Comparison of Feature Selection for Naïve Bayes Classification Method in A Case Study of The Coronavirus Lockdown

Adian Fatchur Rochim
Departement of Computer Engineering,
Faculty of Engineering
Diponegoro University
Semarang 50275, Indonesia
adian@ce.undip.ac.id

Ratri Kusumastuti
Departement of Computer Engineering,
Faculty of Engineering
Diponegoro University
Semarang 50275, Indonesia
ratrik@student.ce.undip.ac.id

Ike Pertiwi Windasari
Departement of Computer Engineering,
Faculty of Engineering
Diponegoro University
Semarang 50275, Indonesia
ikepertiwi@gmail.com

***Abstract*—Classification in sentiment analysis often involves less relevant features for the modeling process. This causes the accuracy obtained to be not optimal. Therefore, a feature selection method is needed to sort out features that have high relevance to the dataset. This research aims to compare the accuracy between three different methods. They are Naïve Bayes Classification without using any feature selection, using Information Gain, and using Chi-Square feature selection. The datasets used are sentiments related to the lockdown as a policy for the Coronavirus pandemic from Twitter. Feature selection methods affected the accuracy by filtering features and sorting the most relevant features based on its algorithm. The results showed that the average accuracy of Naïve Bayes without feature selection, using Information Gain, using Chi-Square based on both Indonesian and English datasets were 63.2%, 64.2%, and 65%, respectively.**

Keywords— sentiment analysis, naïve bayes classification, feature selection, information gain, chi-square

I. Introduction

Coronavirus disease (COVID-19) that began in December 2019 has become a global pandemic. By the middle of 2021, a surge of the third wave of COVID-19 was happening. There were 178,868,783 confirmed cases including 3,880,649 deaths per June 23rd 2021 [1]. In Indonesia, there were 15,308 new cases and 303 new deaths on June 23rd 2021.

As the impact of increased number of cases, the government has imposed an emergency restriction toward community activities policy to suppress the spread of virus. The terms of policy in limiting community activities vary, depending on each country and region. Some advocate social distancing, and some partially or totally limit community activities, which became known as the term lockdown. In Indonesia, terms of the restrictions are well known as Large-Scale Social Restrictions (LSSR/PSBB), Enforcement of Restrictions on Community Activities (ERCA/PPKM), and lockdown itself.

The policies reaped the pros and cons in various circles of society around the world. Public sentiments of pros and cons are expressed in various kinds of social media, especially Twitter as a popular text-based social media. Therefore, the diversity of opinions regarding the coronavirus lockdown can be used as a case study for this sentiment analysis research.

Sentiment analysis is a process of finding user opinions on certain topics or texts under consideration or commonly known as opinion mining. Sentiment analysis is carried out to classify whether an opinion is included in a negative or positive opinion [2]. In machine learning, the sentiment analysis process can be carried out with various classification methods.

In general, the classification uses all the features contained in the data to build a model, even though not all of these features are relevant to the classification results, thus feature selection needs to be added [3]. This is because of the amount of all features from a large dataset could degrades the classification accuracy of the built model [4]. In addition, adding feature selection to the classification process can affect the accuracy. Classification methods with feature selection will generally produce better accuracy [5] [6]. But it is also difficult to simultaneously reduce the number of features and maintain classification accuracy [7].

Based on the problems above, this study is conducted to examine how feature selection affects accuracy by comparing Naïve Bayes Classification without feature selection, using Information Gain and using Chi-Square feature selection. This aims to find out the best feature selection in making a sentiment classification related to the coronavirus lockdown. The best model that has been obtained is used to analyze sentiment towards the lockdown policy that occurred in Indonesia.

This paper divides into five sections. First section is introduction, second is literature review, third is research methodology, forth is results and discussion and fifth is conclusion.

II. Literature Review

This research is conducted based on several previous studies with related methods. Pratama, et al. (2018) examined the comparison of accuracy between Naïve Bayes with Chi-Square feature selection and Naïve Bayes without feature selection. In their research, the accuracy of Naïve Bayes with Chi-Square was higher than Naïve Bayes without feature selection [8].

Other than that, Sari (2016) examined sentiment analysis classification using Naïve Bayes without feature selection and Naïve Bayes with Information Gain. Her research showed that the implementation of feature selection techniques could increase the level of accuracy. It was because features that were not relevant to the classification target were reduced. The Information Gain feature selection technique completed by selecting ten features on the top ranking showed the best results in her study [9].

