Jurnal Teknik Informatika Vol. 15 No. 2, 2022 (110-118)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v15i2.24996

JURNAL TEKNIK INFORMATIKA

Homepage: http://journal.uinjkt.ac.id/index.php/ti

Comparison of Support Vector Machines and K-Nearest Neighbor Algorithm Analysis of Spam Comments on Youtube Covid Omicron

Sudianto Sudianto[1*], Juan Arton Masheli[2], Nursatio Nugroho[3], Rafi Wika Ananda Rumpoko[4], Zarkasih Akhmad[5]

[1,2,3,4,5] Informatics Engineering, Telkom Institute of Technology Purwokerto
[1,2,3,4,5] Jl. DI Panjaitan No.128, Purwokerto, Indonesia
E-mail: [1]sudianto@ittelkom-pwt.ac.id, [2]18102199@ittelkom-pwt.ac.id, [3]18102208@ittelkom-pwt.ac.id, [4]18102209@ittelkom-pwt.ac.id, [5]18102217@ittelkom-pwt.ac.id

ABSTRACT

Article:
Accepted: July 10, 2022
Revised: June 08, 2022
Issued: November 15, 2022

***Correspondence Address:**
sudianto@ittelkom-pwt.ac.id

Every time a new variant of Coronavirus (Covid-19) appears, the media or news platforms review to find out whether the new variant is more dangerous or contagious than before. One of the media or platforms that is fast in presenting news in videos is YouTube. YouTube is a social media that can upload videos, watch videos, and comment on the video. The comment field on YouTube videos cannot be separated from spam comments that annoy other users who want to follow or participate in the comment's column. Indication of spam comments is still done by observing one by one; this is very inefficient and time-consuming. This study aims to create a model that can classify spam on YouTube comments. The classification method uses the SVM (Support Vector Machines) algorithm and the KNN (K-Nearest Neighbor) algorithm to identify spam comments or not with comment data taken from Omicron's Covid-19 news video on national news channels. The classification results show that the SVM method is superior inaccuracy with the Linear SVC algorithm of 75.12%, SVC of 76.11%, and Nu-SVC of 77.11%. While the KNN algorithm with k=2 is 65.67%, k=4 is 64.51%, k=6 is 62.35%.

Keywords: *Comment, Covid-19, KNN, spam, SVM, YouTube*

I. INTRODUCTION

Coronavirus (Covid-19) in the last two years has become news in the print and electronic news era. Covid-19 is a virus infection from human to human [1]. The development of Covid-19 has entered a new phase, including the emergence of a more infectious variant, the Omicron variant. The Omicron variant was first reported in Botswana, South Africa, in early November 2021.

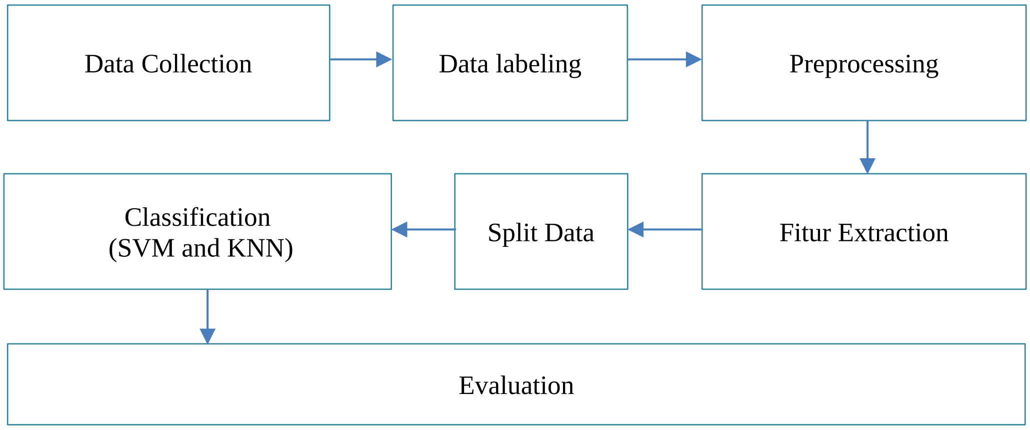


Figure 1. Diagram of research flow

In the enthusiasm for news coverage, especially on social media sourced from videos such as YouTube, netizens often provide opinions in the comment’s column. However, many social media users often harm others by adding irrelevant and unrelated comments to the shared items or posted or so-called spam. Spam is an act that is detrimental to other readers who want to follow the news flow, especially indicating that spam is still often done manually. Therefore, automatic spam classification is needed because classifying comments as spam can improve experience and knowledge in analyzing essential and unimportant information by identifying spam [2], [3].

