

Marketplace Sentiment Analysis Using Naive Bayes And Support Vector Machine

Muhamad Azhar ^{1,*}, Noor Hafidz ¹, Biktra Rudianto ², Windu Gata ¹

¹ Master of Computer Science; STMIK Nusa Mandiri; Jl. Jatiwaringin No 2 Cipinang Melayu Makasar Jakarta Timur, (021)8005722;

² Informatics Engineering Department; STMIK Nusa Mandiri; Jl. Jatiwaringin No 2 Cipinang Melayu Makasar Jakarta Timur, (021)8005722;

E-mail: 14002310@nusamandiri.ac.id; 14002298@nusamandiri.ac.id; biktra.brd@nusamandiri.ac.id; windu@nusamandiri.ac.id

* Correspondence: e-mail: 14002310@nusamandiri.ac.id

Received: 8 August 2020; Review: 22 August 2020; Accepted: 5 September 2020; Published: 25 September 2020

Abstract

Technology implementation in the marketplace world has attracted the attention of researchers to analyze the reviews from customers. The klikindomaret application page on GooglePlay is one application that can be used to get information on review data collection. However, getting an information on consumer's opinion or review is not an easy task and need a specific method in categorizing or grouping these reviews into certain groups, i.e. positive or negative reviews. Sentiment analysis study of a review application in GooglePlay is still rare. Therefore, this paper analysis the customer's sentiment from klikindomaret app using Naive Bayes Classifier (NB) algorithm that is compared to Support Vector Machine (SVM) as well as optimizing the Feature Selection (FS) using the Particle Swarm Optimization method. The results for NB without using FS optimization were 69.74% for accuracy and 0.518 for Area Under Curve (AUC) and for SVM without using FS optimization were 81.21% for accuracy and 0.896 for AUC. While the results of cross validation NB with FS are 75.21% for accuracy and 0.598 for AUC and cross validation of SVM with FS is 81.84% for accuracy and 0.898 for AUC, while there is an increase when using the Feature Selection (FS) Particle Swarm Optimization and also the modeling algorithm SVM has a higher value compared to NB for the dataset used in this study.

Keywords: Naive Bayes, Particle Swarm Optimization, Support Vector Machine, Feature Selection, Consumer Review.

1. Introduction

Digitalization has consistently changed human habits and culture in almost every daily activity. This technology also affects people's shopping habits and helps to provide human needs. It has begun to lead human civilization towards digitalization and the use of personal computer and mobile based applications. The definition of the application itself is a program that is built and developed to meet the needs and desires of the user, with the aim of making it easier for users to do various activities (Gunawan, Fauzi, and Adikara 2017).

Nowadays, some small markets near the residential areas are important, especially in pandemic situation. Indomaret, a walaraba company, now has a digital-based application called Klikindomaret, an online shopping place for Indomaret which is one of the innovations of PT Indomarc PrismaTama which provides various products on one site or application that aims to fulfill all kinds of needs of each consumer. Indomaret is a retail company in Indonesia that has been around for more than 27 years. It currently has more than 12,000 outlets spread across several regions of Java, Madura, Bali, Lombok, Sumatra, Kalimantan and Sulawesi, of which 40% are franchise outlets. A large numbers of merchandise supplies for all outlets were sent from 27 Indomaret Distribution Centers throughout Indonesia (Kusnawati, Rokhmawati, and

Rachmadi 2018), and the application is already available on googleplay and has downloaded more than 15,000 downloads (Indomarco Prismatama 2020).

Technology in the marketplace, especially klikindomaret app, is the concern of many study on sentiment analysis. Klikindomaret application page on GooglePlay contains the information of review data collection.

To get an information from an opinion or a review of the application, there are several things that need to be done, including sentiment analysis. Sentiment analysis is one of the computational sciences of a variety of opinions, sentiments and emotions expressed in text or writing (Nugroho, Chrisnanto, and Wahana 2015). The purpose of the sentiment analysis is to extract "sentiment" from a text on a particular topic (Santosa and Umam 2018). We can extract the reviews in the comments column into information that we want or have defined separately, one example of a sentence from the review is whether it contains negative sentences or positive sentences, which later can be used as material for evaluating products or services the.

A study used the Particle Swarm Optimization-Based on Naive Bayes Algorithm. This study discussed the analysis of usage comments on the OVO application on Google Play and the results obtained are at a rate of 82.30% without Feature Selection and get an Accuracy rate of 83.60% with Feature Selection (Saputra et al. 2019).

