Advances in Social Science, Education and Humanities Research, volume 585
Proceedings of the 1st UMGESHIC International Seminar on Health, Social Science and Humanities (UMGESHIC-ISHSSH 2020)

Comparison of Naive Bayes and Random Forests Classifier in the Classification of News Article Popularity as Learning Material

Utomo Pujianto, Ilham Ari Elbaith Zaeni, Khalida Izdihar Rasyida

Universitas Negeri Malang, East Java, Indonesia
Correspoding author email: utomo.pujianto.ft@um.ac.id

Abstract. Tweentribune.com is a website that provides news texts with Lexile level information for each text. The Lexile level information feature is useful as a consideration for visitors to choose text that has a level of difficulty that matches their age. With such information, any news text on the site can be used as attractive teaching material for both teachers and students. However, the number of visits to each text varies widely. This study assumes that the popularity of each text is influenced not only by Lexile level information but also by other text characteristics. This research produces a number of engineering features that are extracted from the text to be used as a predictive attribute in classifying the popularity of the texts in question. Naïve Bayes and Random Forest are two classifiers used together with two popularity cluster scenarios based on k-means clustering. The results of the testing and evaluation phase show that the Random Forest algorithm has the best performance, with an accuracy value of 99.75%, an average recall value of 99.7%, and an average precision value of 98.7%.

Keywords: *Naïve Bayes, Forest, Learning, Material*

1. INTRODUCTION

The age of children is an age that has not been able to choose good reading material for themselves[1]. In the stages of children's psychological development has different characteristics where this means that there are differences in children's responses to reading books in each development. Therefore, proper reading will play a role in supporting the growth and development of children's mental health including aspects of thinking, language, personality, and morality[2].

The selection of reading load in accordance with the ability of students is one of the factors supporting the search for good learning resources. Because the purpose of reading is understanding[3], if the reading load is not in accordance with the child's ability, the information contained in the reading cannot be conveyed to the maximum[4]. In the US there is a mechanism for classifying reading loads called lexiles. This is intended to find the level of reading load in accordance with the intended ability of the reader. The higher the reading load given, the higher the reading ability needed.

According to data from the Pew Research center in 2005-2015, 65% of adults in the US use social media, where most traces are left in the form of comments about feelings, opinions, and ideas about individual, social, product, or even events that occur in around[5]. The distribution of data more and more every day raises the problem of the difficulty of monitoring the distribution of news or important events that occur and deserve to be used as literacy material [6].

News articles are reports about an event or a recent occurrence; reports of facts that are actual, interesting attention, considered important or extraordinary so it needs to be known by human readers / listeners / viewers[7]. News articles on the internet can be used as literacy or learning material if the contents, delivery methods, and reading load are in accordance with the intended object. Its newest, actual, and not limited geographical nature can be used as a material to stimulate students' mindset in responding to events that occur in the environment and the world.

The way of writing news articles actually cannot be used as a standard predi ction of attractiveness and quality of content[8]. This happens because in addition to the journalists writing topics that are trending with a long narrative, the length of time a news article is distributed also affects the number of responses to the news article. Good writing can affect the quality of remembering readers so that the contents delivered can be understood[4] and the reader captures the impression that the journalist wants to convey. However, until now there has been no specific research that discusses the comparison or characteristics of good article writing by overriding the topics discussed[9][10]. So there needs to be a

study of the similarity between news article data that is of interest to the reader to get characteristics that can be used as literature and learning material.

Naïve Bayes and Random Forest are data mining algorithms that have good performance in the application of text mining. Algorithm naïve Bayes algorithm is in the process of classification, based on the probability of dataset. While random forest algorithm is in the process of classification, based on the majority result klasfikasi process every tree.

Based on the elaboration, comparative performance of naïve bayes and random forest algorithm will be performed to classify the popularity of good reading material characteristic based on sentence complexity, vocabulary, and lexail level values.

2. METHODS

In this study, there are tree stages: (1) Data Collection, (2), Preprocessing, (3) Selection and Addition, (4) Classification, (5) Evaluation.

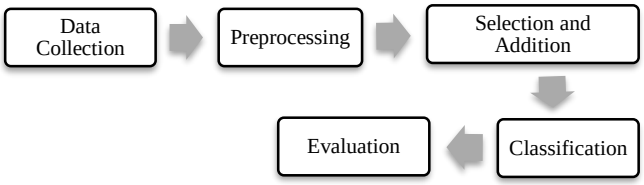

FIGURE 1. Research Stage

2.1. Data Collection

The data was obtained from the scraping process at www.tweentribune.com . This is online educational websites daily news provider for children and adolescents. Provided the article is equipped with a value lexile levels corresponding to four grade age group grade k-4, grade 5-6, grade 7-9, grade 9-12. The scrapping attributes are grade, article content, lexile level, assigned, comment, and date the article was uploaded with detailed explanation which can be seen in TABLE II.

2.2. Preprocessing Data

Before entering the stage of data classification, the need for the preprocessing of data including: (1) Calculate total number of words, (2) Transformation, (3) Calculate the total number of unique words, (4) Calculate the age of the article.

2.2.1. Calculate Total Number of Words

To calculate total number of words in the article. This process is done to look at the complex itas text articles were interesting to read[10] based on the value lexile level.. Because the more complex text of the article content, the higher memory needed to remember and understand the contents of the article.