Previous research on the identification of spam comments was carried out by sentiment analysis about the new variant of Covid-19, the algorithm used with SVM and Naive Bayes to identify spam text messages [4]. Detect spam comments on the Instagram social media platform using the Support Vector Machine (SVM) algorithm [5]. Identification of Indonesian SMS spam using Support Vector Machines [6]. Optimization of K-values for classifying spam and non-spam KNN algorithm [7]. Knowing the accuracy of the Distance Weight K-Nearest Neighbor algorithm for identifying spam comments on Instagram posts [8]. This study uses the SVM and KNN algorithms. SVM is often used to find the one with the best global attributes. Only a small amount of training data is stored by the SVM for predictions. Optimal hyperplane search in the input space functions as a separator of two data classes. Besides that, K-Nearest Neighbor (KNN) is an instance-based learning group. The KNN algorithm is also a form of lazy learning technique. KNN works to find a group of *k* objects in the training dataset that are most comparable or similar to the objects in the new or test data.

Therefore, this study aims to classify spam comments by comparing two different algorithms, namely SVM and KNN. The classification was carried out to identify YouTube comment spam about Covid-19 Omicron. In contrast to previous studies [4], [5], this study divides several tunings on the two algorithms to be compared.

II. METHODOLOGY

This research method consists of several stages described in a flow diagram in Figure 1.

2.1 Data Collection

In this study, the dataset was obtained by scraping comments on YouTube video comments to discuss the spread of the Omicron variant of the COVID-19 virus from national news channels. Data retrieval is done by scraping on YouTube. Based on the scraping of the comment data, data obtained as many as 3014 comments. The comment data from videos and channels, as in Table 1.

2.2 Data Labelling

The comment data that has been obtained is then labeled manually. The data class is divided into two classes. The first class is a class that contains YouTube comments that are not categorized as spam; this class is given a value of "0". The second class is a class that contains YouTube comments that are categorized as spam; this class is given a value of "1". The label data will be a dataset with targets and labels [9], [10].

Table 1. Data source from YouTube channel

YouTube Channel	Video Tittle
KOMPAS TV	Virus Corona Varian Omicron Terdeteksi di Indonesia, Perhatikan Gejalanya
CNN Indonesia	Omicron Menyebar di Jawa Timur, Pemerintah Siapkan Skenario Terburuk
KOMPAS TV	Omicron Masuk Indonesia, Jokowi: Jangan Panik, Taat Prokes, Dapatkan Vaksinasi Covid-19!

2.3 Prepossessing

1. Data Cleaning
   The initial stage in the preprocessing is to do data cleaning. Data cleaning improves datasets by replacing, deleting, or modifying irrelevant or valid data [11]. Benefits of data cleansing: (1) eliminate inconsistencies that appear on datasets; (2) facilitate the processing of data as needed; (3) help map different data functions. The data cleaning carried out on the research dataset includes checking the dataset, deleting null values, and removing brake lines.

2. Case Folding
   Case-folding is the stage of changing all uppercase letters from 'a' to the letter 'z' to lowercase. Deletion of characters other than letters is also carried out at this stage. Some of the ways that case folding is done in this study: (1) changing the text to lowercase; (2) removing numbers; (3) removing punctuation; (4) removing whitespace or blank characters.

3. Stopword Removal
   Stopword removal is the stage of removing words that often appear and are quite common but do not significantly affect the meaning of a sentence or text [12]. Examples of stopwords in Indonesian are “yang”, “dan”, “di”, and “dari”.

4. Tokenization
   Tokenization is breaking or dividing a sentence into a word or token [13]. Tokenization separates text consisting of Words, numbers, symbols, punctuation marks, and other important entities that can be considered tokens. In addition, words, numbers, symbols, punctuation marks, and other important entities can be considered tokens. e.g., "I am a student." Each unit that is contested is usually referred to as a token. into one unit. "I", "am", "a", "student". These units are commonly referred to as tokens.

