Recursive Journal of Informatics
Vol. 1, No. 1, March 2023
p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN XXXX-XXXX

Implementation Data Mining with Naive Bayes Classifier Method and Laplace Smoothing to Predict Students Learning Results

Dany Pradana[1], **Endang Sugiharti**[2]
[1,2]Department Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang

Abstract. The application of information technology in the field of education produces big data. It retains information that can be treated as useful. Having data mining, can be used to model highly useful student performance for educators performing corrective actions against weak students.
Purpose: The study was to identify the application and accuracy algorithm *Naive* Bayes *Classifier* to predict students' study results.
Methods: The prediction system for student learning outcomes was built using the *Naive* Bayes *Classifier* and Laplace *Smoothing* methods using a combination of two *Information Gain* and *Chi Square* feature selections. The experiment was carried out 2 times using different *dataset* comparisons.
Result: In the first experiment using a *dataset* of 80:20, the accuracy *Naive* Bayes *Classifier* method with *Laplace Smoothing* and without Laplace *Smoothing* showed the same results as 94.937%. On the second experiment to equate dataset 60:40 results of the *Naive* Bayes *Classifier* accurate method without *Laplace Smoothing* only 86.076%, then score a 91.772% accuracy using the Laplace *Smoothing*. The improvement is caused by a probability of zero that can be worked out with Laplace *Smoothing*.
Novelty: The selection feature process is very important in the classification process. Thus, in this study, *information gain* and *chi square* double selections of such features as information gain and so promote accuracy.

Keywords: *Data Mining, Naive Bayes Classifier, Laplace Smoothing.*
Received December 26, 2022 / **Revised** January 05, 2023 / **Accepted** January 13, 2023

INTRODUCTION

Information has now become one of the principal needs of people besides food and clothing. This has resulted in rapid growth in information services such as television, newspapers, radio and the Internet. Developments in information technology have had a profound impact on people's lives to carry out daily activities. Information technology enables people to obtain information from far away in short periods and relatively inexpensive costs.

The progress of information technology is inevitable, for the advancement of science will allow technology to flourish. A technology was created to benefit in a positive way to various aspects of life. The purpose of information technology is to solve a problem and to increase its effectiveness and efficiency in doing all the work. These benefits can be used as a better basis for life. However, in reality information technology can also have a negative effect if it is not used properly.

Information technology is a combination of computer and telecommunications technology with other technologies such as hardware, software, databases, network technology, and other telecommunications equipment [1]. Information technology can be divided into two main components of communication and computer technology. Computer technology is technology related to computers and various kinds of equipment related to computers. Meanwhile, information technology is a technology related to remote communication devices [2]. Information technology can be summed up simply as the science needed to manage information so that the information can be recovered easily. Currently, information technology has been used in various fields of human life. In the field of health, for example, there are many sophisticated

[1]*Corresponding author.
Email addresses: pradanadany@students.unnes.ac.id (Pradana)
DOI: 10.15294/rji.v1i1.63964

tools that can help doctors do their job easily. In the banking sector, information technology allows customers to make transactions very easily. Then, information technology also provides many benefits in the fields of business and education.

Education is one of the factors that is the key to economic progress. One of the programs designed to measure the academic performance of school children in each country was the Program for International Student Assessment (PISA). The Program for International Student Assessment (PISA) is an assessment program extended by several developed countries that are included in the Organization for Economic Cooperation and Development (OECD) which is held every 3 years which aims to determine the level of students' abilities in reading, mathematics and science in various country [3]. In 2018, the PISA score placed Indonesia in the lowest rank. The test results showed reading, math, and science scores were 371, 379 and 396 respectively. These scores had decreased compared to 2015, where reading, math, and science scores were 397, 386, 403 [4].

Indonesia's PISA score in 2018 can be said to be very worrying because it is far from the international average score. International average scores for reading, mathematics, and science each is 487, 489 and 489. Indonesia did not even make it pass 400 for all three. This decline in quality surely indicates that there is some homework to be done, if PISA is still the standard for our government for education development. The results obtained in 2018 placed Indonesia at sixth from bottom, a result that is clearly disappointing. When compared to countries in Southeast Asia, Indonesia is still under Thailand and Singapore. Based on the PISA score obtained by Indonesia, it is important to know the factors that influence Indonesia's PISA score. Indonesia's PISA score decreased in 2018 compared to 2015.

