ILKOM Jurnal Ilmiah Vol. 14, No. 1, April 2022, pp. 32-38
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Comparison of support vector machine and XGBSVM in analyzing public opinion on Covid-19 vaccination

Rahmaddeni [a,1,*]; M. Khairul Anam [a,2]; Yuda Irawan [b,3]; Susanti [a,4]; Muhamad Jamaris [a,5]
[a] *STMIK Amik Riau, Jl. Purwodadii Indah Km 10 Panam, Pekanbaru and 28294, Indonesia*
[b] *Universitas Hang Tuah Pekanbaru, Jl. Mustafa Sari No. 05 Tangkerang Selatan, Pekanbaru 28000, Indonesia*
[1] *rahmaddeni@sar.ac.id;* [2] *khairulanam@sar.ac.id;* [3] *yudairawan89@gmail.com;* [4] *susanti@stmik-amik-riau.ac.id;* [2] *muhamadjamaris@sar.ac.id*
** Corresponding author*

Article history: Received January 06, 2022; Revised March 26, 2022; Accepted March 27, 2022; Available online April 30, 2022

Abstract

The coronavirus has become a global pandemic and has spread almost all over the world, including Indonesia. The spread of COVID-19 in Indonesia causes many negative impacts. Therefore, the government took vaccination measures to suppress the spread of COVID-19. The public's response to vaccination was quite diverse on Twitter, some were supportive, and some were not. The data used in this study came from Twitter which was taken using the emprit drone portal by using the keyword, "vaccination." The classification is conducted using the SVM and hybrid methods between SVM and XGBoost or what is commonly called XGBSVM. The purpose of this study is to provide an overview to the public on whether the Covid-19 vaccination tends to create positive, neutral, or negative opinions. The results of the sentiment evaluation show that SVM has the highest accuracy of 83% with 90:10 data splitting. On the other hand, the XGBSVM produces 79% accuracy with 90:10 data splitting.

Keywords: sentiment analysis; vaccination; covid-19; SVM; XGBSVM.

Introduction

Sentiment analysis is a process of automatically extracting, processing and understanding data in the form of unstructured text to retrieve sentiment information in a sentence of opinion [1]. Sentiment analysis can be applied to opinion in all fields, such as economics, politics, society and law. The social media Twitter allows researchers to study emotions, moods, and public opinion through sentiment analysis [2]. To perform sentiment analysis, several methods can be used, including the Support Vector Machine (SVM) Algorithm, Naïve Bayes Classifier (NBC), K-Nearest Neighbor (KNN), Decision Tree, and so on [3].

Previous researchers used sentiment analysis to assess the opinion and tendency of an opinion on a topic, whether negative, neutral or positive [4][5]. The data obtained to perform sentiment analysis is usually obtained from social media Twitter [6][7][8]. Twitter is used because most popular or trending news comes from it [9]. Furthermore, one of the news that has been trending in Indonesia is about the covid 19 vaccination.

Currently, several types of COVID-19 vaccines are available in Indonesia, including Astra Zeneca, China National Pharmaceutical Group Corporation (Sinopharm), Moderna, Pfizer-BioNTech, and Sinovac Biotech Ltd [10]. Covid-19 vaccination in Indonesia still becomes a pro and con [11]. In this study, we will analyze sentiment on Twitter regarding the covid 19 vaccination in Indonesia.

The method used is the Support Vector Machine (SVM) Algorithm and a hybrid between SVM and XGBoost, commonly called XGBSVM. These two methods are often used to perform sentiment analysis because they have fairly high accuracy. Research on the use of the XGBoost method conducted by Kaddafi, Rohmat, and Vandha [12] resulted in an accuracy of 90.10%. Another study [13] using the support vector machine algorithm resulted in an accuracy of 92.67%.

http://dx.doi.org/10.33096/ilkom.v14i1.1090.32-38

This comparative research is to get the best accurate results from the two methods. To improve this accuracy, this research combines it with the XGBoost algorithm. XGBoost is a decision tree or regression tree-based boosting algorithm [14]. Previous research [15] performed a hybrid SVM with XGBoost and named it XGBSVM.

To produce the best accuracy of the method used, his study conducted several trials by comparing test data and training data, 70:30, 80:20, and 90:10 respectively. It can be seen from some of these trials that the method produces the highest accuracy in sentiment analysis of the COVID-19 vaccination in Indonesia.

Method

This study uses a research methodology flow to facilitate research work shown in **Figure 1**.

