IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.15, No.3, July 2021, pp. 245~254
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.66965

Comparison of Filter and Wrapper Based Feature Selection Methods on Spam Comment Classification

Amalia Nur Anggraeni*[1], Khabib Mustofa[2], Sigit Priyanta[3]
[1]Master Program of Computer Science, FMIPA UGM, Yogyakarta, Indonesia
[2,3]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: *[1]**amalia.nur.anggraeni@mail.ugm.ac.id**, [2]khabib@ugm.ac.id,
[3]seagatejogja@ugm.ac.id

Abstrak

Pertumbuhan internet menyebabkan penggunaan media sosial untuk berbagai kepentingan meningkat. Beberapa pihak yang tidak bertanggung jawab memanfaatkan fitur komentar di media sosial untuk merugikan orang lain dengan memberikan komentar yang tidak relevan dengan objek yang dibagikan. Komentar tersebut termasuk dalam salah satu jenis spam. Salah satu pendekatan untuk menyelesaikan permasalahan spam yaitu dengan content base filtering. Filterisasi dilakukan menggunakan teknik klasifikasi teks. Variasi komentar menyebabkan jumlah fitur yang harus diproses besar sehingga dapat memberikan pengaruh terhadap performa suatu algoritme klasifikasi. Metode yang digunakan untuk mengatasi masalah tersebut adalah seleksi fitur. Seleksi fitur dilakukan untuk mendapatkan fitur terbaik. Penelitian membandingkan metode seleksi fitur filter, wrapper dan kombinasinya untuk klasifikasi komentar spam. Berdasarkan hasil pengujian dengan data latih sejumlah 4944 komentar dan data uji sejumlah 100 komentar maka didapatkan akurasi terbaik MNB sebesar 96%, precision sebesar 100%, recall sebesar 92% dan f-measure sebesar 95,8%. Akurasi terbaik dicapai menggunakan hasil seleksi fitur metode kombinasi seleksi fitur Chi Square dan Sequential Forward Selection dengan subset 500 fitur. Peningkatan akurasi pada klasifikasi MNB mencapai 8% sedangkan pada klasifikasi SVM mencapai 4%. Penelitian ini menyimpulkan bahwa kombinasi seleksi fitur mampu meningkatkan performa klasifikasi komentar spam berbahasa Indonesia.

Kata kunci*—Komentar Spam, Seleksi Fitur, Naïve Bayes, Support Vector Machine, Klasifikasi teks*

Abstract

The continuous growth of the internet has led to the use of social media for various purposes increase. For instance, some irresponsible parties take advantage of the comment feature on social media platforms to harm others by providing spam comments on the shared object. Furthermore, variation of comments creates many features to be processed, thereby negatively impacting the performance of a classification algorithm. Therefore, this study aims to solve the problem associated with spam comments by comparing filter and wrapper based feature selection using text classification techniques. Data collected from training and test data of 4944 and 100 comments showed that the best accuracy, precision, recall, and f-measure of MNB are 96%, 100%, 92%, and 95.8%. The best accuracy is achieved using feature selection by combining Chi-Square and Sequential Forward Selection methods with a subset of 500 features. Furthermore, the accuracy increase in the MNB and SVM classifications are 8% and 4%. This research concludes that the combination of feature selection improves the classification performance of Indonesian language spam comments.

Keywords*—Comment Spam, Feature Selection, Naïve Bayes, Support Vector Machine, Text Classification*

Received June 23th,2021; Revised July 28th, 2021; Accepted July 31th, 2021

1. INTRODUCTION

There is an increase in the use of social media for various purposes, along with the rapid rise in internet development. Comments sections are one of the features provided by various social media platforms, such as Instagram. This section allows users to share their individual opinions or reviews on a shared status. However, some irresponsible parties take advantage of the feature to harm others by providing comments that are irrelevant to the shared object [1]. Spam refers to unwanted information and is a directive for all social media users to a website with no relation to the content [2]. The information contained in spam is generally intended to market, promote, advertise, and carry out fraudulent activities [3]. Spam interfered with user's comfort in using social media quickly and accurately by disrupting the flow of discussion in a status [4].