2.2.2. Transformation

Transformation is a stage that is carried out to change the text that is form of a document into a token index. The steps taken are case folding, tokenization, and stemming.

Stage case folding is a step change all the letters in the dataset into small letters (lowercase). Tokenisation is the process of splitting a series of sentences into a dataset or piece of word that stands alone (tredword). Stemming is a process carried out to eliminate affixes to words so that all words turn into basic words. This stemming process uses the lovins stemmer algorithm.

2.2.3. Counting Nunber of Unique Words

To count the number of basic words that exist in the article, where the basic words that have repeated use still counted one. Proses is done after the process of transformation that converts text into a document in the form of tokens already count of index frequency of words.

This process is done to see how much of the vocabulary used to prepare articles of interest to the reader based on lexile value pre-determined levels. Because the diversity of vocabulary used affects the child's understanding of the contents of the article in terms of language skills and illustrates the breadth of insights conveyed[11].

2.2.4. Calculate the Age of Article

Calculate how long the life of articles on a daily basis, ranging from articles uploaded to experience the process of scraping.

2.3. Selection and Addition

The first process in stages selection and addition is clustering using K-Means Clustering algorithm to determine the grade based on attributes comments, assigned, and the age of the article.

Clustering is a method used to group statistical data with many fields or attributes into groups of data with the same characteristics. Discretization or grouping of data was done in order to get a semblance of data in the cluster or differences from

one other[12]. Or it can be said with a method that aims to minimize the diversity of characteristics in a group of data.

K-Means Cluster Algorithm is an iterative clustering algorithm that isolates a group of data into several predefined K- clusters. Each instance of the K-Means Cluster is included in a group, but can change depending on the number of iterations performed.

The results of the cluster using the K-Means Cluster method are very dependent on the initial centroid value given. Giving different initial centroids will probably result in different groups. There are several methods to determine the initial centroid, by taking an initial sample of an object, then looking for the centroid value, or by providing a random initial value.

2.4. Classification

Classification is the process of finding a model or function that distinguishes concepts or classes of data, with the aim of being able to estimate the class of an object whose class is unknown. To achieve this goal, the classification process forms a model that can differentiate data into classes using certain rule or function models. The rule model can be an "if-then" rule, a decision tree or even a mathematical formula[13]. To understand the concept of classification, classification process flow can be seen in FIGURE 2

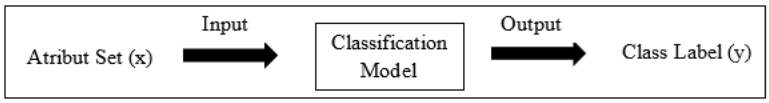

FIGURE 2. Block Classification Model Diagram

Testing the performance of the algorithm for the article classification process, this research uses two classification algorithms, Naïve Bayes and Random Forest.

2.4.1. Naïve Bayes

Naïve Bayes is a simple probabilistic classification that calculates a set of probabilities [14] by adding up the frequency and combination of values from a given dataset [15]. The application of the naïve bayes algorithm is based on the Bayes theorem, where the attribute value in a class does not depend on the value of other attributes [13].

The advantage of using the Naïve Bayes algorithm is that there is little need for training data to determine the estimated parameters needed in the classification process. Naïve Bayes often performs better in most complex real-world situations than expected [16]. To clarify the calculation of Naïve Bayes algorithm can be seen in equation 2.1

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (2.1)$$

Where:

- X : data with unknown classes
- H : class hypothesis that has been determined
- P(H|X) : probability of H occurrence if X occurs
- P(X|H) : probability of event X if H occurs
- P(H) : probability of H
- P(X) : probability of X

2.4.2. Random Forest

The Random Forest method is a development of the CART method which applies the bootstrap aggregating (bagging) method and random feature selection[17]. This method is used to build a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the provisions in force[18].

Decision tree starts with a way to calculate the value entopy as a determinant of the level of impurities attributes and values of information gain. To calculate the entropy value , use Equation 2.2[19].

$$Entropy\ (Y) = -\sum_i p(c|Y) log_2 p(c|Y) \quad (2.2)$$

Where Y is the case set and p (c | Y) is a proportion of the value of Y to class c.

Meanwhile, to see the value of the gain information used Equation 2.3[20]

$$= Entropy(Y) - \sum_{v \in Values(a)} \frac{|Y_v|}{|Y_a|} Entropy\ (Y_v) \quad (2.3)$$

Where Values (a) are all possible values in the case set a. Yv is a subclass of Y with class v relating to class a. Yes are all values that correspond to a.

The flow chart of the Random Forest algorithm is as follows[6]:

1. Divide data into several trees by taking samples in a random way using the bootstrap method
2. Each node in the tree will continue to divide to find the value of m that corresponds to each class.

2.5. Evaluation

After applying the two classification algorithms, the algorithm performance testing is done using the cross validation method with a K- setting of 10 fold. The greater the K value, the difference between the size of the training data and the recurrence of the sample set will be smaller. When this difference decreases, the technical bias becomes smaller. Selection of K value of 10 is recommended to select the best model because it further increases the value of accuracy and reduces the possible bias[21].