5. Lemmatization
   Lemmatization is identifying and removing affixes or prefixes and suffixes in a word. Lemma is the basic form of a word with a primary meaning [14]. e.g., the word from "swim", "swimming", "swims", "swam", is all forms of "swim". Well, so the lemma of all those words is "swim".

2.4 Feature Extraction

Feature extraction is the stage to get the characteristics of an object to get an overview of the characteristics of the object. The Bag of Word algorithm [15] was used in this study. Bag of Words is a model that can reflect objects globally. A document or text sentence is a bag (multiset) of words regardless of word order and grammar to maintain word diversity [16]. Bag of Words simplifies text data into vectors for computers to understand by calculating the frequency with which words appear.

2.5 Data Split

Furthermore, the dataset is divided into two parts randomly: training and test data with a ratio of 80:20. Training data is data used to train and create models. At the same time, the test data is the data used to test the model that has been built using the data train [17].

Split is one of the methods that can be used to evaluate the performance of a model. This model evaluation divides the dataset into two parts: the part used for training data and testing data with a certain proportion.

Train data is used to fit the model. Test data is used to evaluate the results of the fit model. An illustration of the split data can be seen in Figure 2. In addition, trains and tests can be used for classification problems [18], [19].

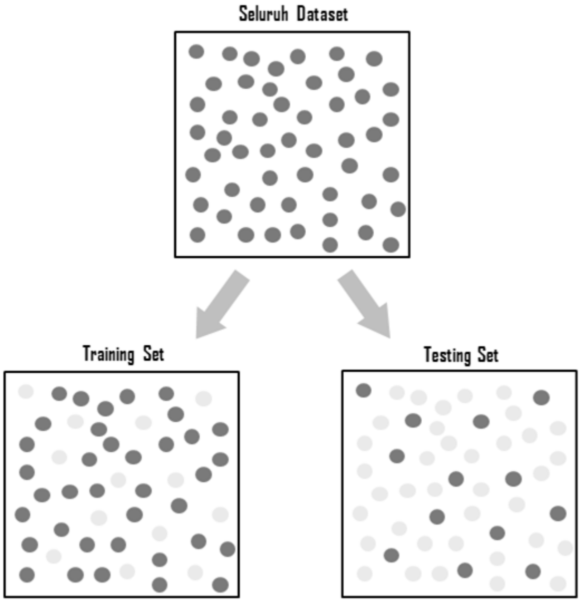


Figure 2. Illustration of data split

2.6 Classification

After passing the preprocessing stage, the next stage is to create a model using two different algorithms. The two algorithms are the SVM (Support Vector Machine) algorithm and the KNN algorithm.

The Support Vector Machine algorithm aims to identify the hyperplane with the most significant margin. The hyperplane is defined as the separator between two classes. In contrast, the margin is defined as the distance between the hyperplane to the nearest data from each class [20]–[22].

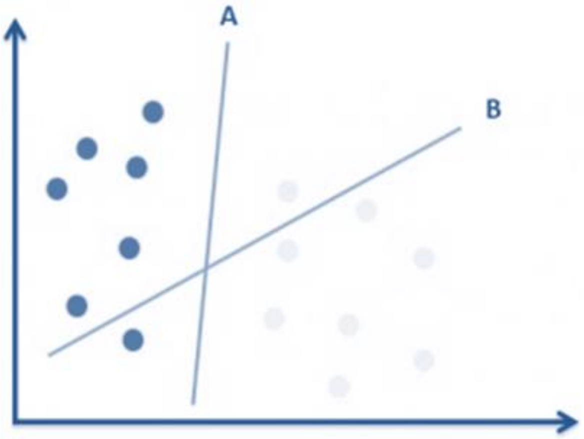


Figure 3. Principle of SVM

It can be seen in Figure 3 that there are two hyperplane lines, namely A and B. It can be concluded that the best hyperplane is hyperplane A. The hyperplane is in the middle between the two classes. The distance between the hyperplane and the data objects is different from the adjacent (outermost) class marked empty and positive round. In SVM, the outer data object closest to the hyperplane is called the support vector. Objects called support vectors are most difficult to classify because of the almost overlapping positions with other classes. Given its critical nature, only this support vector is considered to find the most optimal hyperplane by SVM. Determining the farthest distance is repeated until it finds the best hyperplane [23]. So, optimization is needed on the SVM to find the maximum length of the hyperplane [9], [24], [25]. On the other hand, the SVM algorithm is reliable for conducting classification analysis [26].

