JOIN (Jurnal Online Informatika)
Volume 5 No.1 | June 2020: 61-72
DOI: 10.15575/join.v5i1.505

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Analysis and Implementation Machine Learning for YouTube Data Classification by Comparing the Performance of Classification Algorithms

Riyan Amanda[1], Edi Surya Negara[2]
[1,2]Department of Informatics, Data Science Interdisciplinary Research Center, Bina Darma University

Article Info

Article history:

Received Feb 4, 2020
Revised Jun 21, 2020
Accepted June 24, 2020
Published July 15, 2020

Keywords:

Data Mining
Experimental Method
Machine Learning
YouTube Data Classification

ABSTRACT

Every day, people around the world upload 1.2 million videos to YouTube or more than 100 hours per minute, and this number is increasing. The condition of this continuous data will be useless if not utilized again. To dig up information on large-scale data, a technique called data mining can be a solution. One of the techniques in data mining is classification. For most YouTube users, when searching for video titles do not match the desired video category. Therefore, this research was conducted to classify YouTube data based on its search text. This article focuses on comparing three algorithms for the classification of YouTube data into the *Kesenian* and *Sains* category. Data collection in this study uses scraping techniques taken from the YouTube website in the form of links, titles, descriptions, and searches. The method used in this research is an experimental method by conducting data collection, data processing, proposed models, testing, and evaluating models. The models applied are Random Forest, SVM, Naive Bayes. The results showed that the accuracy rate of the random forest model was better by 0.004%, with the label encoder not being applied to the target class, and the label encoder had no effect on the accuracy of the classification models. The most appropriate model for YouTube data classification from data taken in this study is Naïve Bayes, with an accuracy rate of 88% and an average precision of 90%.

Corresponding Author:

Edi Surya Negara,
Department of Informatics,
Data Science Interdisciplinary Research Center,
Bina Darma University,
Jl. Jenderal Ahmad Yani No. 3 Palembang, Indonesia
Email: e.s.negara@binadarma.ac.id

1. INTRODUCTION

With internet innovation developing rapidly, all levels of society today use the internet in their daily lives with different activities and needs ranging from the ranks of society, businesses, communities, and even government. Dr. Suyanto in his book, quoted the arguments of John Gantz and David Reinsel in the IDC investigation which said that the volume of data in 2011 reached 1.8 zettabytes or 1.8 trillion gigabytes in 2012 it increased more than 50% to 2.8 zettabytes. In 2013 the volume of data had become 4.4 zettabytes and will continue to increase rapidly until it is estimated to reach 44 zettabytes in 2020 [1]. This gave birth to a new term called Big Data. The resulting data is unstructured, sometimes semi-structured, and also unexpected. This data is mostly generated in real-time from social media websites, which is increasing exponentially every day [2]. One of which is YouTube. More and more data are stored, so it can be said that there has been very large data buildup. To dig up information on large-scale data, a technique called data mining can be a solution.

Data mining is the process of finding interesting patterns/information on selected data using certain techniques or methods. In data mining methods, techniques and algorithms vary greatly. The method or algorithm chosen depends on the goal itself [5]. On the other hand, data mining is a set of activities that include gathering, using previous data to find rules, patterns, or relationships in large data sets. The output of data mining can be applied to improve future decision making [6]. Data mining is divided into several groups based on their objectives [7], namely Description, Estimation, Prediction, Classification, Clustering, and Association.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

By using data mining techniques, you can find information in the form of patterns, features, and rules known as knowledge. In the data mining process, there are several data processing methods. One of which is the classification method. For most YouTube users, when searching for video titles do not match the desired video category. Therefore, this research was conducted to classify YouTube data based on its search text.

Text classification or can also be called the classification of text when it is needed, with the emergence of the phenomenon of Big Data. There are two ways in the classification of text, namely text cluster and text classification. Text clusters relate to the discovery of an unsupervised group structure of documents [15]. Text classification is the process of grouping documents into different classes, in that each document has a certain class, so a process is needed to find information from the document. So that the document can represent from the class so that every word that has appeared in the document has a value [16]. There are several ways in text processing, among others: information retrieval, document classification, document clustering, etc.

Text classifications are usually applied to processed datasets because raw text data can contain high levels of noise, such as typos that are often observed on social media. The most commonly applied steps include tokenization, case transformation, stop-word removal, term weighting, stemming, and n-gram development. The aim is to eliminate all non-informative features, which do not contribute to the task of underlying text classification [17].

Fitri conducted a study by comparing several classification methods, namely Naive Bayes, Lazy-ibk, Zero-r, and Decision Tree-j48 [3]. The research focused on knowing the performance of the proposed model based on aspects of prediction accuracy, and speed/efficiency using the WEKA Version 3.7.7 application. The test results show that the Naive Bayes algorithm has the best accuracy of 85.12% in the cross-validation test mode, but the ZeroR algorithm has the best speed for all test modes, and all data sets in his research.

Saputra et al. conducted a sentiment analysis research using the Support Vector Machine and Naive Bayes method of President Jokowi's data in the form of comments taken from social media and political blogs [4]. In his research, comparing the results of the best level of accuracy with normalization and stemming. The results of this study show that the accuracy generated by the SVM method is not always superior to the Naive Bayes method, and vice versa. But for the highest accuracy method in each experiment is the SVM method with an accuracy of 89.2655%.