One of the approaches used to solve this problem is content based filtering as the process of learning content using machine learning algorithms [5]. A common problem associated with identifying spam and non-spam comments are the varied comments. Variation of text significantly impacts a large number of features that need to be processed and classification algorithms. Meanwhile, not all features in the text are relevant or useful, and when used, it aggravates the computation process [6]. Therefore, it is necessary to select features using appropriate methods to improve classification performance.

Several studies have been conducted in handling spam with various feature selection methods and classification algorithms. For instance, the research carried out by [7] integrated the Principal Component Analysis (PCA) feature selection method and the Correlated Naïve Bayes Classifier (CNBC) algorithm, while [8] used the ranking method, and [9] applied the Genetic Algorithm as feature selection to improve the accuracy of Naïve Bayes results. Furthermore, [10] compared three classification algorithms, namely Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Network (ANN), as well as four feature selection algorithms, including information gain, chi-square, forward selection, and backward elimination for analysis of movie review sentiment.

In general, the attributes of the text classification are large, therefore it has the ability to reduce the classification performance when all attributes are used. However, the comparison of filter and wrapper feature selection and their ability to improve the classification performance of spam filtering comments is not yet determined. Therefore, this research focuses on selecting features to enhance the classification performance of Indonesian Instagram spam comments.

This research is organized as follows. Chapter 2 describes the proposed method, while chapter 3 describes the research results. Furthermore, the conclusions and suggestions are presented in chapter 4.

2. METHODS

This section provides a detailed description of the proposed method and an explanation of the data used in the research and models for spam detection.

2.1 Data Collection

Data were collected from the comments column of seven accounts of Indonesian artists and public figures with more than 1 million followers. The comments were collected using web scraping techniques and libraries in the Python programming language, namely BeautifulSoup. The collected data were re-sorted to obtain a collection of Indonesian comments. Furthermore, each comment in the document selected is manually labeled in the spam or non-spam category by paying attention to their characteristics. Keywords used to carry out spam labeling are related to advertisements and promotions [3] and negative and vulgar content [11]. Table 1 is an example of the contents of the comment dataset.

Table 1 Examples of comments and labels

No	Comments	Label
1.	yang sedang mencari produk kecantikan boleh konsultasikan via DM/whatsapp tersedia pemutih glowing wajah dan seluruh tubuh para seleb banyak diskon loh	Spam
2.	perut bunciiit ? susah turun berat badan ? gak suka olahraga ? tidak usah khawatir ada solusinya nih silahkan cek igku semua dapat teratasi, hasil sesuai keinginan kan	Spam
3.	kami mencari pemimpin yang mengayomi seluruh rakyat	Non-spam
4.	kami khawatir akan turun hujan, tlng solusinya pak dan cepat teratasi	Non-spam

2.2 Dataset

Figure 1 shows that the dataset consists of 2383 spam and 2661 non-spam comments. The comments were further categorized into 4944 and 100 training and test comments.