The bad score of PISA was particularly heavy for Indonesian educators. The application of information technology in education offered fertile land for data-mining applications, as there are many different data sources and interest groups. For example, there are interesting questions in this domain that can be answered using data mining techniques, which are possible to predict student performance and what factors influence student achievement. Student performance modeling is an important tool for educators and students as it can help better understand what is lacking and ultimately correct it. For example, the school can take corrective actions for weak students, such as repair classes.

The use of information technology in education can produce enormous data. The resulting data will retain valuable information, such as trends and patterns that can be used to help make decisions and optimize every opportunity, thus giving childbrith data mining. Data mining is a process of discovering knowledge (knowledge discovery) mined from a collection of data having a high volume [5]. This data collection is then analyzes is then analyzed in various angles and summarized into useful information. Data mining is a process that analyzes data in various angles and summarizes the results to useful information [6]. Information technology is used to facilitate the extracting of information that is difficult to process by hand, when the data contained is vast.

The existence of data mining can be used to model student performance which is highly useful for educators performing corrective action on weak students. The model developed would use the Naive Bayes Classifier method. The advantage of the Naive Bayes Classifier method is that it can work quickly when applied to large and varied data [7]. Additionally, several studies have shown that the results obtained with the Naive Bayes Classifier can compete with other techniques [8]. Naive bayes classifier is a simple probability classification that calculates the set of probabilities by quantifying the frequency and value combination of the given datasset. The appiplication of naive baisa classifier on a data sometimes also causes a misclassification if the training data is so small or insufficient so that the testing data is not found in the training data, and causes the probability results to be worth 0 (zero) and lead to an error in the classification process.

The naive bayes classifier's lack can be minimized by using the smoothing method, some recent studies proving that the smoothie method can enhance the performance of naive bayes classifier. One of the smooth methods that can be used is the Laplacian smoothing, or what it is commonly called add-one smoothing. The working principle of the Laplacian smoothing can be said to be very simple by adding a small positive value to each of the cognitive value so that it can avoid the zero probability value on the model [9].

There were earlier studies linked to the research to be done. As for example the naive bayes classifier's prediction of stekom semarang student graduation rate produced an accuracy of 95.14% [10]. Then, another study suggested the use of naive bayes classifier's method to identify the success rate of schools preparing students for national exams resulted in 95% accuracy [11]. So the study regarding the implementation of the naive bayes classifier method for predicting hepatitis shows the value of accuracy reaching 76.77% [12]. On the study of the smooth performance analysis of naive bayes classifier for the exam's classification shows that Laplace smoothing actually increased naive bayes classifier's performance because it can minimize the classifier's classification value by zero [13].

Based on the above description, this research focuses on the system's accuracy to predict students' study results by using a naive bayes classifier algorithm with Laplace smoothing. The purpose of this study is to identify the application and accuracy algorithm naive bayes classifier and Laplace smoothing to predict student study results.

METHODS

The study was conducted using the naive bayes classifier and Laplace smoothing with a combination of information gain and chi square selection features. The level of the classification process conducted is addressed in Figure 1.

