = Tf_{i,j} * idf_i \quad (2.2)$$

Term Frequency (Tf) is the frequency in which a term (word) appears in a document. To calculate Tf you can use the formula as follow:

$$Tf = \frac{n}{\sum n_k} \quad (2.3)$$

Where n is the number of occurrences of terms that will be weighted in a document and nk is the amount of words in the document.

Inverse Document Frequency (Idf) is a reduction in the dominance of the terms (words) in a document. This Idf calculation is necessary because in every document, terms (words) that appear in many documents will be considered as general terms [10]. To calculate IDF you can use the formula:

$$Idf = log\frac{n}{df} \quad (2.4)$$

Where :

Idf = inverse document frequency
n = number of document
df = the number of documents that have the word (term)

2.7. System Analysis

The system that will be built consists of 2 stages, namely the training and testing stages where each stage is divided into several stages. The general description of the system that is going to be built can be seen in Figure 1.

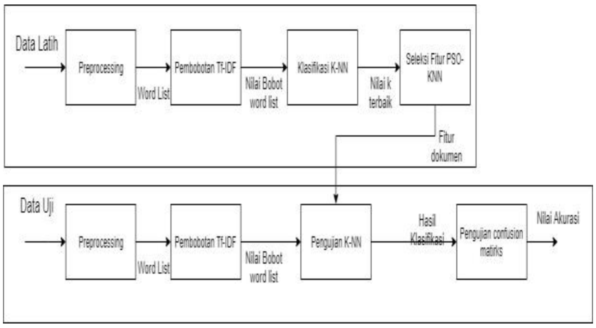


***Figure 1 System** Overview*

1. The first step taken is the training process. In the training process the training data will go through 4 main processes, namely the preprocessing process, TF-IDF weighting, K-NN classification, and PSO-KNN feature selection. The preprocessing process includes case folding, filtering, tokenizing, and stopword removal. In the TF-IDF weighting phase, weighting will be carried out on each word in the document. At the KNN classification stage a classification process will be carried out using the k-nearest neighbor method to find the best k value. The best k value obtained will be used as a parameter in feature selection using pso-knn. After getting the best k value, the feature selection process will be performed to look for features from the training data document that will be used as features during the testing proces.
2. The next stage taken is the testing process. In the testing process, it will go through 4 main processes, namely the pre-processing, weighting, KNN classification testing phase, and confusion matrix testing. The results of this testing phase are test documents that have been grouped using the k-Nearest Neighbor method. After testing the k-Nearest Neighbor classification, the system testing is done using a confussion matrix. The results of these tests are the accuracy of the k-Nearest Neighbor and Particle Swarm Optimization methods.

2.8. Data Input Analysis

Data input that will be used in the system that will be built is news article data that has the format *. Txt. Many data will be used as many as 250 documents 25 documents are used as training data and 225 documents as test data. In the input data will be classified into 5 classes, namely travel, techno, sports, health, automotive.

2.9. Training Stage Analysis

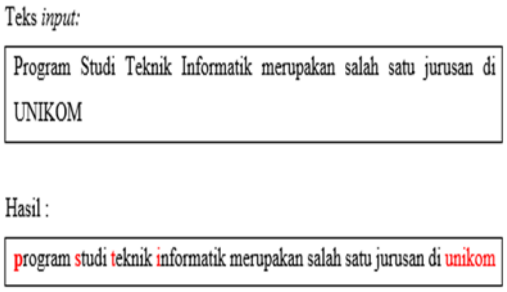
Analysis of the training stage is the stage for finding informative features that will be used in the process of classifying news articles. This stage consists of pre-processing, Tf-Idf weighting, KNN Classification, and PSO-KNN feature selection.

2.9.1. Preprocessing Phase

Preprocessing stage is the initial stage in the classification process. This stage is an important stage in the classification process. At the pre-processing stage there are 4 stages which are case folding, filtering, tokenizing, and stopword removal.

a) Case folding

Case folding is the stage for uniforming the shape of characters in a document. Uniformation can be done in the form of a lower case and upper case. An example in the case folding stage can be seen in Figure 2.



