Jurnal TRANSFORMATIKA,
Vol.20, No.1, Juli 2022, pp. 44 - 56
P-ISSN: 1693-3656, E-ISSN: 2460-6731
https://journals.usm.ac.id/index.php/transformatika/

The Application of Naïve Bayes Classifier Based Feature Selection on Analysis of Online Learning Sentiment in Online Media

Ryanda Satria Putra[1], Wirta Agustin[2], M. Khairul Anam[3*], Lusiana[4], Saleh Yaakub[5]

[1] STMIK Amik Riau Jl. Purwodadi Indah Km. 10 Panam, pekanbaru (0761) 7047091, e-mail: 1610031802101@stmik-amik-riau.ac.id

[2] STMIK Amik Riau Jl. Purwodadi Indah Km. 10 Panam, Pekanbaru (0761) 7047091, e-mail: wirtaagustin@stmik-amik-riau.ac.id

[3] STMIK Amik Riau Jl. Purwodadi Indah Km. 10 Panam, Pekanbaru (0761) 7047091, e-mail: khairulanam@sar.ac.id

[4] STMIK Amik Riau Jl. Purwodadi Indah Km. 10 Panam, Pekanbaru (0761) 7047091, e-mail: lusiana@sar.ac.id

[5] Fakultas Sains dan Teknologi Universitas Muhammadiyah Jambi Jl. Kapten Pattimura, Simpang IV Sipin, Kec. Telanaipura, Kota Jambi (0741) 60825, e-mail:saleh@umjambi.ac.id

ARTICLE INFO

History of the article :
Received 26-05-2022
Received in revised form 20-07-2022
Accepted 20-07-2022
Available online 28-07-2022

Keywords:
Sentiment Analysis, Online Learning, Naive Bayes Classifier (NBC), Feature Selection, XGBoost

*** Correspondence:**

E-mail:
khairulanam@sar.ac.id

ABSTRACT

There are problems that still exist in online learning including limited-reach networks, inadequate facilities and infrastructure, and others. This study discussed the analysis of sentiment which used the Naïve Bayes Classifier (NBC) method with XGBoost feature selection as a performance improvement that took data from news portals. The results of this study showed that graph data on the application of online learning forms in Indonesia had a "Negative" opinion. Performance testing of the NBC method based on XGBoost feature selection was conducted four times. The first experiment resulted in an accuracy value of 60.18% with 50/50 split data. The next experiment had an accuracy value of 56.92% with 70/30 split data. After that, the third experiment resulted in an accuracy value of

65.90% with 80/20 split data. The result of the last experiment was an accuracy value of 63.63% with 90/10 split data. After using XGBoost feature selection, it produced an accuracy of 60.18%, 67.69%, 70.45%, and 77.27%. The study also produced the highest average score at 10-Fold Cross-Validation in the second trial with a score of 65.62%.

1. INTRODUCTION

Online learning is a form of educational activity without face-to-face interaction with a relationship between teachers (teachers/lecturers) and students to achieve a goal online. This learning uses the internet network so that it can interact online from one to another. Online learning is also learning that requires an internet network with accessibility, connectivity, flexibility, and the ability to produce different types of learning interactions [1]. At the level of implementation, online learning requires the support of mobile devices such as *smartphones* or android phones, laptops, computers, tablets, and iPhones that can be used to access information anytime and anywhere [2].

At the level of its implementation, it cannot be separated from the various problems that exist. Many people have an opinion about the problem in this form of learning. Some common problems that often occur during online learning include unstable network connections and lack of interacting directly or face-to-face with teachers (teachers/lecturers) and friends [3].

The current *coronavirus disease* 2019 (COVID-19) pandemic has become a global outbreak and infected citizens all over the world very quickly, including Indonesia [4]. The spread of COVID-19 was initially very impactful on the economic world that began to fall, but now the impact is also felt by the world of education [5]. With this pandemic and some common problems in online learning above, new problems arise in the learning, one of which is like a limited reach network. According to [6], problems exist also in inadequate facilities and infrastructure, limited internet quotas for schoolchildren and especially parents with limited resources and education.

With this pandemic, the government appealed to all teaching and learning activities to be carried out online / remotely. This statement is reinforced based on the Circular Letter (CL) issued on March 24th, 2020, the Minister of Education and Culture of the Republic of Indonesia who issued the CL Number 4 of 2020 concerning the Implementation of Education in the Emergency Period of the Spread of COVID-19 which states that learning activities are carried out remotely/online to provide a meaningful learning experience for students at home and distance learning is focused on improving students' knowledge in the COVID-19 outbreak. With this pandemic, teaching and learning activities are carried out online or at home. A form of a lecture that can be used as a solution during the COVID-19 pandemic is online learning [7].

