NAÏVE BAYES ALGORITHM OPTIMIZATION USING *PARTICLE SWARM OPTIMIZATION* (PSO) FOR COVID-19 VACCINE SENTIMENT ANALYSIS ON TWITTER

Rivan Adi Nugraha[1], Teguh Iman Hermanto[2], Imam Ma'ruf Nugroho[3]
[1,2,3] Sekolah Tinggi Teknologi Wastukancana
Jl. Cikopak No. 53, Sadang, Purwakarta - Jawa Barat - Indonesia
E-mail : rivanadi82@wastukancana.ac.id

ABSTRACT- The Covid-19 vaccine is a vaccine that is quite popular, because it is the most needed and most discussed vaccine. There are 5 types of vaccines that are very popular including AstraZeneca, Moderna, Pfizer, Sinopharm and Sinovac. Sentiment analysis is a branch of text classification with computational linguistics and natural language processing that refers to a broad field, and text mining has a function to analyze opinions, judgments, sentiments, attitudes, evaluations and emotions of a person regarding an individual, organization, certain topics, services and other activities. This study aims to classify public sentiment towards the type of Covid-19 vaccine on social media Twitter, whether the opinion is positive or negative by using the Naïve Bayes algorithm based on *Particle Swarm Optimization* (PSO). The conclusion of this study is that the results of testing the *Naïve Bayes* algorithm with PSO using RapidMiner software are 79.17% accuracy, 87.69% precision, 85.07% recall for AstraZeneca vaccine, 68.82% accuracy, 92.29% precision, 71.72% recall for Moderna vaccine, 67.54% accuracy, precision 77.83%, recall 62.95% for Pfizer vaccine, accuracy 93.33%, precision 91.67%, recall 100.00% for Sinopharm vaccine, and accuracy 74.93%, precision 82.61%, recall 70.90% for Sinovac vaccine. It can be concluded that with the help of optimization PSO, the resulting confusion matrix value is greater and is proven to be more accurate.

Keywords : Vaccine; Covid-19; Sentiment Analysis; *Naive Bayes*; *Particle Swarm Optimization.*

1. PRELIMINARY

On March 11, 2020, the international health organization World Health Organization or WHO, officially determined the corona virus (Covid-19) to be a pandemic level which was initially still at the outbreak level [1]. At the end of 2019, even though the center of the spread of this virus was originally in China, precisely Wuhan City [2]. Currently, the virus is still infecting and spreading to all regions in Indonesia [3].

One solution to anticipate the spread of this virus is by developing a vaccine [4]. Vaccines are useful for protecting people who are vaccinated, and can reduce the rate of disease spread [5]. To stop and prevent the spread of disease, it is very important for us to develop effective and safe vaccines [6].

At this time the Covid-19 vaccine is a vaccine that is quite popular, because it is the most needed and most discussed vaccine. Coming from abroad and domestically, there are quite a number of types of Covid-19 vaccines that make people confused because they have many choices. So that people will seek information about the Covid-19 vaccine, possibly by reading opinions or reviews.

Opinions or reviews from the public can certainly provide considerable benefits for other communities, because the public can get information from other community reviews, the majority of which share their opinions and personal experiences about the Covid-19 vaccine through social media networks, especially on the Twitter platform.

Twitter is one of the social media platforms that can provide information in the form of images or reviews to the public [7]. There are so many opinions or reviews on Twitter, it will take a lot of time to read all the reviews one by one to find that information. Sentiment analysis is one way to solve the problem by grouping opinions or reviews into positive sentiments or negative sentiments.

Sentiment analysis is a branch of text classification with computational linguistics and natural language processing which refers to a broad field [8]. Text Mining has a function to analyze opinions, judgments, sentiments, attitudes, evaluations and emotions of a person regarding an individual, organization, topic, service and certain other activities [9]. Sentiment analysis is useful for determining whether the opinion or review has a positive or negative tendency [10]. Classification methods that are often used in sentiment analysis for text categories are *Support Vector Machine* (SVM) and *Naïve Bayes* [11].

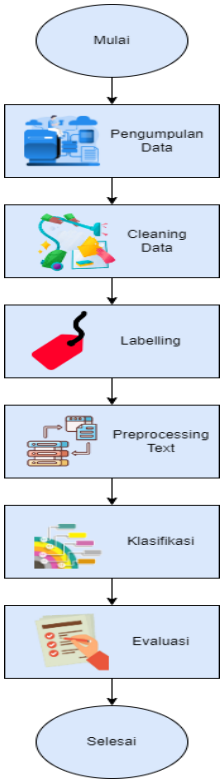
The *Naive Bayes* algorithm or method is very suitable to be implemented on data whose scale is very large and is able to overcome missing values (empty data values) [12]. The *Naive Bayes* method also has drawbacks, where the magnitude of the classification accuracy cannot be calculated from the probability [13]. Therefore, it is necessary to optimize with the help of the *Particle Swarm Optimization* (PSO) algorithm by assigning a weight value to each attribute to produce increased accuracy and assist the public in determining the best type of

Covid-19 vaccine based on the results of a review from Twitter.

