# Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review



# Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review

**Dwi Andini Putri[1]*,Dinar Ajeng Kristiyanti[2], Elly Indrayuni[3], Acmad Nurhadi[4], and Denda Rinaldi Hadinata[5]**

[1,2]Informatics Engineering, STMIK Nusa Mandiri, Indonesia
[3]Accounting Information System, Universitas Bina Sarana Informatika, Indonesia
[4]Computer Technology, Universitas Bina Sarana Informatika, Indonesia
[5]Management Science, IPB University, Indonesia

E-mail: dwiandini@nusamandiri.ac.id

**Abstract.** Cashless payment habits have been widely applied to the transportation system, restaurants and shops in the mall area. So, it is normal if the growth of mobile payment services is currently very rapid. The ease of doing transactions and promotional offers in the form of points and cashback in digital wallet applications (e-wallets) is very beneficial for its users. One of the most popular e-wallets is OVO. With so many reviews about OVO customer opinions on social media, there has also been a lot of public opinion. These opinions can produce negative or positive statements. Sentiment analysis is the mining of opinions or text to classify opinions or user reviews, of a brand reviews, product reviews, or service reviews into the category of positive or negative opinion. The methods used in this research are Naive Bayes and SVM. Both of these algorithms are the best algorithms widely used in text classification research. However, both of these algorithms have weaknesses in several parameters. So, in this study Feature Selection is used to improve its performance. The evaluation was carried out using 10-fold cross validation. Measurement accuracy is measured by confusion matrix and ROC curves. This study uses 500 positive reviews and 500 negative reviews as data training. The results of this study indicate that the use of PSO-based Naive Bayes algorithm produces an accuracy value of 93.10 percent with an AUC value of 0.750. While the results of research from the PSO-based SVM algorithm are 91.30 percent with an AUC value of 0.970. Based on these results the accuracy value generated by the Naive Bayes algorithm is classified as Fair Classification and SVM is classified as Excellent Classification. The AUC value generated by the Naive Bayes algorithm is also smaller than SVM. Therefore, in this study found that SVM is the best algorithm in classifying text.

## 1. Introduction

Cashless payment habits have been widely applied to the transportation system, restaurants and shops in the mall area. Most residents in urban areas are now accustomed to transacting using digital wallet (e-wallet). So, it is normal if the growth of mobile payment services is currently very rapid. The ease of doing transactions and promotional offers in the form of points and cashback in digital wallet applications (e-wallets) is very beneficial for its users not only for buyers but also for sellers. The seller does not need to bother preparing change because the

nominal amount of payment is appropriate. E-wallets in Indonesia come in the form of mobile applications [1].

Nowadays, one of the most popular e-wallet applications in Indonesia is OVO. With so many reviews about OVO customer opinions on social media, there has also been a lot of public opinion. These opinions can produce negative or positive statements. Therefore, OVO can use these opinions as material for evaluating the company's strategy in maintaining customers' trust or comfort when using their applications. Classification techniques that are often used and have high accuracy in sentiment analysis are (NB) and (SVM) [2].

Several studies on sentiment analysis have been carried out including research [3] about the Go-Pay e-wallet which is part of the Gojek application and one of the most popular fintechs in Indonesia. The study uses the lexicon-based method to provide an opinion label into positive and negative classes. The classification methods used are linear kernel SVM and polynominal kernel. Then research conducted by [1][4][5]. The algorithm used in the study is Naive Bayes Classifier (NB), with the optimization of the use of PSO Feature Selection (FS). The result is that the accuracy improvement is very significant with the use of PSO Feature Selection. Furthermore, in research conducted by [6] concluded that the KNN algorithm has a higher accuracy than NB. In this study Naive Bayes classifier will be compared with SVM classifier using PSO Feature Selection method to classify a review of e-wallets so that it can be seen which classifier is better. NB has several advantages such as simple, fast, and possess high accuracy. Many researchers have classified sentiments using NB. However, this classification has a limitation that is unable to fulfill the assumption of independence between attributes. This can have an effect on the level of accuracy [7]. However, (SVM) also has a disadvantage, that is in the selection of appropriate parameters [5]. Based on the weaknesses in the two algorithm methods, we need Feature Selection to improve optimal analysis. Feature selection has been proven to make classifiers better and optimal to be able to identify in the machine learning process [8]. In this study, the PSO algorithm will be used as a FS to review OVO e-wallet sentiment analysis by comparing the NB and SVM.

