

Moderna's Vaccine Using the K-Nearest Neighbor (KNN) Method: An Analysis of Community Sentiment on Twitter

Marlyna Infryanty Hutapea^{1*}, Arina Prima Silalahi¹

¹ Universitas Methodist Indonesia, Jl. Hang Tuah No.8, Madras Hulu, Kec. Medan Polonia, Kota Medan, Sumatera Utara, Indonesia.

Received: February 16, 2022

Revised: May 18, 2023

Accepted: May 28, 2023

Published: May 31, 2023

Corresponding Author:

Marlyna Infryanty Hutapea
marlynahtpea@gmail.com

DOI: [10.29303/jppipa.v9i5.3203](https://doi.org/10.29303/jppipa.v9i5.3203)

© 2023 The Authors. This open access article is distributed under a (CC-BY License)



Abstract: The COVID-19 is still in Indonesia. The government has made efforts to stop the COVID-19 virus, by moving vaccination program. There are various types of vaccines, one of which is moderna vaccine or MRNA-1273 that applied intramuscularly. The vaccination programs using modern vaccines creates different opinions in public, especially among Twitter users. The opinion uploaded will be the data on Public Sentiment Analysis on Twitter About Moderna Vaccines Using K-Nearest Neighbor Method research. In this study, TF-IDF method is used for weighting the words and KNN for classifying the sentiment into two groups of sentiments, namely positive and negative. The tools used in this research are Rapid miner to collect tweet data and Python for sentiment classification and evaluation. From the test results Based on 50 training data when $k = 3$ it is known that the accuracy value is 80%, precision is 80%, recall is 100% and F-Measure is 89%.

Keywords: Evaluation Measure; K-Nearest Neighbor; Moderna Vaccines; TF-IDF; Twitter

Introduction

Currently the world is being hit by an outbreak of covid-19 which is very disturbing the activities of people around the world and until now it cannot be predicted when it will end (Baj et al., 2022). This virus can spread very quickly through direct touch or through the air. Various efforts have been made by the government to stop the spread of this virus. One way is to carry out a vaccination program using moderna vaccines. The responses of the Indonesian people regarding this moderna vaccine also varied (Harapan et al., 2020). There are those who welcome it and some who oppose this vaccination program. In the current era of digitalization, people are more inclined to express themselves and their views through social media, one of which is Twitter (Ramadhani & Wahyudin, 2022). Public opinion on Twitter will become data for sentiment analysis research on modern vaccines.

Sentiment Analysis is a stage of text analytics to obtain various data sources from the internet and several social media platforms. To obtain opinions from users who are on the platform (Alshuwaier et al., 2022).

Sentiment analysis is the process of understanding and classifying emotions (positive or negative) contained in writing using text analysis techniques (Veritawati et al., 2015).

Several related studies are used as references in this research such as the research entitled "Analysis of Sentiments to Astra Zeneca Vaccination on Twitter Using the Naïve Bayes and KNN Methods" which discusses public opinion on Twitter about the Astra Zeneca vaccine, in this case it is grouped into three, namely sentiment positive, neutral and negative (Ramadhani & Wahyudin, 2022). Research with these data has an accuracy value for the Naïve Bayes method of 88.56% +/- 4.71% (micro average: 88.62%) while for the KNN method the results obtained from sentiment analysis are: 74.78% +/- 3.74% (micro average: 74.77%). Another study entitled "Classification of Tweets on Twitter using the K-Nearest Neighbor (KNN) Method with TF-IDF weighting" conducted sentiment analysis on Kompas and detik news media, then classified them into technology, health, economics, sports and automotive groups (Satrio & Fauzi, 2019). Based on the results obtained from this study, the smaller the k value

How to Cite:

Hutapea, M.I., & Silalahi, A.P. (2023). Moderna's Vaccine Using the K-Nearest Neighbor (KNN) Method: An Analysis of Community Sentiment on Twitter. *Jurnal Penelitian Pendidikan IPA*, 9(5), 3808–3814. <https://doi.org/10.29303/jppipa.v9i5.3203>

used, the more accurate the KNN method. Another study entitled "Application of Sentiment Analysis on Twitter Users Using the K-Nearest Neighbor Method" discusses sentiment analysis of the DKI Pilkada 2017 which is then classified into two classes, namely positive and negative. The results obtained from this study are an accuracy of 67.2% for the value of $k = 5$.