Based on the circular issued by the government above, the implementation of online learning is in order to reduce direct contact with many people. Online learning encourages the emergence of *social distancing* behavior and minimizes the emergence of student crowds so that it is considered to reduce the potential spread of COVID-19 in the school environment. According to [8], *social distancing* behavior is a good solution to prevent the spread of COVID-19.

With many problems existing in online learning, the statement has a wide influence on the public with various comments through the mass media where people can comment and respond to what they think [9]. Mass media becomes one of the media to break their minds and respond to the information they get. Mass media is a place to convey information directly related to society at large and an invention of advanced technology that allows people to be able to communicate such as the radio, television, and newspapers [10]. One of the mass media that is often used to

DOI : https://doi.org/ 10.26623/transformatika.v20i1.5144

see and comment on the application of online learning systems is Kompas.com, Suara.com, and Tribunnews.com. The existence of internet users who participate online in the mass media making the implementation of the online learning system in Indonesia get a variety of comment speculations, including suggestions, praises, criticisms, satire, and hate speech. With the *feedback* from the community, it is expected to make it easier for the government to determine what steps should be taken in the development of online learning systems in Indonesia. With so many people giving these opinions, it can be used for analysis.

Sentiment analysis is used to analyze people's representations, opinions, evaluations, judgments, attitudes, and emotions towards something such as a product, service, organization, individual, event or topic [11]. With the sentiment analysis, word processing can be performed to track the *mood* of the community from comments obtained in the mass media. To find out public opinion on the application of online learning systems in the form of comments ranging from negative, neutral, and positive, it is necessary to analyze using sentiment analysis. In this study, sentiment analysis was conducted to see public opinion on online learning systems in the mass media. In conducting sentiment analysis, the comments analyzed were then classified into negative, neutral, and positive classifications, so as to make it easier for the government to monitor the development of online learning and assist in decision making for what steps are appropriate in developing the online learning system.

Many methods can be used to implement sentiment analysis, such as *Support Vector Machine (SVM)*[12], C4.5, and *Naive Bayes Classifier* [13]. Based on the 3 methods above, the research from [14] obtained accuracy of 81%, the research from [10] resulted in an accuracy of 84%, and the result of research from [15] was accuracy of 85.98%. The study used the Naïve Bayes Classifier method because it is simpler, has good accuracy when applied in large databases as well as diverse data. According to [16], the *Naïve Bayes Classifier* method is a simple probability classification method that applies *Bayes' Theorem* assuming high independence and has the advantage that the number of training data needed is simpler, with similar implementation to other approach methods and with high accuracy.

The research on sentiment analysis has been conducted by [16], on the classification of public opinion to measure opinion tendencies towards the 2019 Presidential Election in Indonesia through online media (Kompas.com and Detik.com). Research on *Naïve Bayes Classifier* has been conducted by [17], about the election of governors expressed through online media and research from [18], on the classification of public opinion in universitas Sjakhyakirti based on opinions from Twitter and even research conducted by [19], also discussing the results of poverty predictions with *Naïve Bayes* based on *XGBoost*. Research on online learning has been conducted by [20] on the sustainability of online learning after the COVID-19 pandemic and research from [21] to get information about the effectiveness of the online learning process during the COVID-19 pandemic.

According to [22], the *Naive Bayes* method is a method with a simple and effective classification, but Naive Bayes has a simple classification and is very sensitive in the selection of features [23]. In this study feature selection was used to improve accuracy in the Naive Bayes method. According to [24], feature selection is an optimization process to reduce a large set so that the feature subnet becomes relatively small to improve the accuracy of classification quickly and effectively. According to [25], the feature selection process is useful to reduce the size of the data and allow learning algorithms to work faster and more effectively. *XGBoost* can be used well in handling small and large cases using fewer resources. In addition, *XGBoost* makes system learning faster and can also adjust flexibility to a case [26]. *XGBoost* is one of the *machine learning* techniques to overcome regression and classification problems [27]. Many studies utilized *XGBoost*, such as research from [28]. The results of the study resulted in an acceptance of 89.52% and research from [29] resulted in an accuracy value of 73.17%.

The problem of learning has become the basis in this research by knowing how public opinion on the implementation of online learning systems in Indonesia. This research was conducted by taking data on people's comments in the mass media and the data taken only related to online learning systems in Indonesia. The data obtained will be processed and classified by the *Naïve Bayes* method so that it becomes a negative, neutral, and positive opinion. Then feature selection was carried out on Naive Bayes as a technique to improve the accuracy of sentiment analysis so that it can ease the government to make decisions for what steps are appropriate in evaluating the online learning system. The results of this study also resulted in an average *score* of *10-Fold Cross Validation* from four split data on the comparison of training data and test data. The four split data were split data 50/50, 70/30, 80/20, and 90/10 splits.

