ISSN: 2477-5126
e-ISSN: 2548-9356

Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis

Styawati[1*)], Auliya Rahman Isnain[2], Nirwana Hendrastuty[3], Lili Andraini[4]
[1,3]Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
[2]Program Studi Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
[4]Program Studi Teknik Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
[1,2,3,4]Jl. ZA. Pagar Alam No.9 -11,Kota Bandar Lampung, Lampung 35132, Indonesia
email: [1]styawati@teknokrat.ac.id, [2]auliyarahman@teknokrat.ac.id, [3]nirwanahendrastuty@teknokrat.ac.id, [4]lili_andraini@teknokrat.ac.id

***Abstract* – Twitter is a social media that is widely used by the public. Twitter social media can be used to express opinions or opinions about an object. This shows that there is a huge opportunity for data sources, so they can be used for sentiment analysis. There are many algorithms for performing sentiment analysis, including Support Vector Machine (SVM) and Naive Bayes (NB). Because of the many opinions regarding the performance of the two methods, the researcher is interested in classifying the data using the SVM and NB methods. The data used in this study is data on public opinion regarding the Covid-19 vaccination policy. The first classification process is carried out by the SVM method using various kernels. After getting the highest accuracy result, then the accuracy result is compared with the accuracy value from the NB method classification results.**

Keywords—SVM, NB, Twitter, Sentiment Analysis, Covid-19 vaccination

I. INTRODUCTION

Today many people express their opinions through social media. Opinions conveyed through social media are more interactive than print media. One of the social media that is widely used today is Twitter. According to We are Social sources in 2020, social media Twitter is ranked fifth in the category of social media that is often used with a 56% percentage of users after Youtube, Whatsapp, Facebook and Instagram. This shows that there is a very large opportunity for data sources, so that it can be used for analyzing one's sentiments towards an object. There are many algorithms to perform sentiment analysis, including: Support Vector Machine (SVM) and Naïve Bayes (NB)[1][2].

Support Vector Machine (SVM) is a superior classification method compared to other classification methods. This method produces 88.52% accuracy when classifying twitter data [3]. In addition, research conducted by [4] also said that SVM can perform data classification very well compared to conventional methods such as artificial neural network methods. The SVM concept can help in finding the best hyperplane that serves as a separator between the two data label classes. The next research was conducted by [5] SVM can work very well in classifying data. This is evidenced by the accuracy value of the results of Twitter data classification with two class labels of 86%. In addition to the SVM method, the Naïve Bayes (NB) method is also said to have good classification capabilities compared to SVM and K-NN. This is evidenced by the accuracy of tweet data using the NB method of 75.58%, SVM of 63.99%, and K-NN of 73.34%[6]. In addition, research conducted by Dinar [7] Twitter data classification using SVM and NB methods produces the highest accuracy between the two methods, namely the NB method, with an accuracy value of 94%. Based on previous research, this study will compare the accuracy values of the results of Twitter data classification. The Twitter data used is public opinion data regarding the Covid-19 vaccination. This data was chosen because the Covid-19 vaccination had become a trending topic on Twitter.

The purpose of this study is to compare the accuracy of the SVM method with various kernels. In addition, this study also aims to compare the accuracy of the SVM method with the NB method. This comparison was conducted to determine the performance of the two classification methods.

II. METHODOLOGY

In this study, the classification process of public opinion data related to Covid-19 vaccination was carried out. The data used is 5000 tweets. The data was taken from March 30 to April 30, 2021. Examples of data can be seen in the Figure 1

Number	Tweet
1	@Reisa_BA Berarti semakin banyak orang yang divaksin mandiri semakin untunglah pemerintah, dana yangg sudah dianggarkan dikurangi jumlah vaksin mandiri = laba. Kemana nanti duitnya ?? +62 ???
2	Baguslah semoga dengan banyaknya masyarakat divaksin vaksin pemerintah hcovid selesai #vaksinasi masal
3	Udah lumayan tenang sekarang setelah divaksin #VaksinUntukKita https://t.co/QGn2tGP25j
4	Pemerintah mendorong kesetaraan akses vaksin #SemangatPemulihanNegeri
5	Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunanan aja #vaksin gagal

