Journal of Applied Science, Engineering and Technology
e-ISSN: 2722-8363 p-ISSN: 2722-8371
DOI: https://doi.org/10.47355/aset.v4i2.71

J. ASET

Optimization of Thesis Topic Classification Using Support Vector Machine

Rizky Hadi [1], Puput Budi Wintoro [2*], Meizano Ardhi Muhammad [3], Hery Dian Septama [4]

[1,2,3,4] Teknik Informatika, Engineering Faculty of Universitas Lampung; Prof. Dr. Ir. Sumantri Brojonegoro Street No. 1, Bandar Lampung City, Lampung 3514

* Correspondence: budi.wintoro@eng.unila.ac.id

Received: 20.07.2024; Accepted: 14.11.2024; Published: 31.12.2024

Abstract: The Technical Implementation Unit (UPT) of the University of Lampung Library is responsible for archiving student scientific work (thesis). Filing is done manually, using the DDC or Dewey Decimal Classification system, thus allowing for the tendency of inaccurate Subject selection and long grouping durations. This study aims to apply the Support Vector Machine (SVM) algorithm in classifying thesis subjects stored in the Unila student scientific article repository. The application of the SVM algorithm uses the machine learning life cycle method which includes the data collection process, data pre-processing, data splitting, model training, to the model evaluation process. The data used are Unila student thesis titles totaling 1707 data. The results of this study are a thesis subject classification model based on the title with the SVM model where the accuracy of the training data in the model training process is 0.95, and the evaluation model process with an accuracy rate of 0.65.

Keywords: Classification; Dewey Decimal Classification; Support Vector Machine, Machine Learning

1. Introduction

Artificial Intelligence (AI), commonly referred to as artificial intelligence, is a branch of computer science that studies the development of smart machines to solve complex problems quickly and assist human needs [1]. Some of the elements that characterize artificial intelligence include: the presence of an expert system, natural language processing, recognition of human and machine language, reasoning with robotics and sensors for logic, the ability to interpret images and objects via computers, the capacity to guide or tutor humans, and, lastly, the facilitation of deep learning.

Machine learning is a branch of computer science that studies how to create computer programs that learn through experience [1]. Machine learning attempts to solve problems by establishing general rules for various tasks through the use of statistical techniques, such as artificial neural networks. In machine learning, there are two main types of approaches: supervised learning and unsupervised learning. Supervised learning is a technique commonly used to solve classification problems, where the goal is to learn the expected output from inputs that have already been identified, using datasets that are labeled [2].

Classification is a form of data analysis that extracts a model to describe the classification or category of data [3]. Text classification has two approaches: text clustering and text classification. Text clustering aims to find an unsupervised group structure from a set of sentences in a document. Meanwhile, text classification aims to group documents based on a previously known (supervised) group structure. Based on this explanation, it can be seen that there are several steps to perform text

classification automatically, including pre-processing, feature extraction/selection, choosing a model using machine learning techniques, and training and testing the classifier [4].

Research on classification based on supervised learning has been conducted previously. A classification and comparison using six different machine learning algorithms, with SVM achieving the highest accuracy followed by Random Forest, was reported [5]. Another study used SVM with an RBF kernel to classify text with three attributes, resulting in 93.9% accuracy [6]. Based on the results obtained, classification in previous studies has proven effective in solving prediction problems.

Therefore, this study aims to use SVM for subject classification multi-label data from essay using Support Vector Machine (SVM). SVM is widely used because it has good performance in text classification that use multi-dimensional data. SVM is a learning system that uses a hypothetical space in the form of a linear function in a high-dimensional feature space, which is trained with a learning algorithm based on optimization theory by applying machine learning derived from statistical learning theory. The concept of classification with SVM is to find the best hyperplane to separate the two data classes and use a support vector approach. [7] There are three key ideas needed to understand SVM: maximizing margins, dual formulation, and kernels.

2. Materials and Methods

In this study, the subject classification process used the machine learning life cycle. This method has several stages in the research flow, as shown in Figure 1.