Sofiana et al., (2012) compared Information Gain feature selection with Chi-Square using Naïve Bayes classifier. Her research showed that using Chi-Square feature selection with 20% features selected of the overall features gave the best accuracy, while Information Gain gave the best accuracy at 80% features selected of the overall features. This case made Chi-Square better than Information Gain because 20% features selected would take less time to calculate than 80% [10].

Also, in 2021 Subagio examined the comparison between Information Gain and Chi-Square feature selections on movie genre classification using Naïve Bayes classifier. His result showed that Chi-Square gave the best accuracy. Separation of training and testing data also affected accuracy. The bigger the training data, the better the accuracy [11].

Meanwhile, in 2018, Soemantri, et al. examined Support Vector Machine classification using Information Gain and Chi-Square feature selections of restaurant service sentiment analysis. The result showed that using Information Gain gave the higher accuracy than Chi-Square [12].

III. Research Methodology

Based on considerations regarding Information Gain and Chi-Square feature selections that could increase accuracy for sentiment analysis classification, this study aims to compare Information Gain and Chi-Square feature selection in getting the best accuracy using Naïve Bayes classifier. Different from other research before, datasets used in this case are lockdown sentiments as a policy of the coronavirus pandemic in Indonesian and English languages. This research uses two different datasets intended as variations of modelling scenarios to obtain accuracy from different feature selection methods, which is then taken the average of the two scenarios.

In this research methodology, section the sequence of research steps, datasets, and methods that were used for classification will be explained.

A. Data Collection

Data of public sentiments regarding the COVID-19 lockdown from Twitter in this study were taken from March 1st, 2020 to May 1st, 2020 and from January 1st, 2021 to March 1st, 2021 by a scraping technique using Jupyter Notebook with Python programming language. The data taken were in Indonesian and English of 1200 tweets each, then grouped into two datasets. Each dataset was labeled into positive and negative sentiments of lockdown, PSBB, and PPKM. Table 1 and 2 are example of labeled data from original tweets.

TABLE 1. EXAMPLE OF LABELED TWEET IN INDONESIAN DATASET

Tweet	Label
Jadi terngiang opsi lockdown/karantina wilayah di kondisi seperti ini. Andai saja semua siap untuk itu, niscaya penyebaran virus dapat ditekan.	1
Jadi gini loh..kita yang punya warung itu bangkrut karena PSBB dan sejenisnya itu yang batasin jam buka sampe jam 7 malem doang..lah dikira corona itu jam kerjanya cuman malam doang?	0
Psbb dan ppkm sepertinya ga efektif PERCUMA diperpanjang juga. Krn hanya membatasan waktu saja sedang pergerakan serta hilir mudik manusia tak diperhatikan. Penerapannya pun hanya getol di awal loyo setelahnya.	0
Positif COVID-19 di Jakarta Melambat, Fahira Idris Minta Formulasi PSBB Diperkuat. â€œJika kasus positif Covid-19 di Jakarta melambat artinya kita harus semakin disiplin. Ini agar ke depan benar-benar tidak lagi ditemukan kasus baru". https://t.co/YVhzPpXJI0	1

TABLE 2. EXAMPLE OF LABELED TWEET IN ENGLISH DATASET

Tweet	Label
The fact that I think Lockdown is idiotic and retarded policy amounting to an admission of failure and intellectual bankruptcy doesn't mean I don't think COVID-19 is serious, but that's just par for the course in the false-choice binary thinking galaxy-brains have foisted on us.	0
@HelenBranswell Lockdown goal was to avoid swamping healthcare. Any state/country can choose to loosen up, while watching healthcare capacity closely. Loosening will likely cause uptick in cases. And states may need to cycle thru loosening &amp; tightening. But we donâ€™t need one-size-fits-all.	1
Only thing ima do during this lockdown is learn new drills and improve daily on being a better coach !	1
This lockdown got my sleep schedule all the way messed up	0

B. Flowchart

The steps of this research from the beginning until classification result and analysis could be seen in Fig.1.

