Indonesian Journal of Applied Research IJAR

COMPARISON OF FEATURE SELECTION BASED ON COMPUTATION TIME AND CLASSIFICATION ACCURACY USING SUPPORT VECTOR MACHINE

Salmun K. Nasib[1], Fadilah Istiqomah Pammus[1]
Nurwan[2] , La Ode Nashar[1]

[1]Statistics Study Program and Gorontalo State University, Indonesia
[2]Mathematics Study Program and Gorontalo State University, Indonesia

Volume 4 Issue 1
(April 2023)
e-ISSN 2722-6395
doi: 10.30997/ijar.v4i1.252

ARTICLE INFO

Article history:
Received: 20-12-2022
Revised version received: 28-12-2022
Accepted: 29-03-2023
Available online: 18-04-2023

Keywords:
chi-square; feature selection; mutual information; sentiment classification; support vector machine.

How to Cite:
Nasib, S. K., Pammus, F. I., Nurwan, & Nashar, L. O. (2023). COMPARISON OF FEATURE SELECTION BASED ON COMPUTATION TIME AND CLASSIFICATION ACCURACY USING SUPPORT VECTOR MACHINE. *Indonesian Journal of Applied Research (IJAR)*, *4*(1), 63-74. https://doi.org/10.30997/ijar.v4i1.252

Corresponding Author:
Salmun K. Nasib
salmun@ung.ac.id

ABSTRACT

The goal of this research to compare Chi-Square feature selection with Mutual Information feature selection based on computation time and classification accuracy. In this research, people's comments on Twitter are classified based on positive, negative, and neutral sentiments using the Support Vector Machine method. Sentiment classification has the disadvantage that it has many features that are used, therefore feature selection is needed to optimize a sentiment classification performance. Chi-square feature selection and mutual information feature selection are feature selections that both can improve the accuracy of sentiment classification. How to collect the data on twitter taken using the IDE application from python. The results of this study indicate that sentiment classification using Chi-Square feature selection produces a computation time of 0.4375 seconds with an accuracy of 78% while sentiment classification using Mutual Information feature selection produces an accuracy of 80% with a required computation time of 252.75 seconds. So that the conclusion are obtained based on the computational time aspect, the Chi-Square feature selection is superior to the Mutual Information feature selection, while based on the classification accuracy aspect, the Mutual Information feature selection is more accurate than the Chi-Square feature selection. The recommendations for further research can use mutual information feature selection to get high accuracy results on sentiment classification.

1. INTRODUCTION

The tendency for someone to respond in assessing something by using opinions is more accessible through social media, especially Twitter because social media is experiencing rapid development worldwide (Mujilahwati, 2016). The rise of news in the community resulted in the emergence of positive or neutral opinions and negative ones. Analyzing the comments as a whole can be time-consuming and vice versa. If only a few words are understood, it will produce biased evaluation results (Khairunnisa et al., 2021). Therefore sentiment analysis aims to classify user comments into positive, negative, or neutral opinions automatically. In this condition, a classification method is needed that can produce sentiment information with a high degree of accuracy, namely the Support Vector Machine (SVM) method. Syamsiah (2014) states that Support Vector Machine is a Supervised Learning method. In the training stage, you want to find two parameters between the lines and the bias after finding the optimal parameters through quadratic programming optimization. In this case, the classification using Twitter data will encounter obstacles in conducting sentiment analysis. Generally, the comments obtained still contain non-standard words, such as writing abbreviated words and using contemporary language. Twitter has a maximum writing limit of 140 characters for one upload. Therefore, the preparatory stages need to be carried out in the classification of comments to produce sentiments in the form of appropriate sentences to be used at the next stage of the process (Khairunnisa et al., 2021).

There are areas for improvement in sentiment classification in the number of dataset features or attributes used. Therefore feature selection is needed to optimize sentiment classification performance (Tsani et al., 2020). The use of feature selection was also explained in previous research that features selection can select important features in documents to increase the accuracy of the classification results (Irene, 2017)—chi-Square feature selection and Mutual Information (MI), including feature selection in the guided category. Hi-Square is widely used to produce a good performance in classification accuracy (Eduardo T. Ueda Route Terada, 2009) in Tsani et al. (2020). According to Sun, Wang, and Xu in Tsani et al. (2020), Chi-Square can remove many irrelevant features without reducing the level of accuracy that will result in the classification results.