SIGNIFICANCE OF STUDY

Literature

1. Sentiment Analysis

Sentiment analysis or opinion mining is a research branch of text mining. Opinion mining is a sentiment to get the information contained in an opinion sentence by understanding, extracting, and processing textual data automatically. Opinion mining is done to see a person's view or opinion about a problem or object, whether they tend to have a positive or negative opinion [30].

Sentiment Analysis is part of the Components of Natural Language Processing (NLP) and Machine Learning. The way sentiment analysis works is to analyze people's opinions, sentiments, evaluations, assessments, attitudes, and emotions towards something such as products, services, organizations, individuals, problems, events, or a topic [11].

2. Naïve Bayes Clasifier (NBC)

Naïve Bayes Classifier is a simple probability classifier who applies Bayes' Theorem to the assumption of high independence (independent) [16]. An algorithm uses Bayes' theorem and assumes all independent or non-interdependent attributes to be assigned to the value of class variables and the advantage of using this method is that it requires only a simple amount of training data in determining the estimated parameter required in the classification process [18].

Naïve Bayes Classifier in classifying, there are two stages carried out, namely training data (training) and test data (testing). This study used training data from public opinion and the categories had been known. Test data was used for classification predictions or data that was not yet known its classification.

Common forms of this method are:

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

$$P(H|X) = (P(H|X)\,P(H))/(P(X)) \qquad (1)$$

Information:

X = Data with unknown classes
H = Hypothesis of data X is a special class
P (H| X) = Probability of hypothesis H is based on condition X
P (H) = Probability H
P(X|H) = Probability of hypothesis X is based on condition H
P (X) = Probability X

The stages of the Naïve Bayes Classifier method process are as follows:

a) Count the number of classes/labels.
b) Count the same number of cases with the same class.
c) Multiplies all class variables.
d) Compare the results of all variables.

3. XGBoost (Extreme Gradient Boosting)

DOI : https://doi.org/ 10.26623/transformatika.v20i1.5144

XGBoost was first introduced by [31], in his research linking boosting and optimization in building gradient boosting machines (GBM). Building a new model to predict errors from the previous model is used in the boosting method. The addition of a new model is done until no more error repairs can be made. By using fradient descent to minimize errors when creating a new model, the algorithm is called gradient boosting.

According to [32][33], XGBoost is an algorithm that is improved based on gradient boosting decision trees and can form boosted trees efficiently and work parallel. XGBoost is also one of the machine learning techniques to overcome regression and classification problems based on gradient boosting decision tree (GBDT) [27]. XGBoost is basically an ensemble method based on a gradient boosting tree [33]. In the regression tree, the inner nodes represent the values for the attribute test and leaf nodes with the score representing the decision [28]. According to [19], XGBoost is also a powerful and fast machine method and also provides two categories in sampling on columns, namely *consample_bytree* and *colsample_bylevel* and XGBoost needs to calculate hessian, thus requiring double objective functions.

Methodology

The stages in this study were adopted from research [34], presented in Figure 1, namely:

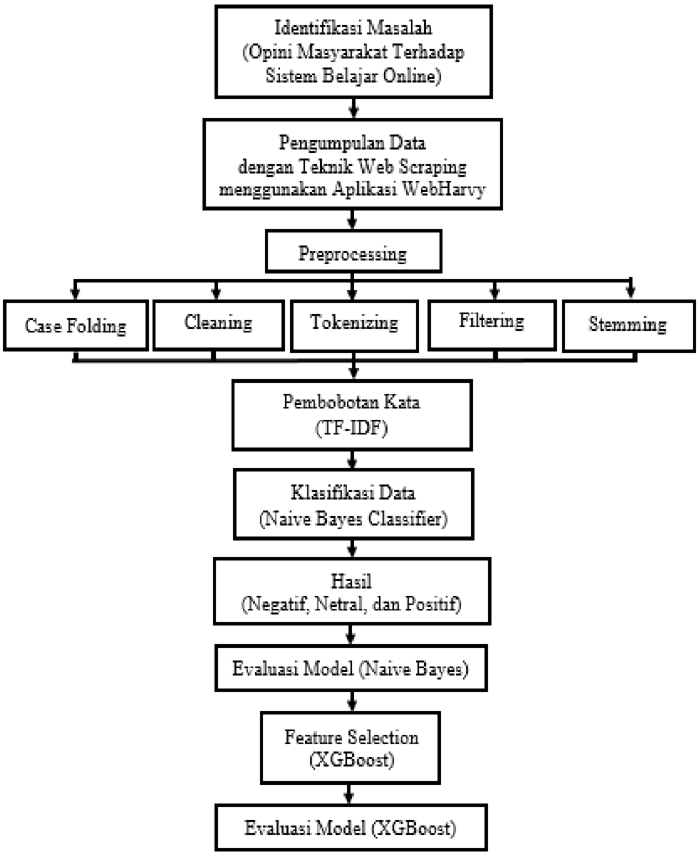

Figure 1 Research Methodology

The following is an explanation of the steps of this research methodology as follows:

a. Identification of Problem

The problem in this study was about public opinion toward the online learning system in the mass media in Indonesia.