### *1.1. Sentiment Analysis*

SA can be considered as a text classification process that has 3 stages namely, the document stage, the sentence selection stage and the aspect stage [9]. The steps are commonly found in sentiment analysis text classification namely (i) Define the dataset domain, (ii) Pre-processing, and (iii) Transformation.

### *1.2. Naive Bayes Algorithm (NB) and Support Vector Machine(SVM)*

NB is one classification algorithm which is simple but has high accuracy. NB has the disadvantage of being very sensitive in feature selection so that it can affect its accuracy. Feature-based classification methods developed in these studies produce good accuracy [10]. Naive Bayes has several advantages namely simple, fast and possess high accuracy. Many researchers have classified sentiments using Naive Bayes [11]. However, this classification has a limitation that is unable to fulfill the assumption of independence between attributes. This can have an effect on the level of accuracy [6]. SVM has the advantage of other algorithms that are able to identify separate hyperplanes and maximize margins [12].

### *1.3. Particle Swarm Optimization(PSO)*

Several studies have used PSO to improve algorithm optimization [11][13][9][13]. PSO has the advantage of being easy to implement and adjusting parameters. The PSO approach can also be used to find a solution [16].

### *1.4. Related Work*

This section shows some research on sentiment analysis using NB and SVM The reasearch conducted by [17] using the Naive Bayes Classifier and PSO Feature Selection with an accuracy of 92.8%. Then research conducted by [1] using the Naive Bayes Classifier and PSO Feature Selection with an accuracy of 83,60% and the research conducted by [18] using the SVM Classifier dan PSO Feature Selection with an accuracy of 77%.

**Table 1.** Comparasion of Related Research

| Title | Preprocessing | Feature Selection | Classifier | Accuracy |
|---|---|---|---|---|
| Particle Swarm Optimization Based Naïve Bayes Data Algorithm for Disease Detection Heart [20] | - | Particle Swarm Optimization | Naive Bayes | 92,86% |
| Sentiment Analysis Analysis of E-Wallet Sentiments on Google Play Using Naive Bayes Algorithm Based on Particle Swarm Optimization [1] | Tokenizer, stemming, stopwords, tranform case | Particle Swarm Optimization | Naive Bayes | 83,60% |
| Opinion Mining of Movie Review using the Hybrid Method of Support Vector Machine and Pasticle Swarm Optimization[21] | Filter Data, Data Cleansing, Extract to text file | Case Normalization, Tokenization, Stemming (Snowball), Generate n-Grams | SVM, SVM-PSO | 77% |

## 2. Methods

### *2.1. Research Framework*

The method in this study uses Naïve Bayes and SVM based on PSO to compare the accuracy and the best AUC value. The dataset used in this study was obtained from the comments of OVO customers via Google Play. This study uses 500 positive reviews and 500 negative reviews as data training. The data is still in the form of separate sets of text in the form of documents. At the beginning of the study, preprocessing datasets were conducted with tokenize, stopwords removal, stemming, and generate n-gram. Then proceed with the classification using the Naive Bayes method and Support Vector Machine based on PSO. then in this study will test 10-fold cross-validation and measure the accuracy of the algorithm using the confusion matrix and the ROC curve.

### *2.2. Data Collection*

This research has conducted a collection of datasets through reviews on the Google Play store about OVO e-wallet. The positive review data entered into a folder that is named "post". While the negative review data entered into a folder that is named "neg". Each document has a .txt extension that can be opened using the Notepad application.

### *2.3. Data Processing*

This study uses 500 positive reviews and 500 negative reviews as data training. In this study, dataset will use preprocess which functions to eliminate syntactic features that are not used.

The preprocessing process carried out in this research, there are 4 processes namely:

1. Tokenization, eliminating punctuation or symbols that are not letters
2. stopwords removal, irrelevant words like "the", "of", "with", and so on will be deleted.
3. Stemming, words that have the same basic words will be grouped.
4. Generate n-grams, n is the order in which feedback is given.
   weighting is done using TF-IDF on each word by counting the appearance of words in the document. then the words that appear in the document are used as calculations from the document data being tested.