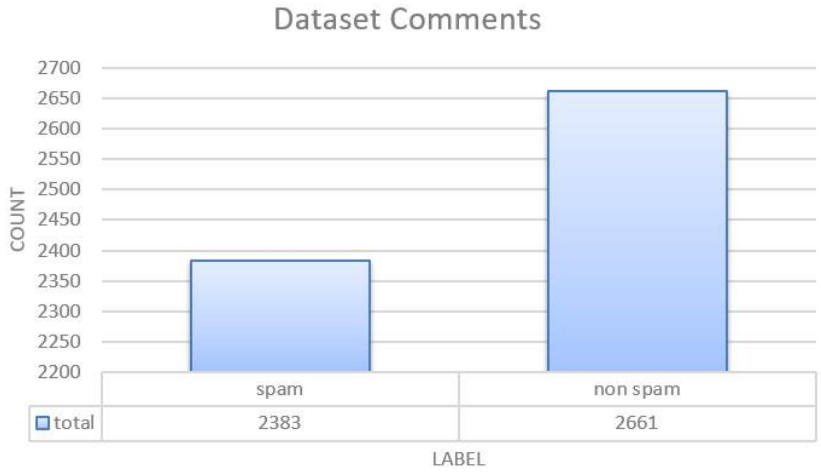

Figure 1 Visualization for the number of comment datasets

2.3 Spam Classification Design Using Feature Selection

The general research design includes the stages shown in Figure 2 below.

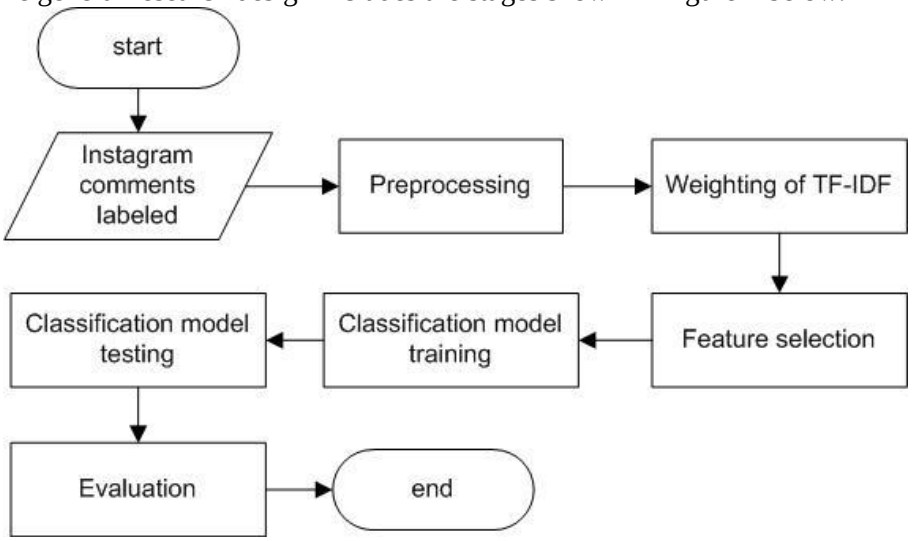

Figure 2 General research design

A detailed explanation of each stage is as follows:

2.3.1 Preprocessing

The content of comments is in the text form and tends to be irregular, while the structured data in accordance with the classification process. The text processing or preprocessing stage consists of the following:

a. Case folding: This is the process of changing all uppercase characters to lowercase. It only accepts Latin letters from 'a' to 'z' and also removes numbers, punctuation marks, symbols, and emoticons. This process is generally responsible for cleaning text from unnecessary characters.
b. Tokenizing: This is the process of truncating the input string based on each word that makes it up. The string truncation is carried out every time a separator or delimiter is found in the form of a space punctuation mark (whitespace).
c. Normalization: This is an approach used to change non-standard words into a standard. Furthermore, it detects abnormal words by matching them with the normal ones stored in a corpus.
d. Stemming is the process of returning a word to its basic form. This process works by removing all affixes in each word consisting of prefixes, suffixes, and confixes in derived words. The dictionary used in this research is Sastrawi.
e. Stopword removal eliminates common words that do not significantly affect the classification process and often appear in text documents such as conjunctions, articles, and pronouns. This research uses a stopword that lists the least important words that appear most frequently in the corpus.