In the KNN algorithm, the working principle of KNN is to measure the distance between the data we want to classify and the existing dataset. The KNN takes several *k* data nearby (its neighbors) to determine the new data class. This algorithm classifies data based on similarity or proximity to other data [27].

In general, how the KNN works [7], [28]: (1) Determine the number of neighbors (*k*) for class determination considerations. (2) Calculate the distance from the new data to each data point in the dataset. (3) Take several K data with the closest distance and determine the new data class.

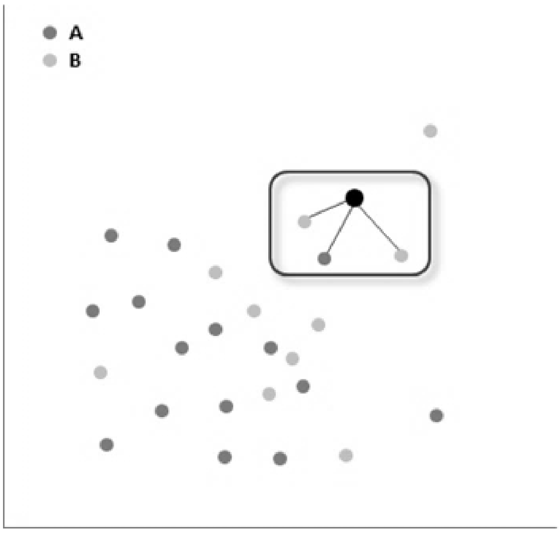


Figure 4. Principle of K-Nearest Neighbor

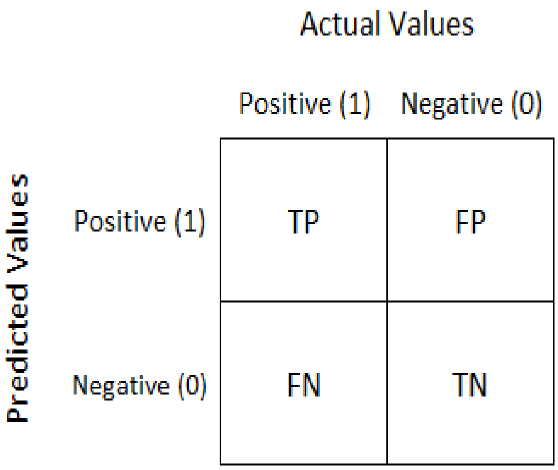
Calculating the distance between two points in the KNN algorithm uses the Euclidean

Distance method, which can be used in 1-dimensional space, 2-dimensional space, or multi-dimensional space. 1-dimensional space means distance calculation using only one independent variable. 2-dimensional-space means there are two free variables, and multi-dimensional space means more than two variables.

In Figure 4, some data is divided into two classes: blue and yellow. If we want to classify new data, it is shown in black in Figure 4, the value of *k*=3. After calculating the distance between the new data and the nearest surrounding data, the three closest data consist of blue and yellow data. Yellow data is more dominant because of the three closest data, so the new data is classified as yellow data.

2.7 Evaluation

After passing through several stages, the last stage is model evaluation. Evaluation model using a confusion matrix. The evaluation aims to find out how well the performance of the model that has been built is. The Confusion Matrix represents the predictions and actual conditions of the data generated by the algorithm. Confusion Matrix can specify Accuracy, Precision, and Recall [29].



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 5. Confusion Matrix

Figure 5 is a form of the confusion matrix. Confusion matrix has four terms: (1) True Negative (TN), the model predicts the data is in the Negative class, and the actual data does exist in the Negative class; (2) True Positive (TP), the model predicts the data is in a Positive class, and the actual data does exist in the Positive class; (3) False Negative (FN), the model predicts the data is in the Negative class, but the actual data is in a Positive class; (4) False Positive (FP), the model predicts the data is in a Positive class, but the actual data is in the Negative class. In addition, to make it easier to understand the confusion of the matrix, that is, if it starts with TRUE, the prediction is correct. If it starts with FALSE, then Positive and Negative are the results of predictions stating that the predictions are wrong.