Another study used the Naive Bayes algorithm for analysing the Indonesian Public Opinion Sentiment on Taman Mini Indonesia Indah (TMII) Tourism, which discusses sentiment reviews on TMII tourist attractions by producing an Accuracy level of 70% without using Feature Selection. and 94.02% using Feature Selection (Hayuningtyas and Sari 2019).

A Sentiment Analysis study also has been conducted to Corruption Eradication Commission (KPK) using SVM, NB, and Particle Swarm Optimization. From twitter data (78 positive tweets and 78 negative), it produces an accuracy of 80.75% and AUC 0.867 for the SVM (Support Vector Machine) algorithm and for the Naive Bayes algorithm it produces 76.92% and AUC 0.729. has an accuracy difference of 3.3% and after being optimized with the Weight Partical Swarm Optimization operator for SVM, it produces an accuracy of 83.79% and AUC of 0.910, while NB produces an accuracy of 80.13% and 0.771 (Hernawati and Windu 2019).

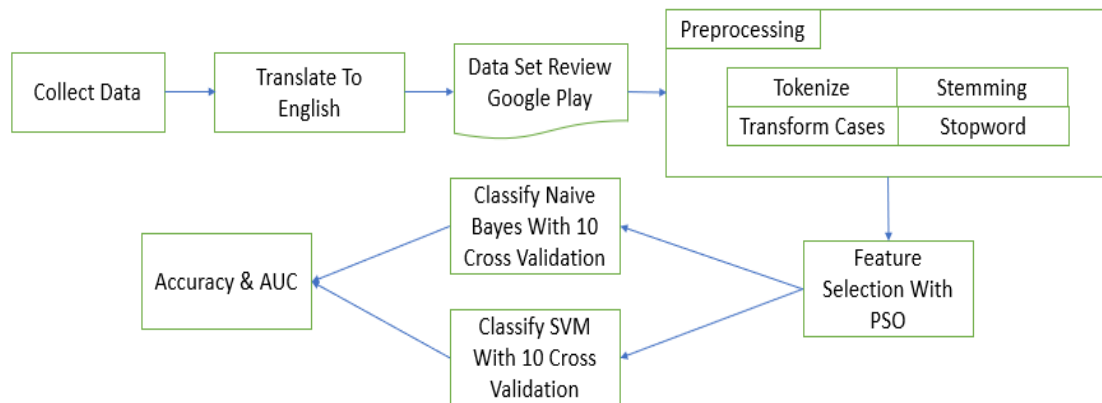
This study differs from previous research in which a klikindomaret application was used as case study. In addition, the Naive Bayes will be compared with the Support Vector Machine, Naive Bayes which is a fundamental statistical approach in recognizing patterns (Pettern Recognition). This method is based on the qualification of trade-offs between various classification decisions using probability techniques and the costs that may arise in these decisions (Santosa and Umam 2018). Feature selection based on Particle Swarm Optimazion (PSO) was used which is a fairly popular and bionic algorithm based on social behavior related to the parable of birds in groups for optimization problems (Wati 2020). In addition, this research used Rapidminer application. This tool has been widely used as a tool in measuring or calculating the accuracy of experimental data carried out in research (Aryanti et al. 2019).

This research discusses the stages to process the sentiment analysis of the reviews on the Klikindomaret application on GooglePlay. This stage will begin with the preprocessing stage until the analysis stage using the Naive Bayes Classifier method which will be compared with the Support Vector Machine using PSO Feature Selection optimization. This study also display the calculation results without using Feature selection and will be compared with the results using Feature selection.

This research aims to determine the best method for labeling a sentence, be it a positive label or a negative label obtained from reviews in the comments column of the Klikindomaret application on Googleplay using comparison of algorithmic modeling between Naive Bayes and the Support Vector Machine which has been done optimization or before optimization.

2. Data and Methods

In this study we tried to use the PSO-based Naive Bayes method to get the best accuracy in analyzing customer reviews or application users for that company on GooglePlay which will be compared with the Naive Bayes method without PSO. The framework for this research is depicted in Figure 1.