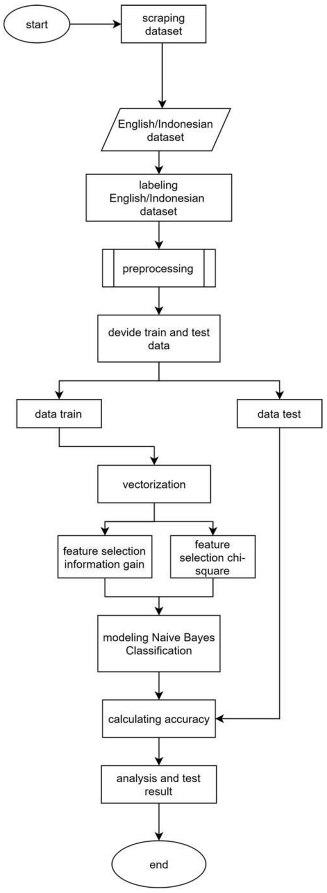
start
scraping dataset
English/Indonesian dataset
labeling English/Indonesian dataset
preprocessing
devide train and test data
data train
data test
vectorization
feature selection information gain
feature selection chi-square
modeling Naive Bayes Classification
calculating accuracy
analysis and test result
end

Fig. 1. Research methodology of classification comparation study

C. Preprocessing

Before classifying the datasets, preprocessing was done to avoid less perfect data, data interruptions, and less consistent data [13]. Datasets that were used in this study were Indonesian and English datasets. Each dataset contained 1200 data and was clustered by being labeled manually into positive or negative sentiment. The distribution of dataset labels can be seen in Fig. 2 and Fig. 3.

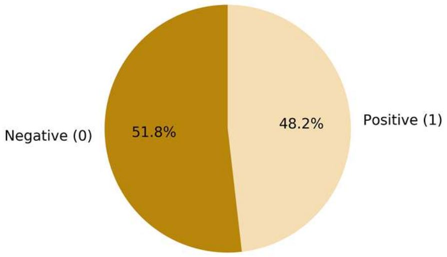

Fig. 2. The distribution of Indonesian dataset label

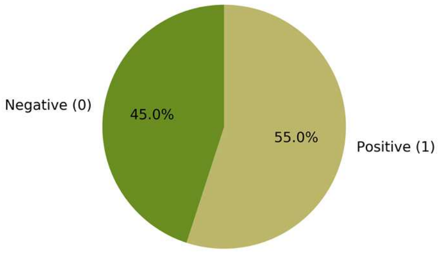

Fig. 3. The distribution of English dataset label

The label given were positive and negative only. In this research, neutral sentiments were grouped into positive label. Manual labeling was performed by one author's perspective on both datasets. The results of the distribution of labels in the Indonesian dataset showed more negative sentiments, on the other hand, the English dataset showed more positive sentiments.

It seems that restrictions policy in Indonesia less acceptable in society. However, the regional distribution of the English dataset cannot be determined because the data were taken randomly based on English language only.

After being labeled, each dataset was preprocessed to remove unnecessary characters and symbols. The steps of preprocessing are as follows.

- Case folding to make letters all lowercased.
- Dataset cleaning for removing URLs, mentions, punctuation, white spaces, numbers, and symbols.
- Normalization to replace non-standard words into standard words.
- Tokenization is for breaking up a sequence of strings into pieces such as words.
- Stop word elimination to clean the low-level information from the data.
- Stemming to remove suffixes and prefixes to get the root words.

D. Information Gain Feature Selection

Information Gain is a feature selection method that is used to determine the limits of the importance of an attribute. The Information Gain value is obtained from the entropy value before the separation minus the entropy value after the separation. This value measurement is only used as an initial stage for determining features that will later be used or discarded [14]. The calculation of the Information Gain value can be seen in the following equation below. Equation (1) is an entropy value before the separation, on other words, an entropy of all data, whereas (2) is an entropy value after the separation [15].

$$Entropy\ (S) = \sum_i^c -p_i \log_2 pi \tag{1}$$

$$Entropy_A(S) = \sum_{v\in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

S = number of data
c = number of classes
i = class i
pi = ratio between the number of samples in class i to the number of all the entire data
A = Attribute
v = A possible value for the attribute A
Sv = number of samples for the value of v

To calculate the gain value of the features, the entropy value before the separation (Entropy(S)) is subtracted by the entropy value after the separation (Entropy$_A$(S)) as can be seen in (3).

$$Gain\ (S, A) = Entropy(S) - Entropy_A(S) \tag{3}$$

In this research, the datasets were weighted using the Information Gain feature selection. The entropy of the entire datasets were calculated using (1). Then, the entropy value of each feature was calculated using (2). Information Gain value of each feature was then calculated using (3) by subtracting the entropy value of the entire set by the entropy value of each feature. In other words, Information Gain of a term was measured by counting the number of bits of information taken from the predicted category with the presence or absence of a term in a document [16].