Teks *input:*

Program Studi Teknik Informatik merupakan salah satu jurusan di UNIKOM

Hasil :

program studi teknik informatik merupakan salah satu jurusan di unikom

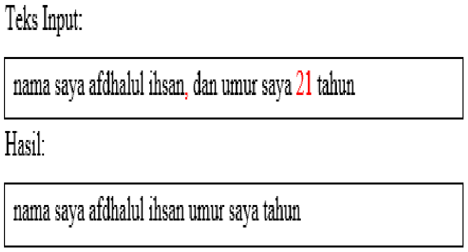
***Figure 2** Case Folding*

b) Filtering

The filtering stage is the stage for removing noise in the document. Noise here can be either punctuation or numbers included in the document. In the research, there are rules in the filtetring process such as:

1. Replacing delimiter character with space.
2. Deleting a number character where at this stage, every number character (0-9) will be deleted.
3. Elimination of spaces due to excessive space from the process of replacing delimiter characters that are converted into spaces, so that it will only leave 1 space separating each word

An example of the filtering stage can be seen in Figure 3.



Teks Input:

nama saya afdhalul ihsan, dan umur saya 21 tahun

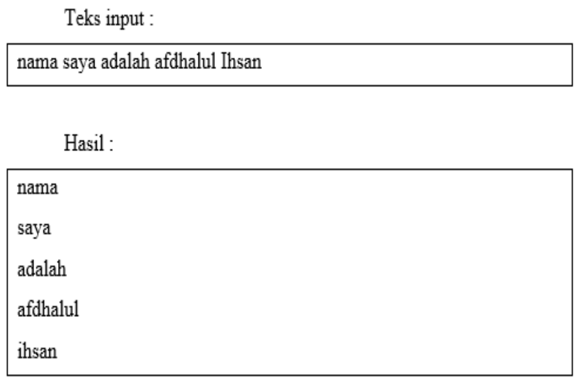
Hasil:

nama saya afdhalul ihsan umur saya tahun

***Figure 3** Filtering*

c) Tokenizing

Tokenizing is the stage of separating documents into terms (words) using a space separator ( ). An example of the tokenizing stage can be seen in Figure 4.



Teks input :

nama saya adalah afdhalul Ihsan

Hasil :

nama
saya
adalah
afdhalul
ihsan

Figure 4 *Tokenizing*

d) Stopword removal

Stopword removal is the stage of removing words that are considered irrelevant (unimportant) based on the stopword list. In this study the stopword list that is used is based on the stopword list found in the Sastrawy python library

.

2.9.2. Tf-Idf Weighting Phase

The Tf-idf weighting step is used to assign weight values to each term (word) contained in the document. Weight calculations are based on Tf and Idf calculations. The Tf calculation is calculated using equation 2.3. Idf calculations are calculated using equation 2.4. The weight calculation (w) is calculated using equation 2.2. The stages of Tf-Idf weighting can be seen in Figure 5.

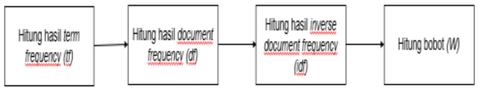

Hitung hasil term frequency (tf)
Hitung hasil document frequency (df)
Hitung hasil inverse document frequency (idf)
Hitung bobot (W)

Figure 5 *Tf-Idf Weighting*

2.9.3. KNN Classification Phase

KNN Classification Stage is the stage of classifying documents using KNN and all features. At this stage, it is aimed to find the value of k that can provide the best accuracy value. The k value that is obtained will be used as a parameter in the PSO-KNN feature selection. The stages of KNN classification can be seen in Figure 6

.

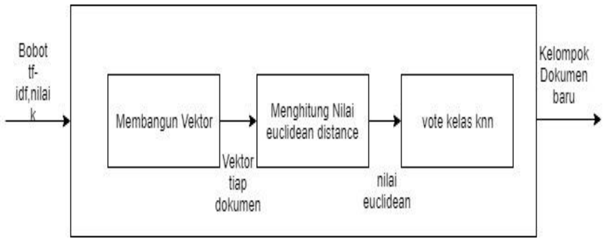


Figure 6 *KNN Classification Phase*

2.9.4. PSO-KNN Feature Selection Phase

The PSO-KNN selection stage is a stage of searching for features that are considered relevant in the classification process. In this study PSS-KNN selection will remove as many as 25% and 50% features. The stages of PSO-KNN feature selection can be seen in Figure 7.