Unlike the research mentioned above, the method used in this study is the Term Frequency-Inverse Document Frequency (TF-IDF) method for word weighting and the K-Nearest Neighbor (KNN) method for classifying sentiment into two classes, namely positive and negative using tools such as Rapid Miner and Python programming (Khalid et al., 2020).

In this study, the data used as an object to be analyzed were taken from Twitter from May 1 to May 16, 2022. If the results are obtained, they will be tested for their truth value using the evaluation measure stage so that they can be sure that the KNN method can be

used effectively. effective in the case of analyzing public sentiment on Twitter regarding the Moderna Vaccine.

Method

Data Mining

Data mining is a science cluster of combining statistical techniques, mathematics, artificial intelligence, machine learning to extract and identify information from complex databases (Aher & Lobo, 2012). The purpose of data mining is to dig up information about the characteristics of the observed data or object, it can also be used as a reference for making decisions or even predicting future conditions based on the data being analyzed (Silalahi & Simanullang, 2022). The process of working on data mining is described in Figure 1.

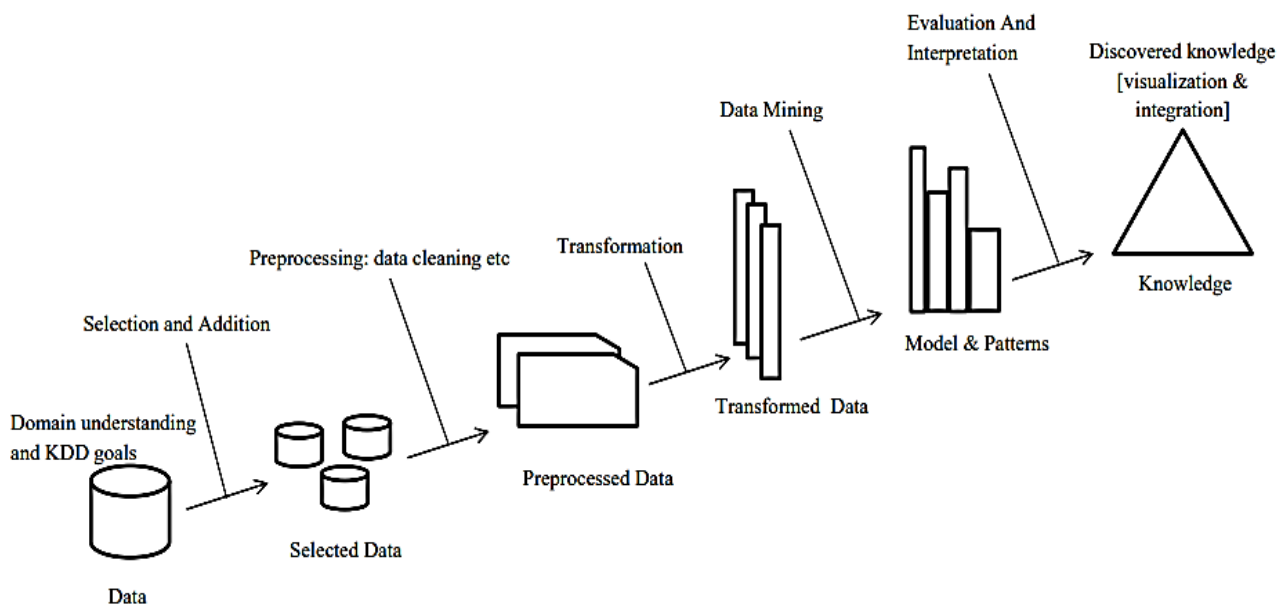


Figure 1. Data Mining Chart

Text Mining

Text Mining is a technique for getting a lot of data that was not previously known or rediscovering information sourced from automatically extracted text information. The purpose of this technique is to get useful information from a collection of documents (Lestari & Saepudin, 2021). There are many methods that can be used to extract text data, but the first step is data preprocessing.

