ILKOM Jurnal Ilmiah Vol. 15, No. 2, August 2023, pp.365-372
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

K-Nearest Neighbors Analysis for Public Sentiment towards Implementation of Booster Vaccines in Indonesia

Ihwana As'ad [a,1,*]; **Muhammad Arfah Asis** [a,2]; **Hariani Ma'tang Pakka** [a,3]; **Randi Mursalim** [a,4]; **Yusnita binti Muhamad Noor** [b,5]
[a] *Universitas Muslim Indonesia, Jl Urip Sumoharjo No.KM 5, Makassar, 90231, Indonesia*
[b] *Universiti Malaysia Pahang Al Sultan Abdullah, Kuantag Pahang, 26300, Malaysia*
[1] *ihwana.asad@umi.ac.id;* [2] *muh.arfah.asis@umi.ac.id;* [3] *hariani.m@umi.ac.id;* [4] *randimursalim04082000@gmail.com;*
[5] *yusnitanoor@ump.edu.my*
** Corresponding author*

Article history: Received December 17, 2022; Revised December 23, 2022; Accepted June 24, 2023; Available online August 16, 2023

Abstract
In order to prevent the spread of COVID-19 in Indonesia, the Government of the Republic of Indonesia has been implementing a booster vaccine program since January 12th, 2022, with priority for the elderly and vulnerable groups as well as those who got the second C-19 vaccine longer than 6 months. The implementation of this program raised many pros and cons among public which were expressed either positively or negatively through social media. Therefore, sentiment analysis is needed to examine these phenomenons. This study aims to determine the positive and negative response from public by employing K-Nearest Neighbor method. A total of 2,000 commentary data were collected to be in turn classified based on positive and negative sentiments. There are 500 comments used as training data and divided equally to positive and negative class, each consists of 250 data. Using the value of K = 9, the results show a positive sentiment of 43% while a negative sentiment of 57%. Based on the validity test using 10-fold cross validation, an accuracy of 82.60% was obtained, a recall value was 82.60% with a precision of 83.89%.

Keywords: K-Nearest Neighbors; RapidMiner; Sentiment Analysis; Vaccine Booster.

Introduction

A new disease outbreak caused by the corona virus (2019-nCoV), commonly called COVID-19, became a pandemic in Indonesia since the first case announced by President Joko Widodo on March 2, 2020, which infected 2 Indonesian from Depok, West Java [1]. Starting from this case, the number of infected cases of Indonesian people by the corona virus continued to increase daily. The government had taken many ways to suppress the spread of COVID-19 in Indonesia, including vaccination.

Vaccination policies, including booster vaccines, which have been carried out since January 12, 2022 with priority on the elderly and vulnerable groups and those who have received a second vaccine within 6 months, have become a public discussion [2]. The implementation of booster vaccines in Indonesia has caused many pros and cons among the public. Various opinions were expressed through social media, both positive and negative opinions. Therefore, sentiment analysis is required. Previous studies can be used as a reference to determine linkages and prevent duplication in current research. The following paragraphs describe several studies on sentiment analysis.

Sentiment analysis or commonly known as opinion mining is a research branch of text mining which aims to determine public (audience) perceptions or subjectivity of a topic, event, or problem [6]. One method that is often used in sentiment analysis is the K-Nearest Neighbor (KNN) method. Recent research has found that the KNN method gets the best accuracy of 94.4% in analyzing sentiment [5]. A study applied the KNN algorithm in sentiment analysis of online learning methods at Wira Wacana Christian University, Sumba [3]. RapidMiner was used with text data through the KNN algorithm with the aim of finding values for accuracy, precision and recall. The research obtained an accuracy value of 87.00% and an AUC value of 0.916.

http://dx.doi.org/10.33096/ilkom.v15i2.1561.365-372

The next research is sentiment analysis using the KNN method on Twitter regarding academic information system services for Brawijaya University students [4]. There were 4 main processes for classifying, namely preprocessing, term weighting, feature selection and classification. The best accuracy result of the classification process was 86%. Accuracy results are obtained using the value of $k = 3$ with 100% features. Another study is sentiment analysis of online transportation service products using the KNN method with the best accuracy of 94.4%. Therefore, the algorithm works well for sentiment analysis [5].

Method

A. Text Mining

Text Mining is a stage for determining the relationship of terms or words in documents that are not well structured. Text mining aims to analyze sentiment so that information could be defined by users. Text mining is commonly used in classification, clustering, information acquisition, and information extraction [7].