b. Data Collection

Data collection was done by taking public opinion data from kompas.com, suara.com, and tribunnews.com sites. The selection of news that was searched is related news regarding online learning systems with the search keyword "online learning". The data collected was community comment data related to this study. Data was collected by using web scraping techniques automatically by utilizing the WebHarvy application whose data will be stored in Microsoft Excel. The categories of data used in this study were divided into three classifications, namely negative, neutral, and positive. The first step was to find the intended site in the application. The second step was to enter the keyword "learn online" in the search field facilitated by the site. The third step was to choose the related news and then take the community comment data. The total data was 216 data consisting of kompas.com (168 data), suara.com (29 data), and tribunnews.com (19 data).

c. Preprocessing Stage

The preprocessing stage is defined as the initial stage to clean up unnecessary words or words that have no meaning. The whole process was carried out using the python3 program, so it was done automatically. There were several stages used in this preprocessing stage, namely Case Folding, Cleaning, Tokenizing, Filtering, and Stemming [35].

d. Word Weighting

Weighting is the process of converting words into numbers (word vectors). Weighting was performed with Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a method that aims to give weight to the relationship of a word (term) to a document/comment. TF-IDF evaluates how important a word is in a document. Tf-IDF calculations use a library in Sklearn python that is Tfidf Vectorizer. The word weighting formula is as follows:

$$W_{ij} = tf_{ij}.idf$$

$$W_{ij} = tf_{ij}.\log\left(\frac{n}{df}\right) \qquad (2)$$

Information:

Wij is the term weight (tj) on the document (dj), tfij is the number of occurrences of terms (tj) in (dj), n is the number of all documents, and df is the number of documents containing terms.

e. Data Classification

Before classifying the data, there are several things that must be known in the data classification, namely Data Splitting. Splitting data is used to divide data into two parts, namely training data (training) and test data (testing). In this process, four experiments were attempted. The first trial of the data was divided by 50% for the training data and 50% for the test data. The second trial of the data was divided by 70% for the training data and 30% for the test data. The third trial of the data was divided by 80% for the training data and 20% for the test data. Finally, the fourth trial of the data was divided by 90% for the training data and 10% for the test data.

f. Results

After the data was analyzed sentiment, it will be obtained results in the form of data that showed negative, neutral, and positive opinions. Furthermore, the data was visualized in the form of diagrams to produce a graph of each opinion. In addition, data is also visualized in the form of *word cloud* which consists of words that appear a lot in research data.

g. Naive Bayes Model Evaluation

To find out the performance of the Naive Bayes method, testing of the model was carried out. The classification results will be displayed in the form of confusion matrix. The confusion matrix table consisted of predicted classes and actual classes.

h. Xgboost Model Evaluation

DOI : https://doi.org/ 10.26623/transformatika.v20i1.5144

To see the performance on the Naive Bayes method based on XGBoost, a model evaluation of the model was carried out. The classification results will be displayed in the form of a confusion matrix. In addition, the model evaluation process also produces values from precision, recall, and f-1 scores. To calculate precision, recall, and f1-score values in this study used python programs. After the results are known, furthermore, the performance of the classification method of each class can be seen through precision, recall, and f1-score values in each class. Precision, recall, and f1-score scores have a value of 0-1. The higher the value is, the better it is.

RESULTS AND DISCUSSIONS

After going through data preprocessing and weighting, it is then created a model that will be used to classify the test data. This process was done by the help of a library in the Python3 programming language called sci-kit-learn for the classification process. In the classification process, four experiments were conducted, namely the division of test data as many as 50% of the overall data, the sharing of test data as many as 30% of the overall data, the division of test data as many as 20% of the overall data and finally the distribution of test data as many as 10% of the overall data. In this time the second experiment was conducted with a 30% distribution of test data. The results of the division were obtained from x_train_shape, x_test_shape, y_train_shape, and y_test_shape. The classification process was carried out by calculating probabilities between sentences to each class in order to clearly produce predictions of the data entered. The test data used was 65 data which will be used to test the classification model created using the Naive Bayes method.