Meanwhile, Mutual Information feature selection focuses on the relationship of words with a class, which can produce the best features with higher classification accuracy. Due to the advantages possessed by each feature selection, this study aims to compare the two feature selections to find out which feature selection is better in increasing the accuracy of classification results using the Support Vector Machine method in the case of public comments on Twitter regarding the Covid-19 Pandemic. Pandemic Covid-19 is an event often discussed in society today, making people express their feelings via Twitter. So that generates the comments used for sentiment classification to determine the polarity of sentiment that appears positive, negative, or neutral (Khairunnisa et al., 2021).

2. METHODS

2.1. System Design

This research will build a system to analyze the sentiment of public comments on Twitter regarding the Covid-19 Pandemic. Comments are classified into three classes, namely positive, negative, and neutral, called datasets. Before entering the classification stage, the dataset will be divided into two parts: training and test data. The training data is used to train the algorithm to find a suitable model, while the test data is used to test the previously trained model's performance (Khairunnisa et al., 2021).

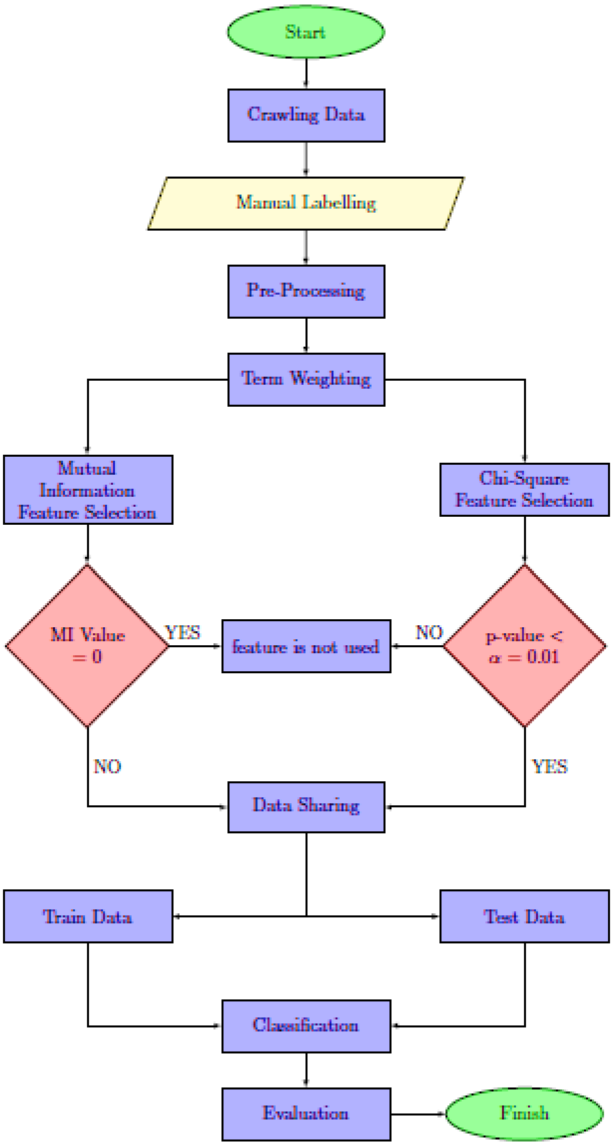

Figure 1 System Design

2.2. Dataset

In this study, public comments on Twitter were taken using the IDE application from Python, namely Anaconda Prompt. The comments taken are uploaded from March 2020 to March 2022 based on the keywords covid19, coronavirus, corona Indonesia, covid19 vaccine, health protocol, wear masks, and pandemicovid19. The results of the data collection were 3,577 tweets. Researchers did the labeling manually and then validated it by Indonesian Language Lecturers from the Faculty of Letters and Culture, State University of Gorontalo. The labeling process is done by giving a code to each comment obtained from data collection, and code 1 is given if the tweets obtained contain meaning that refers to praise, support, or other good things. In contrast, code -1 is given if the tweets contain meaning that refers to insults, satire, or other bad things, and tweets containing informative meanings such as definitions, news, or sentences that do not match the keywords, are given a code of 0.

Furthermore, the labeling results of tweets are called sentiments. The labeling results that have been validated produce 537 positive sentiments, 255 negative sentiments, and 2,785 neutral sentiments. The following is an example of a labeled dataset.