From the two studies, it can be seen that there are differences in the benefits of the classification models. Therefore, this study was conducted to determine the performance comparison of several models in the classification method so that it can be seen which model is most appropriate based on its level of accuracy by applying machine learning. Models to be compared in this study are the Random Forest, Support Vector Machine, and Naïve Bayes. The data used in this research is taken from the YouTube website, which consists of 4 (four) attributes, namely: links, titles, descriptions, and searches where the link is an id, and the search is the target class.

2. METHOD

This study uses an experimental method. This method consists of several stages, namely data collection, initial data processing, the proposed model, testing, and model evaluation. The flow of the research process can be seen in Figure 2.

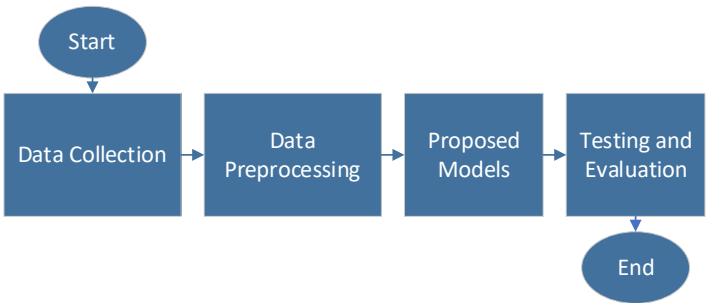

Figure 2. Research Process Flowchart.

2.1. Data Collection

The data to be used in this analysis is derived from YouTube. YouTube is a video sharing website which was founded in February 2005 by three former PayPal employees. This website lets users upload, access , and share photos. The business is based in San Bruno, California, and uses Adobe Flash Video and HTML5 technologies to view a wide range of video content/creators, including movie clips, TV clips, and music videos.

According to Shaila S.G on [2], everyday people around the world upload 1.2 million videos to Youtube, or more than 100 hours per minute, and this number is increasing.

The process of getting this data only uses python programming by browsing the YouTube website HTML documents. Data were taken information in the form of video links, video titles, and video descriptions. Before the scraping process is done, the author uses the search feature on YouTube with the keywords *Kesenian* and *Sains*. These keywords are needed as the target class in the classification. Basically, the keywords used do not have to be *Kesenian* and *Sains*; it is just that this research is related to science. In this study, the data taken was only in the form of videos based on the relevance of search keywords.

Data retrieval is done twice. In the first retrieval of data taken with the keyword *Kesenian* and the second retrieval is carried out with the keyword *Sains* which is then combined into one document. The results of data retrieval can be seen in Figure 3.

Out[13]:

	link	judul	deskripsi	kategori
0	7-WsnhON2	MENGENAL KESENIAN KENTRUNG TULUNGAGUNG,	Banyak di antara kita yang tidak lagi mengenal...	Kesenian
1	QVIhYXkedEQ	Super Minds 1: Unit 2 - The Go-kart Race Story	English for young learners - Super Minds 1\nTh...	Sains
2	Dge3AqV4r88	Kesenian Rampak Buto ~ GK ~ Sinar Siswo Magela...	Didalam Video Berisi Duet Gedruk Sinar Siswo M...	Kesenian
3	SpBTSVNjcL	Kesenian jaranan turonggo budoyo	#kukep	Kesenian
4	nmky1J0Prf4	Pembelajaran Sains SMP Al-Falah Kalibata City ...	Dokumentasi Dan Kegiatan Ini, Terselenggara At...	Sains
...	...	...	...	...
1265	q-06Dsahf_M	HUT Kesenian Jaranan Sarpo Budoyo – Curah Nong...	Warna Warni program hiburan yang disuguhkan de...	Kesenian
1266	2VVVcP6GG9A	ILK - KESENIAN TURUN KE JALAN (23/5/16) 5-1	INDONESIA LAWAK KLUB\n----------...	Kesenian
1267	JnK41CnHHI	SAINS : Tahun 2 - Haiwan 2	Laman web video kurikulum : http://eduwebtv.mo...	Sains
1268	TILbDOw5Psg	Kesenian Dayakan Topeng ireng, Desa wonogiri k...	Kesenian tradisional kab magelang #keseniantra...	Kesenian
1269	LzNuyvZf62Q	Betawi Art: Gambang Kromong Bini Dua song, (Pe...	Subscribe please	Kesenian

1270 rows × 4 columns

Figure 3. Raw Data

In the above of figure 3, it can be seen that as many as 1270 data have been retrieved, this data is still in the form of raw data that must be cleaned.

2.2. Data Preprocessing

At this stage, the data will be cleansed using the NLTK (Natural Language Toolkit) library by removing all punctuation, changing all letters to lowercase, only storing basic words in the title and description attributes and deleting words that are considered to have no meaning. This process is called Preprocessing.

2.3. Proposed Models

At this stage, the proposed model will be applied using Machine Learning with the python programming language. The performance of models to be tested is Random Forest, Support Vector Machine (SVM), and Naïve Bayes. These models will be tested in the ability of the model to recognize positive tuples and negative tuples, so that the accuracy of the model is obtained.

2.3.1 Random Forest

Random Forest is a classification category composed of many decision trees. Each decision tree is built from random vectors. The general approach used for inserting random vectors in tree formation is to select a random F value, such as the input of the F attribute (feature) to be shared at each node in the decision tree to be created. Simply look at the selected F attribute by selecting a random value F then it does not have to search all the available attributes. The parameter used to set the intensity of the random forest when the F value was selected and the number of trees to be built. If the value of F is too small, then the tree has a tendency to have a very small correlation, and it applies equally to the opposite [8]. Therefore, the value of F can be determined by the formula:

$$F = Log2(M + 1) \quad (1)$$

In the above formula, M is the total number of features. In addition to the selection of attributes, it is also done with randomization when selecting training sets.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Random Forest (RF) grouping has been surprisingly successful in various automatic classifications of classification tasks. It has been considered by many to be an algorithm that is monitored top-notch, comparable, and sometimes superior to the Support Vector Machine (SVM) classifier [9].