The results of the sentiment analysis were in the form of categories of negative, neutral and positive opinions. More details can be found in Table 1.

Table 1 Opinion Category Results

Category	Result
Negative	100
Neutral	55
Positive	61

The results found that the category "Negative" was more dominating than neutral and positive. The data is visualized in the form of a bar chart.

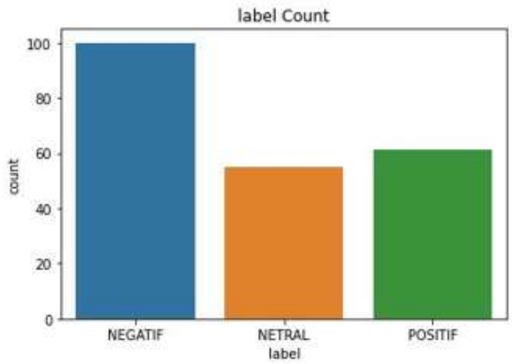

Figure 2. Opinion Category Diagram Results

Naive Bayes Model Evaluation

To find out the performance of the Naive Bayes method, testing of the model that has been made was performed. The results of the classification will be visualized in the form of a confusion matrix. After knowing the Confusion Matrix from the model created, an accuracy value of the Naive Bayes classification model which had been created by using sklearn. Metrics by importing the accuracy score provided by the scikit.learn library was calculated. The result of matrix calculations with python program code was 0.5692307692307692.

Model evaluation was carried out after the model testing process was completed. Model evaluation was performed to calculate the performance of the selected method. In the process of testing the model carried out produced confusion matrix with a size of 3x3. The results of the evaluation of models with confusion matrix are seen in Table 2.

Table 2 Confusion Matrix Results

		Predict Class		
		Negative	Neutral	Positive
Actual Class	Negative	30	0	0
	Neutral	10	2	1
	Positive	17	0	5

As seen in Table 2, the confusion matrix is a matrix with a size of 3x3. Based on the confusion matrix model in the previous chapter or in Table 3.3, each column represents each class, namely positive class, neutral class and negative class. Based on the results of the confusion matrix, it can be explained that the method correctly classifies 5 data as positive, 2 data as neutral and 30 data as negative. In addition, the model is wrong in predicting 17 words into a negative class that should be positive or commonly called False Positive. In addition, if the model incorrectly predicts the data into a positive class that should be negative then it is called False Negative. In the confusion matrix above there is no error in False Negative.

Based on the test results of the Naive Bayes classification model, it produced an accuracy value on the entire system of 56.92%. Figure 3 is the result of the evaluation of the Naïve Bayes model using python.

	precision	recall	f1-score	support
NEGATIF	0.53	1.00	0.69	30
NETRAL	1.00	0.15	0.27	13
POSITIF	0.83	0.23	0.36	22
accuracy			0.57	65
macro avg	0.79	0.46	0.44	65
weighted avg	0.72	0.57	0.49	65

Figure 3. Naive Bayes Model Evaluation

Results Comparison of precision, recall, and f1-score results can be seen in Table 3.

Table 3 Precision, Recall, and F1-Score Values

Classification Type	Precision	*recall*	*f1-score*
Negative	0,53	1,00	0,69
Neutral	1,00	0,15	0,27
Positive	0,83	0,23	0,36

The results of the model evaluation can be seen below the precision and recall values in each class. For precision, it can be said that the level of the system's ability to find accuracy between the information requested by the user is the negative class is 53%, for the neutral class by 100%, for the positive class by 83%. As for recall as the system's success rate in rediscovering information is for the negative class by 1.00%, for the neutral class by 15%, and the negative class by 23%.

FEATURE SELECTION

DOI : https://doi.org/ 10.26623/transformatika.v20i1.5144

In this study, the use of feature selection is to improve the accuracy of the Naive Bayes model that has been made. The feature selection used was XGBoost. To get a good model, the researchers try to adjust some parameters in order to get good accuracy results. The process of the XGBoost-based Naïve Bayes classification model was carried out by using the help of libraries in the Python3 programming language. After that, it drew the results of the model made by Naive Bayes by classifying the combination of MultinomialNB and XGBClassifier.

Xgboost Model Evaluation

To find out the performance of the expansion between the Naive Bayes method and the XGBoost feature selection, a test of the model had been created. The results of the combined classification in terms of increased accuracy would be visualized in the form of a confusion matrix. Similar to the previous evaluation of the Naive Bayes model, the first step started from using the sklearn. Metric python library which has confusion_matrix in it, then visualized by using seaborn. To find accuracy, it is the same thing as the accuracy calculation in the previous Naive Bayes method. From the results of the Confusion Matrix, the accuracy value of the XGBoost-based Naive Bayes was then calculated.