In conducting an evaluation of the confusion matrix [30], [31], performance measurements are divided into several evaluations:

1. Accuracy is the ratio of correct predictions (positive and negative) to the overall data. Accuracy answers the question, "What percentage of conditions are correctly predicted spam and not spam". The Accuracy value is obtained by equation 1.

$$Accuracy=(TP+TN)/(TP+FP+FN+TN) \quad (1)$$

2. Precision is the ratio of correct positive predictions compared to the overall positive predicted results. Precision answers the question, "What percentage of the correct condition is spam from all spammy-predicted comments". The Precision value is obtained by equation 2.

$$Precision = (TP) / (TP + FP) \quad (2)$$

3. Recall is the ratio of correct positive predictions compared to the overall positive correct data. Recall answers the question, "What percentage of comments are predicted to be spam compared to overall comments that are spam". The Recall value is obtained by equation 3.

$$Recall = (TP) / (TP + FN) \quad (3)$$

4. F1-Score is a weighted comparison of precision and recall averages. The value of F1-Score is obtained by equation 4.

$$F1\ Score = 2 \times (Recall \times precision) / (Recall + precision) \quad (4)$$

III. RESULTS AND DISCUSSION

After the datasets that have been labeled, there are 3014 data consisting of 57.9% or 1745 spam comment data and 42.1% or 1268 non-spam comment data. Details of the distribution of data can be seen in Figure 6.

A model is then made using two different algorithms from the dataset, namely SVM and KNN. Three SVM algorithms are used: Linear SVC, SVC, and Nu-SVC. At the same time, the KNN algorithm is carried out three times, with the value of *k* being multiples of 2, 4, and 6. Based on the results of the evaluation of the model, the results are obtained in Table 2.

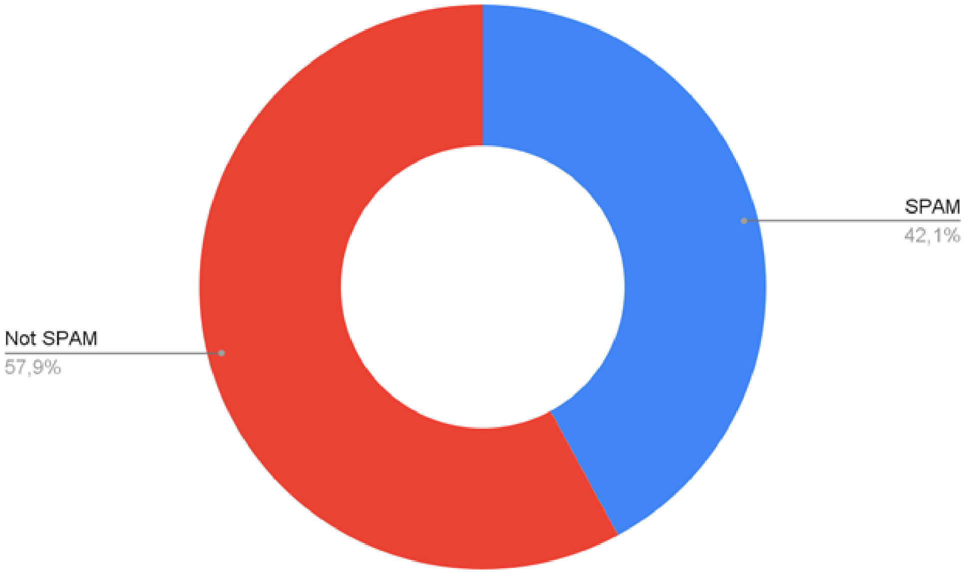


Figure 6. Distribution of data on spam and non-spam classes

Table 2. The evaluation of each model

Method	Precision	Recall	F1-Score	Accuracy (%)	Times (second)
Linear SVM	0.75	0.74	0.74	75.12	0.03
SVM	0.76	0.74	0.75	76.11	0.76
Nu-SVM	0.77	0.75	0.76	77.11	0.84
KNN (*k*=2)	0.67	0.61	0.60	65.67	0.0019
KNN (*k*=4)	0.69	0.59	0.56	64.51	0.0031
KNN (*k*=6)	0.68	0.56	0.50	62.35	0.0011