Table 1 Example of a labeled dataset

Comments	Label
almost all countries are affected by the issue of covid-19, and their children and grandchildren...! only in Indonesia are there official rules from the government so massively changing the rules and procedures for prayer..!	-1
@aldotjahjadi8 Great... I hope that Covid 19, especially Omicron, can be overcome soon. Come on, Indonesia can.	1
President Jokowi responded to the increase in Covid-19 cases in Indonesia. Jokowi has asked the Coordinating Minister for Maritime Affairs, Luhut Binsar Pandjaitan, and the Coordinating Minister for the Economy, Airlangga Hartarto, to evaluate the PPKM level. https://t.co/KP4cCvlv3j	0

2.3. Sentiment Classification

In sentiment classification, a preparatory stage is needed to prepare good data to be processed at the next classification stage. Sentiments processed at this preparatory stage are sentiments validated by Indonesian Lecturers from the Faculty of Letters and Culture, State University of Gorontalo. The following are the preparatory stages in the sentiment classification:

1. Filtering serves to remove characters outside of the az alphabet including punctuation, URLs or links, hashtags, and usernames (Pravina et al., 2019).
2. Tokenize serves to cut sentences into words that aim to facilitate the process of cleaning words at an early stage (Amin F. (Gunawan et al., 2018)).
3. Case Folding functions to change capital letters to lowercase letters which aims to facilitate the search process when processing data (Gunawan et al., 2018)).
4. Stopwords Removal functions to delete words that are not unique or words that often appear but are considered unimportant such as time, conjunctions (Mujilahwati, 2016).
5. It is a steaming function to turn words into essential words (Pravina et al., 2019).

After going through the preparatory stages, sentiment classification also requires word weighting to find ranking information based on word frequency (Gunawan et al., 2018). In this study, the Count Vectorizer was used, a text weighting method applied to sentiment classification which aims to convert text into vector inputs only in the form of Binary numbers, namely 0 and 1 (Ehsan and Hasan (Hakim et al., 2020). Count Vectorizer is a technique of simple weighting and produces relatively relevant results (Reynaldhi & Sibaroni, 2021). According to Reynaldhi & Sibaroni (2021), the process of term weighting from the Count Vectorizer is by counting how many times a word appears in a sentence and using the unit value as its weight or changing it in chunks of words called tokenization and being able to return words in integer form.

In the classification, the results of performance testing are evaluated, namely the Confusion Matrix, which aims to perform accurate calculations on the concept of data mining. There are four terms as a representation of the results of the classification, namely:

1. True Positive (TP) is a positive value that is correctly predicted (Mutawalli et al., 2019).
2. True Negative (TN) is the number of negative data that is correctly predicted (Mutawalli et al., 2019).
3. True Neutral (TNR) is a correct prediction based on actual data (Nurkholis et al., 2022).
4. False positive (FP) is negative data but predicted positively (Mutawalli et al., 2019).

5. False Negative (FN) is negative data predicted as negative data (Mutawalli et al., 2019).
6. False Neutral (FNR) is an error in which actual data labeled neutral is predicted to be positive or negative (Nurkholis et al., 2022).

Table 2 Confusion matrix

Predicted Class	*Actual Class*		
	Positive	*Neutral*	*Negative*
Positive	TP	FNR	FN
Neutral	FP	TNR	FN
Negative	FP	FNR	TN

The classification performance test is as follows Rakhman et al. in Tsani et al. (2020) :

1. Accuracy is the level of documents that are correctly identified.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TNR + TN}}{\boldsymbol{TP + TNR + TN + FP + FNR + FN}} \times \mathbf{100}\% \qquad (\mathbf{1})$$

2. Recall the success rate of the system in retrieving information.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \times \mathbf{100}\% \qquad (\mathbf{2})$$

3. Precision the comparison of the amount of data that matches the requested data.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN + FNR}} \times \mathbf{100}\% \qquad (\mathbf{3})$$

2.4. Chi-Square Feature Selection

According to Wang in Tsani et al. (2020), feature selection is a valuable part of optimizing the performance of the sentiment classification method. Feature selection can also optimize an ample feature space by eliminating less relevant features (Koncz et al. in Tsani et al. (2020)). One of the feature selection methods included in the filter model is Chi-Square. According to IC Negara & Prabowo in Nisa et al. (2019), Chi-Square is a feature selection that pays attention to the dependency of words with their categories, where:

1. There are no cells with an Actual Count value or a reality frequency that is 0 (zero);
2. If the contingency table is in the form of 2x2, then cells with an Expected Count or expected frequency of less than 5 cannot be allowed;
3. If the form of the table is more than 2x2, for example, 2x3, then the number of cells with an expected frequency of less than five cannot be more than 20%.

