Performance analysis of support vector machines with polynomial kernel for sentiment polarity identification: A case study in lecturer’s performance questionnaire

G A Pradnyana¹, I G M Darmawiguna², D K S Suditresna Jaya³, A Sasmita⁴

¹,²,³,⁴Department of Informatics, Universitas Pendidikan Ganesha, Bali - Indonesia

E-mail: gede.aditra@undiksha.ac.id

Abstract. The lecturers’ performance evaluation process can be carried out using an open questionnaire filled out by students at the end of the semester. In this questionnaire, students can provide an assessment in the form of comments, suggestions, and criticism of the lecturer’s performance, which can describe the level of student satisfaction with the lecture process. Conducting assessments or analyses on the open questionnaire entries manually will certainly impact the high costs time and energy. Sentiment polarity identification is a process in sentiment analysis that classifies text into a sentence or document and then determines whether the opinion expressed is positive, negative or neutral. In this research, a sentiment polarity detection system was developed in a lecturer evaluation questionnaire using the Support Vector Machine (SVM) method with a polynomial kernel. The test results showed that the SVM method's performance with the Polynomial kernel was strongly influenced by the value of the learning rate parameter, the maximum iteration, and the degree, with the optimal parameter values, respectively, 0.001, 200, and 0.3. The use of optimal parameter values in the process of identifying sentiment polarity obtained an accuracy value of 84.88%.

1. Introduction

Lecturers are an important and essential element in an educational institution. The roles, duties, and responsibilities of lecturers are very important in realizing the goals of national education, namely to educate the nation’s life, improve the quality of Indonesian people, which includes the quality of faith/piety, noble morals, and mastery of science, technology, and art, as well as realizing Indonesian society which is advanced, prosperous, and civilized. To carry out these very strategic functions, roles, and positions, professional lecturers whose performance can be measured are needed.

One of the lecturer assessment or evaluation processes can be done by using an assessment questionnaire for lecturers filled out by students. An assessment questionnaire is one of the commonly used evaluation tools, containing a list of a number of questions given to respondents to provide an assessment or feedback on an object or activity with a specific purpose [1]. Questionnaire is a data collection technique which is done by giving a set of questions or written questions to respondents to be answered [2]. A closed questionnaire is a questionnaire in the form of questions such as yes, no, multiple-choice, or a checklist, while open questionnaires are in the form of short answers or short descriptions in the form of fields.

One of the challenges in processing an honest assessment questionnaire is the analysis process. This is because, in the open questionnaire, the ‘respondents’ entries are in the form of suggestions,
input, or in the form of opinions. It is not a choice of answers as in a closed questionnaire. Conducting an assessment or one by one analysis of the open questionnaire entries manually will certainly impact high costs, such as time and energy [1]. A large amount of data will also add to the challenge of assessing several lecturers, and the data is very diverse. Therefore, a tool is needed that can be used to assist the open questionnaire assessment process so that the analysis of the results can be carried out more quickly and efficiently.

Sentiment analysis is a technique for detecting opinions on a subject (such as individuals, organizations, or products) in a data set [3–6]. Sentiment analyst or Opinion mining is a system analysis process that converts a particular activity, topic, or product into a system whose process is carried out automatically, including affect analysis, emotion analysis, subjectivity analysis, and sentiment polarity or a positive and negative assessment from the comments on a document or review and the classification of a text [1,7]. Identification of sentiment polarity is one of the fundamental roles in sentiment analysis, namely classifying the text in a sentence or document and then determining the opinion expressed in the sentence or document whether it is positive, negative or neutral. There are two main approaches to extracting sentiment information on a text: the lexicon-based approach and the machine learning-based approach [8]. Several methods are often used in machine learning for sentiment analysis problems, including Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree, and Maximum Entropy Classification [8,9].

The SVM method is a classification method that uses a statistical approach to maximize the distance (margin) between the data and the hyper-plane separator. SVM has been used efficiently in many text classification cases because this method has many advantages, such as easily adjust functions and it works well on high-dimensional data [10]. SVM is a machine learning technique that solves over-fitting problems, local optimal solutions, low convergence ratios, and excellent generalizability in small sample scenarios. Research conducted by Chandani, et al, Arifin and Sasongko, and Supriyatna & Mustika also proved that SVM is a classification method that has better classification accuracy than other classification methods[11–13]. In this study, we developed and analysed the performance of the polarity identification system using the SVM method with the polynomial kernel in lecturer performance evaluation.

2. Method
As shown in Figure 1, to identify sentiment polarity in the lecturer evaluation questionnaire, several steps need to be done to ensure the best identification results are obtained.