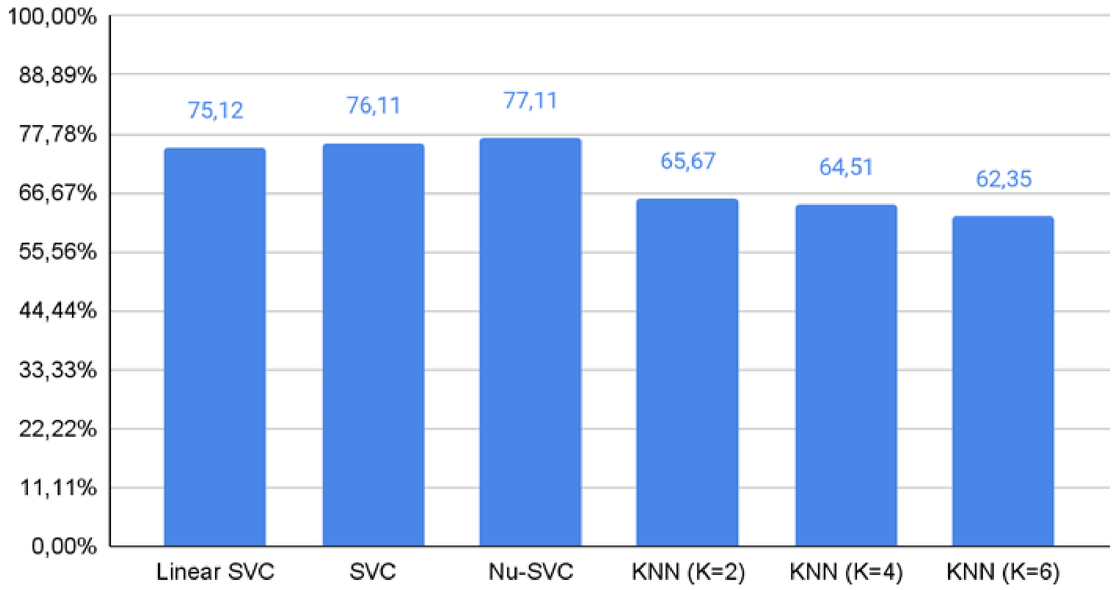


Figure 7. The results of the accuracy of each model

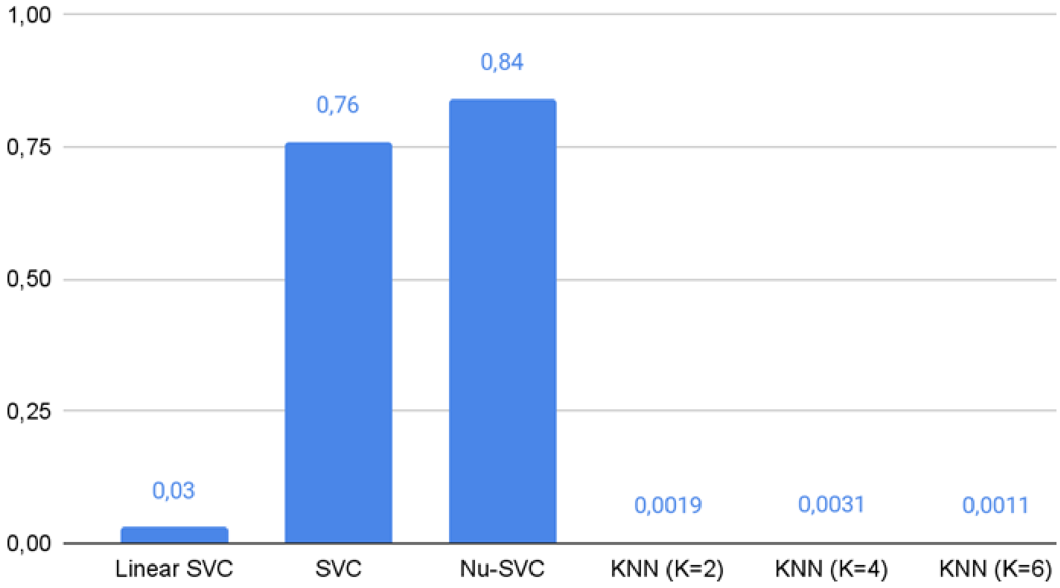


Figure 8. Diagram of computational process times

In Table 2, it can be seen that the three SVM algorithms excel in precision, recall, and F1-Scores. The values of the three variables tend to be stable, with a range of values between 0.74 to 0.76. In contrast, the KNN algorithm has a lower value with an unstable value range between 0.50 to 0.69.

Figure 7 shows that the Nu-Support Vector Classifier method has the highest accuracy of 77.11%. At the same time, the KNN algorithm generates the lowest accuracy with a value of *k*=6. The results in Figure 7 show that the three SVM algorithms are superior in accuracy compared to the KNN algorithm with a considerable distance.

In addition, Figure 8 shows that the KNN algorithm with a value of *k*=6 has the fastest time of 0.0011 seconds. At the same time, the delay time is obtained from the Nu-SVC algorithm with a time of 0.84 seconds. The KNN algorithm is much faster than the SVM algorithm in this test but has lower accuracy. The SVM algorithm has a slower time than KNN but has better accuracy results.

IV. CONCLUSSION

The two algorithms used in this study are quite different. The SVC and Nu-SVC algorithms can produce a reasonably high accuracy value but are relatively slow. In contrast to the KNN algorithm, which can work much faster than the SVC and Nu-SVC algorithms. However, it has a much lower accuracy value. The accuracy of the KNN algorithm continues to decrease as the value of *k* is determined. The Linear SVC algorithm has an accuracy value different from the SVC and Nu-SVC algorithms. However, it has a much faster time than the other two SVM algorithms. However, it is not as fast as the KNN algorithm.

REFERENCES