Table 4. Confusion Matrix Results

		Predict Class		
		Negatif	*Netral*	*Positif*
Actual Class	*Negatif*	25	3	9
	Netral	3	9	3
	Positif	2	1	10

Furthermore, the classification performance of each class can be known through precision, recall and f1-score values in each classification class. The precision value, recall and f1-score on the system can be seen in Figure 4.

	precision	recall	f1-score	support
NEGATIF	0.83	0.68	0.75	37
NETRAL	0.69	0.60	0.64	15
POSITIF	0.45	0.77	0.57	13
accuracy			0.68	65
macro avg	0.66	0.68	0.65	65
weighted avg	0.73	0.68	0.69	65

Figure 3. XGBoost Model Evaluation

The average precision value was 66%, a recall value of 68% and an f1-score value of 65%. Furthermore, the first experiment was conducted with split data sharing, which was 50% for test data and 50% for training data. Data sharing started from Figure 4.2 which was the change in the section "test_size = 0.3" to "test_size = 0.5". Furthermore, there was also the third experiment with 20% split data test data and the fourth experiment with 10% split data test data. This experiment was conducted to see the difference in accuracy of Naive Bayes classification results that were improved by XGBoost feature selection with all experiments. The final experiment was conducted to see if XGBoost's feature selection was successful in improving the accuracy of the Naive Bayes method. The overall test results on all four trials can be seen as the test results of all Naive Bayes accuracy before and after using the XGBoost feature selection which can be seen in table 5 below.

Table 5. Naive Bayes Method comparison results

Data Sharing	Naïve Bayes Without Cross-	Naive Bayes Cross-Validation	Naive Bayes based on XGBoost

		Validation		
1st trial	50% test data	60,18%	58,36%	60,18%
2nd trial	30% test data	56,92%	65,62%	67,69%
3rd trial	20% test data	65,90%	63,36%	70,45%
4th trial	10% test data	63,63%	63,42%	77,27%

Furthermore, *k-Fold Cross Validation* was performed from all experiments. The use of *k-Fold Cross Validation* was done to see the *score* value of the *Naive Bayes* method performance. The use of *k-Fold Cross* Validation was carried out 10 times from each experiment. Furthermore, the *score* in all experiments was 58.36% on the first try. The second experiment, it obtained a *score* of 65.62%. In the third experiment, the *score* was 63.36%. Then on the fourth try, it obtained a *score* of 63.42%.

Based on the results of the evaluation using Confusion Matrix, it proved that XGBoost optimization in the method optimization process could increase the accuracy value of Naive Bayes. The first experiment that had been done obtained an accuracy value of Naive Bayes 60,18% while the accuracy value of Naive Bayes after using the XGBoost feature selection was 60.18%. The second experiment obtained a Naive Bayes accuracy score of 56.92% while Naive Bayes' accuracy value after using the XGBoost feature selection was 67.69%. Furthermore, in the third experiment, Naive Bayes' accuracy value was 65.90% while Naive Bayes' accuracy value after using the XGBoost feature selection was 70.45%. Finally, the fourth attempt obtained an accuracy value of Naive Bayes 63.63% while the Naïve Bayes accuracy value after using XGBoost feature selection was 72.27%.

This accuracy value experienced an increase by 10.77% for the second trial of using Naive Bayes before and after adding the XGBoost feature selection. The third trial of using Naive Bayes before and after adding XGBoost feature selection resulted in an increase in accuracy by 4.55%. The fourth trial of using Naive Bayes before and after adding XGBoost feature selection increased in accuracy also by 13.64%. However, in the first trial of using Naive Bayes before and after adding the XGBoost feature selection, it did not experience an increase in accuracy of 0%.

CONCLUSION

Based on the results of analysis and testing of online learning sentiment analysis using the Naive Bayes method and XGBoost feature selection, it can be concluded as follows:

1. From the data processing that has been done using *feature selection*, namely *XGBoost*, it proves to improve accuracy in *Naive Bayes* classification. In the second experiment, with 70/30 split data, the accuracy value of the *Naive Bayes* model before using the *XGBoost* feature selection reached 56.92%, while after using the *XGBoost feature selection* the accuracy increased to 67.69%. In the third experiment, with 80/20 split data, the accuracy of the Naive Bayes model before using the *XGBoost feature selection* reached 65.90%, while after using the *XGBoost feature selection* the accuracy increased to 70.45%. In the fourth experiment, with 90/10 split data, *Naive Bayes* model accuracy before using *XGBoost* feature selection reached 63.63%, while after using *XGBoost feature selection* the accuracy increased to 77.27%. However, in the first trial of using Naive Bayes before and after adding the *XGBoost feature selection* did not experience an increase in accuracy of 0%.