The results of testing using cross validation will be presented in the form of confusion matrix to facilitate evaluation process. Examples of confusion matrices for binary classification are shown in TABLE I[22]:

TABLE 1. Confusion Matrix

Correct Classification	Classification Results	
	+	-
+	TP	FN
-	FP	TN

True Positive (TP) is data that has "True" value both in the classification results or in the original label class. While True Negative (TN) is a data that has a value of "false" both in the classification results and in the original label class. False Negative (FN) is data that is actually class as " True " but in Classification Results is " False " and the last is False Positive (FP) , False Positive (FP) is data that on the label is actually " false " but on the classification results are " true " . To evaluate the algorithm using a confusion matrix can be done in 3 kinds of ways, namely (1) Accuracy, (2) Precision, (3) True Positive Rate.

2.5.1. Accuracy

Accuracy is used to test how large a percentage classification accuracy results that can show by an algorithm. Accuracy can be calculated using Equation 2.4 2.4:

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} x\ 100\% \quad (2.4)$$

2.5.2. Presisi

Presisi used to test the algorithm in determining the classification of true that correspond to the actual classification. The accuracy of precision can be calculated using Equation 2.5:

$$Presisi = \frac{TP}{TP+FP} \quad (2.5)$$

2.5.3. True Positive Rate (Recall)

Rate TP is used for comparison of the proportion of classification (TP) on all the data that is valuable positive on the actual class. TP Rate can be calculated using Equation 2.6:

$$TP\ Rate = \frac{TP}{TP+FN} \quad (2.6)$$

3. RESULT AND DISCUSSION

3.1. Data Collection

The data taken is all data uploaded up to February 6, 2019 for k-4 grades, January 3, 2019 for grades 5-6, February 07 2019 for grades 7-8, and 08 February 2019 for grades 9-12 on online education sites www.tweentribune.com obtained by scraping using the help of the octoparse application. With a total of 7574 data collected. Lists attributes of the data in the dataset are shown in TABLE 2

TABLE 2. Attribute in Dataset

Attribute Name	**Explanation**	**Data Type**	**Range**
Grade	Age group according to lexile levels	Nominal	k-4, 5-6, 7-8, 9-12
Isi	Fill out the articles obtained from the site www.tweentribune.com	Text	-
Lexile	The lexile level value of the article	Numeric	390-12430
Assigned	The number of times the article was used as an assignment by the teacher	Numeric	0-987
Comment	The number of times the article was commented on	Numeric	0-500
Date	Date of article uploaded	Date	17/04/2014 – 06/02/2019

3.2. Preprocessing

The first process carried out in preprocessing is data cleaning. This process is carried out to clean the dataset from data that does not meet the criteria for processing the research being carried out. Where in this study, the dataset used is only data that has a complete attribute value, and does not contain video elements.

The data cleaning process is done manually by deleting data that does not meet the criteria, as well as cleaning noise that is picked up during the scrapping process. After going through the data cleaning process, the total initial data amounted to 7574 data, reduced to 7061 data. Where the clean data with a total of 7061 data is used as a dataset that is processed in the study.

The second process counts the total number of words in the article content. This process is carried out using Microsoft Excel with Formula 3.1

=(LEN(sel_isi)-LEN(SUBSTITUTE(sel_isi," ","”))+1) *(3.1)*

The formula LEN (sel_isi) functions to count the total number of characters along with the number of matches in the contents of the article. While the formula LEN(SUBSTITUTE (sel_isi, " ", "")) +1 functions to count the total number of characters in the article without counting the existing spaces . So based on these two formulas, you will get the total number of words in the article content by counting the number of spaces plus one where the addition of one is represented as the first word of the article .

The third process is the transformation carried out to change the text that is still in the form of a document into a token index. The steps taken are case folding, tokenization, stopward removal, and stemming.

The case folding stage is the stage of changing all the letters in the dataset to lowercase. Tokenization is the process of splitting a series or series of sentences into a dataset or piece of word that stands alone (tredword). Stopward removal is a phase removal term that is not related to the subject but generally appears in the text. This process uses the Rainbow algorithm. Stemming is a process carried out to eliminate affixes to words so that all words turn into basic words. This stemming process uses the lovins stemmer algorithm.

The fourth process of preprocessing is counting the number of unique words. Counting the number of unique words is done by finding the total number of basic words that appear in the article, where words that have more than one frequency of occurrence, will still be counted one. This process is carried out using Microsoft Excel with Formula 3.2

= ((COUNT(range))-(COUNTIF(range,0))) *(3.2)*

Where the COUNT formula (range) is used to count the number of all the standard words that have been found in the transformation process, while the COUNTIF formula (range, 0) is used to count the standard words that have a frequency value of 0. Based on the two formulas will get the number of standard words from any contents that have a frequency value of greater than 0.

The final process of preprocessing is calculating the age of the article from the article uploaded to the scraping process in units of days. This process is done menggunaka n Microsoft Excel with Formula 3.3.

= DATEIF(tgl_upload,tgl_scrapping,"d") *(3.3)*

The formula is used to calculate the difference between the date the article was uploaded, and the date the article was scrapped in days (d).

Details of the attributes after preprocessing can be seen in TABLE 3

TABLE 3. Attribute in Dataset After Preprocessing

Attribute Name	**Explanation**	**Data Type**	**Range**
Grade	Age group according to lexile levels	Nominal	k-4, 5-6, 7-8, 9-12
Isi	Fill out the articles obtained from the site www.tweentribune.com	Text	-
Jumlah_kata	The total number of words in the article	Numeric	92-2658
Kata_unik	The number of standard words in the article without any repetition of numbers	Numeric	39-303
Lexile	The lexile level value of the article	Numeric	390-12430
Assigned	The number of times the article was used as an assignment by the teacher	Numeric	0-987
Comment	The number of times the article was commented on	Numeric	0-500

Date	Date of article uploaded	Date	17 April 2014 - 06 Februari 2019
Umur_artikel	The age of the article in a matter of days, starting from the date the article was uploaded to the scraping proccess.	Numeric	0-1758

3.3. Selection and Addition

The first stage of the selection and addition process is data clustering. Data clustering is done by grouping the attributes of Assigned, Comments, and Age of the article using the K-Means algorithm.