Table 3 Contingency table

	$Column_1$	$Column_2$	...	$Column_j$	Total
$Line_1$	$\boldsymbol{O_{11}}$	$\boldsymbol{O_{12}}$	...	$\boldsymbol{O_{1j}}$	$\boldsymbol{b_1}$
$Line_2$	$\boldsymbol{O_{21}}$	$\boldsymbol{O_{22}}$	...	$\boldsymbol{O_{2j}}$	$\boldsymbol{b_2}$
...	...	...	...	...	...
$Line_j$	$\boldsymbol{O_{i1}}$	$\boldsymbol{O_{i2}}$	...	$\boldsymbol{O_{ij}}$	$\boldsymbol{b_i}$
Total	$\boldsymbol{k_1}$	$\boldsymbol{k_2}$	...	$\boldsymbol{k_j}$	N

$\boldsymbol{O_{ij}}$ is the value of the two variables studied or the so-called Actual Count. Rows are features, and columns are classes (negative, neutral, positive). The following process in Chi-Square is looking for the Expected Count:

Table 4 Expected count

	$Column_1$	$Column_2$	...	$Column_j$
$Line_1$	$\boldsymbol{E_{11}}$	$\boldsymbol{E_{12}}$	...	$\boldsymbol{E_{1j}}$
$Line_2$	$\boldsymbol{E_{21}}$	$\boldsymbol{E_{22}}$	...	$\boldsymbol{E_{2j}}$
...	...	...	...	...
$Line_j$	$\boldsymbol{E_{i1}}$	$\boldsymbol{E_{i2}}$	...	$\boldsymbol{E_{ij}}$

Look for the Expected Count from the Actual Count, or fill in the cells using the formula according to Pramesti in Nisa et al. (2019)) below:

$$\boldsymbol{E_{ij}} = \frac{\boldsymbol{b_i k_j}}{\boldsymbol{N}}, \qquad (\mathbf{4})$$

with,

$\boldsymbol{E_{ij}}$: Expected count of row i and column j
$\boldsymbol{b_i}$: The sum of the i- line
k_j : Addition of the j- column
N : Total All values

To determine whether existing features can be used for the classification process, seen from the Chi-Square value, the greater the Chi-Square value generated, the more important these features are for use in the classification process (Nisa et al., 2019). The following is the formula for finding the Chi-Square value :

$$\boldsymbol{X^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \frac{(\boldsymbol{O_i} - \boldsymbol{E_i})^2}{\boldsymbol{E_i}} \quad , \qquad (\mathbf{5})$$

with,
$\boldsymbol{X^2}$: *Chi-Square value*
$\boldsymbol{O_i}$: The value of i-Observation
$\boldsymbol{E_i}$: The Value of i-Ekspektasi

2.5. Mutual Information Feature Selection

Feature selection is used after going through a weighting process for each word. The purpose is to select features to be used in the classification process to make it more informative and effective. At this stage, available features will be selected, and which features are relevant to the data class will be determined. In contrast, irrelevant features will be preserved and used in classification (Khairunnisa et al., 2021). Mutual Information Selection, or MI, is a method that has a concept in calculating how much information is contained in a word to determine an appropriate sentiment classification decision for a class (Irham et al., 2019). Determining whether these features can be used in the classification process seen from the MI value, the greater the MI value, the greater an attribute affects a class (Khairunnisa et al., 2021); here's the equation to calculate the Mutual Information value :

$$\boldsymbol{I(U,C)} \sum_{\boldsymbol{ec}} \sum_{\boldsymbol{ec}} \boldsymbol{P(U = et, C = ec) = log_2 \frac{P(U = et, C = ec)}{P(U = et)P\ (C = ec)}}, \qquad (6)$$

with,

U : Features
C : Class
et =1 : Documents that contain the term t in features
et = 0 : Documents that do not contain the term t in features
ec = 1 : Documents that contain the term t in the class
ec = 0 : Documents that do not contain the term t in the class