2.1 Sentiment Data Collection and Labelling
The first stage is the data collection process. In this study, data were taken from the Academic Information System (SIAK) of the Ganesha University of Education. The data used was data from student suggestions and criticisms to lecturers recorded in 2017, 2018, and 2019. The specifications of the data to be used can be seen in the table. In the evaluation process or method performance measurement, 100 data samples from each year will be used, which are taken randomly. After the data have been successfully collected into a dataset, the next step is labelling. In the labelling process, the existing sample data were labelled with the appropriate polarity, namely positive comments or negative comments. In measuring method performance, the sample data were further divided into training data sets and testing data sets.
Figure 1. Flow Chart of the Sentiment Polarity Identification Process Using the SVM Method.

2.2 Text Pre-processing

Before the classification process with the SVM method was carried out, sentiment data should go through the text pre-processing stage. At this stage, the data was prepared to become data ready for [14–17]. The pre-processing stage of this text begins with tokenization. Tokenization (tokenization) is a process carried out to break or cut a sentence into a collection of parts or words. The result of this deduction is also often referred to as a token. In the next process, each token produced from the tokenization was carried out by a case-folding process. The case folding process made the entire characters in the token become lower case. The next stage is the process of getting the root word from an affixed word by separating each word from the root word and its prefix and ending which is known as stemming. The stemming algorithm used was the Nazief & Adriani algorithm. The cleansing process will remove punctuation marks and unnecessary characters from a text. The next process was filtering, which eliminated words that appear in large numbers but were deemed to have no meaning (stop words). The final stage of text pre-processing is the process of calculating the word weight (Term Weighting). Each word was weighted using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF method \( w(t, d) \) is shown in equation 1.

\[
 w(t, d) = tf(t, d) \times idf 
\]

\[
 idf = \log\left(\frac{N}{df}\right) 
\]

Where, \( tf(t,d) \) is the occurrence of the word \( t \) in the document \( d \), \( N \) the number of documents in the document and \( df \) is the number of documents that contain the word \( t \). In this study, a document was a comment written by students for a lecturer.
2.3 Polarity Identification Process using Support Vector Machine (SVM)

SVM is a method that can solve problems linearly or non-linearly. In solving non-linear problems in SVM, the kernel concept was used in high-dimensional workspaces by looking for hyperplanes to maximize margins between data classes. In classifying data using the SVM method, the K kernel function \((x_i, x_d)\) was used. The kernels used in this study are as follows:

\[
K(x_i, x_d) = (X_i^T X_j + C)^d, \gamma > 0
\]

(3)

Processing of training data used a sequential learning algorithm because it is a simple algorithm that 'doesn’t take up much time, with the following calculation stages:

1. Initialize parameters required by SVM
2. Calculate the Hessian matrix obtained from the multiplication between the polynomial and \(y\) kernels, which are 1 and -1 vectors. The equation of the Hessian matrix is:

\[
D_{ij} = y_i y_j (K(x_i, y_i) + \lambda^2)
\]

(4)

3. Perform the following calculations to iterate data i to j
   a. \(E_i = \sum_j a_j D_{ij}\)
   b. \(\delta a_i = \min(\max(\gamma(1 - E_i), a_i), C - a_i)\)
   c. \(a_i = a_i + \delta a_i\)
4. Do the three steps above until you reach the maximum iteration limit.
5. The sequential learning process will get the value from the support vector (SV), where the \(SV = ai > SV\) threshold. After that, it is necessary to calculate the bias value \(b\) as in equation 5.
6. \(b = -\frac{1}{2}(\sum_{i=0}^{N} a_i y_i K(x_i, x^-) + \sum_{i=0}^{N} a_i y_i K(x_i, x^+))\)

(5)

To find out the classification results of comments on a certain sentiment class, the \(f(x)\) function calculation process was carried out. If the result of this function was negative, then the sentiment polarity of the comment was identified as a negative sentiment. If the value of the function was positive, then the sentiment polarity of the comments was identified as a positive sentiment. The function \(f(x)\) is obtained from the following equation:

\[
f(x) = \sum_{i=0}^{N} a_i y_i K(x_i, x) + b
\]

(6)

2.4 Performance Evaluation

System evaluation focuses on evaluating the performance of the SVM method in identifying polarity sentiment from the commentary data of the lecturer evaluation questionnaire. At this stage, the performance of the calculations that have been carried out can be seen from the value of three parameters, namely accuracy, precision, and recall [14].

3. Result

3.1 Polarity Sentiment Identification System Development Using SVM Method with Polynomial Kernels.

In this study, a web-based system was developed using the Laravel framework and the Python programming language. The Python programming language used when developing the SVM method
utilized the Scikit-Learn library. Data communication between Laravel websites and Python was done by API using Flask. Flask is a micro web framework from Python. The results of implementing a web-based system for the identification process are shown in Figure 2.