In this study, samples of 5 types of Covid-19 vaccines will be taken including AstraZeneca, Moderna, Pfizer, Sinovac and Sinopharm. The purpose of this study is to find out the results of the analysis of 5 types of Covid-19 vaccines on Twitter social media using the Naive Bayes algorithm and *Particle Swarm Optimization* (PSO), compare the level of accuracy obtained between Naïve Bayes and Naïve Bayes + PSO, and produce visualizations. data from the research results.

2. RESEARCH METHODS

In this study, there are several stages carried out including data collection, data cleaning, labeling, preprocessing text, classification, and evaluation.



Picture 1. Research Stages

2.1 Data Collection

At this data collection stage, it contains 3 further stages, which include Literature Study, Crawling Data and Tweet Data. In the literature study stage, it is done by searching, analyzing and reading references from various theoretical sources related to sentiment analysis using the *Naive Bayes* algorithm and *Particle Swarm Optimization* (PSO).

After that in the next stage, researchers must get the Twitter API Key first before crawling data on Twitter using Orange software. Then, after the data crawling process is complete, the data can be saved in a csv file format with the extension.

2.2 *Cleaning Data*

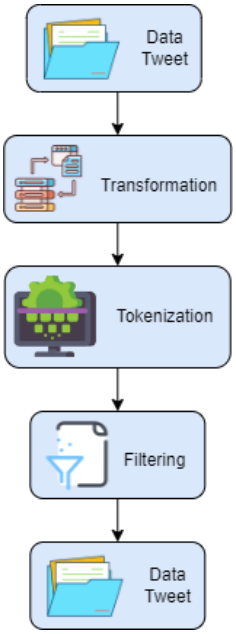
At this stage the process of deleting each line of tweet data containing statements or spam in the dataset, so that the dataset contains purely opinions.

2.3 *Labelling*

At this labeling stage, the data labeling process is carried out to divide and determine the sentiment of the tweet data whether it contains positive or negative sentiment.

2.4 *Preprocessing Text*

At this text preprocessing stage, several techniques are carried out to carry out preprocessing stages including transformation, tokenization, and filtering.



Picture 2. Text Preprocessing Stage

2.5 Klasifikasi

At this classification stage, to produce a text classification that contains positive and negative, the word weighting process is carried out first. The algorithm used in this stage is Naïve Bayes which is optimized with *Particle Swarm Optimization* (PSO).

2.6 Evaluasi

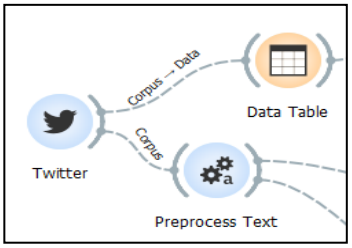
At this evaluation stage, the performance of the *naïve bayes* + PSO algorithm was tested for sentiment analysis of the covid-19 vaccine on the Twitter platform.

3. RESULTS AND DISCUSSION

3.1 *Crawling Data*

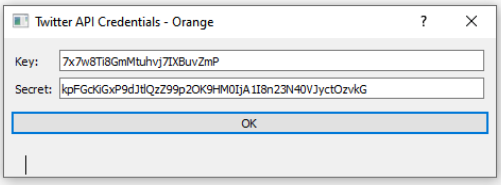
In this stage the data crawling process is carried out with the help of the Twitter API Key, and data crawling is carried out on April 14, 2022 using the Orange software. The keywords used were "astrazeneca", "moderna", "pfizer", "sinopharm" and "sinovac" which resulted in 412 AstraZeneca datasets, 2000 Moderna datasets, 1937 Pfizer datasets, 333 Sinopharm datasets and 1042 Sinovac datasets, with a total of total data is 5724 datasets.

The following are the steps of the data crawling process carried out using the Orange software.