2.6. Support Vector Machine

Support Vector Machine (SVM) is a classification method that uses Machine Learning to predict classes based on models or patterns from the results of the training or training process (Novantirani et al., 2015). SVM includes Supervised Learning which can be applied to classification or regression (Mahdiyah et al., 2015). Supervised Learning is an approach that has previously trained data Train Data and target variables, so this learning aims to group data into existing data (Prayoginingsih & Kusumawardani, 2018). SVM aims to find the most considerable Margin on the dividing line. Margin is the distance between the dividing line and the Support Vector (Irene, 2017). The pattern or value generated from SVM is a dividing line called a Hyperlane. This dividing line separates tweets with opinions with positive, negative, or neutral sentiments (Pravina et al., 2019). In principle, SVM is a Linear Classifier and has now been developed into a Non-Linear SVM with the implementation of Kernel Trick to handle classification on higher dimensions.

The Kernel trick, or the Kernel Function, helps learn SVM because to determine the support vector, you only need to know the Kernel Function used, and you do not need to know the form of the Kernel Function. The kernel function applied in this research is Gaussian RBF. Research by Irene et al. (2017) shows that the Best Kernel RBF classification results are 84.3% using constant values (C) = 1000 and γ= 1.0. The C Constant value is most commonly applied to all SVM kernels, and the function of the parameter γis to determine the level of proximity between two points so that it is more effective in finding Hyperlanes that are consistent with the data (Irene, 2017).

3. RESULTS AND DISCUSSION

In this study, a test scenario focused on comparing Chi-Square feature selection with Mutual Information based on computation time and classification accuracy. In the classification process, this study uses a classification method, namely Support Vector Machine, with the Kernel Function applied, namely Gaussian RBF. From the results of the sentiment classification carried out, the performance evaluation of the classification using the Confusion Matrix will be evaluated to determine the accuracy of the sentiment classification.

3.1. Results

In the first stage, the preparatory stage is carried out to prepare the data to be processed in the classification. The following is an example of the results of the preparation stages obtained from Python:

Table 5 Example of the results of the preparation stages

Process	Input	Output
Filtering	@aldotjahjadi8 Great... I hope Covid 19, specifically Omicron, can be overcome immediately. Come on, Indonesia can.	aldotjahjadi8 Great good luck covid 19 specifically for Omicron can be resolved immediately. Come on, Indonesia can
Tokenize	aldotjahjadi8 Great, I hope that Covid 19, especially Omicron, can be overcome soon. Come on, Indonesia can do it	[”aldotjahjadi8”, “great”, "hopefully," "covid," "19", "special," "omicron," "dapt," "immediately," "overcome," "Come on ", "indonesian," "can"]
Case Folding	aldotjahjadi8 Great, I hope that Covid 19, especially Omicron, can be overcome soon. Come on, Indonesia can do it	aldotjahjadi8 great good luck covid 19 specifically for Omicron can be resolved immediately; let's Indonesia do it
Stopword Removal	aldotjahjadi8 great good luck covid 19 specifically for Omicron can be resolved immediately; let's Indonesia do it	Great, I hope that Covid 19, specifically Omicron, can be overcome by Indonesia
Stemming	Great, I hope that Covid 19, specifically Omicron, can be overcome by Indonesia	cool, I hope that Covid, especially for Omicron, will get Indonesia

In the labeling stage, the number of sentiments obtained was as many as 3577; after being processed at the preparatory stage, it produced a total of 3508 sentiments, which means that there are as many as 69 sentiments that were eliminated at this stage. Sentiments were eliminated. They did not meet the criteria because they only contained symbols and punctuation marks and did not contain unique words that were considered necessary so that at this preparation stage, all sentences that did not meet the criteria would be eliminated. The next stage is selecting features using Chi-Square and Mutual Information feature selection. In this study, Chi-Square feature selection was carried out, which aims to eliminate less relevant features, by looking at the Chi-Square value obtained from the comparison of the p-value with $\alpha = 0.01$ so as to get the Chi-Square value best. According to Sun, Wang & Xu in Tsani et al. (2020), Chi-Square can eliminate many irrelevant features without reducing the level of accuracy that will result in the classification results. The following is an example of the best Chi-Square value :

Table 6 Best Chi-Square value example

Feature	Best Chi-Square Value
punishment	0,001395
stupid	0,000187
smart	0,003282
discipline	0,008361
evaluation	0,000223

Because the rows and columns obtained number in the thousands, determining the Chi-Square value best can be done by comparing the p-value with $\alpha = 0.01$. Words with a p-value of more than $\alpha = 0.01$ will be deleted or not used in the classification process. The number of words obtained from the text weighting process is 5465 after being processed in the Chi-Square feature selection to 970 features, which means there are as many as 4495 features considered irrelevant. The computation time required for the Chi-Square feature selection process is 0.4375 seconds.