![User Interface](image)

**Figure 2.** User Interface.

Identification of sentiment polarity using the SVM method with data sources in vectors containing weighted words using the TF-IDF method. These words have also gone through pre-processing stages such as stemming and stopword removal processes. The following is a snippet of source code in Python for the classification process using the SVM method.

```python
@app.route('/api/run_svm', methods=['GET'])
def run_svm():
    data = DatabaseSentimen(mysql)
    # mengambil vector data train dan labelnya
    X_train, y_train = data.data_train()
    label_feature = data.list_fitur()
    # mengambil vector data test dan ID nya di database
    X_test, id_ks, kata_kata = data.data_kritik_saran() #
    # Klasifikasi SVM dengan Kernel Poly
    classifier = SVC(kernel='poly', degree=1, gamma=0.1, max_iter=200)
    classifier.fit(X_train, y_train)
    y_pred = classifier.predict(X_test)
    X_test_list = X_test.tolist()
    y_pred_list = y_pred.tolist()
    return jsonify({'X_test':X_test_list, 'y_pred':y_pred_list})
```

### 3.2 Method Performance Evaluation

#### 3.2.1 Testing SVM Parameters with Polynomial Kernel

In this study, testing was carried out to determine the effect of parameters on the SVM method's performance and determine the optimal parameters. The SVM parameters tested included the polynomial kernel degree value, the maximum iteration, and the learning rate constant. The observed values of this parameter testing were the value of F-Measure.
The test results shown in Figure 3.a. shows the best learning rate constant is 0.001. This is indicated by an accuracy rate of 87.63%. The constant function of the learning rate is to control the speed of the training process. The Learning Rate constant depends on the number of iterations to achieve convergence. The test results showed that the optimal value of the learning rate constant was 0.01 and occurs when the maximum iteration reached 200 times. Changes in the value of the learning rate in the system affect the value of $\delta_\alpha$. This was because the learning rate was one of the candidate values that influenced the formation of alpha values and the formation of support vector sets. A decrease in the accuracy occurred when the learning rate was too large. Increasing the value of the learning rate had an impact towards the acceleration of the training process but reduced the level of accuracy.

The best accuracy value was obtained from the maximum iteration of 200 times from the maximum iteration test. This is shown in Figure 3.b., the best accuracy was obtained during the 200 iterations of 84.75%. However, increasing the number of iterations continuously did not mean an increase in the accuracy. A decrease in the level of accuracy occurred because the successive learning stages changed the value $\alpha_i$. The change in $\alpha_i$ value was caused by the non-convergent value of $\alpha_i$, which can be proven by the change in the value of $\alpha_i$. High accuracy occurs because iterations give time to reach the convergent $\alpha$ value seen from the change in the $\alpha$ value. While the low accuracy occurred because the $\alpha$ value obtained had not reached convergence, so the $\alpha$ value obtained as not optimal. The small iteration value caused less $\alpha$ value to be used as a support vector.

Testing the parameter value of the Degree Polynomial (lambda) was carried out to get the lambda value that produced the best accuracy. The value of $\lambda$ (lambda) to be tested was 0.1; 0.3; 0.5; 0.7; 0.9; 1; 2; 3; and 4. Based on Figure 3.c., it was found that the parameter $\lambda$ (lambda) affected the accuracy results. This can be seen when $\lambda$ (lambda) = 0.1 to 0.5 produced a high enough accuracy, then when the value of $\lambda$ (lambda) = 0.7 began to decrease in accuracy until $\lambda$ (lambda) = 1. Then
accuracy has increased when $\lambda$ (lambda) = 2. The value of $\lambda$ (lambda) = 0.3 is the best result with an accuracy of 88.86%. A value of $\lambda$ (lambda) that is too large will cause low accuracy. The value of $\lambda$ (lambda) was too large and would affect when calculating the Hessian matrix. The Hessian matrix calculation would be slower because the large value of $\lambda$ (lambda) will result in the speed of reaching convergence being slower and the instability of the learning process. Increasing the degree value in the polynomial kernel affected the results of the Hessian matrix calculation, which functions to find the optimum value in each data document. The hessian matrix results were used to calculate the error rate value for each document and affect the formation of support vector values.