2.3.2 Support Vector Machine

Support Vector Machine (SVM) is a type of non-linear problem method in low dimensional space mapped to high dimensional space so that simple linear classification techniques can be handled according to small sample learning. SVM can effectively overcome the problems of traditional over-fitting methods and neural network learning problems that are commonly seen at a local minimum so that they have strong generalization capabilities [10]. However, not all data can be separated linearly in two dimensions. Therefore, the linear limiting function is then transformed into hyperplanes by using the kernel function so that the hyperplanes can separate data in higher dimensional spaces [11].
In SVM, the separator function aims to determine the class. Supporting separator fields from class +1 and supporting separator fields from class −1. In this case, the separating function sought is a linear function as follows [12].

$$F(x) = sign(wT\ xi + b = 0) \quad (2)$$

With the value W is the weight representing the hyperplane position in the normal plane, X is the input data vector. B is the bias representing the position of the plane relative to the center of the coordinates.

3.3.3 Naïve Bayes

The Naïve Bayes Classifier method has now been developed to calculate the probability measure of each word and provide an assessment for each class [13]. This method is a classification with probability and statistical methods proposed by the British scientist Thomas Bayes, which predicts future opportunities based on past experience so that it is known as Bayes theorem. The theorem is assumed to have independent attributes. The basic theory used in carrying out this classification is the Bayes theorem.

$$P(H|X) = (P(X|H) * P(H))/(P(X)) \quad (3)$$

Where the value of X is unknown class data, H is the hypothesis X on a particular label, P (H | X) is the probability of H based on X (posteriori) conditions, P (H) is the probability of H (prior), P (X | H) is the probability of X in the hypothesis H, P (X) is the probability of X.

3.4. Testing and Evaluation

At this stage, the models that have been applied will be compared using the Confusion Matrix table. A confusion matrix is a table consisting of the number of rows of test data that are predicted to be true and incorrect by the classification model. This table is needed to measure the performance of a classification model [6]. The confusion matrix can be seen in table 1.

Table 1. Confusion Matrix

		Prediction Class		
		Class 1	Class 0	
Actual Class	Class 1	TP	FN	P
	Class 0	FP	TN	N
		P’	N’	P+N

Where True Positives (TP) is the number of positive tuples that are correctly labelled by the classification model. True Negatives (TN) is the number of negative tuples that are correctly labelled by the classification model. False Positives (FP) is the number of negative tuples wrongly labelled by the classification model. False Negatives (FN) is a positive tuple wrongly labelled by the classification model. P' is the number of tuples labelled positive, N' is the number of tuples labelled negative.

3. RESULTS AND DISCUSSION

3.1. Preprocessing

The data preprocessing steps has done with several stages, such as Case Folding, Stemming, and Stopword Removal. This process aims to transform data for the better so that data is not much noise, which

can affect the level of accuracy in classification. In the case of case-folding, the data will be transformed into a lowercase, delete other than letters in the sentence, delete spaces at the beginning and end of the sentence, and divide the sentence into words. In the process of stemming and stopword removal using the Porter Stemming Algorithm library proposed by Martin Porter. This process is carried out twice, the first is based on Indonesian, and the second is based on English because the data taken also includes English. Preprocessing is done on the title and description attributes. The results of preprocessing can be seen in Figure 4.

Out[11]:

	link	judul	deskripsi	kategori
0	7-WsnhON2	mengen kesenian kentrung tulungagung	mengen kentrung salah kesenian dimainkan grup ...	Kesenian
1	QVIhYXkedEQ	super mind unit go kart race stori	english young learner super mind go kart race ...	Sains
2	Dge3AqV4r88	kesenian rampak buto gk sinar siswo magelang l...	didalam video berisi duet gedruk sinar siswo m...	Kesenian
3	SpBTSVNjcL	kesenian jaranan turonggo budoyo	kukep	Kesenian
4	nmky1J0Prf4	pembelajaran sain smp al falah kalibata citi l...	dokumentasi kegiatan terselenggara kerjasama p...	Sains
...	...	...	...	...
1265	q-06Dsahf_M	hut kesenian jaranan sarpo budoyo curah nongko	warna warni program hiburan disuguhkan format ...	Kesenian
1266	2VVVcP6GG9A	ilk kesenian turun jalan	indonesia lawak klub tayang senin selasa wib p...	Kesenian
1267	JnK41CnHHI	sain haiwan	laman web video kurikulum http eduwebtv moe ed...	Sains
1268	TILbDOw5Psg	kesenian dayakan topeng ireng desa wonogiri ke...	kesenian tradis kab magelang keseniantradis ma...	Kesenian
1269	LzNuyvZf62Q	betawi art gambang kromong bini song pecenonga...	subscrib plea	Kesenian

1270 rows × 4 columns

Figure 4. Data after preprocessing

From figure 4, it can be seen that the data is clear of punctuation, numbers, spaces in the beginning and end of sentences, emote, and all letters have become lowercase so that the data is ready to be classified.