The selected features were ranked in KBest scenarios. KBest is the amounts of features according to the K highest score. In this study, KBest is used to limit the number of selected features that will be used for classification. Features that already had Information Gain value were ranked from highest value (KBest). In this study, features for classification were selected in three scenarios: KBest=700, KBest =500, and KBest =200.

E. Chi-Square Feature Selection

Chi-Square is a method used to calculate the level of feature dependency. Chi-Square uses a statistical theory to test the independence of a term with its category. Equation (4) below is to apply Chi-Square feature selection method [17]. In this formula, the feature f means a feature that looked for in class c, and class c means a class target (positive or negative).

$$X^2(f,c) = \frac{N(A\times D - B\times C)^2}{(A+B)\times(C+D)\times(A+C)\times(B+D)} \quad (4)$$

c = class target
f = feature target
N = number of data
A = class c that contains feature f
B = the feature f that is not in class c
C = class c that does not contain feature f
D = not class c and does not contain feature f

In Chi-Square algorithm, the feature occurrence in the expected and unexpected categories is really important. The frequency of a feature becomes less valuable if it often appears in the unexpected category [18].

In the Chi-Square feature selection, the Chi-Square value of a feature was calculated using (4). Based on (4), a contingency table was made as seen in Table 2 [17].

TABLE 3. CHI-SQUARE CONTINGENCY TABLE

Contingency table	***Feature f (f)***	***Not Feature f***
Class c (c)	A	C
Not Class c	B	D

The same as Information Gain, features that already had Chi-Square value were sorted by the largest value and selected into three scenarios for the classification: KBest =700, KBest =500, and KBest =200.

F. Naïve Bayes Classification

Bayesian classification is a simple probabilistic-based prediction technique based on Bayes theorem with the assumption of strong (naive) independence. In Naïve Bayes, strong independence in features means that a feature in a datum has the same important weight, not related to one another [19]. Equation (5) is used to calculate the probability of feature X in class C (P(C|X). Where P(C) is probability of class C based on the number of documents and class, and P(X) is probability of all feature X.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (5)$$

Naïve Bayes calculates a set of probabilities by summing the frequencies and combinations of values from a given dataset [20]. It classified the data by comparing the probability value obtained. The probability value of each feature was measured using (5). If the probability value of positive class P(C_i |X) is higher than the probability of negative class P(C_j |X), then the data will be categorized as a positive and vice versa [14].

In the classification, datasets were split into train set and test set with a ratio of 8:2. Features that had been selected at the feature selection stage were used for classifying the training process to make a classification model. Meanwhile, the test set was used for testing the models that had been made to get the accuracy. Naïve Bayes was used because the features that had been selected will had the same weight for the classification. According to Han and Kamber (2006), the Naïve Bayes classification has been proven to have high accuracy and speed when applied to a large number of databases [21] [22].

G. Confusion Matrix

Confusion matrix is a visualized test formed as a specific table to predict true and false objects. It contains four possible outputs as reference material in comparing the actual events with the predicted events [23]. Confusion matrix table can be seen in Table 1.

TABLE 4. CONFUSION MATRIX

Confusion Matrix		**Prediction Class**	
		Positive	***Negative***
Actual Class	Positive	TP (True positive)	FN (False negative)
	Negative	FP (False positive)	TN (True negative)

The confusion matrix table shows reports of predicted and actual data from classification. These reports then are used for calculating accuracy, precision, and recall. Equation (6)(7)(8) are to calculate the values of accuracy, precision, and recall, respectively.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

IV. Result and Discussion

Accuracy results were obtained from the test set classification then was stored as a classification model. In this section, results and discussion of the accuracy that were modeled from two datasets and three KBest scenarios will be explained. The results of Naïve Bayes Classification without feature selection, with Information Gain, and with Chi-Square feature selection in various KBest from the Indonesian dataset can be seen in Table 3.