Source: Marketplace Sentiment Analysis Using Naive Bayes And Support Vector Machine (2020)

Figure 1. Research Framework

2.1. Data Collection

Data collection methods can be interpreted as a series of methods used by researchers to collect data (Perdana and Irwansyah 2019). Data collection in this study was obtained from user reviews of klikindomaret application products or services on the googleplay web using web scraping technique, which is a technique used to extract large amounts of data from websites where the extracted data is saved to a local file on a computer. or to the database in a table format (spreadsheet) (Rizaldi and Putranto 2017). The process in the rapidminer application to enter the next stage, namely preprocessing and balancing data.

2.2. Translate into English

At this stage, the researcher carried out the process of translating Indonesian-language reviews into English due to the fact that some of the reviews still speak Indonesian and some speak English, so the researchers conducted a uniform language using English.

2.3. Data Preprocessing

Data preprocessing is a step taken to prepare data for modeling (Saputra et al. 2019). At this stage, it includes several activities in shaping data and also sweeping to clean the data so that the data can be processed to the next stage. The following are the steps for preprocessing data.

2.3.1. Tokenizing

The first step of preprocessing text in this study is tokenization, which is a method of breaking text into smaller components (words, sentences, bigrams) (Anggraini and Suroyo 2019).

2.3.2. Stemming

After the tokenization stage, the next step is stemming. A stemming technique was developed for the reason of reducing the term to its basic form. The terms in the document and the query have many morphological variants, so it will be difficult for these terms to be considered equivalent. However, in certain cases the morphological variants of terms have the same semantic interpretation and can be categorized as equivalent. Stemming algorithms for one language are different from stemming algorithms for other languages. The stemming process in Indonesian text is more complicated or complex because there are variations of affixes that must be removed to get the root word of a word.

2.3.4. Transform Cases

In this process, words that contain uppercase letters are converted to lowercase letters (Wardhani et al. 2018).

2.3.5. Stopword

Stopword removal is a process to remove 'irrelevant' words in the parsing results of a text document by comparing them with existing stoplists. Words that appear too often in documents are not necessarily useful in the retrieval process. Words that are useless will later be discarded and not used as index terms (Astuti, Rachmat C., and Lukito 2017).

2.4. Featur Selection with PSO

After the preprocessing process is complete, in the next stage the researcher uses feature selection optimization, which is a data analysis method that aims to select features that have an effect (optimal features) and rule out features that have no effect (Kesuma 2011). And the selection feature used is based on PSO, PSO itself stands for Particle Swarm Optimization, which is the search for optimal solutions globally in the search space through individual interaction in a group of particles by selecting existing attributes (Achyani 2018).

2.5. Classify Naive Bayes with 10 Fold Cross Validation

At the modeling stage, Naive Bayes theorem with the Cross Validation process will be compared to the Support Vector Machine. Naive Bayes classification is a classification method based on probability and the Bayesian Theorem with the assumption that each variable X is independent and has no relation with other variables (Sunardi, Fadlil, and Suprianto 2018). Equation 1 shows the Naïve Bayes formula.

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} \dots\dots\dots(1)$$

Where R represents unknown class, S is a hypothesis on R in the specific class, P(R|S) is probability of R for S condition, P(R) is a probability of R, P(S|R) is a probability of in R condition, and P(S) represents the probability of S.

2.6. Classify Support Vector Machine With 10-Fold Cross Validation

The Support Vector Machine was introduced by vapnik for the first time, and has shown its effectiveness for problems with pattern recognition (Huang et al. 2017). Support Vector Machine has several advantages including: high generalization capability and higher classification accuracy compared to other algorithms, and in this study also uses 10 Fold Cross Validation, which is the process of randomly dividing data into 10 parts (Hermanto, Mustopa, and Kuntoro 2020).

2.7. Accuracy dan AUC Evaluation

The last stage is the calculation of the Accuracy and AUC value. Accuracy is a calculation of the proportion of the total number of correct predictions and is formulated in equation 2 (Kurniawan 2018).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \dots\dots\dots(2)$$

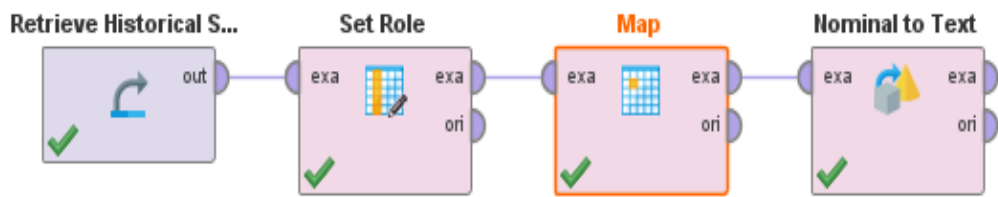
Where, TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative.