3.2. Classification

Before the classification of data, the research has implemented to make Bags of Words. It is intended that the classification model understands keywords against the data applied. In these Bags of Words, each sentence in the document is described as a token, ignoring grammar and even word order but calculating the frequency of occurrences or words appearing from the document. Furthermore, the learning process will be carried out by dividing the data into two parts, namely, training data and testing data. This study will produce a model. Model results from learning training data will then be tested with data testing. This data sharing is done with the library of sklearn that is split_test_train with test data of 20%. This study conducted two experiments, first experiment was conducted by comparing the accuracy of the classification model without transforming the class as numeric or called as label encoder. The second experiment compared the level of accuracy by transforming the class into numeric. The label encoder transforms classes into numerics with values from 0 to n where n is the number of different classes. The number of classes in this study is two, namely *Kesenian* and *Sains*. It means the class has been transformed into 0 for *Kesenian* and 1 for *Sains*.

3.2.1. Random Forest

The first experiment carried out by not transforming the class into numerical obtained an accuracy level of 0.8228 or rounded up to 82%. The evaluation results can be seen in the next figure 5.

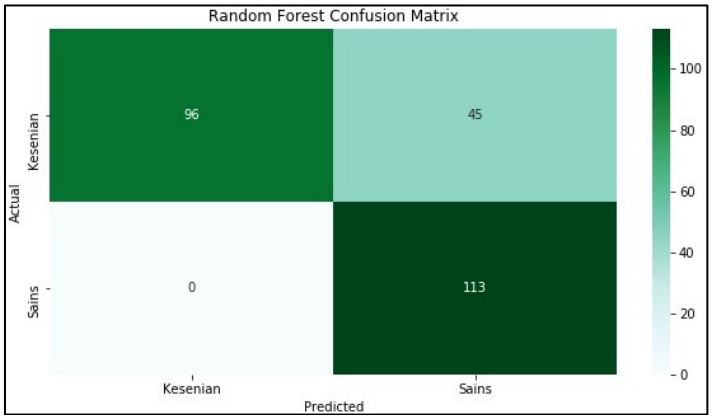

Figure 5. Confusion Matrix Random Forest first experiment table

p-ISSN: 2528-1682
e-ISSN: 2527-9165

In the picture above, it can be seen that the model successfully labels correctly 96 data for all *Kesenian* classes. And in the *Sains* class, this model successfully labels 113 *Sains* class data correctly and 45 labels incorrectly. For further evaluation, see Figure 6.

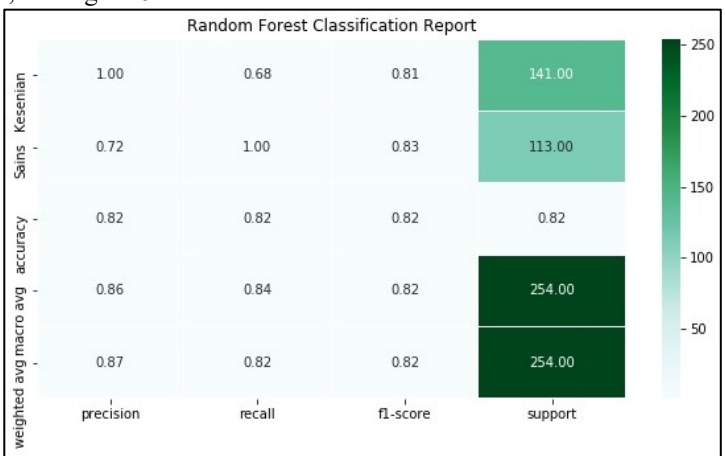

Figure 6. Random Forest Classification Report first experiment

In the picture above, it can be seen that this model produces 100% precision and 68% recall of the *Kesenian* class. For the *Sains* class, this model produces 72% precision and 100% recall. The second experiment carried out by transforming classes into numerical results, obtained an accuracy level of 0.8188 or rounded to 82%. The evaluation results can be seen in Figure 7.

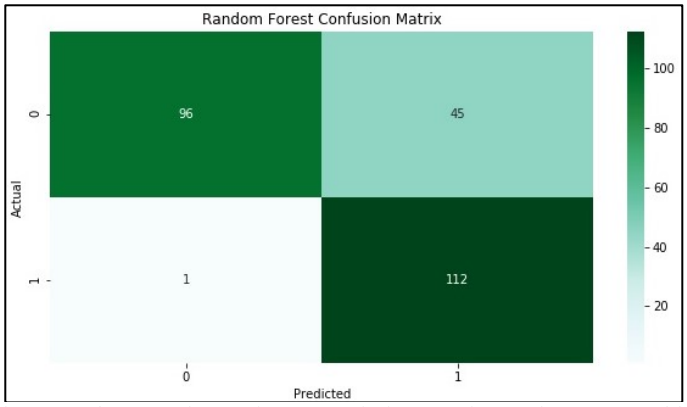

Figure 7. The second Experiment Confusion Matrix Random Forest table

In figure 7 shows that there are differences in the results of the first experiment where the first experiment of the model successfully labelled the *Sains* class correctly at 113 data, but in this second experiment the model incorrectly labelled 1 data correctly as *Kesenian*. For further evaluation, it can be seen in figure 8.