Picture 3. The process of crawling data with Orange

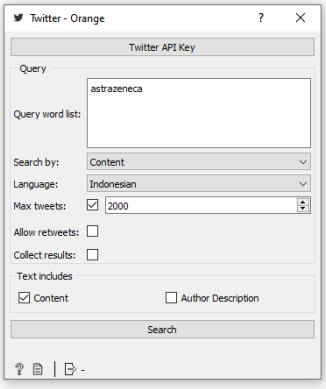
1. Entering *Twitter API Key*



Picture 4. *Twitter API key* and *secret*

2. Enter keyword query

Below is an example of the keyword entered, namely "astrazeneca" for the data search process. The data search is carried out based on content or tweets in Indonesian with a maximum data collection of 2000 tweets.



Picture 5. Enter the keyword "astrazeneca"

3.2 *Cleaning Data*

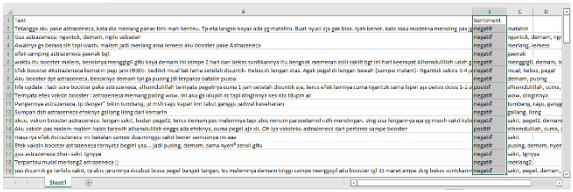
Cleaning data is done to delete each row of tweet data that contains statements or spam, so that the final result of this dataset contains purely opinions. The total number of tweet data after this process is done is 1363 data.

Table 1. Total data after cleaning

Dataset	Total Data
AstraZeneca	87
Moderna	587
Pfizer	425
Sinopharm	29
Sinovac	235

3.3 *Labelling*

After the Data Cleaning process is carried out, then the Labeling process is carried out by adding a computerized "Sentiment" attribute column.



Picture 6. Added the "Sentiment" attribute

The results of the data labeling process resulted in 355 positive sentiment data and 1008 negative sentiment data with the following details:

Table 2. Results of labelling on the dataset

Keywords	Total Data	Positive	Negative
AstraZeneca	87	20	67
Moderna	587	53	534
Pfizer	425	174	251
Sinopharm	29	7	22
Sinovac	235	101	134

3.4 *Preprocessing Text*

Text preprocessing is a process that is carried out with the aim of avoiding inconsistent data, imperfect data, and disturbances contained in the data [14]. In this process, the transformation, tokenization, and filtering stages will be carried out.

1. *Transformation*

In this process, the lowercase and remove url stages are carried out. That is the process to change the letter data of all text into lowercase (lowercase) and delete the url contained in the tweet data line.

Table 3. Before the process Transformation

S3 AstraZeneca, loyal don't switch to the next brand https://t.co/DCJzHhzv8c

Table 4. After the process Transformation

s3 astrazeneca, loyal don't switch to the next brand

2. *Tokenization*

In this process, a sentence is broken or cut per word or a sequence of string pieces, so that the end result becomes a word from sentence fragments.

Table 5. Tokenization Process

s3 astrazeneca, loyal don't switch to the next brand

s3	astrazeneca	loyal
don’t	switch	to
the	next	brand

3. *Filtering*

There are 2 stages in this process: *stopwords* and *regexp*. The process is to delete words that often appear, but do not have a meaningful meaning and remove connecting words that have no effect [15].

Table 6. Before the process Filtering

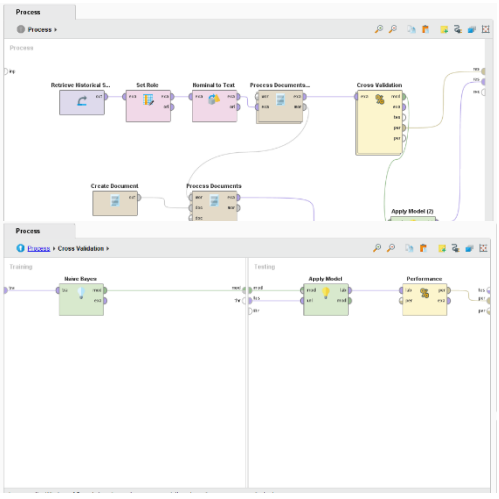
s3	astrazeneca	loyal
don’t	switch	to
the	next	brand