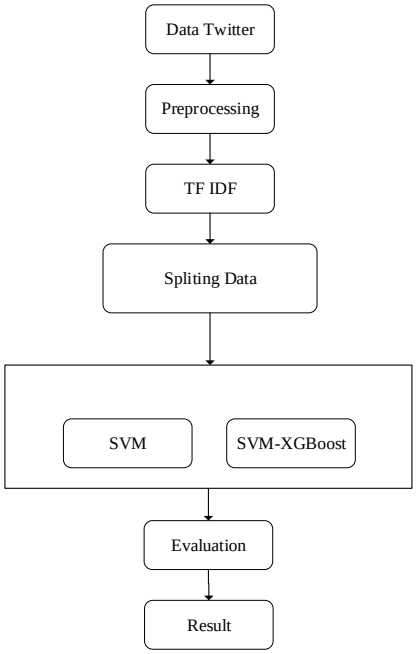

Figure 1. The flow of Research Methodology

The following is an explanation of the research methodology.

A. Preprocessing

The preprocessing stage is the initial stage of cleaning up unnecessary words or words that have no meaning. The whole process is done using the python3 program, so it's done automatically. There are several stages used in this preprocessing stage.

1. Case Folding

Case Folding is a process in text preprocessing that is carried out to uniform the characters in the data? The case folding process is the process of converting all letters into lowercase letters [16][17]. This is done so that all words can be uniform. Examples of case folding applications include the word 'Online' becomes 'online,' 'That' becomes 'that.'

2. Cleaning

Cleaning is a process to remove punctuation marks, numbers, symbols, URL links, and usernames in the text [18]. The frequent appearance of symbols, punctuation marks, and numbers in public comments make the data ineffective and meaningless.

3. Tokenizing

Tokenizing is the process to break the document text into sentences and into words [19]. This stage is the process of grouping the text's contents, which were originally in the form of sentences to become units of words.

4. Filtering

Filtering is the process of removing meaningless words in the document [20]. List of unimportant words like 'and, but, that, why, like, which' and so on.

5. Stemming

Stemming is a process to find the root word of each word in the document by removing affixes, both prefixes, and suffixes [20]. This process returns the word to its basic word by removing the affixes in front and after the words, for example, 'troublesome' becomes 'repot,' 'understanding' becomes 'understand'.

B. TF-IDF

TF IDF is done after the preprocessing stage. In this word weighting, every word that has passed the preprocessing process was parsed first and stored in the database [21].

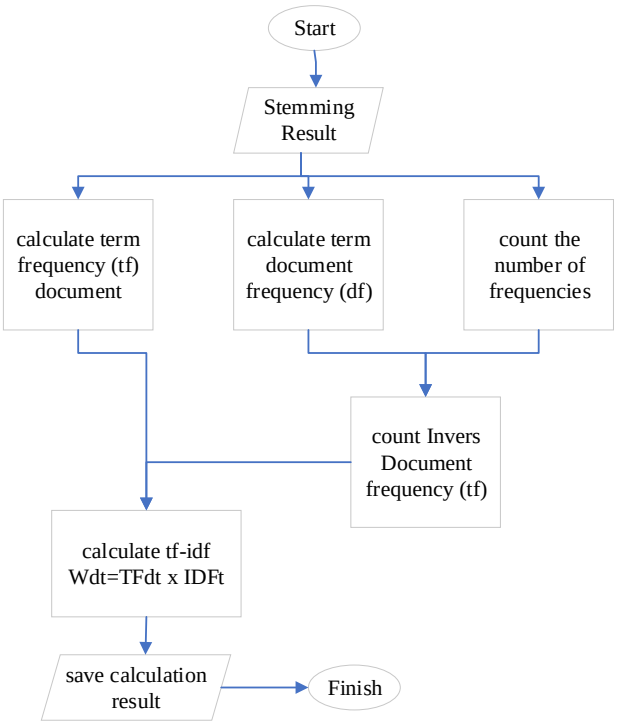

Figure 2. Flowchart TF-IDF [21]

Figure 2 illustrates the word weighting stage using the term frequency-inverse document frequency (TF-IDF) method, where the list of stemming terms was calculated to determine the weight of words by calculating the number of term frequency documents (tf) first, then calculating the value of the number of documents that have a term (df), and then calculate the IDF value with the formula log=N/df, where N is the number of all existing documents. After the TF and IDF values had been obtained, the last step was to determine the word weight by multiplying TF and IDF with the formula Wdt=TFdt x IDFt. The results of this calculation process were stored in the database and were continued with the next stage to calculate the cosine similarity, which was the final stage of the process.

C. Split data

In this study, the data split started from 90:10, 80:20, and 70:30. This data split aimed to perform the best accuracy results.

D. Evaluation

The evaluation was used to see the results of accuracy, recall, and f1-score.

Results and Discussion

The following results were produced in this study using multiple data splits using the Support Vector Machine and XGBoost-Support Vector Machine (XGBSVM) methods.