Area under the curve (AUC) is an area under the receiver operating characteristic (ROC). Receiver operating characteristic (ROC) is a curve that results from a tug of war between sensitivity and specificity at various cut points. The theoretical AUC value is between 0 and 1. The AUC value provides an overview of the overall measurement of the suitability of the model used. The larger the under-curve area, the better the variables studied in predicting events (Maskoen and Purnama 2018).

3. Result and Discussion

3.1. Data Collection

The data collection aims to establish a dataset for this research. The data is taken from google play web reviews on the klikindomaret application page. The data are 1305 review data with 793 positives and 512 negatives. The unbalanced data has been balanced with the SMOTE Upsampling technique which is a technique that can overcome the problem of imbalance in data sets (Hidayati and Arcana 2020). The data is stored and made into a file on Rapidminer for easy processing of data in Rapidminer.



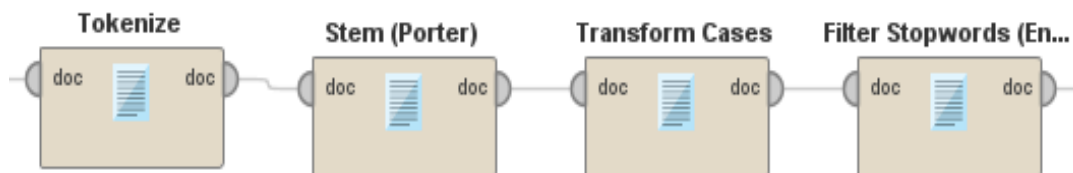
Source: Rapid Miner Modeling Results (2020)

Figure 2. Load dataset process

Figure 2 shows the Rapidminer load data stage. It is followed by determining the attribute name and target role to determine the label in the Set Role process. After that, enter the MAP process to change labels that are true to Positive and False to Negative and in the next stage, i.e. the Nominal to Text to change the nominal value into a text.

3.2. Data Preprocessing

After the data exploration stage or Load data set is complete, the next stage is Data Preprocessing (Figure 3).



Source: Rapid Miner Modeling Results (2020)

Figure 3. Data Preprocessing

In Figure 3, there are several processes, i.e. Tokenize, Stem (Porter), Transform Cases and Stopwords.

3.2.1. Tokenize

At this stage the review will be broken down into words and eliminated punctuation and numbers. Table 1 shows an example of tokenize.

Table 1. Tokenize process	
Review	Tokenize
Easy to use, fast respon, goods arrived at home fastly. Recomend....	Easy to use fast respon goods arrived at home fastly Recomend

Source: Research Results (2020)

3.2.2. Stem (Porter)

The Stemming (porter) process is a process where words that still have affixes will be returned to their basic words in English. Table 2 shows an example of the Stemming process.

Table 2. Stemming Process	
Tokenize	Stemming
Easy to use fast respon goods arrived at home fastly Recomend	Easy to use fast respon good arriv at home fastli Recommend

Source: Research Results (2020)

3.2.3. Transform Cases.

Transform Cases are changing the uppercase letters to lowercase letters. Table 3 shows a change from the review that has been stemming to be processed with transform cases.

Tabel 3. Transform Cases

Stemming	Transform Cases
Easy to use fast respon	easy to use fast respon
goods arrived at home	good arriv at home fastli
fastly Rekomended	recomend

Source: Research Results (2020)

3.2.4. Stopwords

The last stage in the preprocessing was stop words removal. Table 4 shows the changes to the example reviews for using Stopwords:

Table 4. Stopwords Process

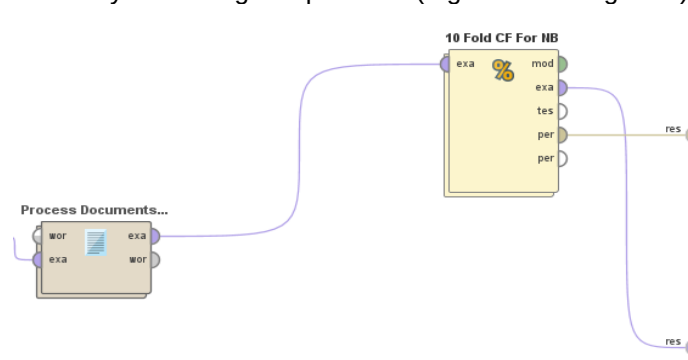
Transform Cases	Stopwords
Easy to use fast respon	easy fast respon good arriv
goods arrived at home	fastli recomend
fastly Rekomended	

