JOURNAL OF DATA SCIENCE | Vol.2023:03
eISSN:2805-5160

Analysis of Feature Selection Methods for Sentiment Analysis Concerning Covid-19 Vaccination Issues

Muhammad Fajar[1], Tri Basuki Kurniawan[1*], Edi Surya Negara Harahap[2]

[1] Magister of Information Technology, University of Bina Darma, Palembang, Indonesia
[2] Computer Science, University of Bina Darma, Palembang, Indonesia

Email: tribasukikurniawan@binadarma.ac.id

Abstract

Sentiment analysis or opinion mining is a computational study of a person's opinions, sentiments, evaluations, attitudes, moods, and emotions. Sentiment analysis is one of the most active research areas in natural language processing, data mining, information retrieval, and web mining. One of the problems identified in the sentiment analysis process is the massive amount of data or text properties. In sentiment analysis, each word or term is collected into properties or dimensions, forming a data table. Due to the vast number of terms, this causes the process to take too long and requires a computer with tremendous power or ability. In addition, this can lead to a decrease in the quality of the model because data that is too large will also provide a significant bias value. Not all terms have contributions or relationships to decisions or labels in the form of positive, negative, and neutral values. For this reason, the feature selection method will be used in this study to select features or terms that contribute more to decisions or labels. It is also hoped that this can increase the quality of the prediction model that will be formed. In this study, the author will continue the research from another researcher by adding a feature selection process, such as two algorithms from the filtered method, chi-square, and information gain, and one algorithm from the wrapped method, which is Genetic Algorithms (GA). The experiment result shows that the GA obtained result has the highest accurate value compared to the other methods.

Keywords

Sentiment Analysis, Feature Selection, Filtered Method, Wrapped Method

Introduction

In the current era, social media, accompanied by advanced information technology (IT), has become a place for public information. It can be seen from the increasing number of social media users every year. It is fascinating to study because most of the existing data on social media contain sentiment opinions. A recent hot issue was Covid-19 immunization, a response to a dangerous virus that has spread worldwide between the end of 2019 and today. It still attracts the world's attention.

Submission: 28 February 2023; **Acceptance:** 21 March 2023

To reduce the transmission of the Covid-19 virus, the government issued a vaccination program to achieve immunity in the community. However, with the problems experienced in the previous vaccination program carried out by the government, some people refused to vaccinate this time. The government is trying to disseminate vaccination information through social media (Instagram), then this is the beginning of what attracts researchers to process and research further, to regain public trust.

Through social media, people can establish social relationships with users remotely and share information, events, or experiences that can be shared on (Negara, Andryani, & Saksono, 2016) their social media. With the establishment of these social relations, it produces data in the form of sentiment. Some things are still challenging to find from a large amount of data, which is why text analysis is needed. An investigation is carried out to produce specific information (Negara et al., 2016).

By conducting further research on public sentiment towards the government's vaccination program through social media, relevant information will be obtained on the response from the public. Does the program get a good response from the community or not? More complete and precise information will also be brought on which issues are well received and which are not accepted by the community.

Sentiment analysis or opinion mining is a computational study of a person's opinions, sentiments, evaluations, attitudes, moods, and emotions (Reyhana, 2018). Sentiment analysis is one of the most active research areas in natural language processing, data mining, information retrieval, and web mining (Gifari et al., 2022). Sentiment analysis is a subset of opinion mining, primarily using natural language processing and information extraction techniques to perform text mining and analysis. The tendency of specific texts is assessed based on the context and polarity obtained and can become a potential argument, opinion, or sentimental state of the text (Septianingrum, Jaman, & Enri, 2021). The more textual data collected in sentiment analysis research, the easier it is to find a significant correlation between text and type of sentiment. The kinds of sentiments are positive, negative, and neutral.

One of the problems identified in the sentiment analysis process is the massive amount of data or text properties (Az-Zahra, 2021). In sentiment analysis, each word or term is collected into properties or dimensions, forming a data table. Due to the vast number of terms, this causes the process to take too long and requires a computer with tremendous power or ability. In addition, this can lead to a decrease in the quality of the model because data that is too large will also provide a significant bias value (Kustiyo, Firqiani, & Giri, 2008). Not all terms have contributions or relationships to decisions or labels in the form of positive, negative, and neutral values. For this reason, the feature selection method will be used in this study to select features or terms that contribute more to decisions or labels. It is also hoped that this can increase the quality of the prediction model that will be formed (Hasibuan, 2019).

Much research has been done on the feature selection process in the classification process, where the data is discrete. However, there is still little research on feature selection for the sentiment analysis process. The features formed in the sentiment analysis process are much more

than those in the discrete data classification process. So there is expected to be a significant reduction in processing time and a much-increased accuracy result.

Several feature selection methods that can be applied to sentiment analysis, as described in research conducted by (Ahmad, Bakar, & Yaakub, 2019), include syntactic, semantic, lexico-structural, implicit, explicit, and frequent methods. Among these methods used in this study is the information gain method which is included in the common method, and one other method, namely the wrappers method.

In this study, the author will continue the research results conducted by Anggraini, Negara Harahap, & Kurniawan (2021). In this case, the author will use the same dataset used in previous research. This study will continue by implementing feature selection and then comparing the process results in terms of processing time and the level of accuracy of the model formed.

Methodology

Stages are the processes taken to research so that the research process can be well structured and systematic to achieve the expected goals. The following will explain the stages of the research methodology in Figure 1 below:

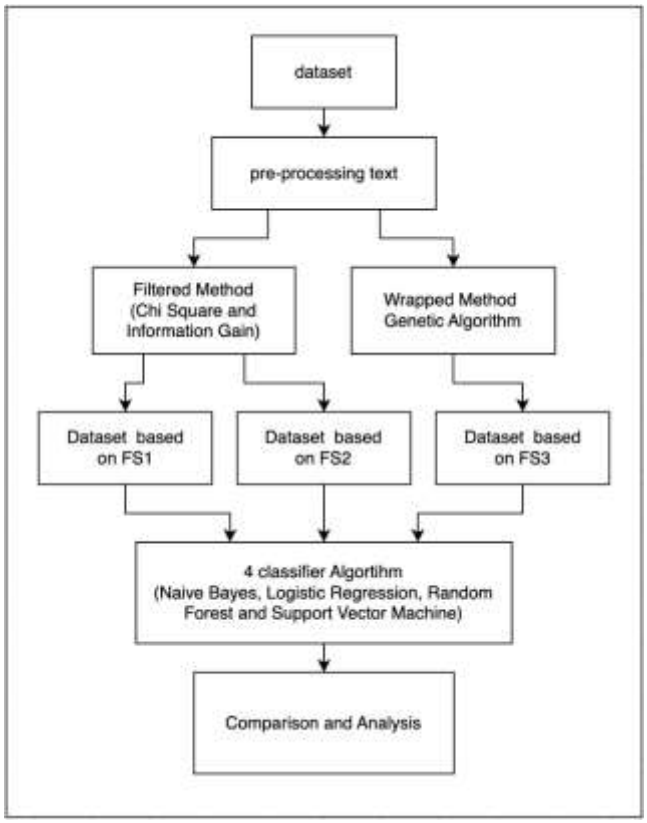

Figure 1. Research Stage

Figure 1 shows the research stage begins with doing pre-processing of the dataset. Later the filtered and wrapped feature selection methods are done to get the feature selection. After that, we create four (4) classifier models based on Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine using the data from the feature selection result. Lastly, the results are compared based on accuracy and execution time.

Dataset Collection

The data collection stage in this study was not carried out directly because it would use existing data from previous research conducted by Anggraini, Negara Harahap, & Kurniawan (2021). The dataset consists of 2,927 rows, and the number of terms generated when n-gram is set to (1, 3) is 37,796 words.

Experiment Setup

Based on Figure 1, we prepare the dataset by selecting the feature based on filtered and wrapped methods. For each dataset, we create a model based on four (4) classifier algorithms: Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine. Comparison between the training and testing datasets split is 70% and 30%, which is 2048 and 879 data. Figure 1 shows the distribution of each class label in the testing dataset.

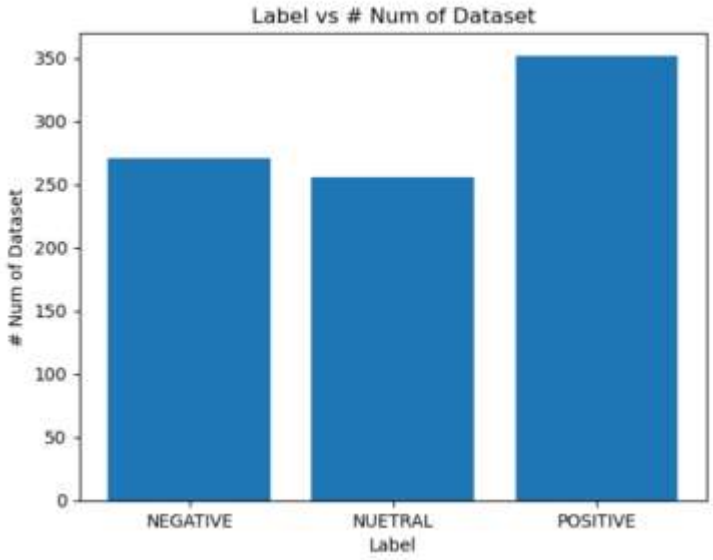

Figure 1. Data distribution for testing dataset.

Figure 1 shows negative, neutral, and positive class labels with 251, 276, and 352 data, respectively.

Feature Selection

The input variables fed into the machine learning algorithm are referred to as features. Each column in the data collection is a feature. To train an optimal model, make sure that it only employs the necessary features. If there are too many features, the model can identify meaningless patterns and learn from noise. Feature selection refers to the process of selecting important characteristics from data (Susetyo et al., 2019).

It automatically selects relevant features for a machine learning model based on the problem you want to solve. The system does this by including or omitting important features without changing them. It helps reduce noise in the data and facilitates data input.

Machine learning models follow a straightforward rule: what goes in, comes in, or gets out. If we introduce garbage into our model, we can expect the outcome to be garbage as well. In this instance, garbage refers to data noise. To train a model, the system must collect a big amount of data. Most of the data gathered is typically noise, and some data set fields may not have a significant impact on model performance. Furthermore, the large quantity of data can slow down training and the model. Models can also learn from this useless data and become inaccurate. (Nugroho & Wibowo, 2017).

Good data scientists distinguish themselves through feature selection. Some people outperform the competitors with faster and more accurate models when they have the same model and computational resources, this is because feature selection. In addition to selecting the appropriate model for the data, selecting the right data to include in the model is required.

In research conducted by (Adnyana, 2019), three feature selection techniques were used: correlation-based, information gain-based, and learner-based. The Learner-based technique shows the highest accuracy of the three methods used, while the Information Gain-based technique gives the lowest accuracy results.

The Correlation Feature Selection (CFS) technique, derived from the Correlation-based approach, is one of the popular techniques for selecting the most significant features from a data set. This technique calculates the correlation between each attribute/feature and the outcome variable, then specifies attributes with moderate to high correlation values (close to 1) and discards attributes with low correlation values (close to 0). CFS uses features' predictive power and cross-correlation to find a good set of features. Experiments with discrete and continuous data sets show that CFS can dramatically reduce the size of data sets while maintaining or increasing the performance of learning algorithms (Hall, 2000).

Another popular technique for feature selection is Information Gain Feature Selection (IGFS), an Information Gain-based technique. This technique calculates each attribute's information gain or entropy based on the output variable. The output value varies from 0 (minimum information) to 1 (maximum information). Features that provide more info will have a higher gain value and can be selected. In comparison, attributes that offer less information will have a low information gain value and can be discarded (Shaltout et al., 2014).

Another technique is Wrapper Based Feature Selection (WBFS) which comes from Learner-based techniques. This technique evaluates the algorithm's performance on datasets with different attribute subsets. The subset that produces the best performance will be the selected subset (Allam & Malaiyappan, 2020).

In addition, Bouchlaghem, Akhiat, & Amjad (2022) and Miao & Niu (2016) in their research summarized and explained many of the feature selection techniques used by researchers

to date in the feature selection field. Bouchlaghem, Akhiat, & Amjad (2022) divide feature selection techniques into three (3) significant groups: filtered, wrapped, and embedded. From each group, then each of the two (2) methods was selected, and then the results were compared on seven (7) datasets that had different characteristics. From the experimental results, it can be concluded that the techniques in the wrappers group, namely the sequential forward selection (SFS) technique and support vector machine-recursive feature elimination (SVM-RFE). Provide the best accuracy results in all datasets, especially the SFS technique gives the best results on six (6) datasets. In contrast, the SVM-RFE technique gives the best results on only 1 dataset.

Whereas the research conducted by Miao & Niu (2016) specifically compared feature selection techniques that are within the scope of unsupervised feature selection. It includes feature selection based on all features, MaxVar, Laplacian Score, and six (6) other techniques on 12 different datasets. From the results of their experiments, it can be concluded that no single process can give the best results for all datasets. Here it can be supposed that the accuracy results obtained for each technique depending on the dataset's type and characteristics to be processed.

This research experiment used Python programming with Jupyter Notebook as a development framework on the 13th Gen Intel Core i5 CPU with 8GB computer RAM. The execution time of the experiment results is collected based on this computer specification.

Results and Discussion

First, we show here the result of the sentiment analysis accuracy of the model before applying a feature selection. In this result, the dataset obtained from (Anggraini, Negara Harahap, & Kurniawan, 2021) is used without any feature selection applied. The results are shown in Table 1. This research uses four classifier algorithms: Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM).

Table 1. The accuracy and execution time of four models with the dataset without FS

#	Classifier Algorithms	Accuracy (%)	Execution Time (HH:MM:SS)
1	Naïve Bayes	90.67	00:00:03
2	Logistic Regression	40.05	00:00:03
3	Random Forest	94.20	00:00:55
4	Support Vector Machine	40.05	00:05:30

Table 1 shows the accuracy and execution time of each model based on four classifier algorithms. The Random Forest gives the highest accuracy from those results, while the Logistic Regression and SVM provide the lowest. In terms of execution time, Naïve Bayes and Logistic Regression give the shortest time, while Random Forest gives the longest execution time.

Next, the feature selection is applied based on the same dataset. In the first experiment, the Filtered methods: Chi-Square and Information Gain, are used, and the results are shown in Table 2 and 3, also Figure 2 and 4. Table 2 and Figure 2 show the results based on the Chi-Square approach, and Table 3 and Figure 4 shows the results based on the Information Gain approach.

Table 2. The accuracy and execution time of four models with the old dataset with Chi-Square

% of filtered features		Naïve Bayes	Logistic Regression	Random Forest	SVM
100 (37,796)	Accuracy (%)	90.67	40.05	94.20	40.05
	Execution Time	00:00:03	00:00:03	00:00:55	00:05:30
90 (34,016)	Accuracy (%)	91.81	40.05	94.99	40.05
	Execution Time	00:00:01	00:00:02	00:00:47	00:07:28
80 (30,236)	Accuracy (%)	93.52	40.05	95.79	40.05
	Execution Time	00:00:01	00:00:02	00:00:56	00:07:15
70 (26,457)	Accuracy (%)	94.77	40.05	95.68	40.05
	Execution Time	00:00:01	00:00:01	00:00:47	00:05:43
60 (22,677)	Accuracy (%)	91.70	40.05	96.93	40.05
	Execution Time	00:00:00	00:00:00	00:00:18	00:01:43
50 (18,898)	Accuracy (%)	93.52	40.05	96.59	40.05
	Execution Time	00:00:00	00:00:00	00:00:16	00:01:14
40 (15,118)	Accuracy (%)	97.61	40.05	96.82	40.05
	Execution Time	00:00:00	00:00:00	00:00:14	00:01:17
30 (11,338)	Accuracy (%)	97.27	40.05	96.59	40.05
	Execution Time	00:00:00	00:00:00	00:00:12	00:00:49
20 (7,559)	Accuracy (%)	96.59	40.05	96.93	40.05
	Execution Time	00:00:00	00:00:00	00:00:10	00:00:25
10 (3,779)	Accuracy (%)	80.89	40.05	88.74	40.05
	Execution Time	00:00:00	00:00:00	00:00:06	00:00:08

Figure 2. Line graph of the accuracy for four classifiers with different % filtered features based on the chi-square approach.

Table 2 and Figure 2 show the accuracy and execution time from four classifiers algorithms with different % filtered features. 100% filtered features mean all features are used, and 10%

denotes only 10% of features are used. The Naïve Bayes gives the highest accuracy when using 40% of filtered terms, 97.61%. The Random Forest algorithm obtained an accuracy quite competitive compared with Naïve Bayes but required much more time to get the results. The Naïve Bayes and Logistic Regression need little time to execution, almost 0 seconds for 60% filtered features and below. Compared with 100% filtered features, 40% filtered features increased accuracy by about 7.65 %, from 90.67 to 97.61.

The Logistic Regression and SVM provide the same accuracy to all % filtered features, 40.05%. It means Logistic Regression and SVM do not care which terms are used or are not sensitive enough about the operated or selected words.

Not only give low accuracy, but the SVM also needs the highest time to execute compared with other algorithms. It shows that by only using 40% and 30% of the terms, the Naïve Bayes got the best result, 97.61%, and 97.27%, respectively. Logistic Regression got relatively fast in terms of the execution time, but it got low accuracy, similar to SVM.

Figure 3 shows the confusion matrix for Logistic Regression and SVM algorithms for all % filtered features.

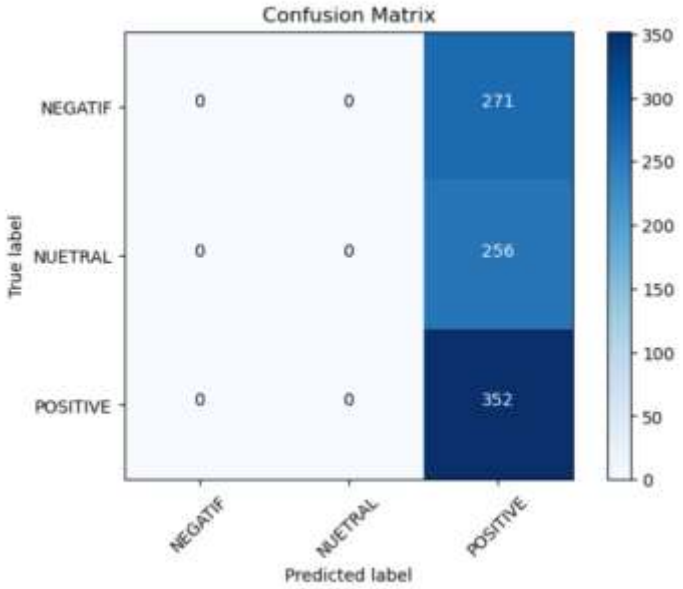

Figure 3. The confusion matrix for Logistic Regression and SVM algorithms.

Figure 3 depicts all data that are predicted as positive. It means the model created based on Logistic Regression and SVM algorithms fail to predict the testing data. These results are because that model cannot differentiate each class label based on available features. Since the positive in the testing dataset is a last class label, all data is predicted as neutral. If we change the order or class label, the predicted class will always be the last class label based on the defined order.

Next, the results based on the IG approach are shown in Table 3 and Figure 4 below.

Table 3. The accuracy and execution time of four models for Information Gain's SF

% of filtered features		Naïve Bayes	Logistic Regression	Random Forest	SVM
100 (37,796)	Accuracy (%)	90.67	40.05	94.20	40.05
	Execution Time	00:00:03	00:00:03	00:00:55	00:05:30
90 (34,016)	Accuracy (%)	91.13	40.05	94.54	40.05
	Execution Time	00:00:01	00:00:01	00:00:24	00:02:04
80 (30,236)	Accuracy (%)	92.50	40.05	95.22	40.05
	Execution Time	00:00:00	00:00:01	00:00:21	00:01:48
70 (26,457)	Accuracy (%)	93.40	40.05	95.22	40.05
	Execution Time	00:00:00	00:00:01	00:00:20	00:01:38
60 (22,677)	Accuracy (%)	95.34	40.05	95.90	40.05
	Execution Time	00:00:00	00:00:00	00:00:19	00:01:27
50 (18,898)	Accuracy (%)	93.17	40.05	96.02	40.05
	Execution Time	00:00:00	00:00:00	00:00:16	00:01:13
40 (15,118)	Accuracy (%)	96.36	40.05	96.59	40.05
	Execution Time	00:00:00	00:00:00	00:00:14	00:01:00
30 (11,338)	Accuracy (%)	97.95	40.05	97.04	40.05
	Execution Time	00:00:00	00:00:00	00:00:12	00:00:39
20 (7,559)	Accuracy (%)	96.82	40.05	96.70	40.05
	Execution Time	00:00:00	00:00:00	00:00:09	00:00:22
10 (3,779)	Accuracy (%)	79.98	40.05	93.74	40.05
	Execution Time	00:00:00	00:00:00	00:00:06	00:00:08

Figure 4. Line graph of the accuracy for four classifiers with different % filtered features based on the IG approach.

Table 3 and Figure 4 show the results based on Information Gain (IG) using the old dataset. Similarly, the IG gives the same pattern for the executing time. Naïve Bayes provides the fastest processing time, followed by Logistic Regression, Random Forest, and the last one, the longest

time, is SVM. For accuracy, Naïve Bayes also gives the highest accuracy on 30% of filtered features, which is 97.95%.

Table 4 and Figure 5 demonstrate the results of the wrapped method based on the Genetic Algorithm and the use of Nave Bayes as a classifier with 30% of the filtered features.

Table 4. The accuracy for each chromosome in every generation based on Genetic Algorithm

Initialization	Generate Initial Population ... unique_words 37796 chromosome_to_feature 11338 5 population was created
Initialization Population	Population: 1 - Accuracy: [0.9681456200] Population: 2 - Accuracy: [0.9522184300] Population: 3 - Accuracy: [0.9374288965] Population: 4 - Accuracy: [0.9146757679] Population: 5 - Accuracy: [0.9601820250] --- Best Accuracy: [0.9681456200]
	Running Genetic Algorithm ...
Generation 1	Generation: 1 Population: 1 - Accuracy: [0.9681456200] Population: 2 - Accuracy: [0.9590443686] Population: 3 - Accuracy: [0.9590443686] Population: 4 - Accuracy: [0.9590443686] Population: 5 - Accuracy: [0.9579067122] --- Best Accuracy: [0.9681456200]
Generation 2	Generation: 2 Population: 1 - Accuracy: [0.9681456200] Population: 2 - Accuracy: [0.9806598407] Population: 3 - Accuracy: [0.9817974972] Population: 4 - Accuracy: [0.9806598407] Population: 5 - Accuracy: [0.9817974972] --- Best Accuracy: [0.9817974972]
. . .	. . .
Generation 10	Generation: 10 Population: 1 - Accuracy: [0.9897610922] Population: 2 - Accuracy: [0.9761092150] Population: 3 - Accuracy: [0.9761092150] Population: 4 - Accuracy: [0.9772468714] Population: 5 - Accuracy: [0.9761092150] --- Best Accuracy: [0.9897610922]
Execution Time	Time: [00:31:58]

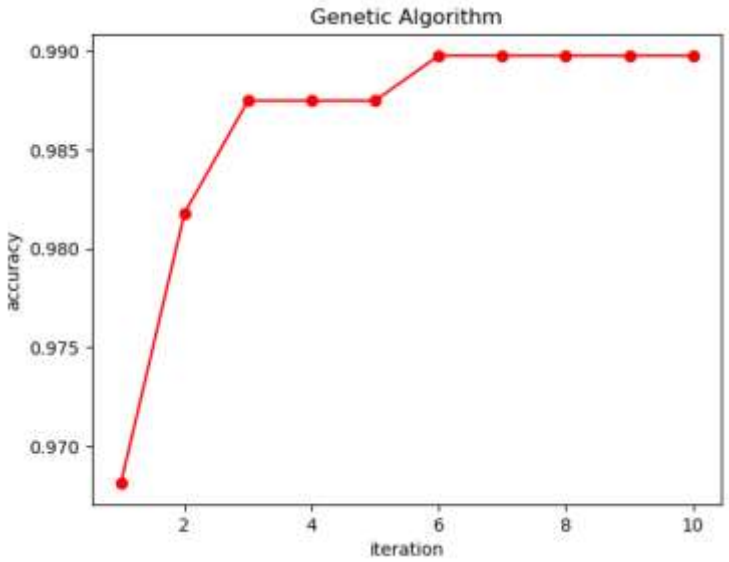

Figure 5. Convergence Curve the results based on Genetic Algorithms

Table 4 and Figure 5 show the Genetic Algorithm, in the beginning, obtained low accuracy and slowly increased it. After six (6) iterations, the solution generated cannot get better accuracy until the end of the iterations. The best result is a solution with the accuracy is 98.98%. Regarding execution time, Genetic Algorithm needs 650 times compared with the Chi-Square approach. Because, in GA, the process is lopping several times. So, the reasonable thing is that the GA needs more time to execute the algorithms to compare the other methods.

Conclusions

To enhance the accuracy of a sentiment analysis model, two feature selection methods are used. The results indicate that the first method, chi-square, can improve accuracy by approximately 7.28% for Information Gain and 6.94% when compared to the non-applied feature selection. The wrapped method based on Genetic Algorithms is tested next, and the results show that it improves accuracy by about 1.03% when compared to the filtered methods. Although the increase is insignificant, the Genetic Algorithm provides a new method for improving the accuracy of filtered methods. Another reason is that the filtered technique already achieved a high enough level of accuracy, and it is difficult to achieve higher levels of accuracy.

References

Adnyana, I. M. B. (2019). Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa. *Jurnal Sistem Dan Informatika*, *13*(2), 72–76.

Ahmad, S. R., Bakar, A. A., & Yaakub, M. R. (2019). A review of feature selection techniques in sentiment analysis. *Intelligent Data Analysis*, *23*(1), 159–189. https://doi.org/10.3233/IDA-173763

Allam, M., & Malaiyappan, N. (2020). Wrapper based feature selection using integrative teaching learning based optimization algorithm. *International Arab Journal of Information Technology*, *17*(6), 885–894. https://doi.org/10.34028/iajit/17/6/7

Anggraini, N., Negara Harahap, E. S., & Kurniawan, T. B. (2021). View of Text Mining - Analisis Teks Terkait Isu Vaksinasi COVID-19 (Text Mining - Text Analysis Related to COVID-19 Vaccination Issues).pdf. *Jurnal IPTEK-KOM (Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi)*, *32*(2), 141–153. Retrieved from https://jurnal.kominfo.go.id/index.php/iptekkom/article/view/4259/1644

Az-Zahra, T. S. (2021). Analisis sentimen terhadap belajar daring menggunakan optimasi naive bayes classifier dengan adaboost. *Tazkia Shabrina*.

Bouchlaghem, Y., Akhiat, Y., & Amjad, S. (2022). Feature Selection: A Review and Comparative Study. *E3S Web of Conferences*, *351*, 01046. https://doi.org/10.1051/e3sconf/202235101046

Dubois, P. F., Oliphant, T. E., Pérez, F., Granger, B. E., & Greenfield, P. (2007). *PYTHON : Guest Editor ' s Introduction Python for Scientific Computing IPython : A System for Interactive Scientific Computing Reaching for the Stars with Python*.

Gifari, O. I., Adha, M., Freddy, F., & Durrand, F. F. S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *Journal of Information Technology*, *2*(1), 36–40. https://doi.org/10.46229/jifotech.v2i1.330

Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. In *Computer Science Working Papers*. Retrieved from University of Waikato, Department of Computer Science website: https://hdl.handle.net/10289/1024

Hasibuan, M. R. (2019). *Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN)*. *3*(11), 3659–3875. Retrieved from http://j-ptiik.ub.ac.id

Kustiyo, A., Firqiani, H., & Giri, E. (2008). Seleksi Fitur Menggunakan Fast Correlation Based Filter pada Algoritma Voting Feature Intervals 5. *Jurnal Ilmiah Ilmu Komputer*, *6*(2), 245184.

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*(Itqm), 919–926. https://doi.org/10.1016/j.procs.2016.07.111

Negara, E. S., Andryani, R., & Saksono, P. (2016). Analisis Data Twitter: Ekstraksi dan Analisis Data G

eospasial. *INKOM Journal*, *10*, 27. https://doi.org/10.14203/j.inkom.433

Notebook, J. (2017). Jupyter. Retrieved October 10, 2022, from jupyter website: https://jupyter.org/

Nugroho, M., & Wibowo, S. (2017). Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes. *Jurnal Informatika Upgris*, *3*. https://doi.org/10.26877/jiu.v3i1.1669

Reyhana, Z. (2018). *Analisis Sentimen Pendapat Masyarakat Terhadap Pembangunan Infrastruktur Kota Surabaya Melalui Twitter Dengan Menggunakan Support Vector Machine Dan Neural Network*.

Septianingrum, F., Jaman, J. H., & Enri, U. (2021). Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier. *Jurnal Media Informatika Budidarma*, *5*(4), 1431. https://doi.org/10.30865/mib.v5i4.3260

Shaltout, N., Elhefnawi, M., Rafea, A., & Moustafa, A. (2014). Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts. *Lecture Notes in Engineering and Computer Science*, *1*, 625–631.

Susetyo, B., Eosina, P., Nurhayati, I., & Indupurnahayu, I. (2019). Model Feature Selection dalam Penentuan Parameter Pengelompokan Kompetensi SDM IG. *Krea-TIF*, *7*, 80. https://doi.org/10.32832/kreatif.v7i2.2696