Figure 1. Research Flow

Based on Figure 1, the thesis and subjects are collected first from the www.digilib.unila.ac.id site. After the dataset is obtained, the next process is data pre-processing, which includes data labeling, data cleaning, and feature extraction. After the data is processed, the next step is data splitting. The SVM model is used in the model training process by applying several architectures. In the final stage, the SVM model is evaluated with a confusion matrix to define the model's performance. A complete explanation will be given in the following sub-chapters.

The first step begins with data collection. The dataset in this study is Lampung university student thesis title and it's subject. The subject is limited in DDC nine main categories. The dataset was taken on July 30, 2022. The dataset has 2 attributes: title and subject. The dataset has 1707 data and is used to predict subjects. The sample of the dataset can be seen in table 1.

Table 1. The sample of the dataset

Title	Subject
APLIKASI METODE DEKOMPOSISI ADOMIAN PADA PENYELESAIAN SISTEM PERSAMAAN DIFERENSIAL PARSIAL LINIER HOMOGEN ORDE 1.	Natural science and mathematics
WEB-BASED INPATIENT HEALTH SERVICE INFORMATION SYSTEM USING FRAMEWORK LARAVEL.	Computer science, information, and general works

ANALISIS PERUBAHAN LUAS HUTAN TAMAN NASIONAL WAY KAMBAS TAHUN 2000-2015 MELALUI CITRA LANDSAT DI KABUPATEN LAMPUNG TIMUR	History and Geography
PROSES ADAPTASI MANTAN NARAPIDANA KASUS PENGEDARAN NARKOBA DI MASYARAKAT (Studi Kasus Di Kelurahan Kaliawi Kecamatan Tanjung Karang Pusat Kota Bandar Lampung)	Social Sciences
PENINGKATAN SIKAP POSITIF TERHADAP PURPOSE IN LIFE DENGAN BIMBINGAN KELOMPOK TEKNIK ROLE PLAYING PADA SISWA KELAS XI SMK NEGERI 4 BANDAR LAMPUNG TAHUN AJARAN 2018/2019	Philosophy and Psychology

This study uses nine subjects based on the amount of data in the unila digital repository site(www.digilib.unila.ac.id). The distribution of the dataset is shown in table 2.

Table 2. The distribution of the dataset

Digit	Category	Amount of data
000	general works	310
100	philosophy and psychology	70
200	religion	12
300	social sciences	424
400	language	117
500	natural sciences and mathematics	240
600	technology	416
700	history, biography, and geography	96
900	the arts	22

The next step is data preprocessing. Data preprocessing is important because it improves accuracy and reliability. Data pre-processing involves handling raw data, which is often unstructured or unclean, to prepare it for further analysis [8]. This process includes activities such as data cleaning, which involves removing punctuation marks, eliminating stop-words, applying case folding, and, finally, performing feature extraction as the last step in data pre-processing.

Data cleaning works by selecting relevant data, redundant and noisy data will usually be deleted which will improve the quality of the input data. The process of matching the title of all data into lower case letters needs to be done to reduce the amount of vocabulary that the computer needs to recognize.

The process of removing stop-words in Indonesian and removing numeric, so that any punctuation and numbers in the thesis title will be removed. Stop-words are conjunctions that appear frequently in textual language. Conjunctions cannot stand alone in a sentence because they have no meaning (for example, "and", "or", and so on in Indonesian), so they are not useful for classification [9].

e-ISSN: 2722-8363 p-ISSN: 2722-8371

This is done so that the computer will only focus on the words contained in each category. Comparison before and after the data cleaning process can be seen in table 3.