[1] K. K. Sahu, A. K. Mishra, and A. Lal, “Comprehensive update on current outbreak of novel coronavirus infection (2019-nCoV),” *Ann. Transl. Med.*, vol. 8, no. 6, pp. 393–393, 2020, doi: 10.21037/atm.2020.02.92.

[2] Arif Siswandi, A. S. Sunge, and R. Y. Wulandari, “Analisa Data Mining dengan Metode Klasifikasi untuk Produk Cacat pada PT Shuangying International Indonesia,” vol. 76, no. 1, pp. 57–58, 2018.

[3] R. Feldman, “Techniques and applications for sentiment analysis,” *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013, doi: 10.1145/2436256.2436274.

[4] G. A. Sandag, R. J. Sambur, and J. Bororing, “Klasifikasi Sms Spam Menggunakan Support Vector Machine,” *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 275–280, 2018, doi: 10.33480/pilar.v15i2.693.

[5] A. T. Syam, A. Fahri, B. Saputra, and R. F. Maulana, “Klasifikasi Komentar Spam pada Instagram menggunakan Metode Support Vector Machine,” *Angew. Chemie Int. Ed. 6(11), 951–952.*, vol. 6, pp. 5–24, 2020.

[6] R. Mahardika, “Identifikasi SMS spam berbahasa Indonesia menggunakan Algortima Support Vector Machine,” *Repos. Institusi USU*, 2018.

[7] E. Laksono, A. Basuki, and F. Bachtiar, “Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 377–383, 2020, doi: 10.29207/resti.v4i2.1845.

[8] A. R. Chrismanto, Y. Lukito, and A. Susilo, “Implementasi Distance Weighted K-Nearest Neighbor Untuk Klasifikasi Spam & Non-Spam Pada Komentar Instagram,” *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 236, 2020, doi: 10.26418/jp.v6i2.39996.

[9] I. Muhammad and Z. Yan, “Supervised Machine Learning Approaches: a Survey,” *ICTACT J. Soft Comput.*, vol. 05, no. 03, pp. 946–952, 2015, doi: 10.21917/ijsc.2015.0133.

[10] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2007, doi: 10.1007/s10751-016-1232-6.

[11] R. Patel and K. Passi, “Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning,” *IoT*, vol. 1, no. 2, pp. 218–239, 2020, doi: 10.3390/iot1020014.

[12] A. F. Hidayatullah, “Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016.

[13] Y. Puspitarani and Y. Syukriyah, “Pemanfaatan Optical Character Recognition Dan Text Feature Extraction Untuk Membangun Basisdata Pengaduan Tenaga Kerja,” vol. 1, no. 3, pp. 704–710, 2020.

[14] M. Yusup, P. Jasman, and D. Renita, “Penerapan Algoritma Lemmatization pada Dokumen Bahasa Indonesia,” *MIND J.*, vol. 3, no. 2, pp. 47–56, 2018.

[15] Satria D dan Mushthofa, “Perbandingan Metode Ekstraksi Ciri Histogram dan PCA untuk Mendeteksi Stoma pada Citra Penampang Daun Freycinetia Comparison of Histogram and PCA as Feature Extraction Methods in Detecting Stoma in Freycinetia Leaf Images Abstrak,” *J. Ilmu Komput. Agri-Informatika*, vol. 2, no. 1, pp. 20–28, 2013, [Online]. Available: http://journal.ipb.ac.id/index.php.jika%0AVolume

[16] T. Mardiana and R. D. Nyoto, “Kluster

Bag of Word Menggunakan Weka," *J. Edukasi dan Penelit. Inform.*, vol. 1, no. 1, pp. 1–5, 2015, doi: 10.26418/jp.v1i1.10145.

[17] J. W. G. Putra, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning," *Tokyo. Jepang*, 2019.

[18] Sudianto, Y. Herdiyeni, A. Haristu, and M. Hardhienata, "Chilli quality classification using deep learning," 2020. doi: 10.1109/ICOSICA49951.2020.9243176.

[19] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *J. Transform.*, vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.

[20] C. X. Zhang, J. S. Zhang, and G. Y. Zhang, "An efficient modified boosting method for solving classification problems," *J. Comput. Appl. Math.*, vol. 214, no. 2, pp. 381–392, 2008, doi: 10.1016/j.cam.2007.03.003.

[21] H. J. Cho and M. Te Tseng, "A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation," *Math. Comput. Model.*, vol. 58, no. 1–2, pp. 438–448, 2013, doi: 10.1016/j.mcm.2012.11.003.

[22] A. Kulsumarwati, I. Purnamasari, and B. A. Darmawan, "Penerapan SVM dan Information Gain Pada Analisis Sentimen Pelaksanaan Pilkada Saat Pandemi," *J. Teknol. Inform. dan Komput.*, vol. 7, no. 2, pp. 101–109, 2021, doi: 10.37012/jtik.v7i2.641.

[23] J. T. Lalis, "A new multiclass classification method for objects with geometric attributes using simple linear regression," *IAENG Int. J. Comput. Sci.*, vol. 43, no. 2, pp. 198–203, 2016.

[24] C. Science, "Efficient Classification Algorithms using SVMs for Large Datasets Master of Technology," no. June, 2007.

[25] A. Amrani, Lazaar, and E. Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *The First International Conference on Intelligent Computing in Data Sciences*. pp. 511–520, 2018.

[26] S. Sudianto, P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "Comparison Of Random Forest And Support Vector Machine Methods On Twitter Sentiment Analysis (Case Study : Internet Selebgram Rachel Vennya Escape From Quarantine) Perbandingan Metode Random Forest Dan Support Vector Machine Pada Analisis Sentimen Twitt," *Jutif*, vol. 3, no. 1, pp. 141–145, 2022.

[27] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[28] Indrayanti, D. Sugianti, and M. A. Al Karomi, "Optimasi Parameter K pada Algoritme K-Nearest Neighbour untuk Klasifikasi Penyakit Diabetes Mellitus," pp. 823–829, 2017.

[29] F. Valencia, A. Gómez-Espinosa, and B. Valdés-Aguirre, "Price movement prediction of cryptocurrencies using sentiment analysis and machine learning," *Entropy*, vol. 21, no. 6, 2019, doi: 10.3390/e21060589.

[30] D. Z. Nathania, Indrianti, and F. A. Bachtiar, "Klasifikasi Spam Pada Twitter Menggunakan Metode Improved K-Nearest Neighbor."

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