Jurnal Mantik, 6 (1) (2022) 43-49

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Classification of Book Types Using the Support Vector Machine (SVM) Method

[1]Fristi Riandari, [2]Hengki Tamando Sihotang, [3]Tarisa Tarigan, [4]Muhammad Rafli

[1234]Computer Engeenering, STMIK Pelita Nusantara, Medan, 20154, Indonesia

E-mail: [1]fristy.rianda@ymail.com, [2]hengki_tamando@yahoo.com

ARTICLE INFO

Article history:
Received: Feb 10, 2022
Revised: Feb 26, 2022
Accepted: March 15, 2022

Keywords:
Data Mining.
Clasification,
Support Vector Machine (SVM)

ABSTRACT

This study aims to create a model that can classify book types based on several categories and analyze the accuracy results of the Support Vector Machine (SVM) method. This research begins with the stages of data collection, namely the dataset of books obtained from the library. Furthermore, the dataset will be categorized into several types. The next stage, after the data is collected, will be carried out in the pre-process stage. This pre-process stage aims to prepare data so that it is ready to be processed in the feature extraction stage. The pre-processing stage consists of text segmentation, case folding, tokenization, stopword removal, and stemming. Next, the feature extraction stage will be carried out which aims to explore potential information and represent words as feature vectors. The next stage is to separate the training data and test data. Then the classification process is carried out using the SVM multiclass method to get the final result of modeling. The resulting classification results will then be evaluated in order to obtain an accuracy value and then will be analyzed whether the resulting classification model is feasible to implement.

1. Introduction

Based on observations, the library becomes a source of information and recreation so that it can be enjoyed by many people. One of them is the library at STMIK Pelita Nusantara. As of October 15, information was obtained that the number of books in this library has reached approximately 8,000 copies with thousands of titles. The large number of books makes it difficult for librarians to organize and search and it takes a long time if visitors search for books that they themselves don't know about the books they need. The results of interviews with librarians provide information that librarians also have difficulty providing information on the location of books needed by visitors. So we need an approach that can classify the types of books based on the title of the book with text classification.

Text classification is a type of supervised learning (guided). Text classification aims to help organize large amounts of information so that it can be understood by users. Several types of text classification algorithms that are often used are Nave Bayes, Support Vector Machine, Decision Tree, and KNN. SVM is a classification method that is now widely developed and applied. This method is derived from statistical learning theory which is promising and gives better results than other methods. SVM works very well on high-dimensional datasets. SVM using kernel technique must map the original data from its original dimension to another dimension which is relatively higher [1]. Of the algorithms mentioned above, SVM is one of the algorithms that produces good accuracy values. This result is evidenced by several previous studies regarding text classification as well, among others, a study in 2015 concluded the results of his research using the SVM algorithm, namely that a good classification model was obtained, the results of model testing using a quadratic function kernel showed an accuracy of 96.2%, and testing using test data shows an accuracy of 98% using a quadratic function kernel [2]. Another study in 2017 that classified

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

complaints using the SVM method concluded that based on the results of testing using a dataset of 1040 data, the results of a comparison of data classification between manual predictions and model predictions, the results of the accuracy of predictions built by the model were 995 data, and the prediction error rate was as much as 45 data [3]. Based on this explanation, a text classification model with the Support Vector Machine (SVM) algorithm will be developed to classify book types.