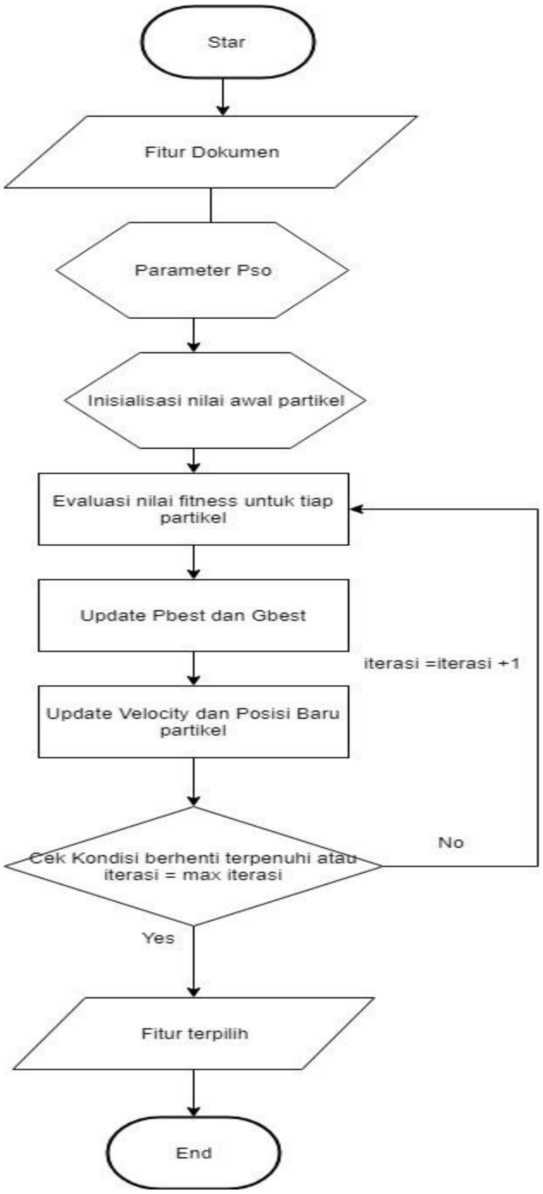

Star
Fitur Dokumen
Parameter Pso
Inisialisasi nilai awal partikel
Evaluasi nilai fitness untuk tiap partikel
Update Pbest dan Gbest
iterasi =iterasi +1
Update Velocity dan Posisi Baru partikel
Cek Kondisi berhenti terpenuhi atau iterasi = max iterasi
No
Yes
Fitur terpilih
End

Figure 7 *Feature Selection PSO-KNN*

At the selection stage of PSO feature, the data input that will be processed is a list of words contained in the documents. Feature selection using particle swarm optimization removes n words based on search results in the particle swarm optimization solution space. The stages in the search solution for particle swarm optimization are:

a) Generate PSO parameters

At this stage, the values that will become the parameter is raised in the search solutions using PSO. The parameters that will be generated *include C1 and C2 values, many particles will be used in each iteration process, the stop condition to be used and many iterations.*

b) *Initializing initial particle values*

At this stage, random initial values will be initialized, where each initial value will be initialized for each generated particle. Particles here are the representation of features in the document as many as n where n is the number of features contained in the dataset [9].

c) *Evaluate fitness value for each particle*

At this stage the particles that have been initialized will be evaluated using the objective function. Where in this evaluation phase, it will be tested whether the features selected by each particle will be evaluated using the objective function. To evaluate each particle it can be calculated using the formula [9]:

$$Fitness\ (xi) = ErrorRate \quad (2.5)$$

$$ErrorRate = \frac{FP+FN}{TP+TN+FP+FN} \quad (2.6)$$

d) *Upadate Pbest dan Gbest*

The Upadate stage pbest and gbest are aimed to update the pbest and gbest values by comparing the current pbest and pbest in the previous iteration. Pbest is the best position of each particle in the iteration process, while gbest is the best position of pbest in each iteration.

e) *Update Velocity*

Velocity update is a step to calculate velocity of each value on a particle. Velocity value is the value that will be used to calculate the position of new particles. To calculate velocity, the following formula could be used.

$$V_{id}{}^{t+1} = V_{id}{}^{t} + c1 * r1 * (pbest_{id} - x_{id}{}^{t}) + c2 * r2 * (gbest_{id} - x_{id}{}^{t}) \quad (2.7)$$

f) *Update Particle Position*

This stage is the stage to calculate the displacement of each particle. To calculate the displacement of each particle, the following formula could be used.

:

$$Xi(t+1) = Vi(t+1) + Xi(t) \quad (2.8)$$

g) *Check Stop Condition*

In the search for solutions using PSO, iteration will stop when the iteration process has met the stop condition or has reached the specified iteration limit [9]. If the process in an iteration meets the stop condition then the process will be stopped, and if it does not meet the stop condition repeat the process starting from the particle evaluation.