2.3.2 TF-IDF Weighting

Term Frequency Inverse Document Frequency (TF-IDF) weighting is the process of calculating the weight of each word or term in a document to determine the availability. TF-IDF assigns a level of importance to words or terms in a document collection. The more often a term appears in a document, the less important it becomes. Term Frequency (TF) shows the appearance frequency of a feature (t) on a document (d), which is mathematically denoted in equation (1).

$$tf_{t,d} = f(t,d) \tag{1}$$

Inverse Document Frequency (IDF) is a calculation used to determine widely distributed terms. IDF is calculated by analyzing the division of a set of N documents against the number of df_t containing t feature. Inverse Document Frequency (IDF) is denoted in equation (2).

$$idf_t = log\left(\frac{N}{df_t}\right) \tag{2}$$

The TF-IDF weight value (W_t) is obtained by multiplying the TF and IDF log values. The TF-IDF (W_t) notation is mathematically shown in equation (3).

$$W_t = f(t,d) \times log\left(\frac{N}{df_t}\right) \tag{3}$$

2.3.3 Feature Selection

Feature selection is carried out by taking a part of all the attributes that exist in the data as a determinant in making decisions. The only attributes relevant to the dataset are selected. The relevance of attributes or features is calculated with and without involving learning algorithms. In this research, feature selection used the filter, wrapper, as well as filter and wrapper combination methods.

a. Filter method

This uses a statistical measure to evaluate features without including learning algorithm. The value of each feature is sorted and selected with the Chi-Square method used to measure the relationship between terms (words) and class (category). The calculation of the Chi-Square value for each term t for class c is shown in equation (4).

$$x^2(t,c) = \frac{N\,x\,(AD-CB)^2}{(A+C)\,x\,(B+D)\,x\,(A+B)\,x\,(C+D)} \quad (4)$$

Description:
$x^2(t,c)$ is the Chi-Square value of term t for class c
N is the total of all documents.
A is the number of documents in class c and contains the term t
B is the number of documents not in class c and contains the term t
C is the number of documents in class c without the term t
D is the number of documents not in class c without the term t

b. Wrapper method

This method uses a learning algorithm to evaluate the combination of features. The work process is carried out by conducting subset selection first, then evaluating the attributes using a classification algorithm. The wrapper feature selection method used in this research is Sequential Feature Selection (SFS). The attribute selection uses forward, forward floating, backward, and backward floating strategies. The forward selection method works by adding one feature at a time to each step before selecting the one with the best value. Furthermore, the next stage combines the features selected in the previous step with the remaining. This is repeated until it uses all the model features and the best is selected from a combination to provide the best performance value. The backward selection method works in the reverse way of forwarding selection by reducing one feature at each step. The floating variation is the development of a forward and backward strategy with backtracking capabilities. For example, floating variations are accommodated in the forward strategy to compensate for the weakness of forwarding selection, which cannot remove features after its addition.

c. Filter and wrapper combination method

The combination method of filter and wrapper feature selection performs two stages of feature selection with the carried using the Chi-Square. This feature is then used as a feed for SFS selection as a wrapper method.

2.3.4 Classification Model

Building a classification model is carried out in two stages, namely the feature selection stage using the wrapper method and the classification model training stage. The classification model at the feature selection stage is needed to evaluate the model's performance against the selected feature subset, which is the best (optimal) combination that produces optimal value for a particular learning algorithm according to the final criteria. Multinomial Naïve Bayes was used as the evaluator algorithm for the wrapper feature selection method in this research. The classification model at the training stage is carried out using the results of the TF-IDF weighting process. The model is validated using k-fold cross-validation with k value of 10, also known as 10-fold cross-validation. The training results form a model used to predict the test data that do not have class. In this research, the learning algorithm used is Multinomial Naïve Bayes and Support Vector Machine.

a. Multinomial Naïve Bayes

Multinomial Naïve Bayes (MNB) is one of the developments of the Bayes method. Meanwhile, MNB is a classification method obtained using the number of words (terms) occurrences in a document. The number of word occurrences is calculated using the Bayes assumption that each word is not related to others in a document. With this assumption, the value of $P(c|d)$ the probability of a document d being included in class c is written as in equation (5).