In this stage there are two scenarios performed on the clustering process attributes. The first scenario is directly inputting three attributes that have been determined in the clustering process. While the attributes of the second scenario is obtained by calculating the average assigned and comments on age, which is done by dividing the number of assigned and comment to the age of the article.

For details and explanations of the attributes of the clustering process, see TABLE 4 and TABLE 5

TABLE 4. Attribute Scenario 1

Attribute	Explanation	Data Type
Assigned	The number of times the article was used as an assignment by the teacher	Numeric
Comment	The number of times the article was commented on	Numeric
Umur_artikel	The age of the article in a matter of days, starting from the date the article was uploaded to the scraping process.	Numeric

TABLE 5. Attribute Scenario 2

Attribute	Explanation	Data Type
Prob. Assigned	Average of assigned every single day, which is obtained by dividing the amount assigned to the age of the article	Numeric
Prob. Comment	Average of comments every single day, which is obtained by dividing the number of comments to the age of the article	Numeric

The clustering process using K-Means Clustering is done by determining the initial centroid value at random, using euclidean distance as a distance measure, and setting a maximum iteration of 100 times.

The result of clustering based on two scenarios treatment appointed in TABLE 6

TABLE 6. Clustering Result

Scenario	Instance Cluster_0	Instance Cluster_1	Instance Cluster_2	Iterasi
1	3217	803	3041	10
2	6871	18	172	10

Visualization clustering results of TABLE VI can be seen in FIGURE 3, FIGURE 4 and FIGURE 5

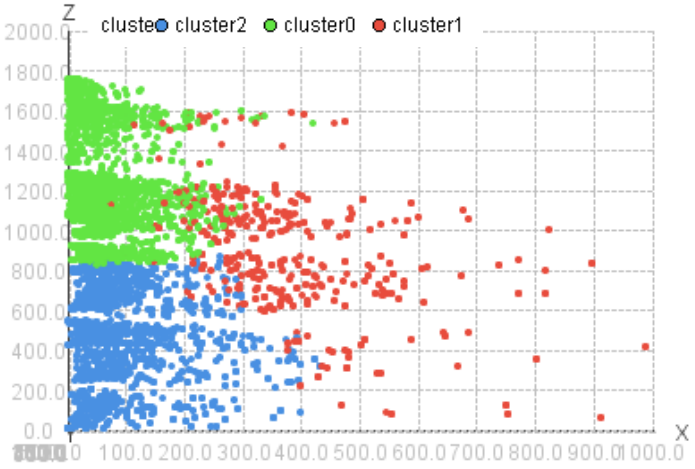

FIGURE 3. Visualization Of Clustering Result Using First Scenario With X-axis Viewpoint

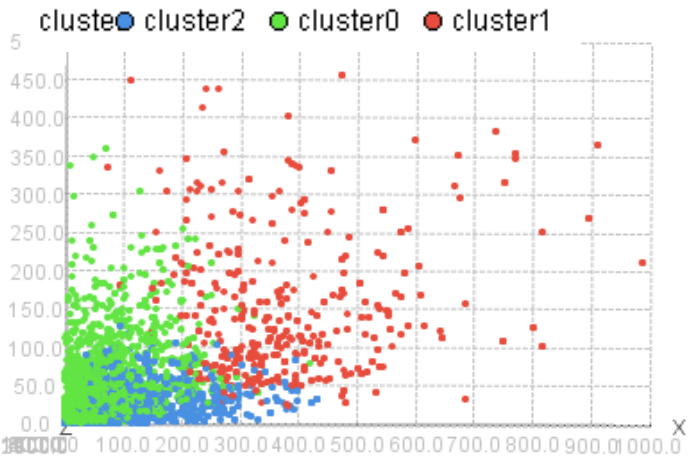

FIGURE 4. Visualization Of Clustering Results Using First Scenario With Z-Axis Viewpoint

Details of the distribution of clustering attributes using scenario 1 can be seen in FIGURE 5

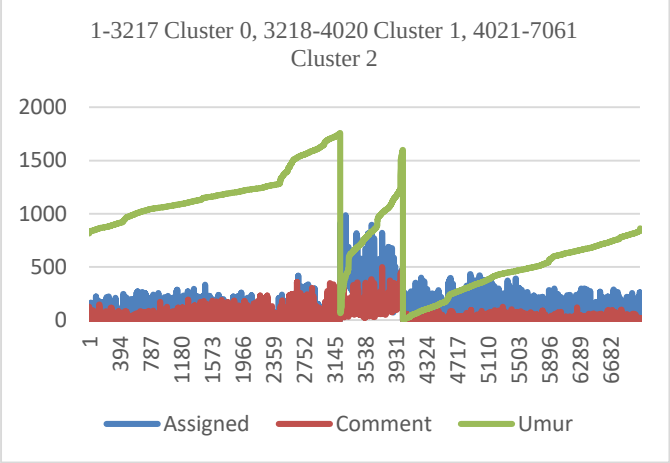

FIGURE 5. Graph Of The Distribution Of Clustering Attribute Using Scenario 1 In Detail

FIGURE 3 and FIGURE 4 are visualizations of the results of the first clustering scenario, where the x axis represents the number of assigned, the y axis represents the number of comments, and the z axis represents the age of the article.