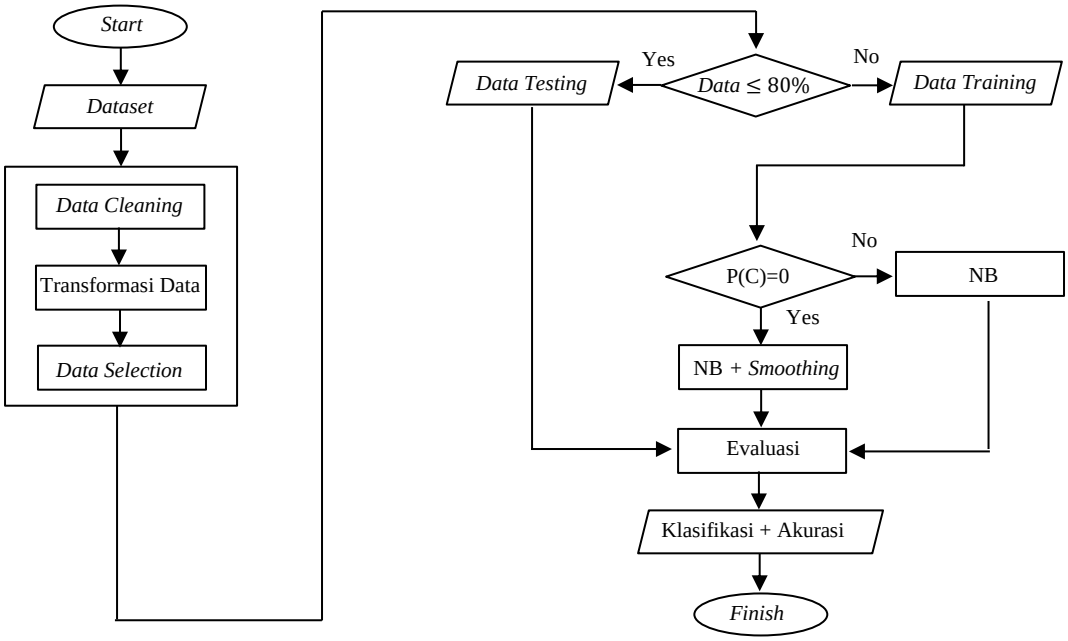

Figure 1. Program Flow

Preprocessing Data

Preprocessing data is done by simplifying data so it becomes data with corresponding characteristics that it can be accepted into training data and testing data. There are three processes involved in data cleaning, data transformation, and data selection. Cleaning data is used to clean data of incorrect values. The transformation of data does by changing value on each attribute has an acceptable range by the program to be created. Data selection is done by selecting significant attributes using the combination of information gain and chi square selection features. Where, it will be taken 10 of the best features of information gain and for chi square selection features will pick up the best feature with alpha 0.05.

Information gain is one of the most used selection techniques of the features [14]. Selection of features will need to be made to select the best features of available data. Information gain USES simple attribute rantings and reduces noise caused by irrelevant features [15]. To count information gain is based on equation 1.

$$Gain(A) = Entropy(S) - \sum_{i=1}^{n} \frac{|s_i|}{|s|} * Entropy(S_i) \quad (1)$$

As for calculating entropy is indicated in equation 2.

$$\text{Entropy(S)} = \sum_{\text{i=1}}^{\text{n}} - \text{pi} * \log_2 \text{pi} \quad (2)$$

Chi square is a method of selecting features that uses statistical theories to test indepities that serve to test the dependency of a class on a feature [16]. Chi square can evaluate irrelevant attributes in the classification process [17]. This chi square model uses the basic statistical theory to test independent features in the classroom. To calculate chi square indicated on equation 3.

$$x^2 = \sum_{i=1}^{n} \frac{\left(o_{ij-e_{ij}}\right)^2}{e_{ij}} \quad (3)$$

Data Analysis

The data analysis phase is done by gathering the datasets to be used on the research. The datasets to use do preprocessing in advance. The preprocessing stage is done to clear the datassets to be used of incomplete or unimportant attributes. After that, it will be divided into two parts: training data and testing data. Training data is used to train systems to predict students' learning results. Whereas, data testing is used to assess the performance of the system being created.

The next was to file a classification by the naive bayes classifier method and be optimized Laplace smoothing. Naive bayes classifier is a simple probability classification that calculates the set of probabilities by quantifying the frequency and value combination of the given datasset [18]. Naive bayes classifier has a competitive performance compared to other classification methods and can be applied to many complex real-world situations [19]. The naive bayes classifier algorithm is a test-machine learning algorithm that USES the bayes theorem and depends on conditional probability [20]. The naive bayes classifier theorem can be identified in equation 4.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (4)$$

Smoothing method is a method for avoiding the results of the classification is zero because the testing data was not found on the training data [21]. Laplace smoothing is a method that adjucts a maximum probability estimate to correct probability instead of zero for data that is invisible and increases model accuracy because of frequency of data [22], [23]. The naive bayes classifier theorem if combined with Laplace smoothing can be seen in equation 5.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i) + 1}{P(X) + K} \quad (7)$$

RESULT AND DISCUSSION

The study aims to identify application and to know the results of the naive bayes classifier and Laplace smoothing's accuracy in predicting students' study. The testing process in the study uses the naive bayes classifier and Laplace smoothing thing by applying a combination of two features found information gain and chi square. The data used are students performance of UCI's learning repracticed. The number of data from datasset students mance is 395 data, where the data has 30 attributes.