Table 3. Comparison before and after the data cleaning process

Step 1 : Case Folding	
Before	After
APLIKASI METODE DEKOMPOSISI ADOMIAN PADA PENYELESAIAN SISTEM PERSAMAAN DIFERENSIAL PARSIAL LINIER HOMOGEN ORDE 1.	aplikasi metode dekomposisi adomian pada penyelesaian sistem persamaan diferensial parsial linier homogen orde 1.
Step 2 : Stopwords removal	
Before	After
aplikasi metode dekomposisi adomian pada penyelesaian sistem persamaan diferensial parsial linier homogen orde 1.	aplikasi metode dekomposisi adomian penyelesaian sistem persamaan diferensial parsial linier homogen orde 1.
Step3 : numeric removal	
Before	After
aplikasi metode dekomposisi adomian penyelesaian sistem persamaan diferensial parsial linier homogen orde 1.	aplikasi metode dekomposisi adomian penyelesaian sistem persamaan diferensial parsial linier homogen orde

The final process of data pre-processing is feature extraction. Feature extraction refers to the process of transferring attribute operations, selecting attributes, subsets of attributes can be combined or can contribute to creating replacement attributes from data sets. To select attributes from data in the form of text, a method is used using TF-IDF (term frequency-inverse document frequency). TF-IDF is a method that integrates term frequency (TF) and inverse document frequency (IDF) which is useful for calculating the weight of each word used in a document so that you can find out how often a word appears in a document. TF is a value that indicates how often a word appears in a document. The more often the word appears the more important it is in the document. However, if a word appears too often in the corpus then the word is too common in that document. This is likened to the word "atom" which does not usually appear in normal documents, but this word usually appears in physics documents. This situation will be resolved by the IDF [10]. To calculate the IDF, the commonly used formula is n⁄df, where n is the total number of documents that appear in the corpus. The TF-IDF used by scikit considers n as n+1, which calculates to ((n+1))⁄df, and adds 1 as the final result [11].

$$TF = fd(i)fd(j) \quad (1)$$

$$IDF = \log Ndfi+1 \quad (2)$$

Where fd (i) is the number of terms (i) in document (j), fd (j) is the number of terms in the document (j).

The first stage of pre-processing data is starts with labeling the data subject using the Label-Encoder which will change the subject to a numeric.

 doi: 10.47355/aset.v4i2.71

e-ISSN: 2722-8363 p-ISSN: 2722-8371

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method used to evaluate the importance of a word in a document relative to a collection of documents (corpus). The method combines two key components: term frequency (TF), which measures how often a term appears in a document, and inverse document frequency (IDF), which reduces the weight of common terms across the corpus. This ensures that unique or context-specific words are given higher importance. The resulting numerical values are often used in text analysis, information retrieval, or machine learning models to represent textual data effectively.

The third step is data splitting. The data that has become numeric after the previous feature extraction stage will then be split into two subsets, training data and test data. The training data is used for model training while the testing data is used to evaluate the model from the training model being trained. Most of the data will be used in the training data, this is done so that the computer gets a lot of data when carrying out the training process which will increase the accuracy of data predictions. To do a split, the data is divided by a ratio of 80:20, where 80% of the data is for training data while the rest is for testing data.

With the data training and data testing has been obtained, the last step in making this classification model is to enter the model and do the testing. This model uses SVM from the Sklearn library with the kernel used is "rbf". The RBF kernel or also called Gaussian kernel is the most widely used kernel concept to solve linearly inseparable data classification problems. The RBF kernel locates a radial basis function centered at each point, then performs linear manipulation to map the points to higher-dimensional spaces that are more easily separable. The RBF kernel function equation is:

$$K(x, xi) = \exp\left(-\frac{\|x1 - x2\|^2}{2\sigma^2}\right)$$

Where σ is the variance of the hyperparameter, and $\|x1 - x2\|$ is the Euclidean rule which is the distance from x1 to x2. Exp is an exponential function in mathematics and has a value of 2.71828183. The rbf kernel is used when more than two data labels are used, because this kernel is able to separate data that exceeds two labels, by repeating 1500 times, and the error tolerance is 1/10000, with a training duration of 0.6 seconds.