$$P(c|d) \propto P(c) \prod_{1 \le k \le n} P(t_k|c) \tag{5}$$

$P(t_k|c)$ is the probability of a term *term* t_k appearing in document *d* which is known to have *c* class. $P(c)$ denotes the prior probability of a document *d* included in class *c*. $P(c)$ is calculated based on the number of class *c* documents (N_c) divided by the total number of documents (N), which is calculated as shown in equation (6).

$$P(C) = \frac{N_c}{N} \tag{6}$$

The probability of conditional events or $P(t_k|c)$ is estimated as the relative weight of the term t in the document belonging to class c, which is calculated as shown in equation (7).

$$P(t_k|c) = \frac{W_{tc} + 1}{\sum_{t \in V} W_{tc} + V} \tag{7}$$

W_{tc} is the number of terms t weights in the training document that is in class c. $\sum_{t \in V} W_{tc}$ denotes the total weight of all terms contained in all documents in class c, including terms that appear repeatedly. Variable V or vocabulary is the number of unique words contained in all training documents with the highest score determined using the Naïve Bayes classification.

b. Support Vector Machine

The Support Vector Machine (SVM) classification concept is the search for the best hyperplane that functions as a separator of the two data classes in the input space. Hyperplane is the best separator between the two classes, which is found by measuring the margin and determining its maximum point. Margin is the distance between the hyperplane and the closest data from each class (support vector).

The data contained in the training data set is denoted as $x_i \in R^D$ while the class label is expressed as $y_i \in -1, +1$ for i=1, 2, 3…, N where N is the number of training documents. Both class -1 and +1 are perfectly separated by hyperplane dimension D in SVM. [12], stated that the hyperplane in the SVM is denoted by equation (8).

$$w.x_i + b = 0 \tag{8}$$

Where w and b are model parameters, and $w.x_i$ is the inner product between w and x_i. Mathematically, the SVM optimization formulation for linear classification cases in primal space uses inequality terms (9).

$$min\ \tau(w) = \frac{1}{2} \|w\|^2 \tag{9}$$

and limited to the following equation

$$y_i(w.x_i + b) \ge 1, \qquad i = 1, 2, 3, \ldots, N$$

It is computationally difficult to solve nonlinear classification where the objective function is quadratic and requires longer time. Nonlinear SVM generally takes a kernel approach to dataset features. This research uses the Radial Basis Function (RBF) kernel, as shown in equation (10).

$$K(x, y) = \exp\left(-\frac{1}{2\sigma^2} \|x - x_i\|^2\right) \tag{10}$$

2.3.5 Testing

Testing is carried out using test data separated from the training data before carrying out the training stage. It uses two classification algorithms, namely Naïve Bayes with the Multinomial Naïve Bayes (MNB) approach and Support Vector Machine (SVM) with the RBF kernel. Both classification algorithms were tested with and without feature selection. The filter and wrapper feature selection methods used are Chi-Square and Sequential Feature Selection (SFS). Testing was also carried out using the combination feature selection method of Chi-Square and SFS. Classification model testing produces predictions that are evaluated and compared.

2.3.6 Evaluation

Classification performance evaluation is carried out using confusion matrix containing information on the actual and predicted classes. The configuration matrix contains True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) values which are calculated to produce accuracy, precision, recall, and f-measure values.

3. RESULTS AND DISCUSSION

This research conducted several tests using the feature selection method and different classification methods. The test focuses on the effect of feature selection on the performance of the two classification models, namely Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM). The evaluation process makes use of training and test data that were previously labeled in the class manually. Performance is evaluated in each test treatment using accuracy, precision, recall, and f-measure values. In addition to recording classification performance, the test also records computation time.

The preprocessing of training data before the feature selection process resulted in 3402 unique tokens or vocabulary. The preprocessing also succeeded in eliminating 19859 features from the previous total of 23261. Furthermore, the feature selection process was carried out on the initial 3402 features obtained from the preprocessing training data. The feature selection treatment is carried out by performance testing using the Chi-Square, Sequential Feature Selection (SFS), and the combination of Chi-Square and SFS. The total selected features are determined using the cut limit of the best k features, namely 30, 50, 100, 150, 200, 500, and 1000.