3.2.2 Performance Test of Sentiment Polarity Identification Using the SVM Method. In this test, testing the accuracy of the classification results and processing time by the system was carried out. This test used the 5-Fold cross-validation method with 1000 comments of data. By using the optimal parameter values, the measurement accuracy and processing time values can be seen in Table 1.

| Fold | Accuracy (%) | Running Time (Second) |
|------|--------------|-----------------------|
| 1    | 83.35        | 134.65                |
| 2    | 85.47        | 166.54                |
| 3    | 84.81        | 148.42                |
| 4    | 86.03        | 165.79                |
| 5    | 84.76        | 170.24                |
|      | **Average**  | **84.88**             | **157.13**            |

4. Conclusions
From the results of this study, it can be concluded that the SVM method with a polynomial kernel can be used to identify sentiment polarity in a lecturer performance evaluation questionnaire. The test results revealed the value of the learning rate parameter, maximum iteration, and degree in the polynomial kernel significantly affected the quality of the identification results of sentiment polarity. The optimal parameter values of the learning rate, maximum iteration, and degree found were 0.001, 200, and 0.3, respectively. Using these parameters produced an average accuracy of 84.88% with a processing time of 157.13 seconds.

Acknowledgments
We express our gratitude for the support provided by Lembaga Penelitian dan Pengabdian Kepada Masyarakat, Universitas Pendidikan Ganesha for funding assistance for this research so that this research can be carried out and completed.

References
[1] Riany J, Fajar M dan Lukman M P 2016 Penerapan Deep Sentiment Analysis pada Angket Penilaian Terbuka Menggunakan K-Nearest Neighbor J. Sisfo06 147–56
[2] Risdianto E 2008 Pengembangan Multimedia Interaktif (MPI) pada Praktikum Fisika Dasar I J. Exacta06 9–16
[3] Sunardi, Fadlil A dan Suprianto 2018 Analisis sentimen menggunakan metode naïve bayes classifier pada angket mahasiswa SAINTEKBU J. Sains dan Teknol.10 1–9
[4] Willana W, Hanasbey S H, Awinero M R, Modouw J V A dan Sitokdana M N N 2018 Analisis Sentimen Terhadap Opini Masyarakat Indonesia Mengenai Bukalapak Semin. Nas. Teknol. Inf. dan Multimed. 2018 Univ. AMIKOM Yogyakarta 1–6
[5] Haryanto D J, Muflikhah L dan Fauzi M A 2018 Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya2 2909–16
[6] Haryanto B, Ruldeviyani Y, Rohman F, T.N. J D, Magdalena R dan Muhamad Y F 2019
Facebook Analysis of Community Sentiment on 2019 Indonesian Presidential Candidates from Facebook Opinion Data *Procedia Comput. Sci.* **161** 715–22

[7] Somantri O dan Apriliani D 2018 Feature Weights Menggunakan Particle Swarm Optimization Untuk Sentiment Analysis Penilaian Kepuasan Pelanggan Makanan Kuliner *Semin. Nas. Sains Teknol. Inf.* 697–704

[8] Song C, Wang X K, Cheng P fei, Wang J qiang dan Li L 2020 SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis *Knowledge-Based Syst.* 105572

[9] Adi S, Wulandari M, Mardiana A K dan Muzakki A 2018 Survei: Topik Dan Tren Analisis Sentimen Pada Media Online *Semin. Nas. Teknol. dan Multimed.* 55–60

[10] Basari A S H, Hussin B, Ananta I G P dan Zeniarja J 2013 Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization *Procedia Eng.* **53** 453–62

[11] Chandani V, Wahono R S dan Purwanto 2015 Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentemen Review Film *J. Intel. Syst.* **1** 56–60

[12] Arifin O dan Sasonoko T B 2018 Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine Dan Naive Bayes Classifier *Semin. Nas. Teknol. Inf. dan Multimed.* 2018 67–72

[13] Supriyatna A dan Mustika W P 2018 Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil *J-SAKTI (Jurnal Sains Komput. dan Inform)* **2** 152

[14] Hadna M S, Santosa P I dan Winarno W W 2016 Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter *Semin. Nas. Teknol. Inf. dan Komun.* **2016** 57–64

[15] Darmawiguna I G M, Pradnyana G A dan Santyadiputra G S 2019 The Development of Integrated Bali Tourism Information Portal using Web Scrapping and Clustering Methods *J. Phys. Conf. Ser.* **1165** 0–9

[16] Permana A A J dan Pradnyana G A 2019 Recommendation Systems for internship place using artificial intelligence based on competence *J. Phys. Conf. Ser.* **1165** 0–7

[17] Pradnyana G A, Setiawan G A dan Wirawan I M A 2019 Development of Non-QWERTY Balinese Script Keyboard Through Tamiang Keyboard Optimization with Letter Frequency Concept *J. Phys. Conf. Ser.* **1387** 0–6