Based on the appearance of FIGURE 3 and FIGURE 4 as well as the distribution details of clustering attributes in FIGURE 5, it can be concluded that cluster 1 is a group of articles that are *populer* class, cluster 0 is a group of articles that are class *sedang*, and cluster 2 is a group of articles class *tidak populer*.

This can be seen from the distribution of data plots in groups based on the number of assigned and comments with the influence of the amount of time they have. Cluster 1 has the distribution of data plots with the highest assigned values and comments among other cluster of medium age which means the article is considered good by educator. Because it often use as teaching material, and gives the impression to the reader based of the high comment value. Cluster 0 has a distribution of data plots with assigned values and comments that are straight-line, wich means that article is considered good by educator and is responded well by the reader even though the age is fairly old. While cluster 2 has the distribution of data plots with high assigned values but low comment values compared to other cluster this show that article is often use as a task but does not give the impression to ther eader.

Visualization results of *clustering* using the scenario of both can be seen in FIGURE 6

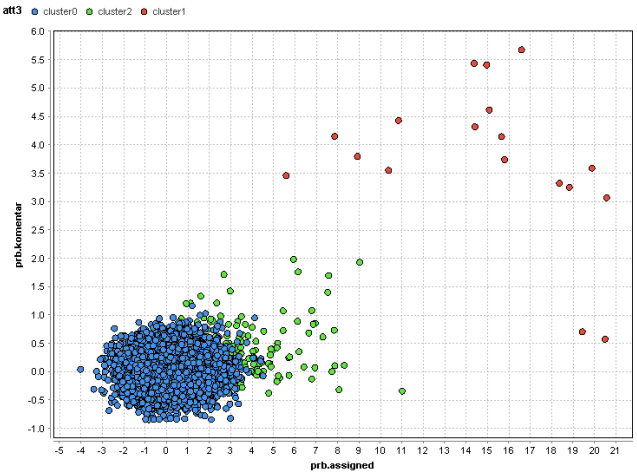

FIGURE 6. Visualization The Result Of The Second Scenario Clustering

Details of the distribution of clustering attributes using scenario 2 are sorted by age of articles from small to large in each class can be seen in FIGURE 7

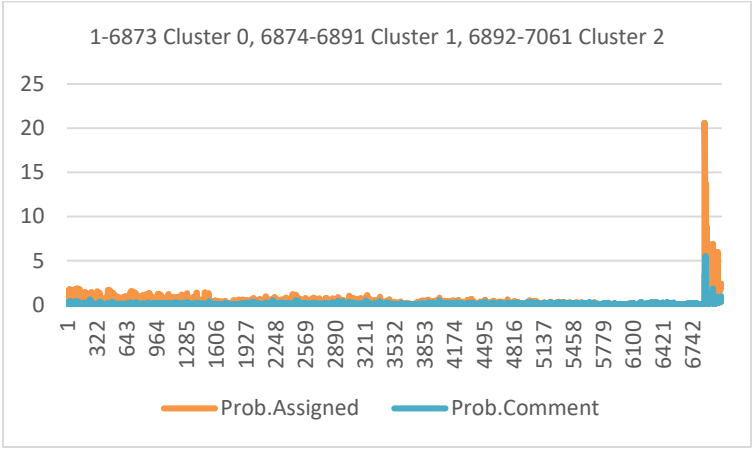

FIGURE 7. Graph Of The Distribution Of Clustering Attribute Using Scenario 2 In Detail

Based on the visualization of FIGURE 6 and the elaboration of the distribution of clustering attributes using scenario 2 in FIGURE 7, it can be concluded that cluster 0 is an article group with a *tidak populer* class, cluster 2 is an article group with a *sedang* class, and cluster 1 is an article group with a *populer* class.

This can be concluded based on the results of the distribution of plot data, where cluster 0 based on upload age has the possibility of being assigned and comments are relatively small, whereas cluster 2 based on upload age has the possibility of being assigned and comments greater than cluster 0, and cluster 1 based on upload age has the possibility the assigned and comments are bigger than the other two clusters. For details on the proportion of class members the results of the two clustering scenarios can be seen in TABLE 7

TABLE 7. Proportion Of Dataset Class Membership

Scenario	Instance Name	Number of Instances	Percentage
1	*Tidak Populer*	3041	43.06%
	Sedang	3217	45.56%
	Populer	803	11.37%
2	*Tidak Populer*	6871	97.30%
	Sedang	172	2.44%
	Populer	18	0.25%

Then the grade attribute is removed because the lexile levels attribute represents the age target of the article reader. Where the growing age of the reader, the need for reading loads is also higher. The article content attribute is removed and replaced by the number of words and the number of unique words that result from the preprocessing process. Deleting the assigned attribute, comment, and article age (date) is also done because this attribute is no longer needed in the classification process and is replaced with the class attribute resulting from the clustering process. List of dataset attributes after the selection and addition process can be seen in TABLE 8