Figure 1 Examples of data tweet

The classification process uses two methods, namely Support Vector Machine (SVM) and Naïve Bayes (NB). The stages of the research can be seen in Figure 2

*) **penulis korespondensi**: Styawati
Email: styawati@teknokrat.ac.id

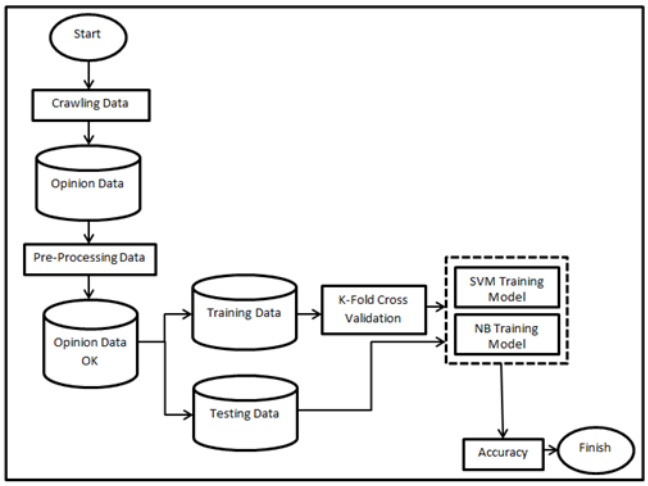

Figure 2 The stages of the research

A. Crawling Data

Crawling data in this study aims to collect data from the Twitter server. Data collection is done by utilizing the Application Programming Interface (API) facility provided by Twitter. the keywords used in the crawling process are "Covid 19 Vaccination" and "Sinovac Vaccines". The data obtained is then stored in a document in the form of .csv. An example of crawled data can be seen in Figure 3.

Baguslah semoga dengan banyaknya masyarakat divaksin vaksin pemerintah hcovid selesai #vaksinasi masal
Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunanan aja #vaksin gagal
Udah lumayan tenang sekarang setelah divaksin #VaksinUntukKita https://t.co/QGn2tGP25j

Figure 3 Example of Crawled Data

B. Preprocessing Data

The purpose of data preprocessing is to clean data, data integration, data transformation, and data reduction. Preprocessing in this study uses five techniques, namely cleansing, tokenization, case folding, stopword removal, and stemming.

1. Cleansing
   Cleansing data is an activity to analyze data quality. This can be done by modifying, changing, or deleting data that is considered unnecessary, incomplete, inaccurate, and has the wrong data or file format in the database. Example "Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja #vaksin gagal" becomes "Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja vaksin gagal".
2. Tokenization
   Tokenization serves to break comments into words. The tokenization process is done by looking at each space in the comment, then based on the space the comments can be split. Example "Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja vaksin gagal" becomes "[Program,vaksin, pemerintah, kacau, dimana, mana, pada, antri, bikin, kerumunan, aja, vaksin, gagal]".
3. Case folding
   Case folding is the process of converting text data sentences into uniforms. Case folding is done by converting text into a standard form, usually lowercase letters or also called lowercase. Example "Program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja vaksin gagal" becomes "program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja vaksin gagal".
4. Stopword
   Stopword is the process of deleting words that are included in the stopword list. Stopwords are common words that appear in large numbers that have a function but have no meaning. The words included in the stopword example are "yang", "atau", and others.
5. Stemming
   Stemming data is a process to filter words that contain conjunctions, pronouns, prepositions, into basic words by eliminating prefixes or suffixes. Example "program vaksin pemerintah kacau dimana mana pada antri bikin kerumunan aja vaksin gagal" becomes "program vaksin pemerintah kacau dimana mana pada antri bikin kerumun aja vaksin gagal".