TABLE 5. CLASSIFICATION RESULT OF INDONESIAN DATASET

Dataset	**Method**	**Accuracy**			
		KBest= 700	***KBest= 500***	***KBest= 200***	***Avg***
Indo-nesian	Without feature selection	63%	65%	63%	63.6%
	Information Gain	66%	65%	66%	65.6%
	Chi-Square	68%	66%	65%	66.3%

The model created have got an accuracy of 63.6% for Naïve Bayes Classification without feature selection, 65.6% with Information Gain, and 66.3% with Chi-Square. The result for the Indonesian dataset shows that feature selection can increase the accuracy although not too significant, this is possible because the feature weighting in the feature selection process is not too different. The highest accuracy is classification with Chi-Square feature selection.

In Indonesia dataset, the model without feature selection have the same accuracy in KBest=700 and KBest=200, that is 63%. And have hagot the best accuracy in KBest=500, that is

65%. It is because sometimes too many features cause the boundaries between classes are getting blurry, causing performance to decrease [17]. In this method, the top 500 features selected are the most relevant features for classifying the data. The opposite result showed from the models with information gain feature selection where KBest=500 gives the lowest accuracy, while in KBest=700 and KBest=200 give higher accuracy that is 66%. It means in 500 features selected there are some features that make it irrelevant to the target class, therefore it reduces the accuracy. In chi-square, we can see that the more features selected, the higher the accuracy obtained. It means, in this model, Chi-Square can give appropriate weight to the relevance of the data.

The results of Naïve Bayes Classification without feature selection, Information Gain feature selection, and Chi-Square feature selection with various KBest in the English dataset can be seen in Table 4.

TABLE 6. CLASSIFICATION RESULT OF ENGLISH DATASET

Dataset	**Method**	**Accuracy**			
		KBest= 700	***KBest= 500***	***KBest= 200***	***Avg***
English	Without feature selection	63%	63%	62%	62.6%
	Information Gain	63%	62%	63%	62.6%
	Chi-Square	63%	63%	65%	63.6%

The average accuracy shows 62.6% for Naïve Bayes Classification without feature selection and with Information Gain feature selection and 63.6% with Chi-Square. From the table above, it can be seen that there are no differences in accuracy between Naïve Bayes without feature selection and with Information Gain in average, but there is 1% increasing accuracy of Chi-Square feature selection in average. It means in the models observed, Chi Square could filter more relevant features comparing to Information Gain and no feature selection. Especially in KBest=200, Chi Square could weight features that are really relevant to the model of English dataset. In Naïve Bayes without feature selection method we can see that it has similar accuracy in KBest=700 and KBest=500. It shows that the relevancy of 700 and 500 features against class target has the same value.

By the KBest=700, it can be seen that the accuracy does not increase or decrease. Meanwhile, by the KBest=500, there is 1% decrement in Information Gain. This small difference of accuracy could be caused by the distribution of features that are found in all target classes (positive & negative). So that the weighting of features against class does not produce a high difference. In this study, it shows that feature selection is not always increasing accuracy. The result of the English dataset shows that the highest average accuracy is classification with Chi-Square feature selection.

By the result obtained, we can see that the amount of the features cannot determine the accuracy. This result is in accordance with the research that had been carried out by Wibowo & Indriyawati (2020) which got the best accuracy on the top 9 features of the 19 features examined [24]. Listiowarni (2018) in her research had got the best result in the top 5 features from 15 feature examined [17]. Novaković, et.al, (2011) examined that the number of features applied to different classification methods could result different accuracy [25].

Therefore, the average accuracy of both datasets taken are 63.2% in Naïve Bayes without feature selection, 64.2% Naïve Bayes with Information Gain, and 65% Naïve Bayes with Chi-Square. Fig. 4 is the chart of the average accuracy obtained.

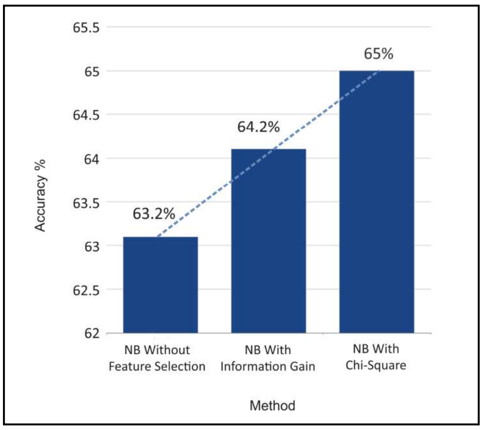

Fig. 4. Average Accuracy of Indonesian and English datasets

In the average of all accuracy results, the highest accuracy is Naïve Bayes classification method with Chi-Square; second is Naïve Bayes with Information Gain; and the lowest accuracy is Naïve Bayes without feature selection. Chi-Square worked by testing the independency of a term with its category [26]. In this lockdown sentiment analysis, Chi-Square method could filter the features better for classification. This result is in accordance with the research that had been done by Sofiana and Subagio [10] [11]. The insignificant differences occurred because the distribution of features in existing classes (positive & negative) tends to be evenly distributed.