Each method produces a different feature subset used in making classification models. This research compares the performance of two classification models, namely Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM). The test data consisted of 100 comments with 50 spam and 50 non-spam comments. Evaluation is carried out by comparing the classification (prediction) process with the actual class of the test data. Meanwhile, testing is carried out on test data that has gone through the preprocessing process using a trained classification model.

Testing starts with classifying the test data without selecting features. Table 2 shows the results of the classification evaluation without using feature selection, which led to an accuracy of 88% and 96% using the MNB and SVM classifier model. Furthermore, the value of the MNB and SVM precision models are 86.5% and 94.2%, with recall values of 90% and 98%. The F1 value of the MNB model is 88.2%, and the SVM model is 96.1%.

Table 2 Results of classification evaluation without using feature selection

Classifier	acc	prec	rec	f1
MNB	0.88	0.865	0.9	0.882
SVM	0.96	0.942	0.98	0.961

Furthermore, testing is carried out on the same test data using features selected using different methods, namely the Chi-Square selection result feature, the Sequential Feature Selection, and a combination of the Chi-Square and Sequential Feature Selection. The accuracy comparison of each feature selection method using the MNB classification is shown in Figure 3.

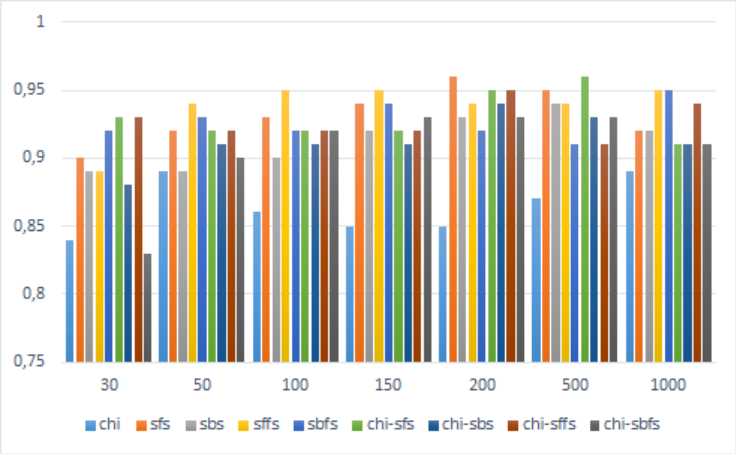

Figure 3 Comparison of MNB classification accuracy with the feature selection method

As in the Multinomial Naïve Bayes, the Support Vector Machine (SVM) classification model also tests the same data using features leading from several feature selections. The accuracy comparison of each feature selection method using the SVM classification is shown in Figure 4.

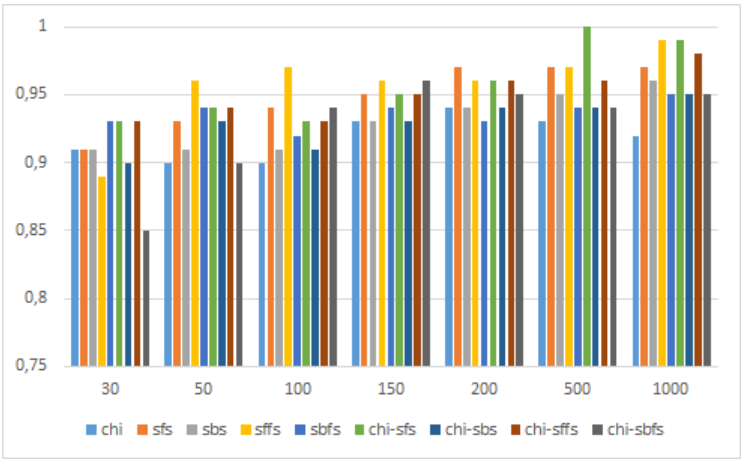

Figure 4 Comparison of SVM classification accuracy with the feature selection method

The test showed that the feature selection of the Chi-Square filter method produces the lowest level of accuracy compared to others. The best accuracy of MNB classification with Chi-Square feature selection is achieved when the feature selection chooses the best 50 features, which is 89%. Meanwhile, the best accuracy of SVM classification with Chi-Square feature selection of 94% is generated when the best 200 features are chosen. The selection of Chi-Square features towards MNB and SVM failed to contribute to increasing accuracy sufficiently.