DOI : https://doi.org/ 10.26623/transformatika.v20i1.5144

2. This study proves that the comparison of split data between 50/50, 70/30, 80/20 and 90/10 greatly affects improved accuracy. The study resulted in an average *score* at 10-Fold *Cross Validation* across all trials. The highest *score* was in the 2nd try with a value of 65.62%.
3. This research proves that the *Naive Bayes* classification method with *XGBoost feature selection* can be used in analyzing public opinion on the application of online learning in Indonesia in the mass media kompas.com, suara.com, and tribunnews.com.
4. Data that has been collected as many as 216 data. The results of the sentiment analysis showed that the negative category was 100 data, neutral amounted to 61 data, and positive amounted to 50 data.
5. Graph data shows that the application of online learning systems in Indonesia has a "Negative" opinion. So it is recommended for the government to review or reorganize policies in the implementation of online learning systems.

REFERENCE

[1] J. L. Moore, C. Dickson-deane, and K. Galyen, “e-Learning , online learning , and distance learning environments : Are they the same ?,” *Internet High. Educ.*, pp. 1–7, 2010, doi: 10.1016/j.iheduc.2010.10.001.

[2] J. Gikas and M. M. Grant, “Mobile computing devices in higher education : Student perspectives on learning with cellphones , smartphones & social media,” *Internet High. Educ.*, vol. 19, pp. 18–26, 2013.

[3] L. Handayani, “Keuntungan , Kendala dan Solusi Pembelajaran Online Selama Pandemi Covid-19 : Studi Ekploratif di SMPN 3 Bae Kudus Lina Handayani,” *J. Ind. Eng. Manag. Res. (JIEMAR)*, vol. 1, no. 2, pp. 15–23, 2020.

[4] M. K. Anam, Rahmaddeni, M. B. Firdaus, H. Asnal, and Hamdani, “Sentiment Analysis to analyze Vaccine Enthusiasm in Indonesia on Twitter Social Media,” *JAIA – J. Artif. Intell. Appl.*, vol. 1, no. 2, pp. 23–27, 2021.

[5] Agus Purwanto *et al.*, “Studi Eksploratif Dampak Pandemi COVID-19 Terhadap Proses Pembelajaran Online di Sekolah Dasar,” *J. Educ. Psychol. Couns.*, vol. 2, no. 1, pp. 1–12, 2020.

[6] R. H. S. Aji, “Dampak Covid-19 pada Pendidikan di Indonesia: Sekolah, Keterampilan, dan Proses Pembelajaran,” *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 5, p. 402, 2020, doi: 10.15408/sjsbs.v7i5.15314.

[7] A. Sadikin and A. Hamidah, “Pembelajaran Daring di Tengah Wabah Covid-19,” *BIODIK J. Ilm. Pendidik. Biol.*, vol. 6, no. 2, pp. 214–224, 2020, doi: 10.22437/bio.v6i2.9759.

[8] R. A. Stein, “COVID-19 and rationally layered social distancing,” *Int. J. Clin. Pract.*, vol. 74, no. 7, pp. 1–3, 2020, doi: 10.1111/ijcp.13501.

[9] M. K. Anam, “Analisis Respons Netizen Terhadap Berita Politik Di Media Online,” *J. Ilm. Ilmu Komput.*, vol. 3, no. 1, pp. 14–21, 2017, doi: 10.35329/jiik.v3i1.62.

[10] A. Rakhman and M. Rifqi Tsani, “Analisis Sentimen Review Media Massa Menggunakan Metode C4.5 Berbasis Forward Selection,” *Smart Comp*, vol. 8, no. 2, pp. 78–82, 2019, doi: 10.30591/smartcomp.v8i2.1491.

[11] Y. Cahyono, “Analisis Sentiment pada Sosial Media Twitter Menggunakan Naïve Bayes Classifier dengan Feature Selection Particle Swarm Optimization dan Term Frequency,” *J. Inform. Univ. Pamulang*, vol. 2, no. 1, p. 14, 2017, doi: 10.32493/informatika.v2i1.1500.