TABLE 8. List of Attribute After Selection and Addition Process

Attribute Name	Explanation	Data Type	Range
Jumlah_kata	The total number of words in the article	Numeric	92-2658
Kata_unik	The number of standard words in the article without any repetition of numbers	Numeric	39-303
Lexile	The lexile level value of the article	Numeric	390-12430
Class	The results of clustering use K-Means based on the assigned attributes, comments, and age of the article	Nominal	*Populer*, *Sedang*, *Tidak Populer*

The distribution of input attribute data from the clustering process uses the first scenario sorted by age of the article can be seen in FIGURE 8

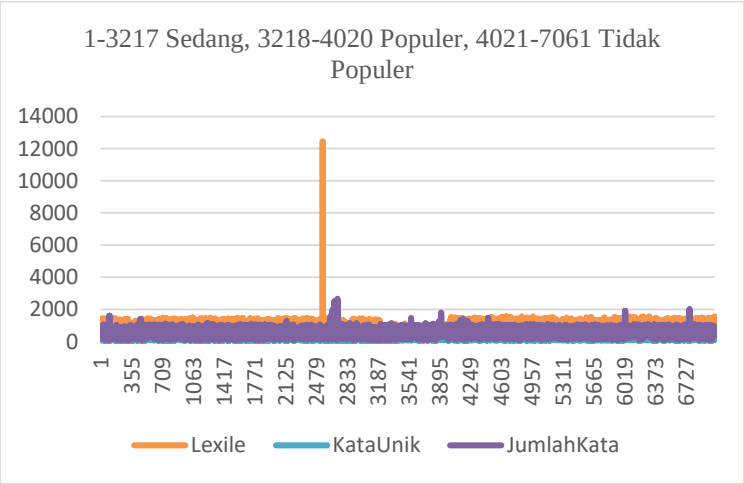

FIGURE 8. Distribution of Input Clustering Process Using Scenario 1

Whereas for the distribution of input attribute data from the clustering process using the second scenario sorted by age of the article can be seen in FIGURE 9

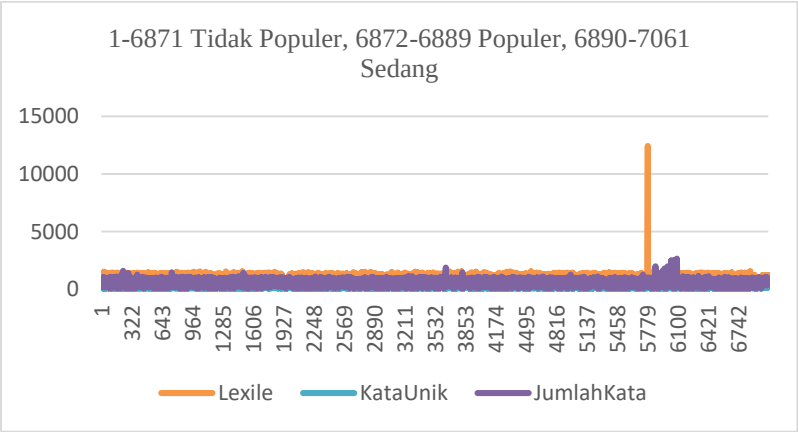

FIGURE 9. Distribution of Input Attribute Clustering Using Scenario 2

3.4. Classification

The evaluation process was carried out with several data experiments. The experiments carried out include: (1) Accuracy of the first scenario using Naïve Bayes algorithm, (2) Accuracy of the first scenario using Random Forest algorithm, (3) Accuracy of the second scenario using Naïve Bayes algorithm, (4) Accuracy of the second scenario using Random Forest algorithm.

Data processing in this study uses the help of Weka application, where the results of each scenario experiment with different algorithms will be compared and analyzed to find the algorithm with the best performance value.

3.5. Evaluation Hasil Klasifikasi

Here are the results of research using the Naïve Bayes and Random Forest algorithms on the article dataset obtained from the page www.tweentribune.com with 7061 data.

3.5.1. First Scenario with Naive Bayes classifier

The Naïve Bayes algorithm test results with the 10 Fold Cross-Validation testing method are shown using the confusion matrix in TABLE 9

TABLE 9. Confusion Matrix Scenario 1 Uses Naïve Bayes

Correct Classification	Classification Results		
	Tidak Populer	*Sedang*	*Populer*
Tidak Populer	1808	1606	452
Sedang	1230	1606	345
Populer	3	5	6

Based on the results of the confusion matrix TABLE, the calculation of accuracy is as follows:

$$Accuracy = \frac{1808 + 1606 + 6}{7061} \; x \; 100\%$$

$$= \frac{3420}{7061} \; x \; 100\% = 48.44\%$$

As for the precision and recall results can be seen in TABLE 10

TABLE 10. Test Performances of Naïve Bayes Algorithm Using Scenario 1

	Presisi	Recall
Tidak Populer	0.468	0.595
Sedang	0.505	0.499
Populer	0.429	0.007

3.6. First Scenario with Random Forests classifier

The results of testing the Random Forest algorithm with the test method in the form of 10 Fold Cross-Validations are shown using the confusion matrix in TABLE 11

TABLE 11. Confusion Matrix Scenario 1 Uses Random Forest

Correct Classification	Classification Results		
	Tidak Populer	*Sedang*	*populer*
Tidak Populer	2891	270	21
Sedang	136	2889	40
Populer	14	58	742