Testing is done by evaluating the training model to determine the value of accuracy, precision, and recall through the confusion matrix. This process can also be done with the sklearn library, by importing the confusion matrix from sklearn.metrics.

```
Confusion_matrix(y_test, pred)
#outpu:array([[16, 0, 0, 0, 6, 7, 0, 0, 8],
              [ 0, 0, 0, 0, 0, 2, 0, 0, 0],
              [ 0, 0,20, 0, 1, 2, 0, 0, 1],
              [ 0, 0, 0,16, 0, 0, 0, 0, 0],
              [ 5, 0, 2, 0,29,16, 0, 0, 8],
              [ 2, 0, 0, 1, 0,63, 0, 0, 9],
              [ 0, 0, 0, 0, 0,14, 4, 0, 2],
              [ 0, 0, 1, 0, 0, 6, 0, 0, 0],
              [ 0, 0, 1, 0, 2, 8, 0, 0,70]]
              ,dtype=int64)
```

To read the results of the confusion matrix with labels form the data title, we will use a heat map. On the heat map, there are already various thesis subjects that have previously been converted to integers with lable.encoder. This heat map can be seen in figure 2.

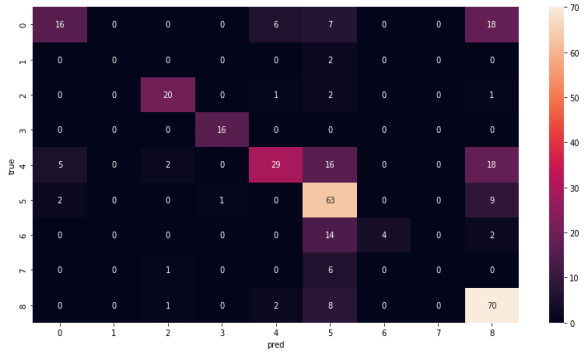

Figure 2. Confusion matrix in heat map

3. Results

A total of 342 data from 20% of total dataset used to evaluate and with the confusion matrix accuracy, precision, recall, can be obtained in the form of a classification report in table 3.

Table 3. Classification report

subjects	**precision**	**recall**	**support**
natural sciences and mathematics	0.70	0.34	47
religion	0.00	0.00	2
language	0.83	0.83	24
philosophy and psychology	0.94	1.00	16
general works	0.76	0.41	70
social sciences	0.53	0.84	75
history, biography, and geography	1.00	0.20	20
the arts	0.00	0.00	7
technology	0.59	0.86	81
accuracy			0.64
avg / total	0.67	0.64	342

4. Discussion

The result from evaluating model with confusion matrix in table 3 is average precision 0,67, recall 0,64 and accuracy 0,64. Based on this result the amount of data used is very influential on the value of the model. As already seen in categories with a small number of support columns or units that do not have precision and recall values.

5. Conclusions

From the result of this study, it can be concluded that the Support Vector Machine algorithm can be used to classify text in the form of thesis titles. The final result of the model made at the model training stage using training data has an accuracy of 0.95, whereas at the model evaluation stage using data testing it obtains an accuracy of 0.65, a precision of 0.54, and a recall of 0.45. The amount of data used is very influential on the value of the model. The future research will propose preprocessing data using stop words that can recognize at least two languages, Indonesian and

English, because among the data quite a lot of languages are used besides Indonesian also using algorithms other than SVM for classification text for comparison of the resulting accuracy levels.

References

1. Mitchell, T.M. *Machine learning*; McGraw-hill, New York,1997.
2. Nasteski, V. An overview of the supervised machine learning methods. Horiz. B 4, 51–62.
3. Han, J., Pei, J., Tong, H., 2022. Data mining: concepts and techniques. Morgan kaufmann
4. Dalal, M.K. and Zaveri, M.A., 2011. Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), 37-40.
5. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.
6. Octaviani, P.A., Wilandari, Y. and Ispriyanti, D., 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. Jurnal Gaussian, 3(4),811-820.
7. Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
8. García, S., Luengo, J. and Herrera, F., Data Preprocessing in Data Mining. Intelligent Systems Reference Library. 2015. doi, 10, 978-3.
9. Dalal, M.K. and Zaveri, M.A., 2011. Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), 37-40.
10. Yoo, J.-Y., Yang, D., 2015. Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier. Presented at the Computer and Computing Science 2015, 263–266
11. Lavin, M., 2019. Analyzing documents with TF-IDF