Select the Mutual Information feature aims to calculate the dependency value between words or the reciprocal relationship of a word. Words that have a lower MI (Mutual Information) value or can be said to have a value of 0 will be deleted. According to Irham et al. (2019), Mutual Information feature selection focuses on the relationship of words with a class, which can produce the best features with higher classification accuracy. The following is an example of the results of the Best MI score:

Table 7 Best MI value example

Feature	Best MI Value
apathetic	0,002333
onus	0,001035
smart	0,003111
peace	0,006139
education	0,000831

At the text weighting stage, the number of features obtained was 5465. After processing the Mutual Information feature selection, it changed to 2648 features, which means that 2817 features were deleted because they were deemed irrelevant. The computation time required for the Mutual Information feature selection process is 252.72 seconds. The next step is classifying sentiment using the Support Vector Machine method with the kernel parameter RBF (Radial Basis functions). Classification aims to classify sentiments based on positive, negative, and neutral classes in SVM using the Kernel Function in this study using the RBF Kernel. Research by Irene et al. (2017) shows that the Best Kernel RBF classification results are 84.3% using constant values (C) = 1000 and $\gamma = 1.0$. Constant C is the most commonly applied value for all SVM kernels, and the function of parameter γ is to determine the level of proximity between two points so that it is more effective in finding Hyperlane consistent with the data (Irene, 2017). The results of sentiment classification using both feature selections are presented in the Confusion Matrix table below:

Table 8 Confusion matrix uses Chi-Square feature selection

Predicted Class	Actual Class		
	Positive	Neutral	Negative
Positive	5	47	0
Neutral	3	529	12
Negative	0	84	17

Table 9 Confusion matrix uses mutual information feature selection

Predicted Class	Actual Class		
	Positive	Neutral	Negative
Positive	2	54	0
Neutral	1	547	9
Negative	0	78	11

In this study, the accuracy results are preferred because we want to see how precise the classification results predicted by the algorithm are with the overall results on the target variable (validated Initial Labeling). The SVM accuracy generated using Chi-Square feature selection is 78%, meaning that the SVM algorithm successfully predicts positive, negative, and neutral sentiments by 78% of the target variable (initial labeling results), and there are 22% of the classification results that are incorrectly predicted by the SVM algorithm based on the target variable. Whereas the accuracy of SVM uses mutual information feature selection of 80%, which means that the SVM algorithm successfully predicts positive, negative, and neutral sentiments of 80% of the target variable (the result of initial labeling), and the algorithm fails to predict positive, negative and neutral sentiments of 20% based on on the target variable.

3.2. Discussion

Based on the previously described discussion, the SVM algorithm uses the RBF kernel parameters to classify sentiment with an accuracy of 80% on the Mutual Information feature selection and 78% on the Chi-Square feature selection of the target variable. The target variable is obtained from the labeling process, validated by Indonesian linguists, and passed the text cleaning at the preparation stage; there are 3508 sentiments. Because the results of the preparation stage are in the form of sentences, the step that must be passed is to change the text into a vector that can be processed at the feature selection stage. Text weighting is done to produce input vectors in the form of binary numbers, namely 0 and 1, which results in 5465 features.

After weighing the text, it is processed in feature selection, which aims to eliminate irrelevant features or words used at the sentiment classification stage. Before being classified, the target variable is processed at the data division stage which aims to divide the data into training data to train the algorithm with the appropriate model and test data to test the performance of the model that has been previously trained. The training data obtained is 2807 sentiments, and the test data is 702. After passing through the data sharing stage, the best features obtained from the Chi-Square feature selection stage and the Mutual Information feature selection are used in the sentiment classification stage, which aims to classify sentiment based on positive, negative, and neutral classes. In the Chi-Square feature selection, if the p-value $< \alpha = 0.01$, the resulting value is the best Chi-Square value, that is, there are 970 Best Chi-Square values obtained. In the Mutual Information feature selection, the higher the MI value obtained, the better the accuracy results. There are 2648 best MI values obtained at the Mutual Information feature selection stage. The results obtained from all stages of data processing are to produce accuracy using Mutual feature selection Information is higher than Chi-Square feature selection.

4. CONCLUSION

Comparison of feature selection based on computational time and classification accuracy: Chi-Square feature selection requires as much computation time as 0.4375, while Mutual feature selection Information takes as much as 252.75 seconds. Regarding classification accuracy, selecting Chi-Square features resulted in a classification accuracy of 78%, while the selection of Mutual Information features by 80%. So the comparison is based on the computational time aspect of the Chi-Square feature selection superior to the Mutual Information feature selection, while based on the accuracy of the classification of the Mutual Information feature selection more precisely accurate than the Chi-Square feature selection.

ACKNOWLEDGMENT

Thanks to those who have helped prepare this article, especially the statistics study program and the mathematics and natural sciences *faculties at Gorontalo State University*. It can provide benefits and help in solving problems in classifying sentiment.

REFERENCES

Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *4*(2), 113. https://doi.org/10.26418/jp.v4i2.27526

Hakim, L., Gustina, S., Putri, S. F., & Faudiah, S. U. (2020). Perancangan Chatbot di Universitas Proklamasi 45. *Edumatic : Jurnal Pendidikan Informatika*, *4*(1), 91–100. https://doi.org/10.29408/edumatic.v4i1.2157

Irene, A. F. (2017). *Klasifikasi Sentimen Review Film Menggunakan Algoritma Support Vector Machine Sentiment Classification of Movie Reviews Using Algorithm Support Vector Machine*. *4*(3), 4740–4750.

Irham, L. G., Adiwijaya, A., & Wisesty, U. N. (2019). Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine. *Jurnal Media Informatika Budidarma*, *3*(4), 284. https://doi.org/10.30865/mib.v3i4.1410

Khairunnisa, S., Adiwijaya, A., & Faraby, S. Al. (2021). Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19). *Jurnal Media Informatika Budidarma*, *5*(2), 406. https://doi.org/10.30865/mib.v5i2.2835

Mahdiyah, U., Irawan, M. I., & Imah, E. M. (2015). Integrating Data Selection and Extreme Learning Machine for Imbalanced Data. *Procedia Computer Science*, *59*(Iccsci), 221–229. https://doi.org/10.1016/j.procs.2015.07.561

Mujilahwati, S. (2016). Pre-Processing Text Mining Pada Data Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, *2016*(Sentika), 2089–9815.

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto). *Jurnal Informatika Dan Rekayasa Elektronik*, *2*(2), 43. https://doi.org/10.36595/jire.v2i2.117

Nisa, A., Darwiyanto, E., & Asror, I. (2019). Analisis Sentimen Menggunakan Naive Bayes

Classifier dengan Chi-Square Feature Selection Terhadap Penyedia Layanan Telekomunikasi. *E-Proceeding of Engineering* , *6*(2), 8650.

Novantirani, A., Sabariah, M. K., & Effendy, V. (2015). Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. *E-Proceeeding of Engineering*, *2*(1), 1–7.

Nurkholis, A., Alita, D., & Munandar, A. (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(2), 227–233. https://doi.org/10.29207/resti.v6i2.3906

Pravina, A. M., Cholissodin, I., & Adikara, P. P. (2019). Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2789–2797. http://j-ptiik.ub.ac.id

Prayoginingsih, S., & Kusumawardani, R. P. (2018). Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM). *Sisfo*, *07*(02). https://doi.org/10.24089/j.sisfo.2018.01.002

Reynaldhi, M. A. R., & Sibaroni, Y. (2021). Analisis Sentimen Review Film pada Twitter menggunakan Metode Klasifikasi Hybrid Naïve Bayes dan Decision Tree. *E-Proceeding of Engineering*, *8*(5), 10127–10137.

Syamsiah. (2014). Pemilihan Model Penentuan Kelayakan Pinjaman Anggota Koperasi Berdasarkan Algoritma Support Vector Machine , Genetic Algorithms , Dan Neural Network. *Faktor Exacta*, *7*(2), 141–153.

Tsani, M., Rupaka, A., Asmoro, L., & Pradana, B. (2020). Analisis Sentimen Review Transportasi Menggunakan Algoritma Support Vector Machine Berbasis Chi Square. *Smart Comp :Jurnalnya Orang Pintar Komputer*, *9*(1), 35–39. https://doi.org/10.30591/smartcomp.v9i1.1817