Classification of Thesis Topics Based on Informatics Science Using SVM

Eka Mala Sari Rochman\textsuperscript{1}, Ika Oktavia Suzanti\textsuperscript{2}, Imamah\textsuperscript{3}, Muhammad Ali Syakur\textsuperscript{4}, Devie Rosa Anamisa\textsuperscript{5}, Ach. Khozaimi\textsuperscript{6}, Aeri Rachmad\textsuperscript{7}

Faculty of Engineering, University of Trunojoyo Madura. Raya Telang Road P.O. BOX 2 Kamal, Bangkalan-Madura,Indonesia

\textsuperscript{1}ekamala.sari@yahoo.com, \textsuperscript{2}iosuzanti@trunojoyo.ac.id, \textsuperscript{3}i2munix@gmail.com, \textsuperscript{4}syakuralir@yahoo.co.id, \textsuperscript{5}devros_gress@yahoo.com, \textsuperscript{6}khozaimi@trunojoyo.ac.id, \textsuperscript{7}aery_r@yahoo.com

Abstract. Thesis topic is an inseparable part in the world of tertiary education. Determining the thesis topic becomes a problem for students. The determination of the thesis topic leads to the trend of the topic in the development of computer science. The determination of the topic of thesis for students often ignores their ability to process. Ideally in determining the topic of the thesis, the record of student grades can be an important variable in deciding topics for students, where the student's grade record is contained in the transcript. Therefore, this study uses the Support Vector Machine (SVM) method in recommending thesis topics by classifying selected subject groups that have been taken by students. The Support Vector Machine method is a classification method of supervision because it requires testing data and training data as a training process at the time of prediction. Support Vector Machine provides an optimal model, which provides a solution with a maximum margin to determine the distance of data to the hyperplane. The test results show an accuracy of 80%.

Keywords : Thesis topic, Support Vector Machine (SVM), Students

1. Introduction
Thesis is a scientific work based on a student's independent research activities, compiled within one semester under the guidance of a supervisor and can be assisted by a supervisor. The thesis is carried out independently by students, it is intended that the research design is initiative, the implementation of research and writing the final project report are on the students themselves \[1\]. According to the big Indonesian dictionary, Topics are the subject of discussion in lectures, essays, and so on as discussion material \[2\]. In its scope the topic is the subject discussed in a text.

Data mining is the discovery of useful knowledge and information from data stored in databases in large enough quantities by automatic or practical pattern extraction that represents stored knowledge. Using data mining can answer business questions that usually take a long time to solve. Classification techniques are widely used in data mining which is used to predict group membership for data instances on machine learning \[3\].

Classification is the process of constructing a model that classifies an object based on its attributes. Label classes are already available from previous data so that they focused on how to study existing
data so that the classifier can classify automatically. [4] One statistical method that can be applied to classify is the Support Vector Machine (SVM). SVM is a technique for finding hyperplane that can separate two data sets from two different classes [5][6]. SVM has the advantage of being able to determine distance using support vectors so that the computational process becomes fast (Vapnik, 1995).

Support Vector Machine (SVM) is one of the methods in Machine Learning that is widely used to solve complex problems. In a research journal about the SVM method which is a reference to solve the problem of determining the topic of a thesis by looking at the student's record. In this study the supervised SVM classification method requires training data as machine training and data testing is used to predict new data. It is important for this method to find the best Hyperplane and optimal margins to classify data, so that topics can be obtained that fit the student's grade record.

In Support Vector Machine weights (w) and bias (b) are global optimum solutions from quadratic programming which is a mathematical formulation of Support Vector Machine; so that once running it, it can produce a solution that will always be the same for the same choice of kernels and parameters [7]. The purpose of this study is to implement the SVM method to solve the thesis topic selection problem by making it easy to determine the topic based on the student's grade record.

2. Methodology

2.1. Data Collection

The determination of the topic of this thesis was made for the informatics study program at the Faculty of Engineering, University of Trunojoyo Madura. The data used is the transcript data of students of Informatics Engineering study program class of 2013. The topic used is to adjust the name of the Research Laboratory in the Informatics Engineering study program, namely Software Engineering, Artificial Intelligence, Distributed Systems and Multimedia.