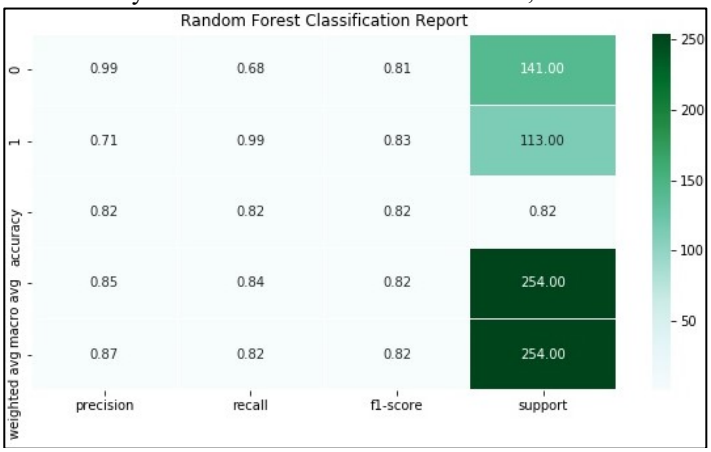

Figure 8. Random Forest Classification Report second experiment

In the picture above, it can be seen that this model produces 99% precision and 68% recall of the *Kesenian* class. For the *Sains* class, this model produces 71% precision and 99% recall.

From the two experiments that have been carried out, the level of accuracy in the first experiment was obtained by 0.8228 and in the second experiment was obtained by 0.8188. The difference in the level of accuracy obtained is only 0.004%, but if rounded together, have an accuracy rate of 82%. This shows that whether or not the label encoder is applied does not affect the performance of the random forest classification model.

3.2.2. Support Vector Machine

The first experiment carried out by not transforming the class into numerical obtained an accuracy level of 0.8228 or rounded up to 82%. The evaluation results can be seen in Figure 9.

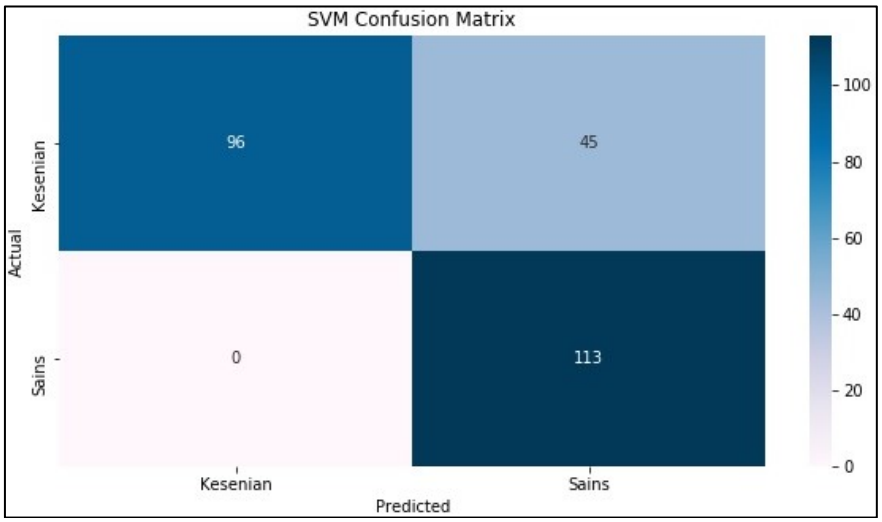

Figure 9. First experiment SVM confusion matrix table

In the above figure 9, it describes that the model successfully labels correctly 96 data for all *Kesenian* classes. And in the *Sains* class, this model successfully labels 113 *Sains* class data correctly and 45 labels incorrectly. For further evaluation, see Figure 10.

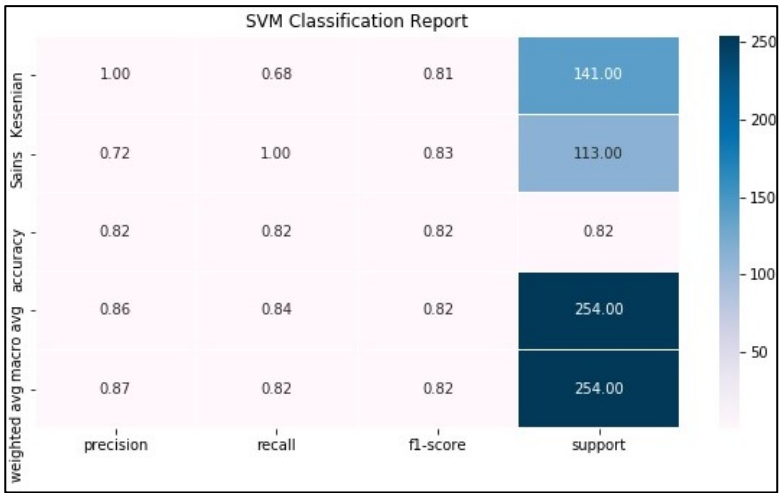

Figure 10. SVM Classification Report first experiment

In the picture above, it can be seen that this model produces 100% precision and 68% recall of the *Kesenian* class. For the *Sains* class, this model produces 72% precision and 100% recall. The second experiment carried out by transforming classes into numerical results obtained an accuracy level of 0.8228 or rounded up to 82%. The evaluation results can be seen in Figure 11.