C. Data validation

The data validation technique used is K-fold cross-validation. K-fold cross-validation is one of the methods used to determine the average success of a system [8]. In this study, the number of folds used is 10 folds.

D. Data Classification based on Support Vector Machine (SVM) algorithm

Support Vector Machine (SVM) is one of the supervised machine learning algorithms that have excellent performance in classifying data [9]. SVM is also said to be a linear classifier that is based on the principle of maximizing margin [10]. SVM uses hyperplane optimally to classify data into two groups in higher dimensional space [11]. Margin is the distance between the hyperplane and the closest data from each class [12][13]. This closest data is called the support vector [14]. The hyperplane is the best separator between two predefined classes [15][16]. The basic principle of SVM is a linear classifier, and then it was developed so that it can work on non-linear problems, namely by incorporating the concept of kernel tricks in high-dimensional workspaces[17]. The SVM kernels used in this research are Linear, Radial Basis Function (RBF), and Polynomial kernels.

The SVM method has the main concept in classifying data, namely finding the best hyperplane to separate between two predetermined classes [18]. The best hyperplane is obtained by maximizing the margin support-vector. The process of maximizing the margin support vector can be done by minimizing the Lagrangian and deriving it from w and b is found in equation 1 with conditions 1 and conditions 2

$$Lp = \|w\|2 - i\, yi\, (w.xi + b) - 1 \quad (1)$$

Condition 1:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (2)$$

Condition 2:

$$b = y_i - w.x_i \quad (3)$$

In the process of maximizing the Lagrangian multiplier, there are still many possible values of w, b, and α. Based on these problems, the process of maximizing the Lagrange multiplier must be transformed to the duality of the Lagrange multiplier in equation 4 with conditions 1 and 2.

$$\text{Max } Ld = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i.x_j} \quad (4)$$

Condition 1:

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \; i = 1,2,...,N \quad (5)$$

Condition 2: (6)

$$0 \leq \alpha i \leq C, \; i = 1,2,...,N$$

E. Data Classification based on Naïve Bayes Method

Naïve Bayes (NB) is a classification method that can predict the probability of a class so that it can produce decisions based on learning data [19]. NB has advantages, among others, simple, fast, and produces high accuracy when applied to large data [19]. In general, the NB classification equation can be seen in equation 7

$$P(Wi|C) = \frac{count(wi\,|c)+1}{count(c)+|V|} \quad (7)$$

III. RESULT AND DISCUSSION

A. Modeling

At this stage, data classification is carried out using the Support Vector Machine (SVM) linear kernel, SVM Polynomial kernel, SVM kernel RBF, and Naïve Bayes (NB) algorithms. The SVM method classification process requires a C parameter and the RBF kernel requires a gamma parameter. Parameters C and Gamma used to refer to research conducted by Styawati [10]. Parameters C and gamma can be seen in table 1.

Table 1 Parameters C and gamma

C	Gamma
2.33	0.45
2.25	0.46
2.13	0.50
1.63	1.08

B. Linear SVM Kernel Modeling

The linear kernel in the SVM method serves to separate data linearly. The source code for the classification process using the Linear kernel can be seen in Figure 4

```
clf = SVC(kernel='linear', C=2.33)
clf.fit(Train_X_Tfidf,Train_Y)
```

Figure 4. Classification of data using SVM kernel Linear

The first line will create a clf variable containing SVC(Support Vector Classifier) with a linear kernel and C=2.33. The results of the classification using the linear kernel SVM method by trying various C values can be seen in table 2

Table 2 Classification results of linear kernel SVM

Kernel	C	Accuracy
LINEAR	2.33	88.3
	2.25	88.2
	2.13	88.2
	1.63	87.8

Based on table 2, it can be seen that the highest accuracy is 88.3. The accuracy is obtained from the use of the value of C=2.33.