*2.4. Proposed Method*

This study proposes a feature selection method using PSO to improve optimization on Naïve Bayes and Support Vector Machines.Measurement of accuracy using confusion matrix with validation using 10 fold cross validation. The ROC curve is used to measure AUC. Accuracy measurements are used to compare the accuracy of Naïve Bayes and Optimized Support Vector Machines plus PSO.

## 3. Result and Discussion

*3.1. Experiment on Model Indicator*

In this study adjustments were made to NB based on the PSO by making adjustments to the value of Population Size from the default value then increasing it by a multiple of 5 to get the highest accuracy. If the Population size indicator is changed in value, it can cause data processing to become longer. Then adjustments are made to Support Vector Machine for the values of the parameters C and Epsilon as controlling parameters with a multiple of 0.1.

In the adjustment on NB + PSO, the best accuracy is obtained at the value of population size = 10, initialization = 1.0, Number of Validation = 5, Generate n-gram = 4 with an accuracy of 93.10%. Then in the adjustment of the Testing Support Vector Machine + PSO obtained at the value of population size = 15, parameter C = 0.1, and epsilon = 0.5 with an accuracy reaching 91.30%. Adjustments to these parameter values are very influential on the length of time data processing through the Rapidminer applications.

From the existing sentiment analysis data, weighting is carried out in order to know which sentiments are positive and which are negative by looking at the highest weights. In this study, the results of testing the model are discussed using a confusion matrix. The value generated for the Naive Bayes + PSO test was 93.10% and the UC value was 0.750. While the value generated for testing Support Vector Machine + PSO is 91.30% and UC value is 0.970. More details can be seen in the table below.

The following is a calculation based on the confusion matrix table. The number of true positive (TP) was 442 reviews, false negative (FN) was 58 reviews. Then 489 reviews for true negative (TN) and 11 reviews for false positive (FP).

## 4. Conclusion

The result of this study indicates that the use of PSO-based Naive Bayes algorithm produces an accuracy value of 93.10% with an AUC value of 0.750. While the research result from the PSO-based SVM algorithm produces an accuracy value of 91.30% with an AUC value of 0.970. Based on this result the accuracy value generated by the Naive Bayes algorithm is classified as Fair Classification and SVM is classified as Excellent Classification. The AUC value generated by the Naive Bayes algorithm is also smaller than SVM. Therefore, in this study it was found that SVM is the best algorithm in classifying text. The selection feature of the PSO can improve

**Table 2.** Indicators and Test Results on Naive Bayes + PSO

| Population Size | Validation Number | Generate n-Gram | NB+PSO (OVO) Accuracy | AUC |
|---|---|---|---|---|
| | | 2 | 84.70% | 0.793 |
| 5 | | 3 | 91.80% | 0.893 |
| | | 4 | 91.80% | 0.664 |
| | | 2 | 84.20% | 0.728 |
| 10 | | 3 | 91.10% | 0.810 |
| | | **4** | **93.10%** | **0.750** |
| | 5 | 2 | 84.40% | 0.739 |
| 15 | | 3 | 91.50% | 0.664 |
| | | 4 | 92.70% | 0.913 |
| | | 2 | 84.40% | 0.737 |
| 20 | | 3 | 91.30% | 0.889 |
| | | 4 | 92.40% | 0.666 |
| | | 2 | 83.50% | 0.718 |
| 5 | | 3 | 91.30% | 0.617 |
| | | 4 | 92.50% | 0.500 |
| | | 2 | 84.40% | 0.733 |
| 10 | | 3 | 91.90% | 0.706 |
| | | 4 | 94.10% | 0.542 |
| | 10 | 2 | 84.50% | 0.736 |
| 15 | | 3 | 92.40% | 0.574 |
| | | 4 | 93.50% | 0.667 |
| | | 2 | 84.10% | 0.729 |
| 20 | | 3 | 94.30% | 0.582 |
| | | 4 | 93.00% | 0.584 |