Source: Research Results (2020)

3.3. PSO and Without PSO

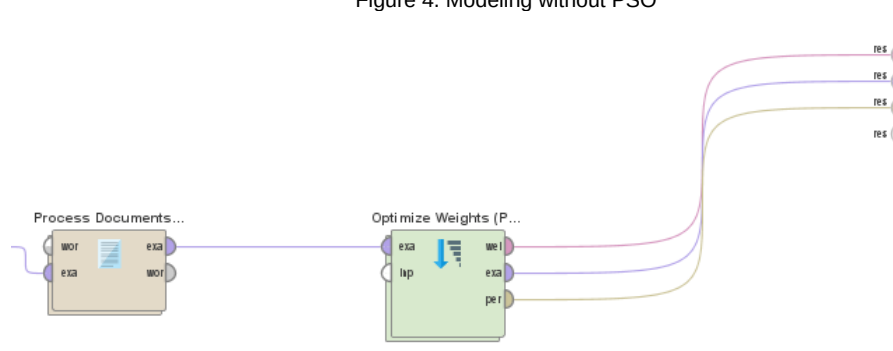
After the preprocessing stage, the optimization process was conducted, including by way of the attribute weight or weighting on all the attributes used.

In this study, a comparison of modeling using PSO optimization and those that do not use PSO optimization was analyzed through Rapidminer (Figure 4 and Figure 5).



Source: Rapid Miner Results (2020)

Figure 4. Modeling without PSO



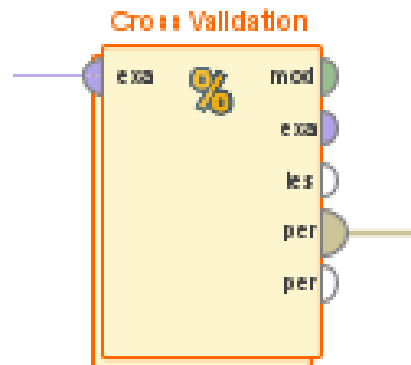
Source: Rapid Miner Modeling Results (2020)

Figure 5. Modeling Using PSO

Figure 4 and Figure 5 show the difference in the process using and without PSO, and for the two results of the modeling, the researcher will explain in the evaluation process.

3.4. Validation

The validation stage used the k-10 Fold cross validation operator as shown in Figures 6 and 7.



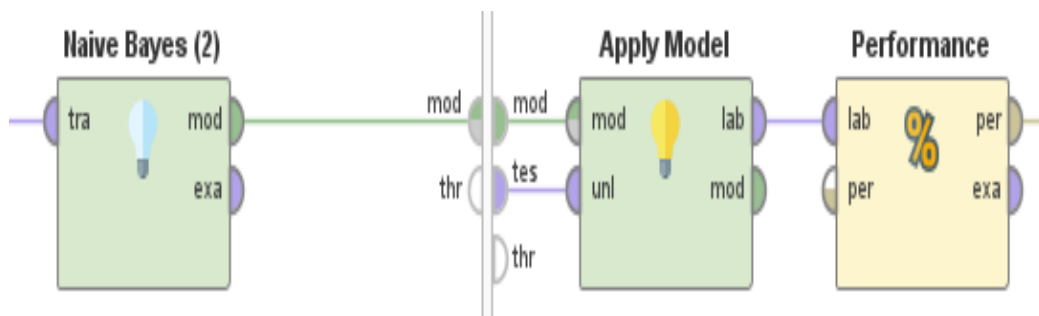
Source: Rapid Miner Modeling Results (2020)

Figure 6. Cross Validation Modeling

Source: Rapid Miner Modeling Results (2020)

Figure 7. Cross Validation Parameters

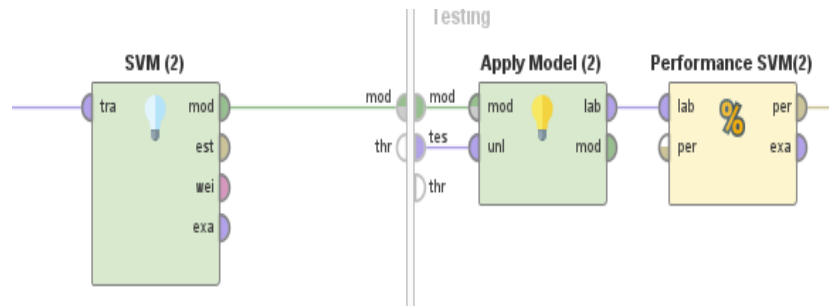
Inside these operators, there are several other operations to generate values from the evaluation (Figure 8) for the Naive Bayes Model and for the Support Vector Machine (Figure 9).