The best model that had been created was then used for clustering sample tweets of 400 sentiments about lockdown, PSBB, and PPKM which were taken in June 1st -30th 2021 near Jakarta. The clustering results obtained as many as 31.1% sentiments with positive labels and as many as 68.9% sentiments with negative labels. Fig.5 is the distribution of tweet labels obtained.

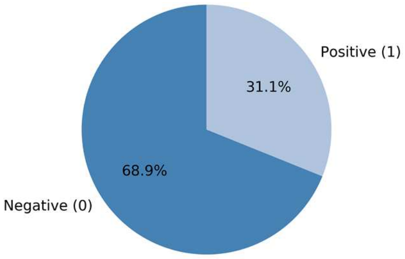

Fig. 5. The average of clustering results on the tweet sample

From the 400 tweets that had been clustered, it can be seen that as many as 68.9% citizens gave negative sentiments about lockdown, PSBB, or PPKM policies. Only 31.1% gave positive sentiment. This is because that many communities are badly affected by the restriction policy. The policy is more difficult for the majority of the community than to have a positive effect in Indonesia.

V. Conclusion

The conclusion obtained in this study is that feature selections tend to increase the accuracy even though not really significant. Reducing irrelevant features could increase the accuracy, but in this study, feature selection is not always increasing accuracy.

The best accuracy of lockdown sentiment classification using Naïve Bayes algorithm in both Indonesian and English datasets is Naïve Bayes with Chi-Square feature selection. The differences in accuracy occurred because of the value of each feature in these two feature selection methods was different. It caused differences in various features for the classification process. In this study, the number of features used did not have an effect on accuracy. Selecting the top rank of feature values (KBest) might reduce the accuracy because unweighted features in certain feature selection methods might be used in the classification process and lower the accuracy. This issue needs to be studied further.

By clustering 400 sample tweets regarding lockdown, PSBB, and PPKM, it can be seen that citizens near Jakarta tended to disagree with lockdown, PSBB, or PPKM policies. It is because there were 68.9% of sample tweets showing negative sentiments of the policies. This research could be one of the considerations in determining the restriction policy in Indonesia. However, this research also still needs to be studied further and more specifically, for example the research based on specific region or the social status of the community, so the results obtained would be more precise in accordance with what is happening in the field.

For further research, it is suggested to implement the feature selection methods using various thresholds and alphas on Information Gain and Chi-Square. Other suggestions are using datasets that have been clustered by several point of view, or using various kinds of clustering library.

Acknowledgment

This research was financially supported by the Ministry of Research and Technology, Republic of Indonesia through Fundamental Research Grant No. 225-98/UN7.6.1/PP/2021.

References

[1] "World Health Organization," 4 Desember 2020. [Online]. Available: https://covid19.who.int. [Accessed 4 Desember 2020].

[2] I. F. Rozi, H. S. Pramono and E. A. Dahlan, "Implementation of opinion mining (sentiment analysis) for data extraction of public opinion in universities," EECCIS Journal Vol. 6, No. 1, vol. 6, no. 1, p. 37, 2012.

[3] H. N. Firqiani, A. Kustiyo and E. P. Giri, "Feature selection using a fast correlation based filter on the voting feature intervals 5 . algorithm," Scientific Journal of Computer Science, vol. 6, no. 2, 2008.

[4] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *Journal of Information Science,* vol. 43, no. 1, pp. 25-38, 2017.

[5] A. Adhiselvam, K. Umamaheswari and J. Ramya, "Feature Selection Based Enhancement Of The Accuracy Of Classificationalgorithms," *Turkish Journal of Computer and Mathematics Education ,* vol. 12, no. 10, pp. 5621-5628, 2021.

[6] J. Cai, J. Luo, S. Wang and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing,* vol. 300, pp. 70-78, 2018.