MNB classification with Sequential Feature Selection produces a higher accuracy value compared to without using selection feature and using Chi-Square. The highest accuracy of 96% was generated by the Sequential Forward Selection, using 200 best features. Using the classification of more than 200 best features the SVM classification results showed improved accuracy compared to without feature selection. The highest accuracy of the wrapper method is generated by the Sequential Forward Floating Selection at 99% using the best 1000 features.

The combination of filter-wrapper feature selection results in higher accuracy values, with the highest from the MNB classification combination of Chi-Square feature selection and Sequential Forward Selection at 96% using the best 500 features. Meanwhile, SVM classification with a combination of Chi-Square feature selection and Sequential Forward Selection using the best 500 features led to an accuracy of 100%. The filter-wrapper combination feature selection works in two steps. Firstly it selects the filter feature with a high level of class relevance, and secondly, it selects the wrapper feature for the subset. In the first stage, the number of features has been selected based on certain cutting limits; therefore, it is less in the second stage, thereby leading to better performance.

In terms of computation time, each classification using the feature selection method has a different execution time. Furthermore, the computing time is calculated by adding the feature selection, training, and testing time. Tables 3 and 4 show the computation time comparison results of MNB and SVM classifications, which are 0.009 seconds and 1.339 seconds.

Table 3 Results of MNB classification computation time

features	30	50	100	150	200	500	1,000
chi	0.023	0.026	0.032	0.031	0.027	0.025	0.239
sfs	776.155	1,298.892	2,596.947	3,904.763	5,206.747	12,769.16	24,196.627
sbs	31,545.882	31,415.955	31,186.321	31,215.695	31,129.057	30,639.204	28,747.897
sffs	823.67	1,327.154	2,662.637	4,100.167	5,497.208	16,775.345	44,823.175
sbfs	152,753.301	151,155.125	130,480.309	113,018.907	84,259.953	57,599.86	44,522.866
chi-sfs	468.941	825.67	1,555.921	2,269.181	3,009.339	7,802.588	12,917.531
chi-sbs	17,673.671	19,035.53	17,453.379	17,265.944	17,237.007	17,279.59	13,936.894
chi-sffs	469.443	772.085	1,541.706	2,348.932	3,392.497	14,085.802	32,393.472
chi-sbfs	92,037.313	94,594.113	73,971.34	59,830.377	38,862.375	29,702.458	18,240.335

Table 4 Results of SVM classification computation time

features	30	50	100	150	200	500	1,000
chi	0.420	0.443	0.498	0.491	0.570	0.720	0.909
sfs	776.619	1,298.525	2,597.145	3,904.938	5,206.958	12,769.617	24,197.349
sbs	31,545.949	31,416.045	31,186.461	31,215.858	31,129.228	30,639.485	28,748.292
sffs	823.816	1,327.288	2,662.773	4,100.377	5,497.428	16,775.698	44,823.631
sbfs	152,753.372	151,155.210	130,480.450	113,019.083	84,260.158	57,600.150	44,523.264
chi-sfs	468.842	825.778	1,556.076	2,269.350	3,009.554	7,803.003	12,918.359
chi-sbs	17,673.765	19,035.696	17,453.643	17,266.237	17,237.419	17,280.126	13,937.797
chi-sffs	469.542	772.179	1,541.710	2,349.091	3,392.701	14,086.120	32,393.950
chi-sbfs	92,037.487	94,594.310	73,971.583	59,830.469	38,862.728	29,702.974	18,241.083