A. Support Vector Machine (SVM)

Figure 3 is the result of a 70:30 data split using SVM.

	precision	recall	f1-score	support
Negative	0.77	0.50	0.61	72
Neutral	1.00	0.25	0.40	8
Positive	0.75	0.93	0.83	130
accuracy			0.76	210
macro avg	0.84	0.56	0.61	210
weighted avg	0.77	0.76	0.74	210

Figure 3. Split Data 70:30 SVM

From the 70:30 data split, the accuracy result is 78%. This result is relatively high, although it is still low compared to other studies such as [22]. Then in **Figure 4** is an 80:20 data split.

	precision	recall	f1-score	support
Negative	0.56	0.59	0.57	46
Neutral	0.40	0.67	0.50	3
Positive	0.79	0.76	0.78	91
accuracy			0.70	140
macro avg	0.59	0.67	0.62	140
weighted avg	0.71	0.70	0.70	140

Figure 4. Split Data 80:20 SVM

Figure 4 shows that the accuracy decreased from 0.2% to 76%. In precision, recall, and f1 scores were decreased compared to the previous split data. The following experiment used a 90:10 data split.

	precision	recall	f1-score	support
Negative	0.76	0.62	0.68	21
Neutral	1.00	0.33	0.50	3
Positive	0.85	0.96	0.90	46
accuracy			0.83	70
macro avg	0.87	0.64	0.69	70
weighted avg	0.83	0.83	0.82	70

Figure 5. Split Data 90:10 SVM

Figure 5 shows an increase of 0.4% accuracy, which is 83%. This indicates that the fewer data set testing is used, the higher the accuracy results produced. **Table 1** is the result of a comparison of split data using the Support Vector Machine (SVM) method. **Table 1** shows that the highest data splitting result in the SVM method is 90:10 with an accuracy of 83%.

Table 1. Comparison of data splitting with the SVM method

Accuracy	
Split Data	**SVM**
70 : 30	78%
80 : 20	76%
90 :10	83%

B. XGBSVM

Figure 6 is the result of a 70:30 data split.

	precision	recall	f1-score	support
Negative	0.51	0.63	0.56	59
Neutral	0.38	0.60	0.46	5
Positive	0.83	0.74	0.78	146
accuracy			0.70	210
macro avg	0.57	0.66	0.60	210
weighted avg	0.73	0.70	0.71	210

Figure 6. data split 70:30 XGBSVM

The resulting accuracy of XGBSVM yielded an accuracy of 70%. This is low compared to SVM alone which produced a higher yield of 0.6. The next step was to do an 80:20 data split.

	precision	recall	f1-score	support
Negative	0.56	0.59	0.57	46
Neutral	0.40	0.67	0.50	3
Positive	0.79	0.76	0.78	91
accuracy			0.70	140
macro avg	0.59	0.67	0.62	140
weighted avg	0.71	0.70	0.70	140

Figure 7. Split Data 80:20 XGBSVM

Figure 7 shows that the results still produced the same accuracy as the 70:30 data split. Then the next experiment used 90:10 data splitting.

	precision	recall	f1-score	support
Negative	0.76	0.62	0.68	26
Neutral	0.67	1.00	0.80	2
Positive	0.80	0.88	0.84	42
accuracy			0.79	70
macro avg	0.74	0.83	0.77	70
weighted avg	0.78	0.79	0.78	70

Figure 8. Split Data 90:10 XGBSVM

Figure 8 shows that using lower testing data resulted in high accuracy compared to the others. The following is a comparison of split data using XGBSVM. **Table 2** shows the same thing as the previous comparison using SVM. Using less testing data result in a fairly significant increase in accuracy, which is around 0.9.

Table 2. Comparison of data splitting with XGBSVM

Accuracy	
Split Data	**SVM**
70 : 30	70%

Accuracy	
Split Data	**SVM**
80 : 20	70%
90 :10	79%

Conclusion

After completing the sentiment analysis stage with the object of Covid-19 vaccination in Indonesia on Twitter with a total of 700 tweets, it can be concluded that three data splits, 70:30, 80:20, and 90:10, have different accuracy. The SVM method with 90:10 data splits has the highest accuracy compared to other data splits, 83%, compared to the other two data splits, such as 80:20 data split at 76%, and 70:30 data splits with 78% accuracy. Then, when the SVM method was combined with the XGBoost method called the XGBSVM method, it produced slightly decreased accuracy for the three data splits. The results for splitting 90:10 received 79% accuracy, splitting 80:20 and 70:30 data produced the same accuracy of 70%. This proves that in this study the SVM method was not be able to show optimal performance even though it had been hybridized with the XGBoost (XGBSVM) method.