K-Nearest Neighbor (KNN)

One simple method for classifying data based on data with the shortest distance is KNN (Akbar & Kusumodestoni, 2020; Na'iema et al., 2022). If this

method is used to classify text, it will produce a more optimal value but first weight each word in a text document that will be processed using Term Frequency-Inverse Document Frequency (TF-IDF). Then to calculate the value of the distance between documents using Euclidean Distance.

Research Stages

In this study, the object under study was the public opinion of Twitter users regarding the moderna vaccine. The data used is in the form of tweets in Indonesian. Here are some steps taken in this research shown in Figure 2.

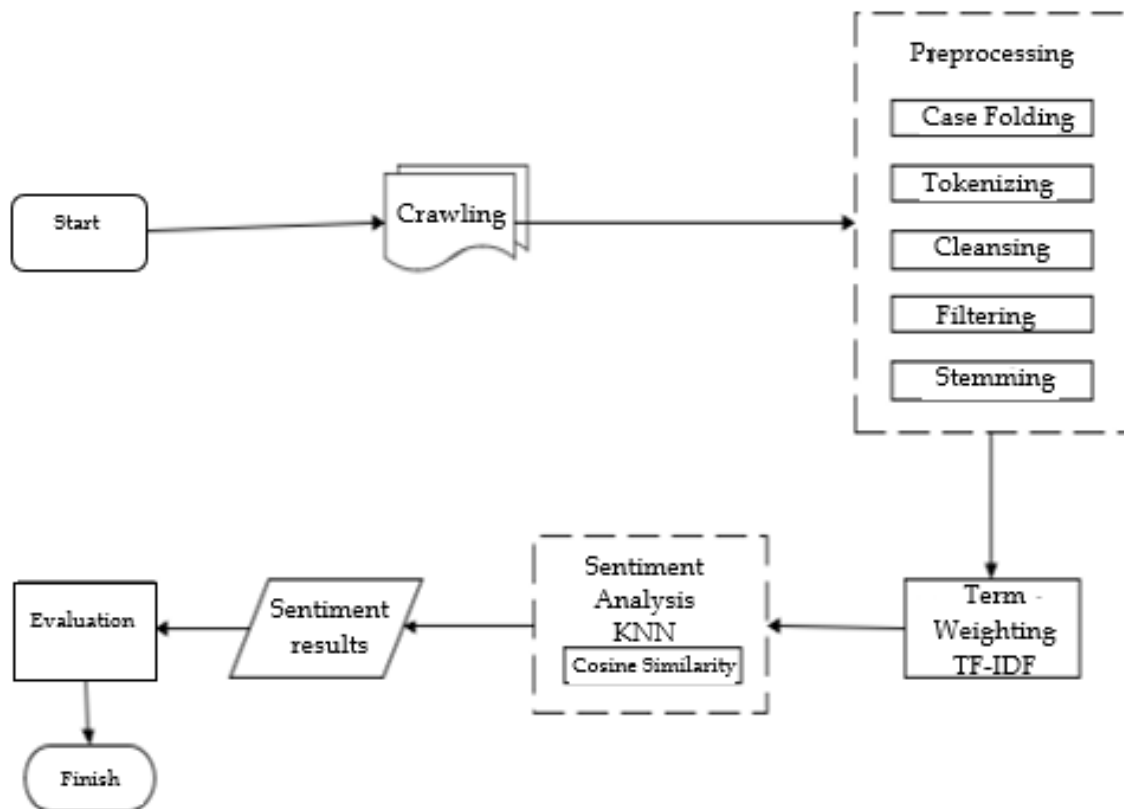


Figure 2. Schema of research

Data Collection

Data collection in this study is a stage of data mining. Data mining is the process of taking data patterns to be processed and then the output is in the form of very important information. The goal is to understand more about the observed data behavior or often referred to as a description and to estimate conditions that will occur in the future or are called predictions. (Nikmatun & Waspada, 2019). The data collection process in this study used Rapidminer tools. The data taken is a tweet about the moderna vaccine in May 2022 which is saved in a csv format file.

Preprocessing Text

Preprocessing is the stage for preparing raw data before other processes are carried out (Naresh & Kiran, 2019). In general, the data preprocessing stage is carried out to eliminate inappropriate data or change data into a form that is easier for the system to process. Some of the processes carried out at this stage are as follows:

Cleansing: Is the stage for cleaning attributes that have no effect such as symbols, numbers and links; *Folding case*: Is the process of changing all the letters in the tweet document to lowercase. Only letters a to z are processed. Characters other than letters will be left; *Tokenizing*: At this stage, sentences are cut or separated based on the specified space; *Filtering*: This is the stage of removing unnecessary words so that the calculation focuses more on words that are far more important; and