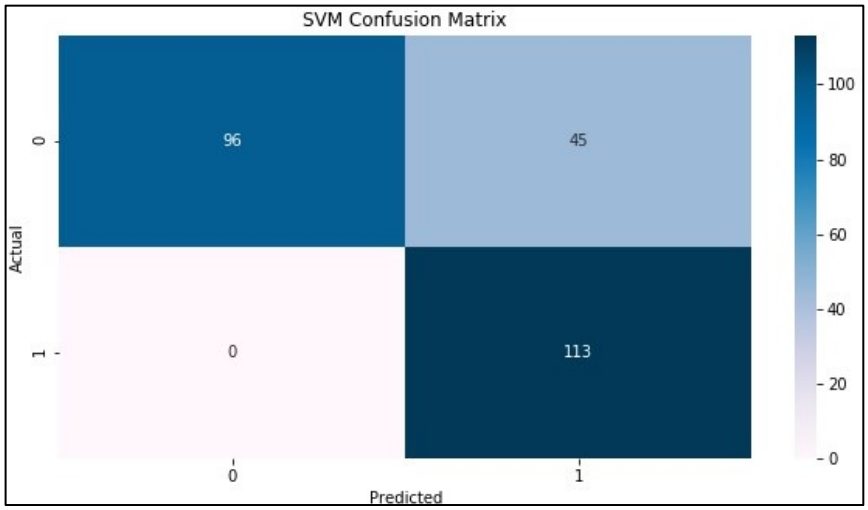

Figure 11. The second experiment table Confusion Matrix SVM

In the picture above, it can be seen that there is no difference in the performance of the SVM classification model by applying the label encoder. For further evaluation, see Figure 11

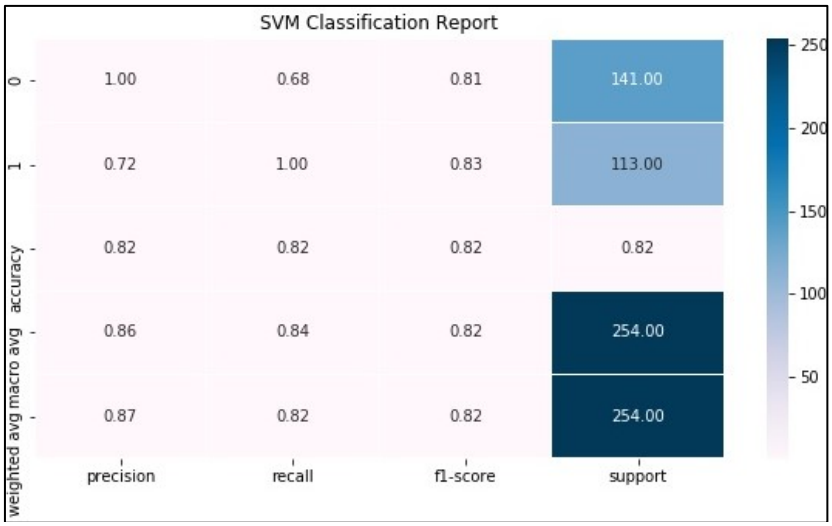

Figure 11. SVM Classification Report second experiment

In the picture above, it can be seen that the results produced by this model are the same as the results of the first experiment, namely 100% precision and 68% recall of the *Kesenian* class. For the precision of *Sains* class by 72% and recall by 100%. From the experiment that has been done shows that applying the label encoder has no effect on the performance of the SVM model.

3.2.3. Naïve Bayes

The first experiment was carried out by not transforming the class into numerical and the accuracy rate was 0.8779 or rounded up to 88% Evaluation results can be seen in Figure 12. It can be seen that the model successfully labeled 141 data for the *Kesenian* class but incorrectly labeled 31 data as a *Sains* class. And in the *Sains* class, this model correctly labels 82 data across all *Sains* classes. For further evaluation, see Figure 13.