Therefore, it is necessary to improve the SVM method by using feature selection such as the filter method, Wrapper method, and embedded method. In addition, it can also be compared with other methods such as nave Bayes, random forest, KNN, and other methods. Eventually, it can be concluded that such a method is the best for conducting sentiment analysis on the object of vaccination in Indonesia.

References

[1] B. Brahimi, M. Touahria, and A. Tari, "Improving sentiment analysis in Arabic: A combined approach," J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 10, pp. 1242–1250, 2019, doi: 10.1016/j.jksuci.2019.07.011.

[2] J. Qiu, Z. Lin, and Q. Shuai, "Investigating the opinions distribution in the controversy on social media," Inf. Sci. (Ny)., vol. 489, pp. 274–288, 2019, doi: 10.1016/j.ins.2019.03.041.

[3] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, "Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah Applications of Naïve Bayes Classifier , K-Nearest Neighbor and Decision Tree to Analyze Sentiment on Netizen and Gove," Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput. ?141, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[4] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," J. Teknol. Inf. dan Ilmu Komput., vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[5] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," J. EECCIS, vol. 6, no. 1, pp. 37–43, 2012.

[6] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," J. Sist. dan Teknol. Inf., vol. 8, no. 2, pp. 44–51, 2020, doi: 10.26418/justin.v8i2.36776.

[7] N. T. Romadloni, I. Santoso, and S. Budilaksono, "Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl," J. IKRA-ITH Inform., vol. 3, no. 2, pp. 1–9, 2019.

[8] M. Nurmalasari, N. A. Temesvari, and S. N. Maulana, "Analisis Sentimen terhadap Opini Masyarakat dalam Penggunaan Mobile-JKN untuk Pelayanan BPJS Kesehatan Tahun 2019," Indones. Heal. Inf. Manag. J., vol. 8, no. 1, pp. 35–44, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/208.

[9] C. Juditha, "Fenomena Trending Topic Di Twitter: Analisis Wacana Twit #Savehajilulung," J. Penelit. Komun. dan Pembang., vol. 16, no. 2, pp. 138–154, 2015, doi: 10.31346/jpkp.v16i2.1353.

[10] R. N. Rahayu and Sensusiyati, "Vaksin covid 19 di indonesia : analisis berita hoax," Intelektiva J. Ekon. Sos. Hum. Vaksin, vol. 2, no. 07, pp. 39–49, 2021.

[11] 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., vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175.

[12] M. K. Nasution, R. R. Saedudin, and V. P. Widartha, "PERBANDINGAN AKURASI ALGORITMA NAÏVE BAYES DAN ALGORITMA," in e-Proceeding of Engineering, 2021, vol. 8, no. 5, pp. 9765–9772.

[13] M. F. Al-shufi and A. Erfina, "Sentimen Analisis Mengenai Aplikasi Streaming Film Menggunakan Algoritma Support Vector Machine Di Play Store," in SISMATIK, 2021, pp. 156–162.

[14] R. Siringoringo, R. Perangin-angin, and M. J. Purba, "Segmentasi Dan Peramalan Pasar Retail Menggunakan Xgboost Dan Principal Component Analysis," METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt., vol. 5, no. 1, pp. 42–47, 2021, doi: 10.46880/jmika.vol5no1.pp42-47.

[15] W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou, and W. Cao, "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," IEEE Access, vol. 7, pp. 175248–175258, 2019, doi: 10.1109/ACCESS.2019.2957367.

[16] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[17] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," JATISI (Jurnal Tek. Inform. dan Sist. Informasi), vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[18] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," in Jurnal Media Informatika Budidarma, 2021, vol. 5, no. 2, p. 406, doi: 10.30865/mib.v5i2.2835.

[19] E. S. Romaito, M. K. Anam, Rahmaddeni, and A. N. Ulfah, "Perbandingan Algoritma SVM Dan NBC Dalam Analisa Sentimen Pilkada Pada Twitter," CSRID J., vol. 13, no. 3, pp. 169–179, 2021, doi: 10.22303/csrid.13.3.2021.169-179.

[20] P. M. Prihatini, "Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia," J. Matrix, vol. 6, no. 3, pp. 174–178, 2016.

[21] R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," J. Tek. Inform., vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.

[22] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," J. Tek. Inform. dan Sist. Inf., vol. 2, no. 2, pp. 244–253, 2016, doi: 10.28932/jutisi.v2i2.476.

[23] F. Romadoni, Y. Umaidah, and B. N. Sari, "Text Mining Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine," J. Sisfokom (Sistem Inf. dan Komputer), vol. 9, no. 2, pp. 247–253, 2020, doi: 10.32736/sisfokom.v9i2.903.

[24] E. Indrayuni, "Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," J. Evolusi Vol. 4 Nomor 2 - 2016, vol. 4, no. 2, pp. 20–27, 2016.