[12] A. N. Ulfah and M. K. Anam, “Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM),” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[13] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, “Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah,” *Matrik J. Manajemen, Tek. Inform. dan Rekayasa*

Komput., vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.
[14] I. Taufik and S. A. Pamungkas, "ANALISIS SENTIMEN TERHADAP TOKOH PUBLIK MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)," *J. "LOG!K@,"* vol. 8, no. 1, pp. 69–79, 2018.
[15] A. Turmudi and K. S. Yasah, "ANALISA SENTIMEN TWEET INDONESIA MENGGUNAKAN FITUR EKSTRASI DAN TEKNIK CROSS VALIDATION TERHADAP MODEL NAÏVE BAYES," *e-Proceeding Eng.*, vol. 10, no. 4, p. 275, 2020, doi: 10.35393/1730-006-002-014.
[16] C. Jaya and K. Muslim, "Analisis Sentimen Pada Media Daring Tentang Pemilihan Presiden Indonesia Tahun 2019 Menggunakan Metode Naïve Bayes," *e-Proceeding Eng.*, vol. 6, no. 2, p. 9698, 2019.
[17] R. Nadia, K. Muslim, and F. Nhita, "Analisis Dan Implementasi Algoritma Naïve Bayes Classifier Terhadapa Pemilihan Gubernur Jawa Barat 2018 Pada Media Online," *e-Proceeding Eng.*, vol. 5, no. 1, pp. 1678–1700, 2018.
[18] Paisal, "Analisis Sentimen Masyarakat Berdasarkan Opini dari Sosial Media Menggunakan Metode Naive Bayes Classifier (Study Kasus : Universitas Sjakhyakirti)," *J. Ilm. Inform. Glob.*, vol. 11, no. 01, pp. 41–46, 2020.
[19] S. Yualinda, D. R. Wijaya, and E. Hernawati, "Aplikasi Berbasis Dataset E-Commerce Untuk Prediksi Kemiskinan Menggunakan Algoritma Naïve Bayes, Xgboost Dan Similarity Based Feature Selection," *e-Proceeding Appl. Sci.*, vol. 6, no. 2, pp. 9–11, 2020.
[20] A. P. Natasuwarna, "Seleksi Fitur Support Vector Machine pada Analisis Sentimen Keberlanjutan Pembelajaran Daring," *Techno.COM*, vol. 19, no. 4, pp. 437–448, 2020.
[21] B. D. C, A. Amelia, U. Hasanah, A. M. Putra, and H. Rahman, "Analisis Keefektifan Pembelajaran Online di Masa Pandemi Covid-19," *Pendidik. Guru Sekol. Dasar*, pp. 28–37, 2020.
[22] W. Duan, Q. Cao, Y. Yu, and S. Levy, "Mining Online User-Generated Content : Using Sentiment Analysis Technique to Study Hotel Service Quality," *Hawaii Int. Conf. Syst. Sci. Min.*, 2013, doi: 10.1109/HICSS.2013.400.
[23] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, 2009, doi: 10.1016/j.eswa.2008.06.054.
[24] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Feature selection and parameter optimization for support vector machines : A new approach based on genetic algorithm with feature chromosomes," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5197–5204, 2011, doi: 10.1016/j.eswa.2010.10.041.
[25] A. Taufik, "Optimasi Particle Swarm Optimization Sebagai Seleksi Fitur Pada Analisis Sentimen Review Hotel Berbahasa Indonesia Menggunakan Algoritma Naïve Bayes," *J. Tek. Komput.*, vol. III, no. 2, pp. 40–47, 2017.
[26] V. Arun, M. Krishna, B. V. Arunkumar, S. K. Padma, and V. Shyam, "Exploratory Boosted Feature Selection and Neural Network Framework for Depression Classification," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 3, p. 61, 2018, doi: 10.9781/ijimai.2018.10.001.
[27] I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," *Wirel. Days*, 2019.
[28] I. M. K. Karo, "Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan," *J. Softw. Eng. Inf. Commun. Technol.*, vol. 1, no. 1, pp. 10–16, 2020.
[29] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no.

1, pp. 45–55, 2021, doi: 10.26555/jifo.v15i1.a20111.

[30] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, pp. 37–43, 2012.

[31] J. H. Friedman, "Greedy function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[32] M. K. Anam, M. I. Mahendra, W. Agustin, Rahmaddeni, and Nurjayadi, "Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA)," *Intensif*, vol. 6, no. 1, pp. 2549–6824, 2022, doi: 10.29407/intensif.v6i1.15870.

[33] T. Chen and C. Guestrin, "XGBoost : A Scalable Tree Boosting System," pp. 785–794, 2016.

[34] Y. T. Pratama, F. A. Bachtiar, and N. Y. Setiawan, "Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 6244–6252, 2018.

[35] B. N. Pikir, M. K. Anam, H. Asnal, Rahmaddeni, and T. A. Fitri, "Sentiment Analysis of Technology Utilization by Pekanbaru City Government Based on Community Interaction in Social Media," *JAIA – J. Artif. Intell. Appl.*, vol. 2, no. 1, pp. 32–40, 2021.