B. Sentiment Analysis

Sentiment analysis is a categorization of opinions on a matter or issue that is expressed in text so that conclusions can be drawn and subsequent reactions can be determined [8]. Sentiment analysis or opinion mining refers to the broad field of natural language processing, computational linguistics and text mining, which has the aim of analyzing a person's opinions, sentiments, evaluations, attitudes, judgments and emotions, whether the speaker or writer is concerned with a topic, product, service, organization, individuals, or certain activities [9].

C. Data collection

1. Data collection technique

Data in this study is the posted comments regarding the implementation of booster vaccines in Indonesia obtained from Facebook account of Indonesian Ministry of Health. Data was collected using the Export Comments website [10]. 2000 comments related to the implementation of booster vaccines were taken. The data is divided equally based on sentiment, 250 for the positive class and 250 for the negative class. Labeling of training data to determine positive and negative classes was then done manually.

2. Data Cleaning

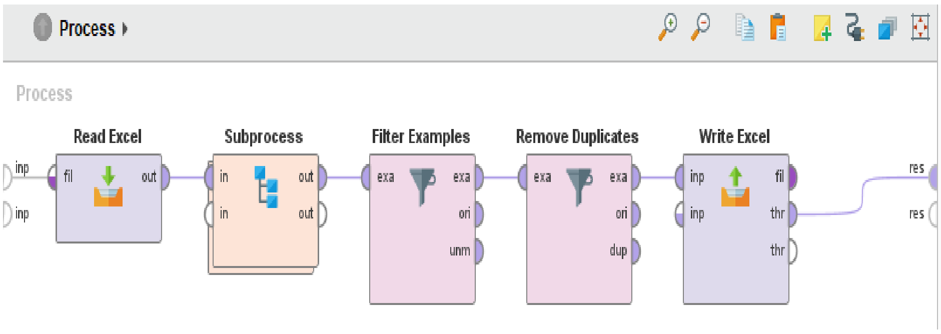

Figure 1. Data Cleaning

The Excel data from the collection stage is then processed at the data cleaning stage which aims to remove duplicate comment data as well as to prevent data similarities and reduce the number of terms. Data cleansing works on all attribute types as follows **Figure 1**. In addition, data cleaning is also carried out to remove data that does not have attribute values, hashtags, links, and mentions in comments.

D. Preprocessing

Preprocessing is a process of changing the form of unstructured data into the structured one. This process removes words or characters that are not needed so that when doing weighting it can run perfectly and it is easy to carry out the classification process. The preprocessing stage is an important stage and the starting point in classification [11]; here, raw documents are prepared to become documents or representative documents that are ready to be processed for the next steps [12]. This stage is the initial stage in the text mining process, as there are seven stages in this study, namely transform case, tokenize, replace abbreviations, filter tokens (by length), stemming, filter stop words and generate n-grams (terms) [13]. The preprocessing stage consists of:

Preprocessing is the process of changing the form of unstructured data into structured. This process removes unnecessary words or characters so that the weighting process can run perfectly and makes it easier for the classification stage. Preprocessing is an important stage which is the starting point in the classification [11]. At this stage, raw documents are prepared to become representative documents ready to be processed for the next step [12]. This stage is the initial stage in the text mining process. This research consists of 7 stages, namely transform case, tokenize, replace abbreviations, filter tokens (by length), stemming, filter stop words and generate n-grams (terms) [13]. The preprocessing stage consists of **Figure 2**.

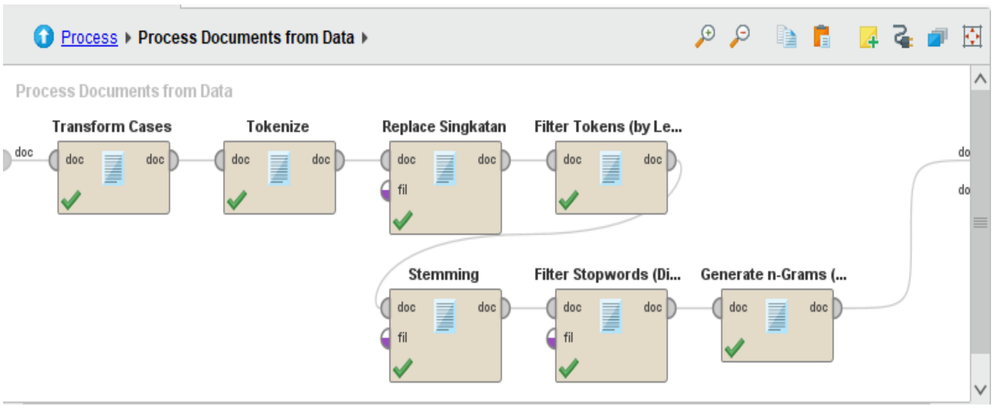

Figure 2. Preprocessing

1. *Transform Case*
   In this stage, all sentences/words in the document will be converted into non-capital letters. This stage is carried out so that the data inputted has the same arrangement/structure of letters, for example, "Vaksin", "vAksin", "vaKsin", "vacSin", "VaksIn", and so on will be changed to the same word, namely "vaccine".
2. *Tokenization*
   In this stage, a group of sentences will be broken down into words, while eliminating special characters / symbols and punctuation which then forms a unique data set.
3. *Replace Abbreviation*
   This stage changes the abbreviation in the comments into a proper word. For example, "*yg*" becomes "*yang*" or "*bkn*" becomes "*bukan*"
4. *Filter tokens (by length)*
   This stage filters tokenize based on length. Token length is a minimum of 3 and a maximum of 25 letters in one word.
5. *Stemming*
   This stage aims to minimize the number of different indexes in the document and to group other words that have the same base word and meaning but have a different form due to different affixes. For example: "*menyuntik*" becomes "*suntik*", "*membela*" becomes "*bela*", and "*mengikuti*" becomes "*ikut*".
6. *Stop word filters*
   The stopword filter is the stage for taking important words from the token results by removing unimportant words or saving important words (wordlist). Stop words are usually common words that appear in large numbers in several sentences including conjunctions, for example: "is", "to", "in", "from", "and". The aims of stop words is to eliminate words that have low information value or have no interest in the contents of the document. The stop words in this study were obtained from the Kaggle site [14].
7. *Generate n-Grams (Terms)*
   This stage aims to solve the problem of document classification into positive or negative sentiments where misclassification is usually caused by a single term [15], for example the word "bad" includes negative sentiment, but the word "not bad" classified as positive sentiment when they are side by side with negation. Therefore, the Generate n-Grams (Terms) operator is expected to be able to handle these problems in order to improve the classification results. The number n at this stage is 2 words (Bigram).

E. Term Frequency Inverse-Document Frequency

Term Frequency Inverse-Document Frequency (TF-IDF) is an algorithmic method used in analyzing a relationship between words (terms) and a set of documents. This method will calculate the Term Frequency (TF) and Inverse Document Frequency (IDF) values for each token (word) in each document in the corpus[16]. Term Frequency is a process used to count the occurrence of a word (term) in a document meanwhile IDF is a process for calculating how important the calculation of terms that are widely distributed in the related document [17]. The following is the TF-IDF equation, named as in Equation 1 [18].

$$wt,d = tf\ t,d \times idft \qquad (1)$$

Where:

tf : number of terms in the document
idf : the number of terms in all documents in the data
t : term
d : document

F. K-Nearest Neighbor

K-NN is an algorithm that classifies objects based on the closest distance to objects or features, the data most commonly used in learning data [19]. The purpose of this algorithm is to classify new objects based on attributes and samples from training data.

The stages of the KNN algorithm are as follows:

1. Determine the parameter k (number of nearest neighbors)
2. the weight for each term using the Term Weighting TF-IDF
3. Calculate the similarity between documents using Cosine Similarity:

$$cosSim(d_j, d_k) = \frac{\sum_{i=1}^{n} (td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} tq_{ik}^2}} \qquad (2)$$

where,

$cosSim(d_j, d_k)$: document similarity level with a particular query
td_{ij} : the i^{th} term in the vector for the j^{th} document
tq_{ik} : the i^{th} term in the vector for the k^{th} document
n : the number of unique terms in the data set

4. Sort the results of the Cosine Similarity calculation from big to small
5. Take as many as k which have the highest similarity with the classified documents, then determine the class.

G. Analysis of Data Validation

At this stage, evaluation is carried out using cross-validation to find out whether the applied model has been produced the expected result. Using 10-fold validation, the operator randomly divides the data into 10 parts. The testing process begins with the model built using the first data group followed by the remaining 9 data. The results of this stage are the values of precision, recall, and accuracy. The cross-validation stage is divided into two parts, namely training and testing of 70% and 30% respectively. The KNN algorithm is applied to the training data, while the testing section consists of the apply model operator and performance. While the former is where the dataset is applied to the model, the latter functions to see the performance of the applied model as in **Figure 3**.

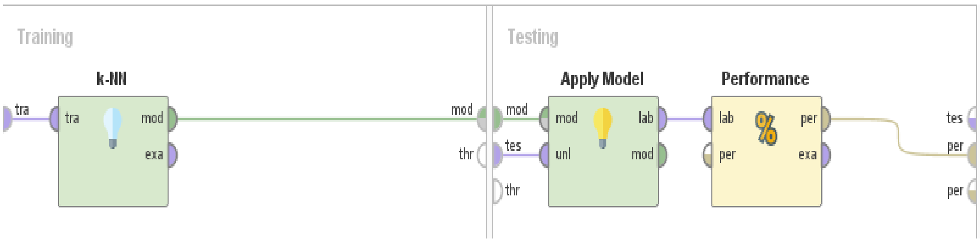

Figure 3. Process validation

Result and Discussion

A. Results of Sentiment Analysis

Based on data taken from official Facebook account of the Ministry of Health of Indonesia, the results of sentiment analysis are described as in **Figure 4**.

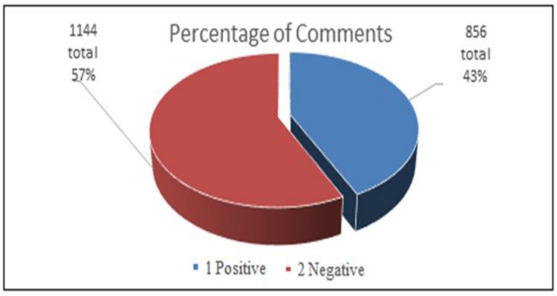

Figure 4. Results of Sentiment Analysis

Figure 4 shows a comparison of public comments regarding the implementation of booster vaccines in Indonesia. The data shows that negative comments dominate the result compared to positive comments. The number of positive comments obtained was 856 (43%) comments while negative comments obtained was 1,144 (57%) comments.

B. Data validity test

Data validity testing was carried out to determine the accuracy, recall and precision values of the model that had been built. The test was carried out 5 times, by entering the values of $k = 3$ to $k = 11$ with the aim of determining the most optimum K value. The k value is taken from odd number with the result as in **Table 1**.

Table 1. Test of k-value for K-NN

K Value	Accuracy	Recall	Precision
3	80,40%	80,40%	82,07%
5	81,40%	81,40%	82,85%
7	82,00%	82,00%	83,34%
9	83,20%	83,20%	84,41%
11	83,40%	83,40%	84,61%

Based on table 1, tests have been carried out starting from $k = 3$ to $k = 11$ to obtain the highest accuracy value, which is at $k = 11$. Thus, $k = 11$ is used in this study.

The outcome of a validity test utilizing 10-Fold Validation to assess how fit the model or algorithm in question performed was as in **Table 2**.

Table 2. 10-Fold Validation

Number of folds	Accuracy
1	83,00%
2	83,00%
3	82,40%
4	83,06%
5	82,40%
6	83,04%
7	82,40%
8	82,21%
9	83,08%
10	83,40%
Average	82,79%

C. Data Visualization

Before creating a word cloud, it is necessary to examine the overall frequency data that has been obtained from the process on RapidMiner, then filtered based on its class, namely positive and negative classes, where each class displays the 20 most frequently occurring words with the details of the table below as follows.

Before creating word cloud, it is necessary to identify all of the frequency data that has been obtained from the process in RapidMiner. The data is then filtered based on its class, namely positive and negative. Each class displays the 20 words that appear most often. The details are shown in the following **Table 3**.

Table 3. Positive sentiment word frequency

	Word	Total
1	*sehat*	225
2	*tidak*	211
3	*covid*	114
4	*alhamdulillah*	97
5	*indonesia*	80
6	*moga*	79
7	*pemerintah*	69
8	*belum*	65
9	*kasih*	59
10	*allah*	55

	Word	Total
11	*pakai*	53
12	*aman*	52
13	*sakit*	52
14	*terima*	48
15	*rakyat*	45
16	*biar*	44
17	*virus*	41
18	*moderna*	34
19	*badan*	33
20	*dosis*	33

After obtaining the frequency of positive sentiment. The next step is to calculate the sentiment frequency for the negative class presented as follows **Table 4**.

Table 4. Frequency of negative sentiment words

	Word	Total
1	*booster*	195
2	*bisnis*	145
3	*kali*	136
4	*corona*	108
5	*bukan*	79
6	*habis*	75
7	*mati*	73
8	*kena*	69
9	*orang*	68
10	*biar*	60
11	*negara*	58
12	*suntik*	55
13	*anak*	54
14	*bodoh*	52
15	*mudik*	45
16	*paksa*	45
17	*takut*	44
18	*masyarakat*	41
19	*cuan*	39
20	*varian*	38

Conclusion

This research has succeeded in conducting a sentiment analysis using the KNN method regarding the implementation of booster vaccines in Indonesia. Based on the implementation results of the KNN evaluation, the study can be concluded as follows.

1. The negative comments are more dominant than positive comments in the study case. There were 856 comments (43%) for positive comments, and 1144 comments (57%) for negative comments. It implies that the implementation of booster vaccines in Indonesia has a negative conversation impact based on the dominant negative results on social media Facebook.
2. