International Journal of Artificial Intelegence Research
Vol. 7, No. 1, June 2023, pp. 64 - 71
ISSN 2579-7298

A Comparison Support Vector Machine, Logistic Regression And Naïve Bayes For Classification Sentimen Analisys user Mobile App

Kiki Ahmad Baihaqi [a,1,*], Iwan Setyawan [a,2], Danny Manongga [a,3], Hendryanto Dwi Purnomo [a,4], Hendry [a,5], Ahmad Fauzi [b,6], Aprilia Hananto [b,7]

[a] Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Jl. Diponegoro No.52-60, Salatiga 50711 , Indonesia
[b] Fakultas Ilmu Komputer, Universitas Buana Perjuangan Karawang, Jl. Ronggowaluyo Telukjambe Timur, Karawang 41352, Indonesia
[1] 982022029@student.uksw.edu.*; [2] iwan.setyawan@uksw.edu; [3] danny.manongga@uksw.edu, [4] hindriyanto.purnomo@uksw.edu, [5] hendry@uksw.edu, [6]afauzi@ubpkarawang.ac.id, [7] aprilia@ubpkarawang.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 09 March 2023
Revised 30 May 2023
Accepted 27 June 2023

Keywords
Data Mining
Scraping
Naïve Bayes
Support Verctor Machine
Logistic Regression

ABSTRACT

Data is the most important thing, the use of data can be useful to get an evaluation from the user of a system or application that is built based on mobile. Not only, the assessment or acceptance results of mobile applications during the trial stage are considered important, assessments and comments from direct users are also important things that can be input for mobile application developers. Data mining, or known in English as data mining, is the answer to the process of retrieving data on any media. In this research, data mining is carried out on the media mobile application download service provider Google Playstore, which provides data in the form of comments and ratings. After scraping the data and obtaining the latest data parameters determined by the latest 2000 comments, the data is pre-processed by removing the emot icon character and eliminating unneeded variables so that the data obtained can be processed to the next stage, namely classification based on ratings and sentiment comments. The algorithms used or compared in this research are Support Vector machine, logistic regression and naïve bayes which are known to be reliable in data mining processing. In this research, the accuracy results are 88% for SVM, 90.5% for Logistic Regression and 91% for naïve bayes.

1. Introduction

There are a lot of mobile app users [1], which also makes many companies try to create mobile-based applications. Therefore, many of the evaluations given by users face these applications in comments and also in satisfaction ratings. Based on this, it is necessary to prove whether the three algorithms used in this research can work optimally. Naive Bayes, support vector machine and logistic regression are the same types of supervised learning [2], [3]. Then the application that is the object of this research is the health service provider that is BPJS, where the data and users are adequate. At the time of this research, its downloads had reached more than ten million.

The text used as a reference to assess the feelings of the user also corresponds to the three algorithms selected [4]. The use of supervised learning algorithms in data mining depends on the presence of "training data" consisting of appropriate input and output. This algorithm studies the relationship between this input and output to create predictive models that can be used to predict unknown outputs based on the given input [5].

Data mining has the advantage of processing information that is still unclear or vague [6]. Where the data is mined with the scrapping process and then will produce the data, the process will then back up to the conclusion or classification stage, after which it can then be seen how the third level of

accuracy of the algorithm And *it could be a further research recommendation to optimize these three fundamental algorithms.*

2. Method

The phase in this research is only through a few stages, which can be described in image 1 of the research course. This is then solved with algorithm performance testing.

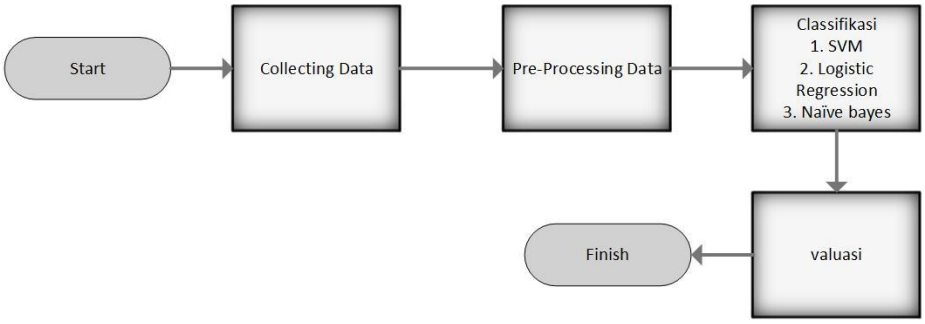

Fig 1. Research flow chart

2.1. Collecting Data

The data collection process used scraping taken from the Google Play Store website after determining the parameters and the amount of data required. The data was collected from 2,000 recent comments due to responses to applications that are already being developed. Figure 2 shows the flow of data mining processes on the website Google Playstore [7] [8].

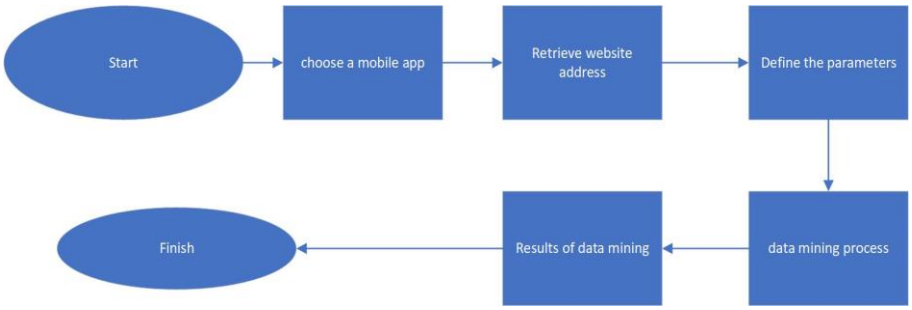

Fig 2. Flow chart of Collecting data

This phase is all conducted using Google Collab, and its implementation only generates from the web.

2.2. Pre-Processing

The next process is to clear the data that initially has many reading marks and then receives icons and others. It is done with case folding, tokenizing, stemming, stopword removal, and TF-IDF [9], [10]. Pre-processing avoids irregular, imperfect, and inconsistent data.

This stage is also a maximum-determining stage of whether or not the algorithm will work, such as the phase of elimination of non-standard languages. The parameters in the Python code must be given the language or word parameters that want to be removed and those that will be retained. This is the result that will optimize the algorithm in the next work.

2.3. Classification

Classification is one of the main tasks of supervised learning in machine learning and data mining. The concept of classification comes from the main purpose of this technique, which is to predict the class of the input data [11], [12]. The purpose of this classification is a computer process that uses a data mining algorithm to process the review set of the Google Play Store Grab application. Several

algorithms are commonly used for data mining classification, including Naive Bayes, Support Vector Machine (SVM), and Logistic Regression.

2.4. Logistic Regression

Logistic regression is a type of regression analysis used to explain the relationship between a dependent variable and an independent variable by linking one or more independent variables to the dependent variable. Class types can be 0 and 1, true or false, major or minor. The type of independent variable is category. This distinguishes logistic regression from multiple or other linear regression [13]–[15]. The logistic regression equation is expressed by Eq.

$$\text{Ln}\left(\frac{p}{1-p}\right) = B_o + B_1 X \quad (1)$$

B_o is a constant, while B_1 is a coefficient of each variable, the value of p is found hiin equation (2).

$$p = \frac{e(B_0+B_1 X)}{(1-e(B_0+B_1 X))} \quad (2)$$

2.5. Naïve Bayes

One of the algorithms that serves to divide classes in the process of classification is this algorithm. Most of the time, these algorithms are used for data mining, making them one of the most popular algorithms. For that, we see what approaches this algorithm uses in data mining research in the formula (3) [16]–[18].

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad (3)$$

X is the unknown class data, H is the separate class x hypothesis data, while P(X|H)is the conditional probability of the H hypotheses. P (H) is the probability of the hypothesis H.

2.6. Support Vector Machine

The vector machine support algorithm is a popular machine-learning technique for text classification and has good performance in many fields. The SVM's ability to detect hyperplanes separately between two different classes is maximized, and the SVM provides the maximum distance between the data that is closest to the hyperplane. In this research, the kernel formula is used, as seen in the formula (4).

$$K(x_i x) = x_i^T x \quad (4)$$

2.7. Evaluation

This step is taken to ensure the validity of the test; the aim of this evaluation is to find the best results from the test results [20]. Measure the accuracy of the model using a confusion matrix. A confusion matrix is a tool for analyzing how a classification model identifies a different set of data [21].

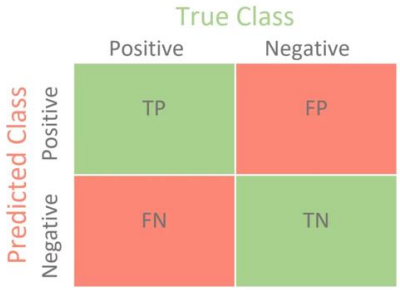

Fig 3. Confusion Matrix

- TP (True Positive) = Positive data is correctly classified
- TN (True Negative) = Negative data is correctly classified

- FP (False Positive) = Negative data is positively classified
- FN (False Negative) = Positive data is negatively classified

3. Results and Discussion

3.1. Result of Clarification

This research found results on the object aimed at, which is a mobile application provider or organizer of national health jamina in Indonesia, namely BPJS. As a result, the sentiment can be seen in Figure 4.

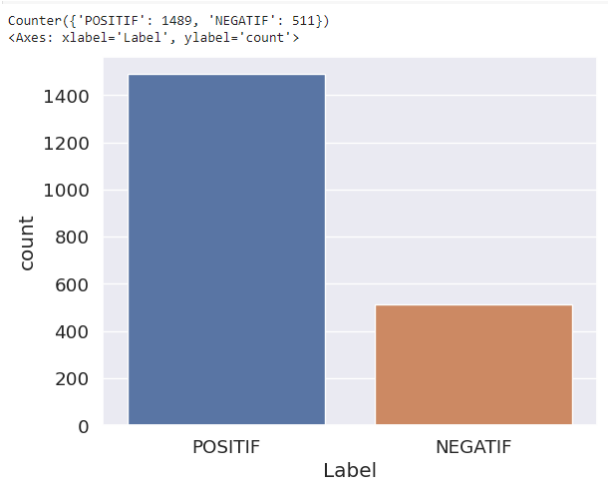

Fig 4. Classification of Positive and Negative Emotions

Results are obtained from data mining using scraping techniques, and after classification, there are more than 1400 positive comments in the data, while comments and negative ratings are only rated for 500 comments. Where these results are more than adequate in terms of mobile application development as assessed by the user directly.

3.2. Comparison of algorithms

Comparison in this research is comparing accuracy, precision and recall using the same datasets and test data. The results can be seen in Table 1.

Tabel 1. Hasil Perbandingan tingkat performa

NO	Algoritma	Akurasi(%)	Precision(%)	Recall(%)
1	Support Vector Machine	88.8	92	94
2	Naive Bayes	91.6	97	92
3	Logistic Regression	90.5	93	94

The explanation can be seen in the results of the encoding performed at the time of the classification process in Figures 5, 6, and 7.

klasifikasi report SVM pada data testing

	precision	recall	f1-score	support
NEGATIF	0.79	0.73	0.76	147
POSITIF	0.92	0.94	0.93	453
accuracy			0.89	600
macro avg	0.86	0.84	0.85	600
weighted avg	0.89	0.89	0.89	600

Fig 5. SVM Report Classification

klasifikasi report Naive Bayes pada data testing

	precision	recall	f1-score	support
NEGATIF	0.78	0.91	0.84	147
POSITIF	0.97	0.92	0.94	453
accuracy			0.92	600
macro avg	0.88	0.91	0.89	600
weighted avg	0.92	0.92	0.92	600

Fig 6. Naïve Bayes Report Classification

klasifikasi report Logistic Regression pada data testing

	precision	recall	f1-score	support
NEGATIF	0.82	0.79	0.80	147
POSITIF	0.93	0.94	0.94	453
accuracy			0.91	600
macro avg	0.87	0.87	0.87	600
weighted avg	0.90	0.91	0.90	600

Fig 7. Logistical Regression Report

3.3. Evaluation

In this research, the evaluation is seen from the confusion matrix, which will explain both the positive and negative aspects of the pedicure and whether the outcome is direct or not.

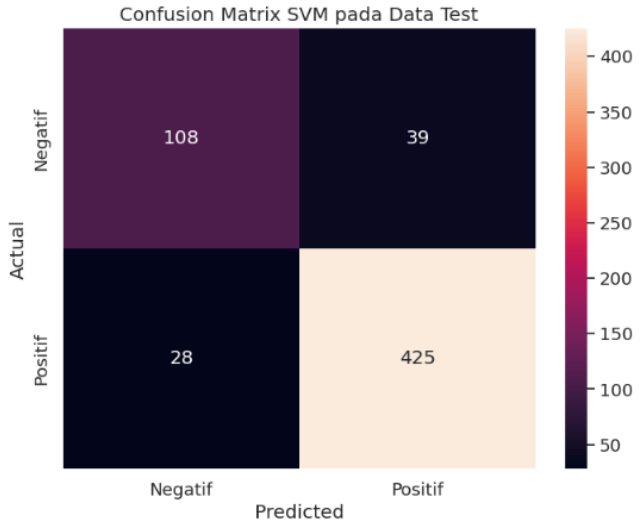

Fig 8. Confusion Matrix SVM

Based on Figure 9 above, the number of TP is 108, FP is 28, FN is 39, and TN is 425. How to calculate it manually is with TP + FP + FN + TN = 600, and next (TP + TN)/600 = 0.888. The result of the calculation of the Confusion Matrix Support Vector Machine is 88.8%.

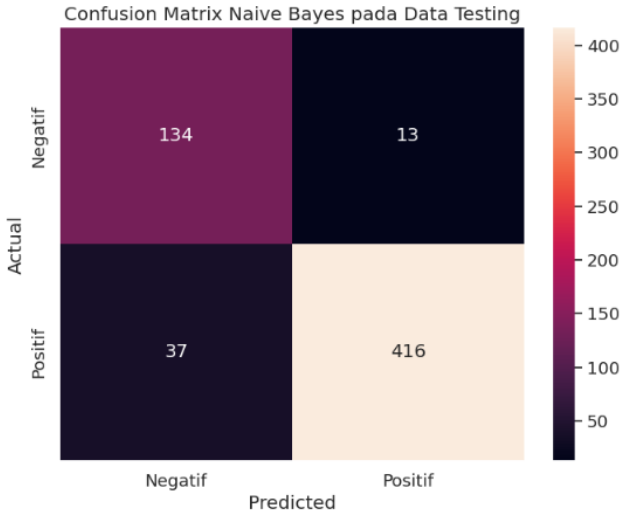

Figure 9. Confusion Matrix Naïve Bayes

Based on Figure 9, the number of TP is 134, FP is 37, FN is 13, and TN is 416. How to calculate it manually is with TP + FP + FN + TN = 600, and next (TP + TN)/600 = 0.916. The result of the calculation of the Confusion Matrix Naïve Bayes was 91.6%.

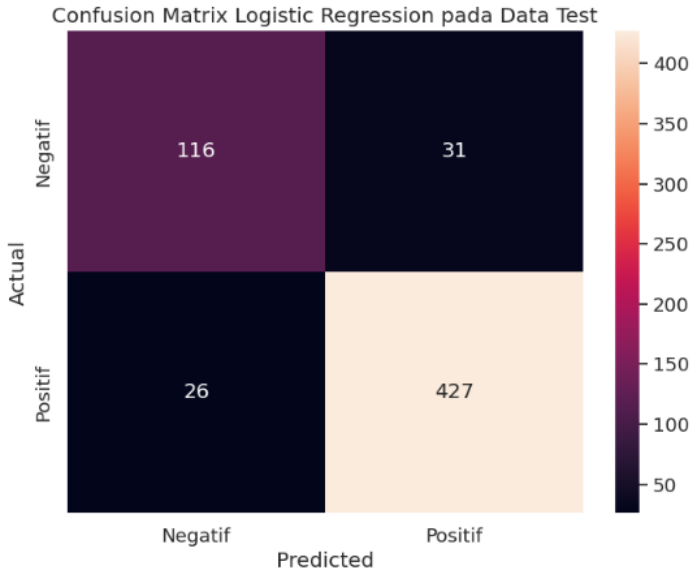

Figure 10. Confusion Matrix Logistic Regression

Based on Figure 9, the number of TP is 116, FP is 26, FN is 31, and TN is 427. How to calculate it manually is with TP + FP + FN + TN = 600, and next (TP + TN)/600 = 0.905. The result of the calculation of the Confusion Matrix Naïve Bayes is 90.5%.

4. Conclusion

The results of this research showed that the accuracy of sentimental analysis was outperformed by Naïve Bayes, followed by logistic regression, and finally by the support vector machine. Thus, for data mining performed on the media, Google Play Store can use the Naïve Bayes algorithm and compare it with other supervised learning algorithms. This research also depends on what mobile applications are used as objects because, in some cases, it fails to obtain application data created based on the.gov extension.

References

[1] T. Li *et al.*, "Smartphone App Usage Analysis: Datasets, Methods, and Applications," *IEEE Communications Surveys and Tutorials*, vol. 24, no. 2, pp. 937–966, 2022, doi: 10.1109/COMST.2022.3163176.

[2] M. Aniche, E. Maziero, R. Durelli, and V. H. S. Durelli, "The Effectiveness of Supervised Machine Learning Algorithms in Predicting Software Refactoring," *IEEE Transactions on Software Engineering*, vol. 48, no. 4, pp. 1432–1450, Apr. 2022, doi: 10.1109/TSE.2020.3021736.

[3] M. Raza, N. D. Jayasinghe, and M. M. A. Muslam, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in *International Conference on Information Networking*, IEEE Computer Society, Jan. 2021, pp. 327–332. doi: 10.1109/ICOIN50884.2021.9334020.

[4] A. H. Espejel and F. J. Cantu-Ortiz, "Data Mining Techniques to Build A Recommender System," in *Proceedings - 2021 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 217–221. doi: 10.1109/ISCSIC54682.2021.00047.

[5] P. Pierleoni, L. Palma, A. Belli, S. Raggiunto, and L. Sabbatini, "Supervised Regression Learning for Maintenance-related Data," in *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, IEEE, Sep. 2022, pp. 1–6. doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927904.

[6] T. T. Chikohora and E. Chikohora, "An Algorithm for Selecting a Data Mining Technique," in *2021 3rd International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/IMITEC52926.2021.9714525.

[7] RVS Technical Campus, IEEE Aerospace and Electronic Systems Society, and Institute of Electrical and Electronics Engineers, "Data Analysis by Web Scraping using Python," *Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]*, 2019.

[8] N. Narayani, P. Kumar, and D. Kumar, "Web Scraping & Automation Bot Using Python : Using Python to automate all the tasks," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, IEEE, Dec. 2022, pp. 1343–1346. doi: 10.1109/ICAC3N56670.2022.10074375.

[9] V. Desai and D. H A, "A Hybrid Approach to Data Pre-processing Methods," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, IEEE, Nov. 2020, pp. 1–4. doi: 10.1109/INOCON50539.2020.9298378.

[10] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, "The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, IEEE, Mar. 2019, pp. 279–283. doi: 10.1109/IEMECONX.2019.8877011.

[11] E. U. Chye, E. I. Glinkin, and A. V. Levenets, "Measurement Data Classification in Information and Measuring Systems," in *2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, IEEE, Mar. 2019, pp. 1–5. doi: 10.1109/ICIEAM.2019.8742926.

[12] M. Zheng, "The Classification and Classification of Big Data Based on the Internet of Things," in *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/ICMNWC56175.2022.10031772.

[13] Z. Aung, I. S. Mihailov, and Y. T. Aung, "Models and Data Mining Algorithms for Solving Classification Problems," in *2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)*, IEEE, Nov. 2019, pp. 532–536. doi: 10.1109/SUMMA48161.2019.8947555.

[14] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, IEEE, Oct. 2019, pp. 135–139. doi: 10.1109/ICCSNT47585.2019.8962457.

[15] I. C. Juanatas and R. A. Juanatas, "Predictive Data Analytics using Logistic Regression for Licensure Examination Performance," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, IEEE, Dec. 2019, pp. 251–255. doi: 10.1109/ICCIKE47802.2019.9004386.

[16] A. Tariq *et al.*, "Modelling, mapping and monitoring of forest cover changes, using support vector machine, kernel logistic regression and naive bayes tree models with optical remote sensing data," *Heliyon*, vol. 9, no. 2, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13212.

[17] W. D. Herlambang, K. A. Laksitowening, and I. Asror, "Prediction of Graduation with Naïve Bayes Algorithm and Principal Component Analysis (PCA) on Time Series Data," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2021, pp. 645–649. doi: 10.1109/ICoICT52021.2021.9527443.

[18] M. C. Kirana, M. Fani, T. S. Kartikasari, and M. Nashrullah, "Downtime Data Classification Using Naïve Bayes Algorithm on 2008 ESEC Engine," in *2020 3rd International Conference on Applied Engineering (ICAE)*, IEEE, Oct. 2020, pp. 1–6. doi: 10.1109/ICAE50557.2020.9350377.

[19] J. Huang, J. Zhou, and L. Zheng, "Support Vector Machine Classification Algorithm Based on Relief-F Feature Weighting," in *2020 International Conference on Computer Engineering and Application (ICCEA)*, IEEE, Mar. 2020, pp. 547–553. doi: 10.1109/ICCEA50009.2020.00121.

[20] D. van Herwerden, J. W. O'Brien, P. M. Choi, K. V. Thomas, P. J. Schoenmakers, and S. Samanipour, "Naive Bayes classification model for isotopologue detection in LC-HRMS data," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, Apr. 2022, doi: 10.1016/j.chemolab.2022.104515.

[21] A. Rojas and G. J. Dolecek, "Evaluation of Supervised Machine Learning Classification Algorithms for Fingerprint Recognition," in *2021 Global Congress on Electrical Engineering (GC-ElecEng)*, IEEE, Dec. 2021, pp. 1–4. doi: 10.1109/GC-ElecEng52322.2021.9788164.