Table 7. After the process Filtering

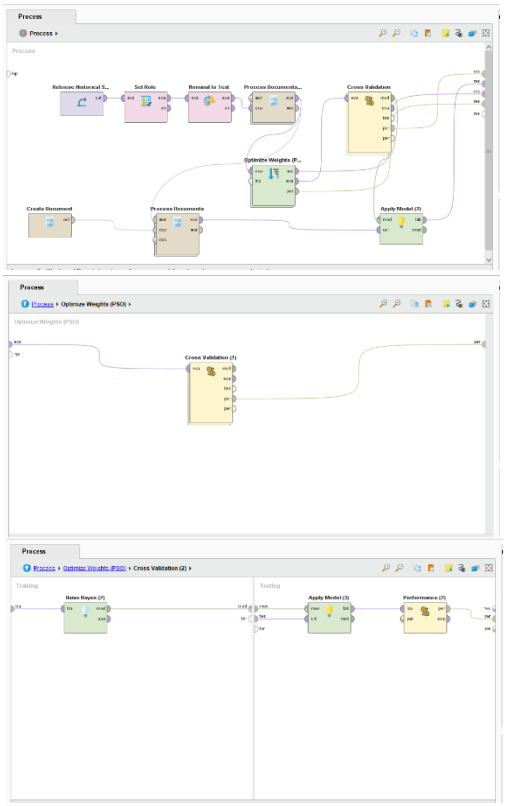
s3	astrazeneca	loyal
switch	next	brand

3.5 Classification

In this classification stage, the method or algorithm used is *Nave Bayes* which is optimized by *Particle Swarm Optimization* (PSO). Calculation of naïve bayes is done by calculating the simple probability of each class from all training data. After that, it is continued with the testing process, which is to determine the accuracy of the model built in the training process. The following is a model of *Naïve Bayes* and *Naïve Bayes* + PSO using RapidMiner software.



Picture 7. *Naïve Bayes* Model



Picture 8. *Naïve Bayes + PSO* Model

3.6 Evaluation

In this stage, a process is carried out to measure the values of accuracy, precision, and recall by using a confusion matrix. The following is the result of the confusion matrix for each type of vaccine.

Table 8. Results of *confusion matrix* AstraZeneca

Confusion Matrix	**NB**	**NB + PSO**
Accuracy	71,39%	79,17%
Precision	80,88%	87,69%
Recall	82,09%	85,07%

Table 9. Results of *confusion matrix* Moderna

Confusion Matrix	**NB**	**NB + PSO**
Accuracy	67,62%	68,82%
Precision	90,76%	92,29%
Recall	71,72%	71,72%

Table 10. Results of *confusion matrix* Pfizer

Confusion Matrix	**NB**	**NB + PSO**
Accuracy	60,94%	67,54%
Precision	69,77%	77,83%
Recall	59,76%	62,95%

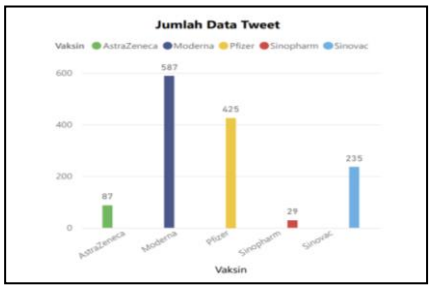
Table 11. Results of *confusion matrix* Sinopharm

Confusion Matrix	**NB**	**NB + PSO**
Accuracy	80,00%	93,33%
Precision	86,36%	91,67%
Recall	86,36%	100,00%

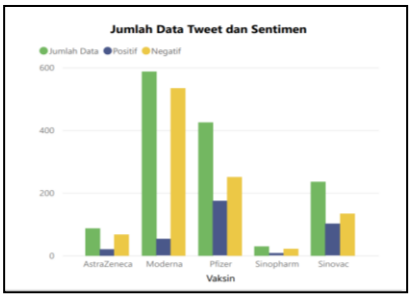
Table 12. Results of *confusion matrix* Sinovac

Confusion Matrix	**NB**	**NB + PSO**
Accuracy	68,04%	74,93%
Precision	72,87%	82,61%
Recall	70,15%	70,90%

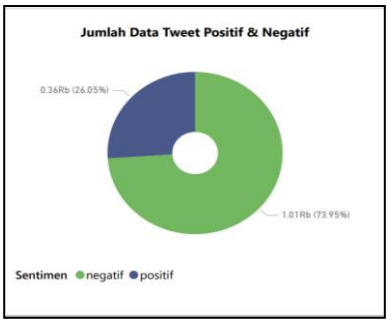
3.7 Visualization



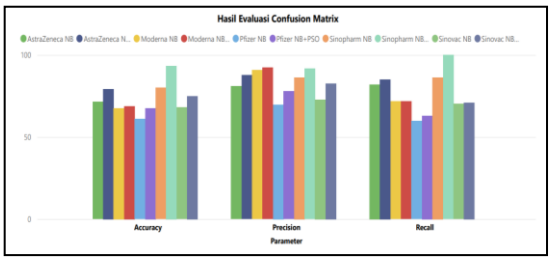
Picture 9. Column diagram of the amount of data for each type of Vaccine



Picture 10. Column chart of the amount of data, positive sentiment and negative sentiment for each type of Vaccine



Picture 11. Circle graph of the total number of positive and negative sentiments on Vaccine data