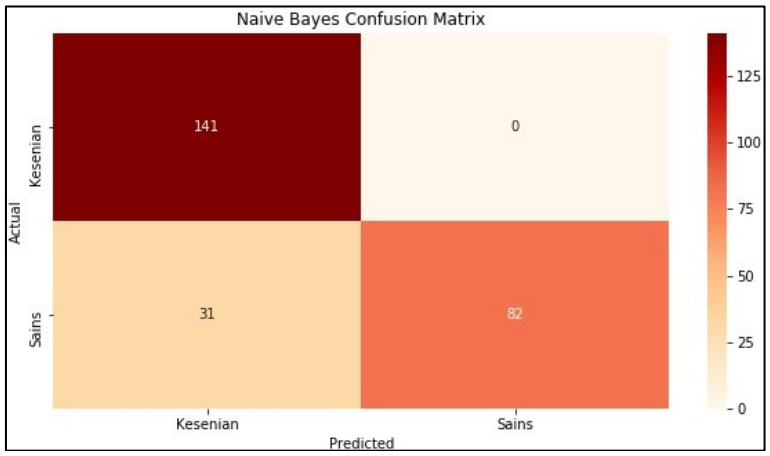

Figure 12. The first Naive Bayes Confusion Matrix experiment table

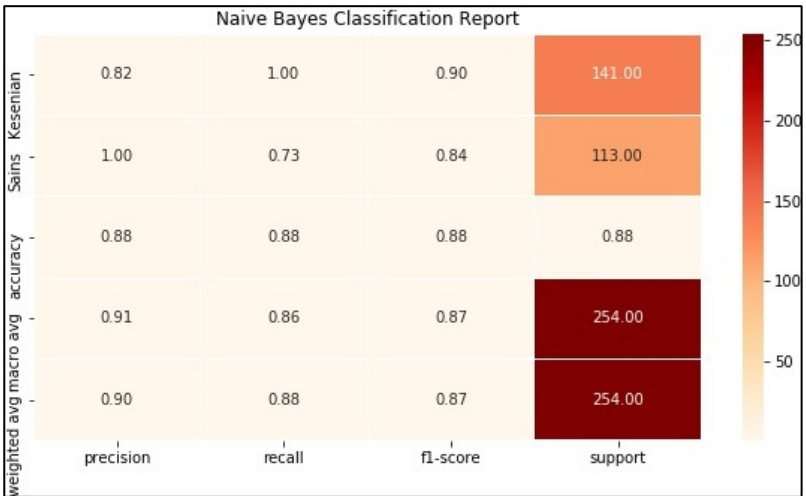

Figure 13. Naive Bayes Classification Report first experiment

In the picture above, it can be seen that this model produces 82% precision and 100% recall of the *Kesenian* class. For the *Sains* class, this model produces 100% precision and 73% recall. The second experiment carried out by transforming classes into numerical results obtained an accuracy rate of 0.8779 or rounded to 88%. Evaluation results can be seen in Figure 14.

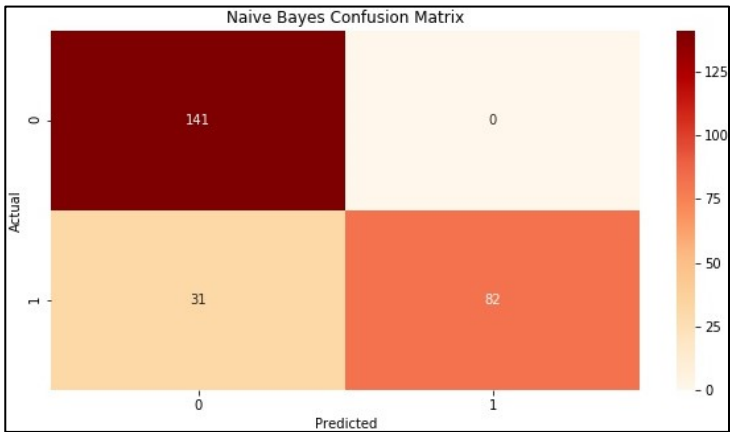

Figure 14. The second Naive Bayes Confusion Matrix experiment table

In the picture above it can be seen that there is no difference in the performance of the Naïve Bayes classification model by applying the label encoder. For further evaluation, see Figure 15.

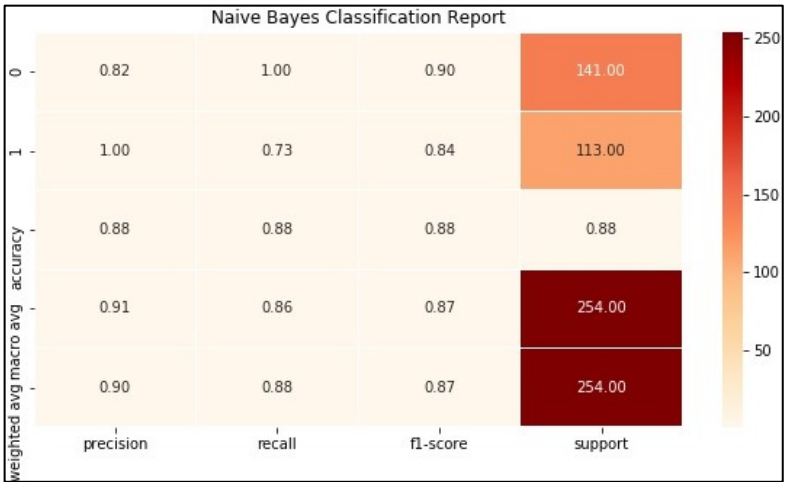

Figure 15. Naive Bayes Classification Report second experiment

In the picture above, it can be seen the results produced by this model are the same as the results in the first experiment, namely 82% precision and 100% recall of the *Kesenian* class. For the precision of the *Sains* class by 100% and recall by 73%. The experiment that has been done shows that the label encoder is applied has no effect on the performance of the Naive Bayes model.

3.3. Comparative Evaluation

After conducting the first and second experiments, the results of the comparison of the performance of classification models applied to YouTube data are obtained. Comparison of the performance results of the random forest classification model, SVM, Naive Bayes by applying the label encoder or not applying the label encoder can be seen in table 2.

Table 2. Model Performance Comparison

Model	Label Encoder	Precision (%)	Recall (%)	Acuracy	Rounded Acuracy (%)
Random Forest	Yes	87	82	0.8188	82
	No	87	82	0.8228	82
SVM	Yes	87	82	0.8228	82
	No	87	82	0.8228	82
Naïve Bayes	Yes	90	88	0.8779	88
	No	90	88	0.8779	88

In the table2, it can be seen that there is a difference in accuracy when the label encoder is applied against the random forest classification model. This accuracy difference is 0.004%, but if rounded off, this model has the same accuracy level of 82% with not applying the label encoder. The table above shows that the overall performance of the random forest classification model and SVM has the same result of 82%, although there is a slight difference in the level of accuracy before rounding. However, the SVM model is superior to the random forest model if the label encoder is applied.

Of the several classification models above, the highest accuracy level produced by the classification model is Naïve Bayes with an accuracy rate of 88% with an average precision of 90% and a recall of 88%. The Naïve Bayes model is better in terms of Accuracy and Sensitivity and Specificity. Sensitivity is the ability of the classification model to recognize positive tuples, and specificity is the ability of the classification model to recognize negative tuples [1]. This is evidenced in the results that have been presented in the previous sub-chapter classification. The results of the comparison can be seen in Figure 16.