Source: Rapid Miner Modeling Results (2020)

Figure 8. Naive Bayes Modeling

Figure 8 shows the process of calculating the performance for Naive Bayes modeling.



Source: Rapid Miner Modeling Research Results (2020)

Figure 9. SVM Modeling

Figure 9 shows the process of calculating performance for Support Vector Machine modeling.

3.5. Evaluation

This study compares the Naive Bayes algorithm to the Support Vector Machine on the evaluation of the Accuracy and AUC (Area Under Curve) values both using PSO optimization and those that do not use optimization can be seen in Table 5 and Table 6. :

Table 5, Accuracy Table without PSO

Evaluasi	NB tanpa PSO	SVM tanpa PSO
Accuracy	69.74%	81.21%
AUC	0.518	0.896

Source: Research Results (2020)

Table 5 describes the Naive Bayes accuracy without PSO optimization and SVM without PSO optimization, where SVM has a better accuracy in this study.

Table 6, Accuracy Table with PSO

Evaluasi	NB + PSO	SVM + PSO
Accuracy	75.21%	81.84%
AUC	0.598	0.898

Source: Research Results (2020)

Table 6 describes the accuracy for Naive Bayes optimization of PSO and SVM without PSO, where SVM also has a better accuracy in this study.

4. Conclusion

After conducting research for the data set reviews of Klikindomaret applications and products on GooglePlay with the Naive Bayes Classifier compared to the Support Vector Machine, it can be concluded that using the Particle Swarm Optimization (PSO) feature selection can increase the value of Accuracy and AUC for the dataset. The increasing accuracy is quite significant in the accuracy value of the naive bayes modeling where the initial accuracy is 69.74% to 75.21% and the AUC value from 0.518 to 0.598. In addition, the comparison between the two Naive Bayes algorithms and the Support Vector Machine either using PSO optimization or not using PSO optimization results in a conclusion that the Support vector Machine has both a higher accuracy and AUC value when compared to Naive Bayes with 11.47% for accuracy and 0.378 for AUC without PSO optimization, and for the difference in comparison using PSO optimization of 6.63% for accuracy and 0.3 for AUC.

References

- Aaputra, S.A., D. Rosiyadi, W. Gata, and S.M. Husain. 2019. Sentiment Analysis Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization. *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi)*. <https://doi.org/10.29207/Resti.V3i3.1118>
- Achyani, Y.E. 2018. Penerapan Metode Particle Swarm Optimization Pada Optimasi Prediksi Pemasaran Langsung. *Jurnal Informatika*. <https://doi.org/10.31311/Ji.V5i1.2736>

- Anggraini, N., and H. Suroyo. 2019. Comparison Of Sentiment Analysis Against Digital Payment "T-Cash And Go-Pay" In Social Media Using Orange Data Mining. *Journal Of Information Systems And Informatics* 1, no. 2: 152–163. <https://doi.org/10.33557/Journalisi.V1i2.21>
- Aryanti, R., A. Saepudin, E. Fitriani, R. Permana, and D.F. Saefudin. 2019. Komparasi Algoritma Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Pengguna Busway. *Jurnal Teknik Komputer*. <https://doi.org/10.31294/Jtk.V5i2.5406>
- Astuti, L.W., A. Rachmat C., and Y. Lukito. 2017. Implementasi Algoritma Naive Bayes Menggunakan Isear Untuk Klasifikasi Emosi Lirik Lagu Berbahasa Inggris. *Jurnal Informatika* 14, no. 1: 16–21. <https://doi.org/10.9744/Informatika.14.1.16-21>
- Gunawan, F., M.A. Fauzi, and P.P. Adikara. 2017. Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes Dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi Bca Mobile). *Systemic: Information System And Informatics Journal* 3, no. 2: 1–6. <https://doi.org/10.29080/Systemic.V3i2.234>
- Hayuningtyas, R.Y., and R. Sari. 2019. Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata Tmii Menggunakan Naive Bayes Dan Pso. *Jurnal Techno Nusa Mandiri*. <https://doi.org/10.33480/Techno.V16i1.115>
- Hermanto, H., A. Mustopa, and A.Y. Kuntoro. 2020. Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*. <https://doi.org/10.33480/Jitk.V5i2.1181>
- Hernawati, and Windu. 2019. Sentimen Analisis Operasi Tangkap Tangan Kpk Menurut Masyarakat Menggunakan Algoritma Support Vector Machine, Naive Bayes Berbasis Particle Swarm Optimization: 14. <https://doi.org/10.30998/Faktorexacta.V12i3.4992>
- Hidayati, I.S., and I.M. Arcana. 2020. Penerapan Chaid Dengan Pendekatan Smote Pada Kematian Balita Di Kawasan Timur Indonesia Tahun 2017. *Seminar Nasional Official Statistics* 2019, no. 1: 357–367. <https://doi.org/10.34123/Semnasoffstat.V2019i1.97>
- Huang, M.-W., C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai. 2017. Svm And Svm Ensembles In Breast Cancer Prediction. Ed. Enrique Hernandez-Lemus. *Plos One* 12, no. 1 (January): E0161501. <https://doi.org/10.1371/Journal.Pone.0161501>
- Indomarc Prismaatama. 2020. Klikindomaret.
- Kesuma, Z.M. 2011. Feature Selection Data Indeks Kesehatan Masyarakat Menggunakan Algoritma Relief. *Statistika*.
- Kurniawan, Y.I. 2018. Perbandingan Algoritma Naive Bayes Dan C.45 Dalam Klasifikasi Data Mining. *Jurnal Teknologi Informasi Dan Ilmu Komputer*. <https://doi.org/10.25126/Jtiik.201854803>
- Kusnawati, W., R.I. Rokhmawati, and A. Rachmadi. 2018. Analisis Pengalaman Pengguna Pada Website E-Commerce (Studi Pada Klikindomaret.Com Dan Alfacart.Com). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer* 2, no. 9.
- Maskoen, T.T., and D. Purnama. 2018. Area Under The Curve Dan Akurasi Cystatin C Untuk Diagnosis Acute Kidney Injury Pada Pasien Politrauma. *Majalah Kedokteran Bandung*. <https://doi.org/10.15395/Mkb.V50n4.1342>
- Nugroho, D.G., Y.H. Chrishanto, and A. Wahana. 2015. Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naive Bayes. In *Prosiding SNST Ke-7 Tahun 2016 Fakultas Teknik Universitas Wahid Hasyim Semarang*, 156–161.
- Perdana, R.P., and I. Irwansyah. 2019. Implementasi Asisten Virtual Dalam Komunikasi Pelayanan Pelanggan (Studi Kasus Pada Layanan Pelanggan Telkomsel). *Jurnal Komunikasi* 11, no. 2 (December): 183. <https://doi.org/10.24912/Jk.V11i2.5491>
- Rizaldi, T., and H.A. Putranto. 2017. Perbandingan Metode Web Scraping Menggunakan Css Selector Dan Xpath Selector. *Teknika* 6, no. 1 (November): 43–46. <https://doi.org/10.34148/Teknika.V6i1.56>
- Santosa, B., and A. Umam. 2018. *Data Mining Dan Big Data Analytics Edisi 2*. Ed. Isa. Edisi 2. Yogyakarta: Penebar Media Pustaka.
- Sunardi, A. Fadlil, and Suprianto. 2018. Analisis Sentimen Menggunakan Metode Naive Bayes Classifier Pada Angket Mahasiswa. *Saintekbu* 10, no. 2 (June): 1–9. <https://doi.org/10.32764/Saintekbu.V10i2.190>
- Wardhani, N.K., Rezkiyani, S. Kurniawan, H. Setiawan, G. Gata, S. Tohari, W. Gata, and M. Wahyudi. 2018. Sentiment Analysis Article News Coordinator Minister Of Maritime Affairs Using Algorithm Naive Bayes And Support Vector Machine With Particle Swarm Optimization. *Journal of Theoretical and Applied Information Technology*.

Wati, R. 2020. Penerapan Algoritma Naive Bayes Dan Particle Swarm Optimization Untuk Klasifikasi Berita Hoax Pada Media Sosial. *Jitk (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*. <https://doi.org/10.33480/Jitk.V5i2.1034>