Stemming: Is the stage to find the basic words. At this stage, the process of taking basic words and removing affixes from existing words is carried out.

Word Weighting

After the data has been cleared, then a value is given to the word per document using the Term Frequency - Inverse Document Frequency (TF-IDF) method using a calculation formula.

$$W_{d,t} = tf_{d,t} * IDF_t \tag{1}$$

Where d is the d document, t is the t word of the keyword, W is the weight of the document and IDF is the inverse of the number of times the word appears.

Clasification

Classification is a method for grouping an object into a particular group or class (Rizki & others, 2019). The classification in this study uses the KNN method whose working system is to classify objects based on data that are closest to the object. Grouping new data based on its nearest neighbors expressed by k. using the following calculation formula.

$$Euclidean\ Distance(A,B) = \sum_{i=1}^t \sqrt{A - B^2} \tag{2}$$

Where A is a testing or testing document, B is a training or training document and t is the number of terms or words.

Evaluation Measure

Evaluation measurement aims to measure the performance that can be achieved by the system (Akhtar et al., 2021). Evaluation in this study is used to determine whether a system has been optimal in detecting pages that are indicated to have semantic similarities to other pages. The evaluations used are precision, recall, accuracy and F-Measure (F1-Score). Using the confusion matrix shown in Table 1.

Table 1. Confusion Matrix

Class	Positive Classification	Negative Classification
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

Result and Discussion

Data collection (Crawling)

Research data collection was obtained using a rapid miner and then stored in a csv format file (Hofmann & Klinkenberg, 2016; Kunnakorntammanop et al., 2019). Before the preprocessing stage is carried out. The data collected will be used as training data as much as 50 data. The stages of crawling data with rapid miner can be seen in Figure 2.