2.10. Analysis of Testing Phase

The testing phase involves the classification and measurement stage of the classification system accuracy using PSO-KNN. At the test stage the test data and training data will pass a similar stages as at the training stage. But at the time the KNN classification vector is built, there will be as many features selected based on the results of the PSO feature selection. Testing is done by testing the news document data which topic categories has not yet been known using the method k - Nearest Neighbor and Particle Swarm Optimization. The test data used in this research includes 25 news documents, while the training data used were 225 news documents. In this study, several methods will be performed in testing the performance of the classification process of the system being built, the test scenarios to be carried out are as follows :

1. *Testing the k - Nearest Neighbor method with the k value to be used is 1 to 10*
2. *Testing the feature selection method, at this stage the feature selection process is carried out in which 25% and 50% features will be removed. The number of particles raised in the feature selection search process using Particle Swarm Optimization is 15, 30, 50.*
3. *Classification testing using the k method - Nearest Neighbor with Particle Swarm Optimization feature selection, in this test all features that have been selected based on feature selection testing will be tested using k from 1 to 10.*

a) *KNN Testing*

In the KNN test the accuracy of the system produced in measuring the classification of documents without feature selection by using a k value of 1-10 will be measured. The results of the accuracy obtained for each k value can be seen in Table 1.

***Table 1** KNN Testing result*

Nilai K	*Accuracy*
1	*60%*
2	*60%*
3	*60%*
4	*44%*
5	*52%*
6	*56%*
7	*56%*
8	*56%*
9	*60%*
10	*60%*

As can be soon on Table , it 1 shows the results of testing with the k-Nearest Neighbor method and produces the greatest accuracy of 60% at k = 1, k = 2, k = 3, k = 9 and k = 10. This shows that the use of k value affects the level of accuracy produced, where the greater value of k decreases the optimacy of the result [1]. In the table above, the optimum accuracy at k = 1.2, and 3 has been obtained, and the greater the k used, the accuracy tends to decrease due to the increase in the value of k, the data also have no similarity to the test data [1].

b) Feature Selection PSO Testing

In the feature selection testing the system will be tested in the feature selection process and measure the fitness value obtained from each testing process. The results of fitness values obtained from each test can be seen in Table 2.

Table 2 Fitness value PSO testing

Partikel	Fitness	
	25% dibuang	50% dibuang
15	0,64	0,68
30	0,76	0,72
50	0,76	0,76

Based on Table 2 it can be seen that the raised number of the particles affecting the results of the fitness value of the optimization to search for the selected features. Where the more particles that are raised, a more optimal selected features can be produced, and it has a greater value of fitness. The number of particles involved in the process of finding features using particle swarm optimization provides many choices in each iteration that is done and from the many choices will then be chosen particles that provide the best fitness value [11].

c) PSO KNN Testing

PSO-KNN testing phase is a stage to measure the accuracy of selected features based on the selection of PSO features. In the PSO-KNN test the K value used is 1-10. The results of each PSO-KNN test can be seen in the following table:

Table 3 PSO-KNN 15 Particle Testing

Nilai K	15 Partikel	
	25% dibuang	50% dibuang
	f =0,64	f =0,68
1	60%	56%
2	60%	56%
3	60%	56%
4	48%	60%
5	56%	56%
6	68%	56%
7	60%	56%
8	56%	60%
9	56%	56%
10	64%	68%

Table 4PSO-KNN 3Particle Testing

Nilai K	30 Partikel	
	25% dibuang	50% dibuang
	f=0,76	f=0,72
1	56%	40%
2	56%	40%
3	32%	40%
4	40%	52%
5	44%	64%
6	52%	68%
7	60%	56%
8	76%	60%
9	60%	72%
10	76%	72%

Table 5 PSO-KNN 50 Particle Testing

Nilai K	50 Partikel	
	25% dibuang	50% dibuang
	f=0,76	f=0,76
1	52%	56%
2	52%	56%
3	44%	56%
4	48%	56%
5	52%	52%
6	56%	52%
7	68%	56%
8	72%	56%
9	80%	80%