The first processing data is done. In the preprocessing stages there are three things to do on the cleaning data, data transformation, and data selection. By taking 10 of the best features of selection information gain and picking up selected features from chi square with alpha 0.005 obtained as many as 11 of the best features used in this study. 11 attributes to be used can be seen at Table 1.

Table 1. Attribute results combination of information gain and chi square

No	Atribut	Description
1.	G2	Second period grade
2.	G1	First period grade
3.	Failures	Number of past class failures
4.	Mjob	Mother's job
5.	Goout	Going out with friends
6.	Studytime	Weekly study time
7.	Walc	Weekend alcohol consumption
8.	Medu	Mother's education
9.	Freetime	Free time after school
10.	Higher	Wants to take higher education
11.	Address	Student's home address

After getting 11 of the best features, the next process is to file a classification by the Naive Bayes Classifier methods and optimize it with Laplace smoothing to avoid a probability of zero value which could lead to a classification error. The test was performed twice by using different datasset comparisons. In the first trial, using a ratio of 80 training data and 20 testing data (80:20), showed if there is no accurate result between using a Laplace smooth thing and not using the Laplace smoothing. The results of the confusion matrix using a data comparison of 80:20 without and with Laplace Smoothing can be seen in Table 2 and Table 3.

Table 2. Confusion Matrix without Laplace Smoothing (80:20)

	Class	
Predicted	pass	fail
pass	34	4
fail	0	41

Tables 2 and 3 show 38 classified data passes with four misclassified data by the system. Then there are 41 classified data with no misclassified data. That means the system successfully predicts 75 data correctly and four misclassified data. Table 3 shows the same results as Table 2. Accuracy results in the first experiment using a dataset 80:20 is 94.937% either with Laplace smoothing or not.

Table 3. Confusion Matrix with Laplace Smoothing (80:20)

	Class	
Predicted	pass	fail
pass	34	4
fail	0	41

A second test is performed by using a 60% of training data ratio and a 40% of testing data (60:40). Confussion matrix results using data comparisons at 60:40 without and with Laplace smoothing shown at Table 4 and Table 5.

Table 4. Confusion Matrix without Laplace Smoothing (60:40)

	Class	
Predicted	pass	fail
pass	73	16
fail	6	63

Table 4 shows 89 classified data pass with 16 misclassified data by the system. Then there are 69 classified data with 6 misclassified data. That means, the system successfully predicts 136 data correctly and 22 misclassified data.

Table 5. Confusion Matrix with Laplace Smoothing (60:40)

	Class	
Predicted	pass	fail
pass	73	7
fail	6	72

Whereas Table 5 shows 80 classified data pass with 7 misclassified data by the system. Then there are 78 classified files with 6 misclassified data. That means the system has successfully predicated 145 data correctly and 13 misclassified data. The accurate results on the second trial as a dataset 3.40, which is 86.076% without Laplace smoothing, then increasing to 91.772% when using the Laplace smoothing.

Discussion

The experiment was carried out 2 times using different dataset comparisons. In the first experiment, using a comparison of 80 training data and 20 testing data (80:20), showed if there is no accurate result between using a Laplace Smoothing and not using the Laplace smoothing. Accuracy results in the first experiment using the ratio of dataset 80:20 can be seen in Figure 2.