Picture 12. NB and NB + PSO confusion matrix column diagram for each type of Vaccine

Picture 13. WordCloud visualization of all Vaccine data

4. CLOSING

4.1 Conclusion

Based on research that has been conducted on sentiment analysis of the Covid-19 vaccine type on Twitter social media using the PSO-based *Naïve Bayes* algorithm, the data used include 412 AstraZeneca data, 2000 Moderna data, 1937 Pfizer data, 333 Sinopharm data and 1042 Sinovac data. The results of testing the *Naïve Bayes* algorithm based on PSO using RapidMiner software include: *accuracy* 79.17%, precision 87.69%, recall 85.07% on AstraZeneca vaccine, accuracy 68.82%, precision 92.29%, recall 71.72% on Moderna vaccine, accuracy 67.54%, precision 77.83%, recall 62.95% on Pfizer vaccine, 93.33% accuracy, precision 91.67%, recall 100.00% on Sinopharm vaccine, and accuracy 74.93%, precision 82.61%, recall 70.90% on Sinovac vaccine. This shows that with the help of optimization PSO is proven to be able to produce more accurate confusion matrix values.

4.2 Suggestions

The suggestions given from the results of this study are expected to be developed for further research, and experiments are carried out using algorithms and optimization such as *Support Vector Machine* (SVM), *Decision Tree*, and so on to get different outputs.

BIBLIOGRAPHY

[1] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter," *Heal. Inf. Manag. J. ISSN*, vol. 8, no. 2, pp. 2655–9129, 2020.

[2] J. Hua and R. Shaw, "Corona Virus (COVID-19) ' Infodemic ' and Emerging Issues through a Data Lens : The Case of China," no. February, 2020.

[3] D. Aditia, D. Nasution, F. S. Sains, U. Pembangunan, P. Budi, and U. S. Utara, "DAMPAK PANDEMI COVID-19 TERHADAP PEREKONOMIAN," vol. 5, no. 2, pp. 212–224, 2020.

[4] C. Liu *et al.*, "Research and Development on

Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases," 2020, doi: 10.1021/acscentsci.0c00272.

[5] I. P. Sari and Sriwidodo, "Perkembangan Teknologi Terkini dalam Mempercepat Produksi Vaksin Covid-19," vol. 5, no. 5, pp. 204–217, 2020.

[6] A. Makmun and S. F. Hazhiyah, "TINJAUAN TERKAIT PENGEMBANGAN VAKSIN COVID – 19 Fakultas Kedokteran Universitas Muslim Indonesia Corresponding author e-mail : armanto.makmun@umi.ac.id COVID-19," vol. 13, 2020.

[7] L. J. Sheela, "A Review of Sentiment Analysis in Twitter Data Using Hadoop," vol. 9, no. 1, pp. 77–86, 2016.

[8] S. Redhu, S. Srivastava, B. Bansal, and G. Gupta, "Sentiment Analysis Using Text Mining : A Review," vol. 4, no. 2, pp. 49–53, 2018, doi: 10.11648/j.ijdst.20180402.12.

[9] F. F. Mailo and L. Lazuardi, "Analisis Sentimen Data Twitter Menggunakan Metode Text Mining Tentang Masalah Obesitas di Indonesia," *J. Inf. Syst. Public Heal.*, vol. 4, no. 1, pp. 28–36, 2019.

[10] N. D. Putranti and E. Winarko, "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine," vol. 8, no. 1, pp. 91–100, 2014.

[11] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," vol. 10, pp. 71–76, 2020.

[12] L. Desyanita and A. Wibowo, "Pemodelan Sistem Prediksi Kelayakan Pengajuan Kredit Kepemilikan Rumah Dengan Metode C4.5 Dan Naive Bayes," vol. 13, no. 2, pp. 10–22, 2020.

[13] T. Arifin and D. Ariesta, "PREDIKSI PENYAKIT GINJAL KRONIS MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER BERBASIS PARTICLE SWARM OPTIMIZATION," vol. 13, no. 1, pp. 26–30, 2019.

[14] A. Fathan Hidayatullah and A. Sn, "ISSN: 1979-2328 UPN "Veteran," *Semin. Nas. Inform.*, vol. 2014, no. semnasIF, pp. 115–122, 2014, [Online]. Available: http://www.situs.com.

[15] E. Dragut, F. Fang, and C. Yu, "Stop Word and Related Problems in Web Interface Integration," 2009.

Contac person Author: Rivan Adi Nugraha
Hp : 083816556289
Teguh Iman Hermanto
Hp : 082298234666
Imam Ma'ruf Nugroho
Hp : 085759005755