2. Method

Several previous studies related to the Support Vector Machine (SVM) approach in classifying: Helena Nurramdhani Irmanda & Ria Astriratma (2020), This study aims to produce a classification model for 3 categories of rhymes. The results of the research on the classification of the types of rhymes with SVM with a maximum feature of 1000 features and the number of datasets of 470 rhymes data consisting of children's rhymes, young rhymes, and parental rhymes (training data 376 records, test data 94 records) can be concluded that SVM can properly classify types of rhymes with an accuracy of 81.91%. In addition, children's rhymes have higher precision, recall, and specificity values than old rhymes and young rhymes, which are 90.63%, 87.88%, 95.08%. However, the values of precision, recall, and specificity for young poems and old poems have good and acceptable values. The highest values of precision, recall, and specificity of children's rhymes are influenced by the amount of data in children's rhymes which are more than young rhymes and parental rhymes. Suggestions for future research are to add more data for the categories of children's poems, youth poems, and parents' poems. Thus, it is expected that the values of accuracy, precision, recall, and specificity will increase [4]. Bayu Sugara & Agus Subekti (2019) The problem in this research is that people with autism are isolated from the outside world so that many parents are embarrassed and lack confidence in the condition of their children and it is recorded that 1 out of 600 children in Indonesia has autism, there must be a way to prevent it. solve the problem . This research on early detection of autism disorder proposes a support vector machine (SVM) algorithm to provide the best accuracy value using a small dataset. The dataset used in this test is 67 by producing the highest accuracy value of 85% in the normalized poly kernel. Two ensemble techniques, namely Ada Boost and Bagging, were also proposed in the testing of this early detection of autism disorder research to improve the classification performance of the support vector machine (SVM) algorithm. Based on the results of experiments that have been carried out, it shows that the ensemble technique shows performance can increase the value of accuracy. SVM model with poly kernel and ensemble bagging technique shows the highest accuracy value of 91% [5]. Indri Monika Parapat, et al (2018) the problem that occurs in this study is that the deviations in child development that are late are known to have long-term consequences and are difficult to repair. The data used in the study were 90 data which were divided into 3 classes. This research class represents 3 types of developmental deviations in children, namely Down Syndrome, Autism, and Attention Deficit Hyperactivity Disorder (ADHD). The SVM algorithm is a linear classification method, so it uses a kernel to deal with nonlinear data . The final result of this research produces the highest average accuracy of 63.11% = 10, C = 1, itermax = 200 and also uses a polynomial kernel. Comparison of the results of the classification of child development with the help of psychologists shows that the system produces poor accuracy. This can be caused by few and unbalanced data used for research [6]. Muhammad Athoillah (2017) In this study, a facial recognition system was built using the multi-kernel SVM method with increased learning. The multi-kernel SVM classification method is carried out by combining the kernels including the Linear Kernel, Polynomial Kernel, and RBF Kernel. While the learning method is increased, it is done by changing the support vector in the previous learning into input data in the next lesson, so that it can reduce the amount of learning data (training) which in the end makes the system run more efficiently. The results obtained from the system test show that the system can recognize facial images well as indicated by the average value of precision, recall, accuracy and good computational time requirements [7].

2.1 Classification

Data classification is a process of finding the same properties in a set of objects in a database and classifying them into different classes according to a defined classification model. The purpose of classification is to find a model from the training set that distinguishes attributes into the appropriate category or class, the model is then used to classify attributes whose class has not been previously known. The classification technique is divided into several techniques, one of which is the Decision Tree [8].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

There are also those who explain that classification is the process of finding a set of models that describe and differentiate data classes. The purpose of classification is so that the resulting model can be used to predict the class of data that does not have a class label. If given a data set consisting of several features and classes, then classification is to find a model of that class as a function of other features. [9].

In the study entitled "The Application of Data Mining to Analyze the Number of Active Customers Using the C4.5 Algorithm" it is stated that the classification technique is a systematic approach to building a classification model from a collection of input data. For example, decision tree techniques, Bayesian (Naive Bayesian and Bayesian Belief Networks), Artificial Neural Networks (Backpropagation), concept-based techniques from mining association rules, and other techniques (K-Nearest Neighbor, genetic algorithm, technique with set approach rough and fuzzy). Classification is a technique of classifying data. The difference with the clustering method lies in the data, where in clustering there is no dependent variable, while in classification there is a dependent variable. [10].

2.2 Text Classification

Text classification is the process of classifying documents into one or more predefined categories [11]. It can also be defined that the text category or text classification is a process that groups a text into a certain category [12]. Text classification is a type of supervised learning (guided). Text classification aims to help organize large amounts of information so that it can be understood by users [13]. Several types of text classification algorithms that are often used are Nave Bayes, Support Vector Machine, Decision Tree, and KNN. Of these algorithms, SVM is one of the algorithms that produces good accuracy values.

2.3 Support Vector Machine (SVM)

SVM is a classification method that is now widely developed and applied. This method is derived from statistical learning theory which is promising and gives better results than other methods. SVM is a learning system that uses a hypothetical space in the form of linear functions in a high-dimensional feature space, trained with a learning algorithm based on optimization theory by implementing a learning bias derived from statistical learning theory. [1]. SVM was developed by Vapnik, Guyon & Boser. SVM was first shown around 1992 at the Annual Workshop on Computational Learning Theory. This method is a learning machine with Juan looking for the best hyperplane that divides the two classes in the input space. The best separator function is to optimize the margin value which is the separating hyperplane in each class and this position can be achieved if the dividing line is in the right position in the middle, dividing between negative classes and positive classes. [14]. The basic principle of SVM is a linear classifier, and then it was developed to work on non-linear problems by incorporating the concept of kernel tricks in high-dimensional workspaces. [15].