Row No.	Id	Created-At	From-User	From-User-Id	To...	Text
1	1524612510...	May 12, 2022	Said Achmad...	1356490089...	?	Vaksin Moderna terbukti melindungi anak usi...
2	1524612288...	May 12, 2022	Yuni_lmut	1356704730...	Yun...	Vaksin moderna terbukti melindungi anak
3	1524610936...	May 12, 2022	Deny Widi Ya...	1036432081	?	Vaksin moderna aman
4	1524608765...	May 12, 2022	Ima Marina	1098539121...	?	BPOM sebut vaksin moderna aman
5	1524603162...	May 12, 2022	Karmilakarmi	1357103967...	?	Vaksin moderna terbukti melindungi anak
6	1524593600...	May 12, 2022	delik jatim	1478265540...	?	Infone maszeeh! Ayo rek sing dunung vak...
7	1524593335...	May 12, 2022	Peppa	1480171541...	?	vaksin moderna di tangerang ada dimana ...
8	1524592509...	May 12, 2022	Bilqiz Ameliani	1484066346...	Ca...	@CahyaBulan Harapan bangsa untuk pa...
9	1524592340...	May 12, 2022	Gandul	1413503905...	?	Vaksin Moderna Terbukti Lindungi Anak U...
10	1524591976...	May 12, 2022	Cahya Itahi	1356295901...	?	Vaksin moderna terbukti melindungi anak
11	1524588925...	May 12, 2022	Putri Adelliazz	1356960647...	?	Vaksin Moderna Terbukti Lindungi Anak U...
12	1524588789...	May 12, 2022	Putri Adelliazz	1356960647...	?	Vaksin Moderna Terbukti Lindungi Anak U...
13	1524588745...	May 12, 2022	Putri Adelliazz	1356960647...	?	Vaksin Moderna Terbukti Lindungi Anak U...
14	1524588711...	May 12, 2022	Putri Adelliazz	1356960647...	?	Vaksin Moderna Terbukti Lindungi Anak U...

Figure 2. Data Collection Using Rapidminer

Preprocessing Results

Data preprocessing is a cleaning stage before the data is further processed. The stages of data preprocessing in this study were carried out using the rapid miner application (Sudarsono et al., 2021). The previously collected data will then be managed using a rapid miner through a series of stages such as cleansing, case folding, tokenizing, filtering, and stemming. The cleaning steps for the 1 example tweet obtained are as follows Table 2.

Table 2. Tweets with the Cleansing Process

Before	After
The POM Agency ensures that Moderna's vaccine is safe to use and does not contain foreign particles in it. Eid Al-Fitr Still Prokes	The POM Agency ensures that the moderna vaccine is safe to use and does not contain foreign particles in it, Eid continues to Prokes

In this stage, the dot symbol is removed. The case folding stage will change all letters to lower case. Examples of tweets that go through this stage can be seen in Table 3.

Table 3. Tweets with Case Folding Process

Before	After
The POM Agency ensures that Moderna's vaccine is safe to use and does not contain foreign particles in it. Eid Al-Fitr Still Prokes	The pom agency ensures that the moderna vaccine is safe to use and does not contain foreign particles in it. Eid al-Fitr is still Prokes

Tokenizing stages to separate sentences into single words can be seen in Table 4 and the filtering stage will remove words that have no effect and can be seen in Table 5.

Table 4. Tweets with Tokenizing Process

Before	After
The pom agency ensures that the moderna vaccine is safe to use and does not contain foreign particles in it. Eid al-Fitr is still Prokes	'agency' 'pom' 'ensure' 'vaccine' 'moderna' 'safe' 'to use' 'and' 'no' 'contains' 'particle' 'foreign' 'in it' 'edulfitri' 'still' 'prokes'

Table 5. Tweets with Filtering Process

Before	After
'agency' 'POM' 'ensure' 'vaccine' 'moderna' 'safe' 'to use' 'and' 'no' 'contains' 'particle' 'foreign' 'in it' 'edulfitri' 'still' 'prokes'	'body' 'pom' 'make sure' 'vaccine' 'moderna' 'safe' 'to use' 'does not' 'contain' 'content' 'foreign' 'particle' 'in it'

Results of Data Weighting

Data weighting is the stage of giving value or weight to data. The data that has been cleaned will then be given a weight or value for each word. In this study, the data weighting process was carried out using the TF-IDF method. The stages of data weighting in this study were carried out using the rapid miner application. The following is the result of weighting the data.