2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a learning system that uses hypothetical spaces in the form of linear functions in a high-dimensional feature space, trained with learning algorithms based on optimization theory by implementing learning bias derived from statistical learning theory. The basic principle of SVM is linear classifiers, and further developed in order to work on nonlinear problems [3][8] by incorporating the concept of kernel tricks in high-dimensional workspaces. This development provides research interest in the area of pattern recognition for the theoretical potential investigation of SVM capabilities both in terms of application [9].

Linearly separable data is data that can be separated linearly [10]. Suppose \(x_i \in \{x_1, \ldots, x_n\}\) is a dataset and \(y_i \in \{+1,-1\}\) is the class label of data \(x_i\). The function used to separate classes is to use a linear function, where the function is defined as follows

\[
g(x) = \text{sign} \left( f(x) \right) \\
dengan \, f(x) = (w^T x + b)
\]

where \(w\) is the normal plane, whereas \(b\) shows the position of the plane relative to the center coordinate. To find the best separator field with the largest margin value, it can be formulated into an SVM constraint optimization problem for the case of linear classification in primal space, where, \(x_i\) shows the input data while \(y_i\) is the output of \(x_i\) shown by equation 3 namely:
In the constraint optimization equation to minimize the objective function $1/2 \| w \|_2$ or maximize $w^T w$ that is by paying attention to the constraint $y_i (x_i \cdot w + b) \geq 1$. If the output data $y_i = +1$, then the constraint becomes $(x_i \cdot w + b) \geq 1$, otherwise $y_i = -1$, then the delimiter becomes $(x_i \cdot w + b) \leq 1$. In some cases, there are some data that cannot be classified correctly (infeasible), then it can be stated through the following equation,

$$\min \frac{1}{2} \| w \|_2^2 + C \left( \sum_{i=1}^{n} \xi_i \right)$$

$$\text{s.t. } y_i (w \cdot x_i + b) \geq 1 - \xi_i$$

Value C (Complexity) is the value chosen before optimization with the Quadratic Programming process. C values range from zero to infinity positive ($0 < C < \infty$). The purpose of the C (Complexity) value is to minimize errors and reduce the value of the variable slack. If the value of C is close to zero, then the width of the margin on the boundary plane becomes maximum and the amount of trained data that is in the margin or in the wrong position will be ignored. This means that it can reduce the level of accuracy in the training process, resulting in test data that cannot be classified properly.

**Fig 1.** SVM tries to find the best hyperplane by separating the two classes -1 and +1

**Fig 2.** Hyperplane formed between class-1 and +1 [2]

The best hyperplane separator between the two classes can be found by measuring the margin of the hyperplane and look for the maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called a support vector. Solid lines in the picture show the best hyperplane, which is located right in the middle of both classes, while the red and yellow dots in the black circle are support vectors. The effort to find the location of a hyperplane is the core of the learning process in SVM.

3. **Result and Discussion**

This section presents the methods used in determining the final assignment topic and steps in implementing the SVM method. Some samples of students were taken with several attributes that influence the determination of the topic.
Before entering the SVM method, the data was labelled. Labeling was done manually in accordance with the thesis topic that had been done by students of class of 2013. Selecting the features was not so influential in determining the topic of the thesis that can be removed in order not too many features are used, then it would be easier to calculate. After that, it can enter into the SVM algorithm starting with finding a hyperplane to separate the class into +1 or -1, then searched the optimal margin. The best hyperplane was a hyperplane with the equation \( w \cdot x - b = 0 \). Data was classified by RBF (Radial Basic Function) kernel.

![Diagram of the SVM process](image)

**Figure 3. Design System**

The topic label initialization in the comparison used the one for all comparison; where one label is considered +1 and the other label is -1. It searched the bias of the hyperplane equation, then it searched maximum and minimum margins to find out the distance of the data to the hyperplane.

| Search Engine Engineering | Distributed system | Digital Image Processing | Visual Programming | Topic |
|---------------------------|--------------------|--------------------------|--------------------|-------|
| 10,5                      | 10,5               | 12                       | 7,5                | CAI   |
| 9                         | 9                  | 12                       | 10,5               | CAI   |
| 10,5                      | 10,5               | 14                       | 12                 | RPL   |
| 10,5                      | 10,5               | 8                        | 7,5                | CAI   |

From the data above, the hyper plane point was found by using the equation

\[
\frac{1}{2} \|w\|^2 = \frac{1}{2} (w_1^2 + w_2^2)
\]

\[
y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, 3, \ldots, N
\]

(4)

| Search Engine Engineering | Distributed system | Digital Image Processing | Visual Programming | Topic |
|---------------------------|--------------------|--------------------------|--------------------|-------|
| 10,5                      | 10,5               | 12                       | 7,5                | 1     |
| 9                         | 9                  | 12                       | 10,5               | 1     |
| 10,5                      | 10,5               | 14                       | 12                 | -1    |
| 10,5                      | 10,5               | 8                        | 7,5                | 1     |
The results of testing and analysis as attached in Figure 4.2.3 and Figure 4.2.5 produced an accuracy calculation of 80% using the Precision Recall method. 80% accuracy was obtained from 24 testing data. Judging from the resulting graph it can be concluded that, if the tested data are 10-40% of the overall data, the accuracy results would be even greater; when the comparison of test data is 50-70%, the accuracy results can be very low; if the test data used is more than 70 % then the accuracy results can be quite good. It can be concluded that the SVM method can provide good results to predict Thesis topics for students.

CONCLUSIONS
Based on the results of research and testing in the application of the SVM (Support Vector Machine) method, the following conclusions can be drawn:

1. The test results went well and with good results
2. The SVM method can help predict the determination of students' Thesis topics based on a recap of grades

ACKNOWLEDGMENT

We would like to thank our colleagues at the University of Trunojoyo Madura - Indonesia, the Laboratory of Research team who helped author in and fellow students who helped assist the authors in completing this research.

References

[1] Mutahar AL-Sayed Muhammed Qassem, *Thesis, Dissertation and Article Writing Preparing Research Paper, Dissertation and Thesis*. 2017.
[2] Kustiyahningsih, Y., Hamid, A. Z., & Rachmad, A. (2015). Rancang Bangun Aplikasi (E-Dms) Electronic Document Management System Dengan Metode Tf/Idf Berbasis Web. *Jurusan Teknik Informatika, Fakultas Teknik, Universitas Trunojoyo*.
[3] H. Bhavsar and M. H. Panchal, “A Review on Support Vector Machine for Data Classification,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 2278–1323, 2012.
[4] E. E. Pratama and R. T. Bambang, “Klasifikasi Topik Keluhan Pelanggan Berdasarkan Tweet dengan Menggunakan Penggabungan Feature Hasil Ekstraksi pada Metode Support Vector Machine (SVM),” *J. Edukasi dan Penelit. Inform.*, vol. 1, no. 2, pp. 53–59, 2015.
[5] A. K. Wardhani, “Implementasi Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Kajen Pekalongan,” *J. Transform.*, vol. 14, no. 1, pp. 30–37, 2016.
[6] D. I. Pusphita Anna Octaviani, Yuciana Wilandari, “Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang,” *J. Gaussian*, vol. 3, no. 8, pp. 811–820, 2014.
[7] R. Rusdah and U. B. Luhur, “Predicting The Suspect of New Pulmonary Tuberculosis Case using SVM, C5.0 and Modified Moran’s I,” no. April, 2018.
[8] A. Basu, C. Walters, and M. Shepherd, “Support vector machines for text categorization,” *Proc. 36th Annu. Hawaii Int. Conf. Syst. Sci. HICSS 2003*, no. January 2003, pp. 1–8, 2003.
[9] Y. Shi, Y. Tian, G. Kou, Y. Peng, and J. Li, “Support vector machines for classification problems,” *Adv. Inf. Knowl. Process.*, no. 9780857295033, pp. 3–13, 2011.
[10] Damayanti, F., & Rachmad, A. (2016). Recognizing Gender Through Facial Image Using Support Vector Machine. *Journal Of Theoretical & Applied Information Technology*, 88(3).