The following is an example based on Figure 2.1 of how SVM tries to find the best hyperplane for separating classes -1 and +1:

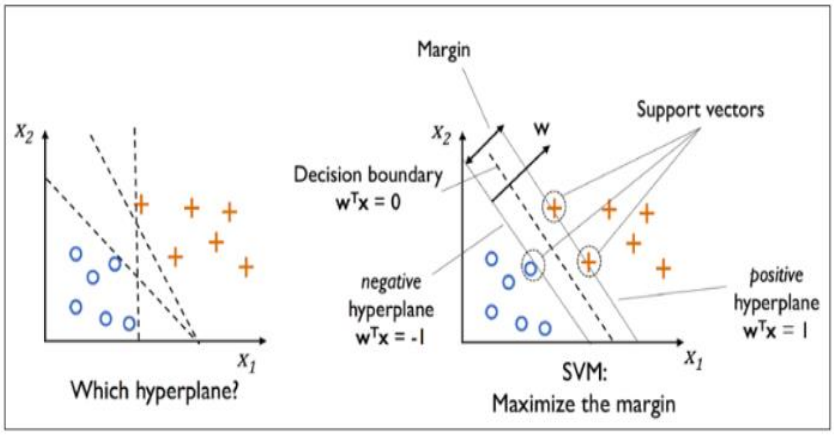

Figure 1. Hyperplane that separates 2 positive classes (+1) and negative classes (-1)

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The hyperplane found by SVM is illustrated as shown in Figure 2.2. Its position is in the middle between the two classes, meaning that the distance between the hyperplane and data objects is different from that of the adjacent (outermost) class which is marked with an empty and positive round. In SVM the outermost data object closest to the hyperplane is called a support vector. Objects called support vectors are the most difficult to classify because of their almost overlapping positions with other classes. Given its critical nature, only this support vector is taken into account to find the most optimal hyperplane by SVM.

Hyperplane is the best dividing line between the two classes. To find the hyperplane, it can be done by looking for the hyperplane margin 11 and looking for the maximum point. Margin is the distance between the closest data between two different classes, which is called the support vector. The solid line in Figure 1-b shows the best hyperplane, because it is located right between the two classes, while the support vector is represented by red and yellow dots inside the black circle.

The SVM linear classification hyperplane is denoted:

$f(x) : w, x + b = 0$

From the above equation, we get the class inequality +1 (negative)

$w.x + b \leq + 1$

Class -1 inequality:

$w.x + b \geq - 1$

3. Result and Discussion

In this study, the data used were book data taken from the STMIK Pelita Nusantara library. Where the data is 600 copies for data testing and has 3 categories that are adjusted to the Dewey Decimal Classification (DDC), namely with the code:

TABLE I
BOOK TYPE

Code	Type
000-199	Computer
200-399	Manajement
400-599	General Knowledge

In this study, one method is applied to data mining, namely by using the Support Vector Machine.

1) Text Preprocessing

Text preprocessing is one of the key components in various text mining algorithms, where data that is initially unstructured becomes structured. The following are the steps used in text preprocessing:

a. *Case Folding*

At this stage, changing the letters of the alphabet in the text which was originally in the uppercase form is converted into lowercase form. Then characters other than letters will be omitted, such as writing punctuation marks and numbers.

b. *Tokenizing*

At this stage, it is a process that is able to break a text into one word or phrase. Where the elements used are other than the letters of the alphabet and a hyphen (-).