Based on the results of the confusion matrix TABLE, the calculation of accuracy is as follows:

$$Accuracy = \frac{2891 + 2889 + 742}{7061}\ x\ 100\%$$
$$= \frac{6522}{7061}\ x\ 100\% = 92.37\%$$

As for the precision and recall results can be seen in TABLE 12

TABLE 12. Test Performance of Random Forest Algorithm Using Scenario 1

	Presisi	Recall
Tidak Populer	0.909	0.951
Sedang	0.943	0.898
Populer	0.912	0.924

3.7. Second Scenario with Naive Bayes classifier

The Naïve Bayes algorithm test results with the 10 Fold Cross-Validation testing method are shown using the confusion matrix in TABLE 13

TABLE 13. Confusion Matrix Scenario 2 Uses Naïve Bayes

Correct Classification	Classification Results		
	Tidak Populer	*Sedang*	*Populer*
Tidak Populer	6871	172	18
Sedang	0	0	0
Populer	0	0	0

Based on the results of the confusion matrix TABLE, the calculation of accuracy is as follows:

$$Accuracy = \frac{6871 + 0 + 0}{7061}\ x\ 100\%$$
$$= \frac{6871}{7061}\ x\ 100\% = 97.31\%$$

As for the precision and recall results can be seen in TABLE 14.

TABLE 14. Test Performance of Naïve Bayes Algorithm Using Scenario 2

	Presisi	Recall
Tidak Populer	0.973	1.00
Sedang	0	0
Populer	0	0

3.8. Second Scenario with Random Forests classifier

The results of testing the Random Forest algorithm on the dataset with a total amount of 7061 data, with the test method of 10 Fold Cross-Validations are shown using the confusion matrix in TABLE 15.

TABLE 15. Confusion Matrix Scenario 2 Uses Random Forest

Correct Classification	Classification Results		
	Tidak Populer	*Sedang*	*Populer*
Tidak Populer	6865	11	1
Sedang	6	161	0
Populer	0	0	17

Based on the results of the confusion matrix TABLE, the calculation of accuracy is as follows:

$$Accuracy = \frac{6865 + 161 + 17}{7061}\ x\ 100\%$$
$$= \frac{7043}{7061}\ x\ 100\% = 99.75\%$$

As for the precision and recall results can be seen in TABLE 16

TABLE 16. Test Performances of Random Forest Algorithm Using Scenario 2

	Presisi	Recall
Tidak Populer	0.998	0.999
Sedang	0.964	0.936
Populer	1.00	0.944

3.9. Perbandingan Seluruh Hasil Evaluasi

After comparing the performance of the Naïve Bayes algorithm and Randmm Forest, the results of the performance comparison in terms of precision are shown in TABLE 17

TABLE 17. Comparison of Precision Algorithm Performance

Scenario	Class	Presisi	
		Naïve Bayes	*Random Forest*
1	*Tidak Populer*	0.468	0.909
	Sedang	0.505	0.943
	Populer	0.429	0.912
2	*Tidak Populer*	0.973	0.998
	Sedang	0	0.964
	Populer	0	1.00

The results of the comparative test performance of the Naïve Bayes and Random Forest algorithms in terms of recall are shown in TABLE 18

TABLE 18. Comparison of Recall Algorithm Performance

Scenario	Class	Recall	
		Naïve Bayes	Random Forest
1	*Tidak Populer*	0.595	0.951
	Sedang	0.499	0.898
	Populer	0.007	0.924
2	*Tidak Populer*	1.00	0.999
	Sedang	0	0.936
	Populer	0	0.944

Comparison of the results of the performance of the two algorithms in terms of average precision, average recall, and accuracy can be seen in FIGURE 10

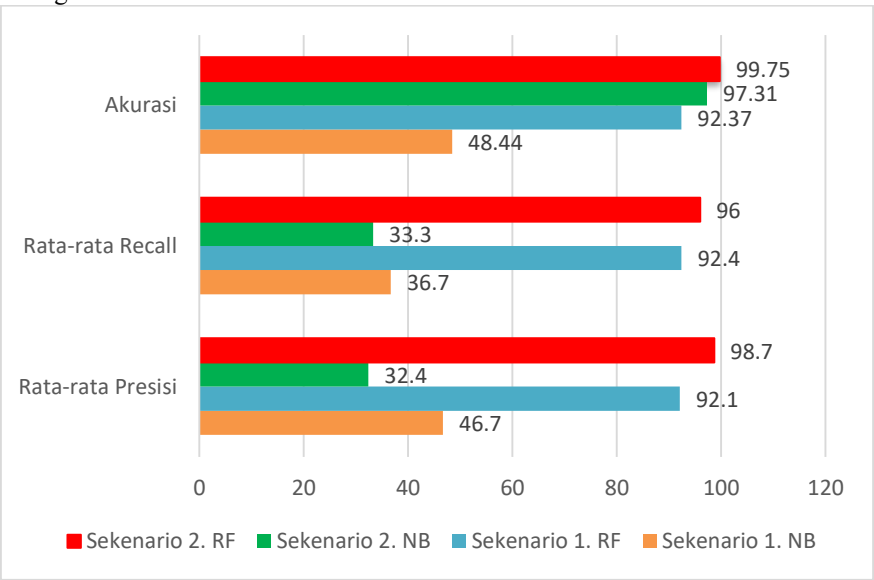

FIGURE 10. Comparison Graph of Algorithm Performance Results

FIGURE 10 shows the very large percentage difference in the accuracy of the Naïve Bayes and Random Forest algorithms. This difference in accuracy is not only influenced by the difference in the lagorithmic process flow, it is also influenced by differences in the clustering scenarios used. Where the proportion of data distribution in scenario 1 of the clustering process is balanced in each class. While the proportion of data distribution in scenario 2 is lame.