Conclusion

The K-Nearest Neighbor (KNN) method is proven to be able to classify sentiments based on the training data that has been provided. The trial was carried out using testing data that had been weighted and inputted into the system then the resulting output was in the form of positive sentiments of 70% and negative sentiments of 30%. After calculating the evaluation measure on the system, with the optimal k value = 3 and using 80% training data and 20% testing data from 50 data sets, the results obtained are accuracy of 80%, precision of 80%, recall of 100% and F1-Score of 89%. Writers of comments in Indonesia do not use good and correct Indonesian when uploading tweets and there is no Indonesian dictionary library available in rapid miner so that it can affect the final result of the classification process.

Author Contribution

The first and second authors work together to collect comment data from tweeters and carry out the preprocessing stages to data analysis using Rapid Miner.

Funding

This research received no external funding.

Conflicts of Interest

It is important to know public sentiment towards the Moderna vaccine given by the government from the various comments that have appeared on Twitter regarding the effects of use and its effectiveness. In addition, by using the KNN method, Twitter comment data can be analyzed to obtain a classification by dividing it into three classes, namely Positive, Negative, and Neutral. If the comments are divided into 3 sentiment classes, then the method is evaluated to get the accuracy, precision and recall values.

References

- Aher, S. B., & Lobo, L. (2012). Course recommender system in E-learning. *International Journal of Computer Science and Communication*, 3(1), 159-164. Retrieved from http://csjournals.com/IJCSC/PDF3-1/Article_35.pdf
- Akbar, A. S., & Kusumodestoni, R. H. (2020). Optimasi nilai k dan parameter lag algoritme k-nearest neighbor pada prediksi tingkat hunian hotel. *Jurnal Teknologi Dan Sistem Komputer*, 8(3), 246-254. <https://doi.org/10.14710/jtsiskom.2020.14007>
- Akhtar, A., Akhtar, S., Bakhtawar, B., Kashif, A. A., Aziz, N., & Javeid, M. S. (2021). COVID-19 detection from CBC using machine learning techniques. *International Journal of Technology, Innovation and Management (IJTIM)*, 1(2), 65-78. <https://doi.org/10.54489/ijtim.v1i2.22>
- Alshuwaier, F., Areshey, A., & Poon, J. (2022). Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review. *Intelligent Systems with Applications*, 200090. <https://doi.org/10.1016/j.iswa.2022.200090>
- Arslan, H., & Arslan, H. (2021). A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Engineering Science and Technology, an International Journal*, 24(4), 839-847. <https://doi.org/10.1016/j.jestch.2020.12.026>
- Baj, A., Dalla Gasperina, D., Focosi, D., Forlani, G., Ferrante, F. D., Novazzi, F., Azzi, L., & Maggi, F. (2022). Safety and immunogenicity of synchronous COVID19 and influenza vaccination. *Journal of Clinical Virology Plus*, 2(3), 100082. <https://doi.org/10.1016/j.jcvp.2022.100082>
- Harapan, H., Wagner, A. L., Yufika, A., Winardi, W., Anwar, S., Gan, A. K., Setiawan, A. M., Rajamoorthy, Y., Sofyan, H., & Mudatsir, M. (2020). Acceptance of a COVID-19 vaccine in Southeast Asia: a cross-sectional study in Indonesia. *Frontiers in Public Health*, 8, 381. <https://doi.org/10.3389/fpubh.2020.00381/full>
- Hofmann, M., & Klinkenberg, R. (2016). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.
- Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 15(2), 121-130. <https://doi.org/10.22146/ijccs.65176>
- Iwendi, C., Mahboob, K., Khalid, Z., Javed, A. R., Rizwan, M., & Ghosh, U. (2021). Classification of COVID-19 individuals using adaptive neuro-fuzzy inference system. *Multimedia Systems*, 1-15. <https://doi.org/10.1007/s00530-021-00774-w>
- Khalid, M., Ashraf, I., Mehmood, A., Ullah, S., Ahmad, M., & Choi, G. S. (2020). GBSVM: sentiment classification from unstructured reviews using ensemble classifier. *Applied Sciences*, 10(8), 2788. <https://doi.org/10.3390/app10082788>
- Kunnakorntammanop, S., Thepwuttisathaphon, N., & Thaicharoen, S. (2019). An experience report on building a big data analytics framework using Cloudera CDH and RapidMiner Radoop with a cluster of commodity computers. *Soft Computing in Data Science: 5th International Conference, SCDS 2019, Iizuka, Japan, August 28-29, 2019, Proceedings 5*, 208-222. https://doi.org/10.1007/978-981-15-0399-3_17
- Lestari, S., & Saepudin, S. (2021). Analisis sentimen vaksin sinovac pada twitter menggunakan algoritma Naive Bayes. *Seminar Nasional Sistem Informasi Dan Manajemen Informatika Universitas Nusa Putra*, 1(01), 163-170. Retrieved from