2) *Fitur Extraction*

After going through the text preprocessing stage, the next step is to go through the feature extraction stage, where at this stage it is necessary to process the text into numeric. This is due to the principle that computers are not able to process data other than numerical data. In addition, feature extraction is used to explore information in presenting words as vectors. One of the techniques used in feature extraction is TF-IDF weighting or called Term Frequency-Inverse Document Frequency. TF-IDF is a weighting method used to calculate the number of matches of words in each document or called Term Frequency (TF), while Inverse Document Frequency (IDF) is the inverse value of Document Frequency (DF). And the final result of this method is an output matrix which includes the unique words and values generated in the TF-IDF of each word in the entire data. The following is the equation value for calculating TF:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$TF_{(t,d)} = f_{(t,d)}$

Where:

$f_{(t,d)}$ = occurrence of the word t in the document d

Meanwhile, DF or Document Frequency is the number of documents owned by term t. The following is the equation value for calculating DF:

$$IDF_{(t)} = \log\left(\frac{N}{nt}\right)$$

When:

N = number of documents

Nt = number of documents containing the word t

So TF-IDF will give the weight of a word taken from the TF value and the inverse DF value. The following is the equation value to find the weights using TF-IDF:

$$\text{TF- IDF} = f_{(t,d)} * \log\left(\frac{N}{nt}\right)$$

TF-IDF is carried out after the data has gone through the preprocessing stage, where the data must be in numeric form. So to change the data, the TF-IDF method is used so that the data can be analyzed using the SVM method.

3) Modeling

The next step is to enter the modeling stage, where the data that has been cleaned and through the preprocessing process will be entered into the model. The model used in this research is the Support Vector Machine (SVM). At this stage the data that has been processed will be built using the SVM model, where this process requires the addition of a package, namely using the NLTK (Natural Language Toolkit) library. Then create an SVM model and enter the test data that has been taken previously. So that the data that has been loaded will display the results of the values for each prediction made on the model. This model is an initial form, where the kernel type and values are specified in the parameters. To get maximum results, then it is necessary to conduct training on the model.

The kernel in SVM has many functions. However the selection of the kernel depends on the problem at hand as it follows with the modeled one. For example, the polynomial kernel makes it possible to model conjunction features finite order to a polynomial. The Radial Basis Function allows to select different groups with linear kernels where it is only possible to select lines (or hyperplanes). In mapping kernel functions, normalized Polynomial, RBF, Linear, and Sigmoid can be used which are used based on application needs. But some kernel functions have proven to work well for a wide variety of applications. Here's an explanation of each kernel:

1) Linear Kernel

Linear kernel is the simplest kernel function. In this kernel the inner product is <x.y> and is added with a constant choice of c.

2) RBF Kernel

The RBF kernel, also known as the Radial Basis Function, is the main kernel. This kernel is called primary because the RBF kernel is a nonlinear kernel that is able to map samples into the highest dimensional space, which is not the case with the Linear kernel. The RBF kernel has fewer hyper parameters than the Polynomial kernel and has less difficulty in the computational (numerical) part.

3) Polynomial Kernel

Polynomial kernel is a non-stationary kernel which is suitable for problems where all training data has been normalized.

4) Sigmoid Kernel

Sigmoid kernel or also called fuzzy sigmoid kernel is a kernel function that models the hyperbolic tangent function through linguistic variables.

When data classification cannot be separated in a linear way, then another method can be used. Functions in the kernel are able to convert data into the highest dimensional space to allow separation. Kernel functions are classes in algorithms intended for analysis or recognition, of which the most well-known element is SVM. The training vectors on xi are mapped into the highest dimensional space (possibly infinite) by the Φ function. Then SVM finds the linear separator hyperplane with the maximum margin in the highest dimension space. C > 0 is a penalty parameter.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

4) Evaluation

At this stage, it enters the evaluation testing stage, where it is intended to determine the quality and value of accuracy that has been produced by the method. So that what is produced at this stage is the calculation of the values of accuracy, precision, and recall using the scikit-learn library.

After carrying out the steps above, the following results will be obtained that is The data that has been collected from the STMIK Pelita Nusantara library is the initial data. Then the data needed in the study are the "title" and "classification" columns. Furthermore, the data goes through the preprocessing stage, where the previous data has been deleted on the duplicate data. In this study, 600 records were used as test data. Where the data will be separated as training data and test data. The determination of training data and test data is done randomly, so as to be able to keep it balanced between data. Then do the division of each proportion between the training data and the test data. This is done to determine the quality of the SVM method. To determine the effect on the amount of training data on the quality and effectiveness of the SVM method, several combinations were made on the amount of training data and test data.