C. Polynomial SVM Kernel Modeling

Kernel Polynomial is a kernel function to use when data cannot be separated linearly. The source code for the classification process using the Polynomial kernel can be seen in Figure 5.

```
poly = SVC(kernel='poly', C=2.33)
poly.fit(Train_X_Tfidf,Train_Y)
```

Figure 5. Classification of data using SVM kernel Polynomial

The results of the classification using the Polynomial kernel SVM method by trying various C values can be seen in table 3.

Table 3. Polynomial kernel SVM classification results

Kernel	C	Accuracy
Polynomial	2.33	85.5
	2.25	85.5
	2.13	85.5
	1.63	85.3

Based on table 3, it can be seen that the highest accuracy is 85.5. The accuracy is obtained from the use of parameters C=2.33, C=2.25, C=2.13.

D. SVM Kernel RBF Modeling

The RBF kernel serves to separate data with higher dimensions. The source code for the classification process using the RBF kernel can be seen in Figure 6

```
rbf = SVC(kernel='rbf', C=2.13, gamma=0.50 )
rbf.fit(Train_X_Tfidf,Train_Y)
```

Figure 6. Classification of data using SVM kernel RBF

The results of the classification using the SVM kernel RBF method by trying various C values can be seen in table 4.

Table 4. Classification results of SVM kernel RBF

Kernel	C	Gamma	Accuracy
RBF	2.33	0.45	88.6
	2.25	0.46	88.7
	2.13	0.50	88.8
	1.63	1.08	87.9

Based on table 4, it can be seen that the highest accuracy is 88.8. The accuracy is obtained from the use of parameters C=2.13 and gamma=0.50.

F. Naïve Bayes Modeling

The Naïve Bayes (NB) method is a text classification method based on keyword probabilities in comparing training documents and test documents. The two are compared through several stages of equations, which ultimately results in the highest probability being assigned as a new document category. The source code for the classification process using NB is shown in Figure 7.

```
from sklearn.naive_bayes import GaussianNB
modelnb = GaussianNB()
nbtrain = modelnb.fit(x_train, y_train)
y_pred = nbtrain.predict(x_test)
nbtrain.predict_proba(x_test)
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))
```

Figure 7. Source code of the classification process with the NB method

The accuracy of the NB method is 82.51%. This accuracy is obtained from the Confusion Matrix technique.

G. Comparison of SVM Accuracy With Various Kernels

Comparison of SVM accuracy values with Linear, Polynomial, and RBF kernels is carried out to determine the highest accuracy value of each SVM kernel. The comparison value is shown in Figure 7.

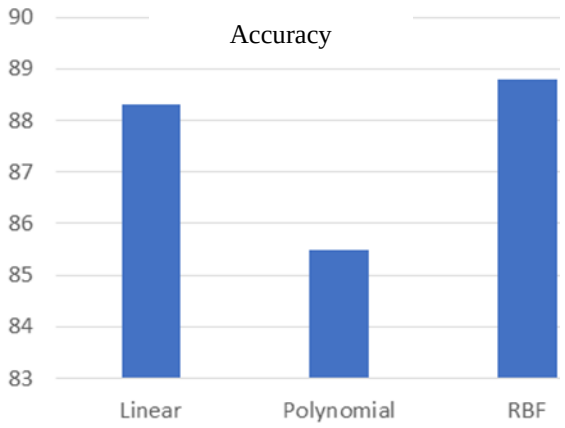

Figure 7. Comparison of SVM accuracy values with various kernels

Based on Figure 7, it can be seen that the highest accuracy is obtained from the SVM method with the RBF kernel. The RBF kernel gets higher accuracy compared to other kernels because the data mapping does not only use the value of the C variable but also considers the value of the gamma variable.

H. Comparison of RBF kernel SVM accuracy with NB

The accuracy value obtained from the data classification results using the SVM kernel RBF method with NB is shown in table 5.