Another measure that can be used to prove that the Naïve Bayes model is better is the F1 Score, with a result of 87%. This measure is basically a harmonic mean of precision and recall [1]. Figure 16 shows that the most appropriate model in the analysis and implementation of machine learning for YouTube data classification for data that was successfully obtained in this study is Naïve Bayes.

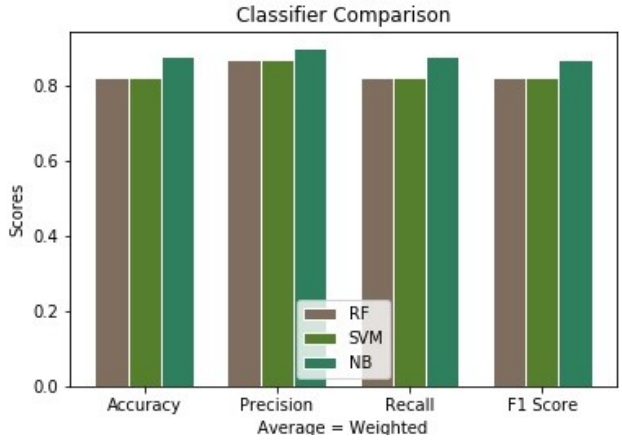

Figure 16. Comparison of Classification Models

4. CONCLUSION

This research was successfully carried out by comparing the performance of several classification methods. The conclusions that can be drawn from this research are the performance of the random forest classification model is better if the label encoder is not applied to the target class. From the results, the performance of the random forest and SVM models gives the same accuracy level of 82%. The label encoder for the target class does not affect the performance of the classification models. Last, the most appropriate model for YouTube data by using Naïve Bayes gives an accuracy rate of 88%, an average precision of 90% and a recall of 88%.

There is some improvement which can implement for further research. The research can improve with using stopwords and stemming of *Sastrawi* in the preprocessing process because there are have more differences between Indonesian and English. Another improvement, the study can be implemented with another classification model to compare levels of accuracy.

5. REFERENCES

[1] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, Edisi Revisi. Bandung: Informatika Bandung, 2019.
[2] Shaila S.G, Prasanna MSM, and K. Mohit, "Classification of YouTube Data based on Sentiment Analysis," *Int. J. Eng. Res. Comput. Sci. Eng. IJERCSE*, vol. 5, no. 6, Art. no. 6, Jun. 2018.
[3] S. Fitri, "Perbandingan Kinerja Algoritma Klasifikasi Naive Bayesan, Lazy-IBK, Zero-R, dan Decision Tree-J48," *J. DASI*, vol. 15, no. 1, Art. no. 1, Apr. 2014.
[4] N. Saputra, T. B. Adji, and A. E. Permanasari, "Analisis Sentimen Data Presiden Jokowi Dengan Preprocessing Normalisasi Dan Stemming Menggunakan Metode Naive Bayes dan SVM," *J. Din. Inform.*, vol. 5, no. 1, Art. no. 1, Nov. 2015.
[5] Y. Mardi, "Data Mining: Klasifikasi Menggunakan Algoritma C4.5," *J. Edik Inform.*, vol. 2, no. 2, Art. no. 2, 2017.
[6] L. Swastina, "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa," *J. GEMA Aktual.*, vol. 2, no. 1, Art. no. 1, Jun. 2013.
[7] D. M. A. Budanis and F. Slamat, "Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree," *J. IPTEK*, vol. 16, no. 1, Art. no. 1, May 2012.
[8] Mambang and A. Byna, "Analisis Perbandingan Algoritma C.45, Random Forest Dengan CHAID Decision Tree Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil," *Semin. Nas. Teknol. Inf. Dan Multimed. 2017*, p. 6, Feb. 2017.
[9] T. Salles, M. Gonçalves, V. Rodrigues, and L. Rocha, "Improving random forests by neighborhood projection for effective text classification," *Inf. Syst.*, vol. 77, pp. 1–21, Sep. 2018, doi: 10.1016/j.is.2018.05.006.
[10] L. Cunhe and W. Chenggang, "A new semi-supervised support vector machine learning algorithm based on active learning," in *2010 2nd International Conference on Future Computer and Communication*, Wuhan, China, 2010, pp. V3-638-V3-641, doi: 10.1109/ICFCC.2010.5497471.
[11] M. Hofmann, "Support Vector Machines — Kernels and the Kernel Trick," *Notes*, vol. 26, no. 3, Art. no. 3, Jun. 2006.
[12] C. Manning, P. Raghavan, and H. Schuetze, "Introduction to Information Retrieval," p. 581, Apr. 2009.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

[13] T. Sutabri, A. Suryatno, D. Setiadi, and E. S. Negara, "Improving Naïve Bayes in Sentiment Analysis for Hotel Industry in Indonesia," in *2018 Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, Oct. 2018, pp. 1–6, doi: 10.1109/IAC.2018.8780444.
[14] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," *Future Gener. Comput. Syst.*, vol. 101, pp. 341–371, Dec. 2019, doi: 10.1016/j.future.2019.06.022.
[15] C. Darujati and A. B. Gumelar, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia," *J. Link*, vol. 16, no. 1, Art. no. 1, Feb. 2012.
[16] I. Destuardi and S. Sumpeno, "Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes," *Semin. Nas. Pascasarj. IX – ITS*, p. 5, Dec. 2009.
[17] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, vol. 36, no. 1, Art. no. 1, Mar. 2019, doi: 10.1016/j.ijresmar.2018.09.009.