	50 Partikel	
	25% dibuang	*50% dibuang*
Nilai K	*f=0,76*	*f=0,76*
10	*76%*	*76%*

Based on Table 3-5, it can be seen that the higher the fitness value that is generated from the selection of particle swarm optimization features, it can produce the higher accuracy value. In Table 4.9 the highest accuracy is obtained in the PSO-KNN test using K = 9 which resulted in 80% accuracy where the results are obtained from the features removed as much as 25% and 50% and as many particles raised as many as 50 with the fitness value obtained at 0.76. In the test of 15 particles generated accuracy only reached 68% for the use of K = 6 for discarded features by 25% and K = 10 for features discarded by 50% with fitness values of 0.64 and 0.68. In testing the particles generated as much as 30 the highest accuracy obtained on the use of K = 10 and K = 9 for 25% of features removed by producing an accuracy of 76% with a fitness value of 0.76, while in testing 50% of the features removed there is a decrease in the level of accuracy where the highest accuracy obtained by 72% for the use of K = 9 and K = 10 with a fitness value of 0.72.

3. CLOSING

3.1. CONCLUSION

Based on the research that has been done, it can be concluded that the implementation of the K-Nearest Neighbor method based on particle swarm optimization as a feature selection resulted in a better accuracy compared to the feature selection that uses only the K-Nearest neighbor method. However, the features resulting from the feature selection process are not as optimal becausethe particle swarm optimization has not met the specified stop conditions. Testing the classification of news documents by using 250 documents, in which 225 documents are used as training data and 25 documents used as test data, produces the highest accuracy of 80% using K = 9 and many particles are raised as many as 50. In the test of 15 and 30 particles that were generated, it is obtained that the highest accuracy value of 68% for 15 particles raised with K = 6, and 76% for 30 particles generated using K = 10 and K = 8. In the classification test without feature selection using the particle swarm optimization method, the result reached only the accuracy level of 60 %.

3.2. Recommendation

In the research that has been carried out, there are still deficiencies that occur, and research can still be developed again in order to get better results. The recommendation that can be given for further research is to use other classification methods as a learning algorithm that has a better level of performance compared to the K-Nerarest Neighbor method for the result of this reserach could only reach the accuracy level of 80%.

REFERENCES

[1] A. A. Irfa, Adiwijaya And M. S. Mubarok, "Klasifikasi Topik Berita Berbahasa Indonesia Menggunakan K-Nearest Neighbor," E - Proceeding Of Engineering , Vol. 5, Pp. 3631-3640, 2018.

[2] C. F. Suharno, M. Fauzi And R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-Square," Systemic, Vol. 3, Pp. 25-32, 2017.

[3] V. Chandani And P. Romi Satria Wahono, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection Pada Analisis Sentimen Review Film," Journal Of Intelligent Systems, Vol. 1, Pp. 56 - 60, 2015.

[4] M. H. Aghdam And S. Heidari, "Feature Selection Using Particle Swarm Optimization In Text Categorization," Jaiscr, Vol. 5, Pp. 231-238, 2015.

[5] H. Februariyanti And E. Zuliarso, "Klasifikasi Dokumen Berita Teks Bahasa Indonesia," Jurnal Teknologi Informasi Dinamik , Vol. 17, Pp. 14-23, 2012.

[6] A. H. Ferdinan, A. B. Osmond And C. Setianingsih, "Klasifikasi Emosi Pada Lirik Lagu Menggunakan Metode K-Nearest Neighbor," E-Proceeding Of Engineering, Vol. 5, Pp. 6187-6194, 2018.

[7] G. Hermawan, "Implementasi Algoritma Particle Swarm Optimization Untuk Penentuan Posisi Strategis Agent Pada Simulasi Robot Sepakbola Dua Dimensi," Jurnal Ilmiah Komputer Dan Informatika (Komputa) , Vol. 1, Pp. 63-70, 2012.

[8] K. W. Mahardika, Y. A. Sari And A. Arwan, "Optimasi K-Nearest Neighbor Menggunakan Particle Swarm Optimization Pada Sistem Pakar Untuk Monitoring Pengendalian Hama Pada Tanaman Jeruk," Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, Vol. 2, Pp. 3333-3344, 2018.

[9] B. Xue, Particle Swarm Optimisation For Feature Selection, Victoria University, 2014.

[10] D. W. Suliantoro, I. Wisnubhadra And Ernawati, "Integrasi Pembobotan Tf-Idf Pada Metode K-Means Untuk Clustering Dokumen Teks," Prosiding Seminar Nasional Manajemen Teknologi, 2012.

[11] D. Ariani, A. Fahriza And I. Prasetyaningrum, "Optimasi Penjadwalan Mata Kuliah Di Jurusan Teknik Informatika Pens Dengan Menggunakan Algoritma Particle Swarm Optimization".