In scenario 1, the process of clustering with *populer* class contains data that has a maximum number of assigned and commented articles that are larger and the maximum number of lives smaller than *sedang* classes. *Sedang* class is containing data with a maximum value assigned and comments smaller and a minimum number of ages greater than *populer* class. *Tidak populer* class contain data with the number of ages approaching the date taken, where the maximum assigned value is greater than the *sedang* class but the maximum number of comments is smaller.

The *populer* class in scenario 2 of the clustering process contains the data with the highest average number of assigned and highest comments among other classes. *Sedang* class contains data with the maximum average number assigned and comments under the *populer* class, and the *tidak populer* class has the lowest maximum value among the other classes.

The results of the average precision and relatively balanced recall on the Random Forest algorithm show that the high accuracy values are true. This can be said because precision and recall are manifestations of the accuracy of data grouped in the correct class.

4. CONCLUSION

Based on the results of the study it can be concluded that to classify the popularity of news articles on learning materials into popularity class, the Random Forest algorithm has a better level of performance than the Naïve Bayes algorithm. This can be proven with a fairly good level of accuracy in the two clustering scenarios, namely 99.75% for scenario 2, and 92.37% for scenario 1 with a relatively balanced average precision and recall above 90%. Whereas the Naïve Bayes algorithm has an accuracy of 97.31% for scenario 2 and 48.44% for scenario 1 with an average precision and recall below 50%.

A good algorithm performance also proved that the number of words, the number of unique words, and lexile levels can be used as one of the characteristics of the level of popularity article, because based on the classification of three attributes have similar patterns on each grade.

REFERENCES

[1] B. Nurgiyanto, "Tahapan Perkembangan Anak dan Pemilihan Bahan Bacaan Sastra Anak," Cakrawala Pendidik., vol. XXIV, no. 2, pp. 197–222, 2005.
[2] F. P. Chew, "Developing children ' s literary resources," Educ. Res. Rev., vol. 7, pp. 155–168, 2012.
[3] K. Nation, "Children's Reading Comprehension Difficulties," pp. 248–266, 2001.
[4] J. Oakhill, "Children ' s Difficulties in Reading Comprehension," Educ. Psychol. Rev., vol. 5, no. 3, pp. 223–237, 1993.
[5] A. Balali, M. Asadpour, and H. Faili, "A Supervised Method to Predict the Popularity of News Articles," Comput. y Syst., vol. 21, no. 4, pp. 703–716, 2017.
[6] D. Liparas and Y. Hacohen-kerner, "News Articles Classification Using Random Forests and Weighted Multimodal Features News articles classification using Random Forests and weighted multimodal features," no. October, 2014.
[7] A. W. W. Wayan Firdaus Mahmudy, "Klasifikasi Artikel Berita Secara Otomatis Menggunakan Naive Bayes Classifier yang Dimodifikasi," TEKNO, vol. 21 Maret 2, pp. 1–10, 2014.
[8] Y. Keneshloo, N. Ramakrishnan, S. Wang, and E.-H. Han, "Predicting the Popularity of News Articles," no. June, 2016.
[9] R. Shreyas, D. M. Akshata, B. S. Mahanand, B. Shagun, and C. M. Abhishek, "Predicting Popularity of Online Articles using Random Forest Regression," Cogn. Comput. Inf. Process., 2016.
[10] T. Uddin, M. Jamshed, A. Patwary, T. Ahsan, and M. S. Alam, "Predicting the Popularity of Online News from Content Metadata," no. October, 2016.
[11] M. Khawaja and F. Chen, "Using Language Complexity to Measure Cognitive Load for Adaptive Interaction Design," no. March 2017, 2010.
[12] B. Rahmat, "Implemetasi k-means clustering pada rapidminer untuk analisis daerah rawan kecelakaan," Semin. Nas. Ris. Kuantitatif Terap. 2017, no. April, pp. 58–63, 2017.
[13] Bustami, "Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi," J. Inform., vol. 8, no. 1, pp. 884–898, 2014.
[14] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI/ICML-98 Work. Learn. Text Categ., pp. 41–48, 1998.
[15] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," Citec J., vol. 2, no. 3, pp. 207–217, 2015.
[16] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naive Bayes," Int. J. Adv. Comput. Math. Sci., vol. 3, no. 3, pp. 2230–9624, 2012.
[17] L. Breiman, "Random Forest," pp. 1–33, 2001.
[18] N. K. Dewi, U. D. Syafitri, S. Y. Mulyadi, M. D. Statistika, and D. Statistika, "Penerapan Metode Random Forest Dalam Driver Analysis," Forum Stat. dan Komputasi, vol. 16, no. 1, pp. 35–43, 2011.
[19] Y. S. Nugroho and N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," J. Tek. Elektro, vol. 9, no. 1, 2017.
[20] K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis."
[21] E. N. Azizah, U. Pujianto, and E. Nugraha, "Comparative performance between C4 . 5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment," 4th Int. Conf. Educ. Technol., no. 1, pp. 18–22, 2018.
[22] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," pp. 233–240, 2006.