- <https://sismatik.nusaputra.ac.id/index.php/sismatik/article/view/23>
- Na'iem, A.-N. S., Mulyo, H., & Widiastuti, N. A. (2022). Klasifikasi penerima bantuan program rehabilitasi rumah tidak layak huni menggunakan algoritme K-Nearest Neighbor. *Jurnal Teknologi Dan Sistem Komputer*, 10(1), 32-37. <https://doi.org/10.14710/jtsiskom.2022.14110>
- Naresh, P. K. M., & Kiran, P. (2019). Preprocessing Methods for Unstructured Healthcare Text Dat. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(2S), 715-719. Retrieved from <https://www.ijtee.org/wp-content/uploads/papers/v9i2S/B10241292S19.pdf>
- Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, 1(4), 7. Retrieved from <https://scholar.smu.edu/datasciencereview/vol1/iss4/7/>
- Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 10(2), 421-432. <https://doi.org/10.24176/simet.v10i2.2882>
- Perez, F., & Granger, B. E. (2015). Project Jupyter: Computational narratives as the engine of collaborative data science. Retrieved September, 11(207), 108. Retrieved from <https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58?gi=fee115e4abfe>
- Ramadhani, S. H., & Wahyudin, M. I. (2022). Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Na{"i}ve Bayes dan K-NN. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, 6(4), 526-534. <https://doi.org/10.35870/jtik.v6i4.530>
- Rizki, M. M., & others. (2019). Analisis sentimen terhadap produk otomotif dari twitter menggunakan kombinasi algoritma k-nearest neighbor dan pendekatan lexicon (studi kasus: mobil toyota). Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah. Retrieved from <https://repository.uinjkt.ac.id/dspace/handle/123456789/48643>
- Satrio, R. H., & Fauzi, M. A. (2019). Klasifikasi Tweets Pada Twitter Menggunakan Metode K-Nearest Neighbour (K-NN) Dengan Pembobotan TF-IDF. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(8), 8293-8300. Retrieved from <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6133>
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, 5, 1-16. <https://doi.org/10.1007/s41133-020-00032-0>
- Silalahi, A. P., & Simanullang, H. G. (2022). Prediksi Jumlah Pasien Covid-19 Di Indonesia Menggunakan Least Square Method Berbasis Android. *Informatika*, 14(1), 86-93. <https://doi.org/10.36723/juri.v14i1.328>
- Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner. *JBASE-Journal of Business and Audit Information Systems*, 4(1). <https://doi.org/10.30813/jbase.v4i1.2729>
- Veritawati, I., Wasito, I., & Basaruddin, T. (2015). Text preprocessing using annotated suffix tree with matching keyphrase. *International Journal of Electrical and Computer Engineering*, 5(3), 409. <https://doi.org/10.11591/ijece.v5i3.pp409-420>
- Zuraimi, M. A. Bin, & Zaman, F. H. K. (2021). Vehicle detection and tracking using YOLO and DeepSORT. 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 23-29. <https://doi.org/10.1109/ISCAIE51753.2021.9431784>