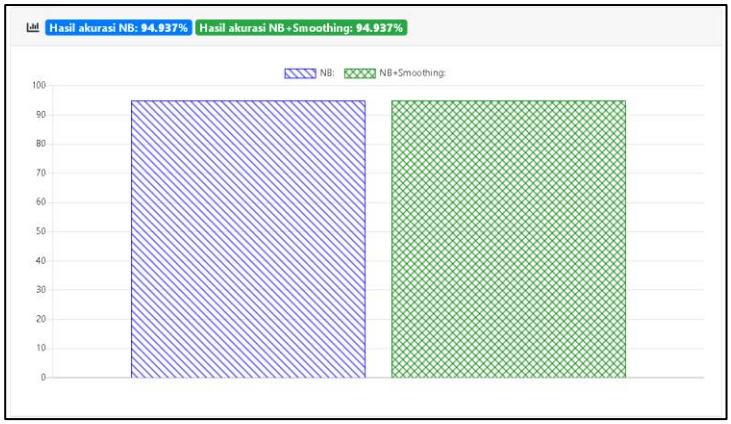

Figure 2. Accuracy Results with 80:20 Dataset Comparison

Figure 2 with a dataset comparison of 80:20 shows the classification method of the Naive Bayes Classifier by applying a combination of two feature selection methods that successfully classifies with an accuracy rate of 94.937% whether with Laplace Smoothing or not. The classification results show the same result because there is no zero-value probability that occurs, so Laplace Smoothing does not have any effect. Laplace Smoothing will only work when there is a zero probability. The second experiment was carried out using a dataset comparison of 60% training data and 40% testing data (60:40). From the second experiment, it was found that there was a difference in accuracy when using Laplace Smoothing and not using Laplace Smoothing. The accuracy results in the second experiment with a dataset comparison of 60:40 can be seen in Figure 3

Figure 3. Accuracy Results with 60:40 Dataset Comparison

Based on Figure 3, showing increased accuracy from 86.076% to 91.772%. The improvement was caused by a zero probability improvement on *Naive Bayes Classifier* using the *Laplace Smoothing*. *Laplace Smoothing* elicited a zero probability and solved the zero-value probability problem that caused a misclassification on the *Naive Bayes Classifier* method. When compared with accuracy in using the 80:20 ratio of *dataset*, there appears to be a decline in accuracy. On the 80:20 *dataset* accuracy ratio is 94.937% either with Laplace smoothing or not, while datlace smoothing equals accuracy 86.076% without Laplace Smoothing and 91.772% using Laplace Smoothing. The decline in accuracy could be caused by a decline in the number of data used as a training data thus affecting the performance of the naive bayes classifier method.

CONCLUSION

The results of the accuracy of the Naive Bayes Classifier method with Laplace Smoothing and without Laplace Smoothing using a dataset comparison of 80:20 show the same result, namely 94.937%. Using the 80:20 comparison data, Laplace Smoothing does not seem to have any effect because there is no zero probability. Whereas when using a dataset comparison of 60:40 the results of the accuracy of the Naive Bayes Classifier method with Laplace Smoothing and without Laplace Smoothing show different results. When using Laplace Smoothing it is not only 86.076%, then you get an accuracy value of 91.772% when using Laplace Smoothing. This increase was due to a zero-value suspicion which was successfully overcome by Laplace Smoothing.

REFERENCES

[1] S. Maharsi, "Pengaruh Perkembangan Teknologi Informasi Terhadap Bidang Akuntansi Manajemen," *Jurnal Akuntansi dan keuangan*, vol. 2, no. 2, pp. 127–137, 2000.

[2] M. Husaini, "Pemanfaatan Teknologi Informasi dalam Audit Investigatif," *Jurnal Mikrotik*, vol. 2, no. 1, pp. 141–147, 2014.

[3] R. A. Maharani and I. N. Aini, "Deskripsi Tahapan Problem Solving Siswa Pada Soal Bertipe Pisa Space and Shape Content," *Jurnal Cendekia : Jurnal Pendidikan Matematika*, vol. 5, no. 2, pp. 1193–1200, May 2021, doi: 10.31004/cendekia.v5i2.608.

[4] L. Lewi and M. Shaleh, "Assesment): Upaya Perbaikan Bertumpu Pada Pendidikan Anak Usia Dini," *Jurnal Golden Age, Universitas Hamzanwadi*, vol. 4, no. 1, 2020.

[5] M. Sattarian, J. Rezazadeh, R. Farahbakhsh, and A. Bagheri, "Indoor navigation systems based on data mining techniques in internet of things: a survey," *Wireless Networks*, vol. 25, no. 3, pp. 1385–1402, Apr. 2019, doi: 10.1007/s11276-018-1766-4.

[6] I. Mahendra, I. Suta, and M. Sudarma, "Classification of Data Mining with Adaboost Method in Determining Credit Providing for Customers," *IJEET Int. J. Eng. Emerg. Technol*, vol. 4, no. 1, pp. 31–36, 2019.