Table 5. Comparison of the Accuracy Value of the SVM Method with NB Method

	SVM Kernel RBF	NB
Accuracy	88.8 %	82.51%

IV. CONCLUSION

The results of this study indicate that SVM with the RBF kernel produces the highest accuracy compared to the Linear kernel and the Polynomial kernel. This is because when mapping data, the RBF kernel considers the value used to find the optimal value in each dataset (gamma). While the results of the comparison of the accuracy of the SVM kernel RBF with NB, the highest accuracy value is obtained from the SVM kernel RBF method, which is 88.8%. This is due to the use of data sets that are not too large. In addition, NB uses probability values in the data classification process.

REFERENCES

[1] D. K. Zala, "A Twitter Based Opinion Mining to Perform Analysis on Network Issues of Telecommunication Companies," pp. 437–441, 2018.

[2] P. Prasetyawan, I. Ahmad, R. I. Borman, Y. A. Pahlevi, and D. E. Kurniawan, "Classification of the Period Undergraduate Study Using Back-propagation Neural Network," in *2018 International Conference on Applied Engineering (ICAE)*, 2018, pp. 1–5, [Online]. Available: file:///C:/Users/CPU/Downloads/Artikel-seminar-Batam-Purwono-Ardiansyah.pdf.

[3] S. P. C. . Sandagiri, B. T. G. . Kumara, and B. Kuhaneswaran, "Detecting Crimes Related Twitter Posts Using SVM based Two Stages Filtering," no. 978, pp. 506–510, 2020.

[4] D. B. Ajipangestu, "Event Classification in Surabaya on Twitter with Support Vector Machine," pp. 482–486, 2021.

[5] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment Analysis on Twitter Posts : An analysis of Positive or Negative Opinion on GoJek," pp. 266–269, 2017.

[6] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler : Twitter," pp. 1–5.

[7] D. Ajeng and L. Marlinda, "Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," no. Citsm, pp. 3–8, 2018.

[8] S. Yadav, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," *2016 IEEE 6th Int. Conf. Adv. Comput.*, no. Cv, pp. 78–83, 2016, doi: 10.1109/IACC.2016.25.

[9] D. A. Kristiyanti, "Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis," pp. 36–42, 2019.

[10] S. Styawati and K. Mustofa, "A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 3, p. 219, 2019, doi: 10.22146/ijccs.41302.

[11] S. Zahoor, "Twitter Sentiment Analysis using Machine Learning Algorithms : A Case Study," pp. 194–199, 2020.

[12] S. Styawati *et al.*, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," vol. 6, no. 3, pp. 150–155, 2021.

[13] R. Joshi, "Comparative Analysis Of Twitter Data Using Supervised Classifiers."

[14] C. F. Chao and M. H. Horng, "The construction of support vector machine classifier using the firefly algorithm," *Comput. Intell. Neurosci.*, vol. 2015, 2015, doi: 10.1155/2015/212719.

[15] A. Kowalczyk, *Support Vector Machines Succinctly*. Syncfusion Inc., 2017.

[16] styawati Styawati, A. Nurkholis, Z. Abidin, and H. Sulistiani, "Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly," *RESTI (Rekayasa Sist. dan Teknol. Inf.)*, vol. 5, no. 10, pp. 904–910, 2021.

[17] B. Luo *et al.*, "Optimized Support Vector Machine Assisted BOTDA for Temperature Extraction With Assisted BOTDA for Temperature," *IEEE Photonics J.*, vol. 12, no. 1, pp. 1–14, 2020, doi: 10.1109/JPHOT.2019.2957410.

[18] W. Jiao, Z. Liu, and Y. Zhang, "Fault Diagnosis of Modular Multilevel Converter with FA-SVM Algorithm," *Chinese Control Conf.*, pp. 5093–5098, 2019.

[19] N. R. Fatahillah, "Implementation Of Naive Bayes Classifier Algorithm On Social Media (Twitter) To The Teaching Of Indonesian Hate Speech," pp. 128–131, 2017.