TABLE 2
COMBINATION OF TRAINING AND TEST DATA

Numb.	Training Data	Test Data
1	300	200
2	150	100
3	150	100

And from the results of the test evaluation, information was obtained that at the evaluation stage the values of accuracy, precision, and recall were calculated, where this was intended to determine the quality of the model made on the SVM method. And the experimental results are obtained as follows:

TABLE 3
TEST PERFORMANCE

Kernel	Akurasi	*Precision*	*Recall*	F1 *Score*
Linear	67,24%	0,78	0,62	0,65
RBF	71,57%	0,76	0,59	0,62
Polynomial	58,56%	0,78	0,39	0,47
Sigmoid	68,02%	0,68	0,59	0,64

4. Conclusion

The Support Vector Machine method can be used to classify the types of books by performing the stages of text preprocessing, feature extraction and followed by the implementation of the support vector machine method. For further research, it is recommended to compare several methods in classifying types of books to determine the level of accuracy of the methods used in classifying types of books and carry out system development.

References

[1] N. Neneng, K. Adi, and R. Isnanto, "Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurrence Matrices (GLCM)," *J. Sist. Inf. Bisnis*, vol. 6, no. 1, p. 1, 2016, doi: 10.21456/vol6iss1pp1-10.

[2] A. S. Ritonga and E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding)," *Ilm. Edutic*, vol. 5, no. 1, pp. 17–25, 2018.

[3] F. Fatmawati and M. Affandes, "Klasifikasi Keluhan Menggunakan Metode Support Vector Machine (SVM) Pada Akun Facebook Group iRaise Helpdesk," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, p. 24, 2018, doi: 10.24014/coreit.v3i1.3552.

[4] H. N. Irmanda and Ria Astriratma, "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machine (SVM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 915–922, 2020,

doi: 10.29207/resti.v4i5.2313.

[5] B. Sugara and A. Subekti, "Penerapan Support Vector Machine (Svm) Pada Small Dataset Untuk Deteksi Dini Gangguan Autisme," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 177–182, 2019, doi: 10.33480/pilar.v15i2.649.

[6] I. M. Parapat, M. T. Furqon, and Sutrisno, "Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3163–3169, 2018, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2577.

[7] M. Athoillah, "Pengenalan Wajah Menggunakan SVM Multi Kernel dengan Pembelajaran yang Bertambah," *J. Online Inform.*, vol. 2, no. 2, p. 84, 2018, doi: 10.15575/join.v2i2.109.

[8] Selvia Lorena Br Ginting; Wendi Zarman; Ida Hamidah, "ANALISIS DAN PENERAPAN ALGORITMA C4.5 DALAM DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA BERDASARKAN DATA NILAI AKADEMIK," *Pros. Semin. Nas. Apl. Sains Teknol. 2014 Yogyakarta*, no. November, pp. 263–272, 2014.

[9] A. S. Sukardi and C. Supriyanto, "Klasifikasi Spam Email Menggunakan Algoritma C4.5 Dengan Seleksi Fitur," *J. Teknol. Inf.*, vol. 10, no. 1, pp. 19–30, 2014, [Online]. Available: http://research.pps.dinus.ac.id/lib/jurnal/Vol10.1 019-030.pdf.

[10] A. I. Jamhur, "Penerapan Data Mining Untuk Menganalisa Jumlah Pelanggan Aktif Dengan Menggunakan Algoritma C4.5," *Maj. Ilm.*, vol. Vol. 23, no. No.2, pp. 12–20, 2016.

[11] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto)," *J. Inform. dan Rekayasa Elektron.*, vol. 2, no. 2, p. 43, 2019, doi: 10.36595/jire.v2i2.117.

[12] O. Somantri, S. Wiyono, and D. Dairoh, "Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," *Sci. J. Informatics*, vol. 3, no. 1, pp. 34–45, 2016, doi: 10.15294/sji.v3i1.5845.

[13] C. Darujati, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa," *J. Link*, vol. 16, no. February 2012, p. 8, 2016.

[14] A. Handayanto, K. Latifa, N. D. Saputro, and R. R. Waliansyah, "Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi," *JUITA J. Inform.*, vol. 7, no. 2, p. 71, 2019, doi: 10.30595/juita.v7i2.4378.

[15] E. H. Harahap, L. Muflikhah, and B. Rahayudi, "Implementasi Algoritma Support Vector Machine (SVM) Untuk Penentuan Seleksi Atlet Pencak Silat," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 10, pp. 3843–3848, 2018.