[7] A. Yasar and M. M. Saritas, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, Jan. 2019, doi: 10.18201/ijisae.2019252786.

[8] P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, "A Collaborative Filtering Approach Based on Naïve Bayes Classifier," *IEEE Access*, vol. 7, pp. 108581–108592, 2019, doi: 10.1109/ACCESS.2019.2933048.

[9] I. A. Musdar, "Aplikasi Prediksi Kerusakan Smartphone Menggunakan Metode Naive Bayes dan Laplace Smoothing," *Jtriste*, vol. 5, no. 2, pp. 8–16, 2018.

[10] Y. Angraini, S. Fauziah, and J. L. Putra, "Analisis Kinerja Algoritma C4.5 dan Naïve Bayes dalam Memprediksi Keberhasilan Sekolah Menghadapi UN," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 5, no. 2, pp. 285–290, Feb. 2020, doi: 10.33480/jitk.v5i2.1233.

[11] I. Listiowarni and E. Rahayu Setyaningsih, "Analisis Kinerja Smoothing pada Naive Bayes untuk Pengkategorian Soal Ujian," *Jurnal Teknologi dan Manajemen Informatika*, vol. 4, no. 2, Jun. 2018, doi: 10.26905/jtmi.v4i2.2080.

[12] D. Novianti, "Implementasi Algoritma Naïve Bayes Pada Data Set Hepatitis Menggunakan Rapid Miner," *Paradigma - Jurnal Komputer dan Informatika*, vol. 21, no. 1, pp. 49–54, Mar. 2019, doi: 10.31294/p.v21i1.4979.

[13] E. Siswanto, "Optimasi Metode Naïve Bayes Dalam Memprediksi Tingkat Kelulusan Mahasiswa Stekom Semarang," *JURIKOM (Jurnal Riset Komputer)*, vol. 6, no. 1, pp. 1–6, 2019.

[14] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection in sentiment analysis using information gain and domain specific ontology," *International Journal of Advanced Computer Research*, vol. 9, no. 44, pp. 283–292, Sep. 2019, doi: 10.19101/IJACR.PID90.

[15] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[16] M. C. Wijanto, "Sistem Pendeteksi Pengirim Tweet dengan Metode Klasifikasi Naive Bayes," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 1, no. 2, Aug. 2015, doi: 10.28932/jutisi.v1i2.378.

[17] S. E. Prasetyo, P. H. Prastyo, and S. Arti, "A Cardiotocographic Classification using Feature Selection: A comparative Study," *JITCE (Journal of Information Technology and Computer Engineering)*, vol. 5, no. 01, pp. 25–32, Mar. 2021, doi: 10.25077/jitce.5.01.25-32.2021.

[18] G. Gustientiedina, M. Siddik, and Y. Deselinta, "Penerapan Naïve Bayes untuk Memprediksi Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademis," *Jurnal Infomedia: Teknik Informatika, Multimedia & Jaringan*, vol. 4, no. 2, pp. 89–93, 2019.

[19] Y.-C. Zhang and L. Sakhanenko, "The naive Bayes classifier for functional data," *Stat Probab Lett*, vol. 152, pp. 137–146, Sep. 2019, doi: 10.1016/j.spl.2019.04.017.

[20] R. V. B. Vangara*, K. Thirupathur, and S. P. Vangara, "Opinion Mining Classification u sing Naive Bayes Algorithm," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 5, pp. 495–498, Mar. 2020, doi: 10.35940/ijitee.E2402.039520.

[21] I. Listiowarni, "Implementasi Naïve Bayessian dengan Laplacian Smoothing untuk Peminatan dan Lintas Minat Siswa SMAN 5 Pamekasan," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 8, no. 2, pp. 124–129, Aug. 2019, doi: 10.32736/sisfokom.v8i2.652.

[22] H. T. Sueno, "Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3937–3944, Jun. 2020, doi: 10.30534/ijatcse/2020/216932020.

[23] Y. N. Ifriza and M. Sam'an, "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *Journal of Soft Computing Exploration*, vol. 2, no. 2, pp. 93–98, 2021, doi: 10.52465/joscex.v2i2.42.