**Table 3.** Indicators and Test Results on SVM+ PSO

| Population Size | Parameter C | Parameter Epsilon | SVM+PSO Accuracy | SVM+PSO AUC |
|---|---|---|---|---|
| 5 | 0.0 | 0.0 | 83.90% | 0.963 |
| 5 | 0.1 | 0.1 | 79.20% | 0.965 |
| 5 | 0.2 | 0.2 | 75.60% | 0.894 |
| 5 | 0.3 | 0.3 | 77.00% | 0.914 |
| 5 | 0.4 | 0.4 | 87.70% | 0.963 |
| 5 | 0.5 | 0.5 | 75.30% | 0.892 |
| 10 | 0.0 | 0.0 | 83.60% | 0.974 |
| 10 | 0.1 | 0.1 | 90.90% | 0.954 |
| 10 | 0.0 | 0.5 | 82.90% | 0.972 |
| 15 | 0.0 | 0.0 | 84.30% | 0.976 |
| 15 | 0.0 | 0.1 | 84.90% | 0.976 |
| **15** | **0.1** | **0.5** | **91.30%** | **0.970** |
| 20 | 0.0 | 0.0 | 84.10% | 0.967 |
| 20 | 0.0 | 0.1 | 81.90% | 0.967 |

accuracy very well on both the Naive Bayes and SVM algorithm classification model, because the accuracy result of both models have high values.

**Table 4.** Confusion Matrix Nauive Bayes+PSO

| **Naive Bayes Accuracy: 93.10%** | | | |
|---|---|---|---|
| | **True Negative** | **True Positive** | **Class Precision** |
| Pred Negative | 489 | 58 | 89,40% |
| Pred Positive | 11 | 442 | 97,57% |
| Class Recall | 97,80% | 88,40% | |

**Table 5.** Confusion Matrix SVM+PSO

| **SVM Accuracy: 91.30%** | | | |
|---|---|---|---|
| | **True Negative** | **True Positive** | **Class Precision** |
| Pred Negative | 445 | 32 | 93,29% |
| Pred Positive | 55 | 468 | 89,48% |
| Class Recall | 89.00% | 93,60% | |

## References

[1] Aaputra S A Didi Rosiyadi Windu Gata and Syepry Maulana Husain, 2019 Sentiment Analysis Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes
[2] Putri D A, 2015 Penerapan Algoritma Support Vector Machine Berbasis Algoritma Genetika Untuk Analisis Sentimen I, 01 P. 1–7.
[3] Mahendrajaya R Buntoro G A and Setyawan M B, 2019 Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine Komputek 3, 2 p. 52.
[4] Kristiyanti D A Umam A H Wahyudi M Amin R and Marlinda L, 2019 Comparison of SVM Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter in 2018 6th International Conference on Cyber and IT Service Management, CITSM 2018.
[5] Kristiyanti D A and Wahyudi M, 2017 Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review in 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017.
[6] Wisnu H Afif M and Ruldevyani Y, 2020 Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes J. Phys. Conf. Ser. 1444, 1.
[7] Dhande L L and Patnaik P G K, 2014 Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier 3, 4 p. 313–320.
[8] Wibowo W S Az-zahra H M and Bachtiar F A, 2018 Evaluasi dan Rekomendasi Tampilan Website E-Complaint Universitas Brawijaya Pada Perangkat Bergerak Menggunakan Metode Heuristic Evaluation J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya 2, 12 p. 7192–7201.
[9] Chou J Cheng M Wu Y and Pham A, 2014 Expert Systems with Applications Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification Expert Syst. Appl. 41, 8 p. 3955–3964
[10] K. Schoefegger, T. Tammet and M G, 2013 A survey on socio-semantic information retrieval Comput. Sci. 8 p. 25–46.
[11] Kristiyanti D A Normah and Umam A H, 2019 Prediction of Indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis Proc. 2019 5th Int. Conf. New Media Stud. CONMEDIA 2019 p. 36–42.
[12] Moraes R Valiati J F and Neto W P G, 2013 Expert Systems with Applications Document-level sentiment classification: An empirical comparison between SVM and ANN Expert Syst. Appl. 40, 2 p. 621–633.
[13] Nur aeni widiastuti S santosa C supriyanto, 2010 Algoritma Klasifikasi Data Mining Naïve Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung Nat. Methods 7, 1 p. 34.