[7] H. Li, Y. Liang and Q. Quan, "A dividing-based many-objective evolutionary algorithm for large-scale feature selection," *Soft Comput,* vol. 24, pp. 6851-6870, 2020.

[8] N. D. Pratama, Y. A. Sari and P. P. Adikara, "Sentiment analysis on consumer reviews using the nave Bayes method with chi-square feature selection for recommendations for traditional food locations," Journal of Information Technology and Computer Science Development, vol. 2, no. 3, p. 2982, 2018.

[9] B. N. Sari, "Implementation of information gain feature selection techniques on machine learning classification algorithms for predicting student academic performance," National Seminar on Information Technology and Multimedia, p. 55, 2016.

[10] I. Sofiana, I. Atastina and A. A. Suryani, "Analysis of the effect of feature selection using Information Gain and chisquare for Indonesian text categorization," Repository Telkom University, 2012. [Online]. Available: https:// repository. telkomuniversity.ac.id/ pustaka/95720/analisispengaruh-feature-selection-menggunakan-information-gaindan-chi-square-untuk-kategorisasi-teks-berbahasaindonesia.html. [Accessed 24 June 2021].

[11] M. M. Subagio, "Comparison of feature selection information gain and chi-square on film classification based on synopsis using nave Bayes classifier method," UMM Institutional Repository, Malang, 2021.

[12] O. Soemantri and D. Apriliani, "Support vector machine based on feature selection for sentiment analysis of customer satisfaction with the services of warungs and culinary restaurants in the city of Tegal," Journal of Information Technology and Computer Science, vol. 5, p. 537, 2018.

[13] Y. Pratama, A. R. Tampubolon, L. D. Sianturi, R. D. Manalu and D. F. Pangaribuan, "Implementation of sentiment analysis on twitter using naïve bayes algorithm to know the people responses to debate of dki jakarta governor election," Journal of Physics: Conference Series, no. 1175 012102, 2019.

[14] Suyanto, Data mining, Bandung: Informatika, 2019.

[15] M. R. Maulana and M. A. Al Karomi, "Information gain to determine the effect of attributes on credit approval classification," Litbang Journal Pekalongan , vol. 9, p. 114, 2015.

[16] I. Maulida, A. Suyatno and H. R. Hatta, "Feature selection in Indonesian text abstract documents using the information gain method," JSM STMIK M, vol. 17, no. 2, pp. 249-258, 2016.

[17] I. Listiowarni and E. R. Setyaningsih, "Feature selection chisquare and k-nn on the categorization of exam questions based on cognitive domain bloom taxonomy," Caltex Politeknik Journal Riau , vol. 4, p. 25, 2018.

[18] J. Ling, I. P. E. N. Kencana and T. B. Oka, "Sentiment analysis using naïve bayes classifier method with chi-square feature selection," E-Journal Matematics, vol. 3, pp. 93-94, 2014.

[19] A. R. Rozzaqi, "Naïve Bayes and filtering feature selection information gain for predicting the accuracy of student graduation," UPGRIS Informatics Journal, vol. 1, p. 30, 2015.

[20] A. Saleh, "Implementation of the nave Bayes classification method in predicting the amount of household electricity use," Citec Journal, vol. 2, no. 3, pp. 207-217, 2015.

[21] R. E. Putri, Suparti and R. Rahmawati, "Comparison of naïve bayes and k-nearest neighbor classification methods in analysis of work status data in Demak district in 2012," GAUSSIAN JOURNAL, vol. 3, no. 4, pp. 831-838, 2012.

[22] J. Han and M. Kember, "Data mining: concepts and techniques," Morgan Kaufman, California, 2006.

[23] Essra, Aulia; Rahmadani; Safriadi;, "Information gain attribute evaluation analysis for intrusion attack classification," ISD (Information System Development) Journal, vol. 2, p. 9, 2016.

[24] R. Wibowo and H. Indriyawati, " Top-k feature selection for hepatitis disease detection using naïve bayes algorithm," *Jurnal of Buana Informatika,* vol. 11, no. 1, pp. 1-9, 2020.

[25] J. Novaković, P. Strbac and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research,* vol. 21, no. 1, pp. 119-135, 2011.

[26] A. Z. Amrullah, A. S. Anas and A. J. Hidayat, "Movie review sentiment analysis using nave Bayes classifier with chisquare feature selection," Journal of BITe, vol. 2, no. 1, pp. 40-44, 2020.