The filter method has the fastest feature selection execution speed, and this is in accordance with the calculation method, which is easier than others. In general, wrapper methods take the longest feature selection time, with its length influenced by the number of features sought and the subset search strategy used. Furthermore, the more the number of forward and forward floating strategies, the longer the computation time. Conversely, the more features are sought using both backward and backward floating strategies, the faster the computation time. This is consistent with the reverse work of forward and backward. Based on the tests conducted, the forward and backward floating methods with 30 features each require a selection time of 776.146 seconds and 152,753.301 seconds, respectively. The chi-forward and chi-backward floating methods with 30 and 50 features require a selection time of 468.726 and 94,594.104 seconds, respectively. The combination filter and wrapper feature selection method need a faster execution time than the wrapper method. This is because the filter feature selection process has reduced the number processed in the wrapper feature selection. Generally, feature

selection results affect the improvement of computation time in the MNB and SVM classifications.

4. CONCLUSION

In conclusion, the wrapper method performs better than the filter despite the long computation time. However, when the filter method is combined with the wrapper method, the feature selection is used to improve accuracy and save computation time. The combination feature selection method of Chi Square and Sequential Forward Selection with a subset of 500 features has the best effect on improving the Indonesian language spam comments classification accuracy using MNB and SVM. The accuracy improvement in the MNB classification reaches 8%, while the SVM classification reaches 4% compared to the accuracy results before using feature selection.

Further research needs to be carried out using feature selection for spam comment classification by comparing the performance of the Naïve Bayes classifier with other classification algorithms as an evaluator algorithm in the wrapper selection method.

REFERENCES

[1] Burhanudin, Y. Musa'adah, and Y. Wihardi, "Klasifikasi Komentar Spam Pada Youtube Menggunakan Metode Naive Bayes, Support Vector Machine, dan K-Nearest Neighbors," *J. Inform. dan Komput.*, vol. 3, no. 2, pp. 54–59, 2018.

[2] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, no. 1, pp. 27–34, 2015.

[3] K. Mathew and B. Issac, "Intelligent spam classification for mobile text message," in *Proceedings of 2011 International Conference on Computer Science and Network Technology*, 2011, vol. 1, pp. 101–105.

[4] A. R. Chrismanto and Y. Lukito, "Identifikasi Komentar Spam Pada Instagram," *LONTAR Komput.*, vol. 8, no. 3, pp. 219–231, 2017.

[5] A. Sharma, M. Sha, D. Manisha, and D. R. Jain, "A survey on spam detection techniques," *Int. J. Adv. Res. Comput. Commun. Eng.*, pp. 8688–8691, Dec. 2014.

[6] Z. Zhen, H. Wang, L. Han, and Z. Shi, "Categorical Document Frequency Based Feature Selection for Text Categorization," in *International Conference of Information Technology, Computer Engineering and Management Sciences*, 2011, vol. 2, pp. 65–68.

[7] E. Zuviyanto, "Integrasi Metode Principal Component Analysis untuk Meningkatkan Performa Correlated Naive Bayes Classifier pada Klasifikasi SMS Spam Berbahasa Indonesia," Universitas Gadjah Mada, Yogyakarta, 2018.

[8] C. A. Sugianto and T. H. Apandi, "Pengaruh Tokenisasi Kata N-Grams Spam SMS Menggunakan Support Vector Machine," *CITISEE*, Jan. 2018.

[9] O. Somantri and M. Khambali, "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 3, 2017.

[10] V. Chandani and R. S. Wahono, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film," *J. Intell. Syst.*, vol. 1, no. 1, pp. 56–60, 2015.

[11] C. Radulescu, M. Dinsoreanu, and R. Potolea, "Identification of spam comments using natural language processing techniques," in *Proceedings - 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing, ICCP 2014*, 2014, pp. 29–35.

[12] Suyanto, *Data Mining untuk Kalsifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.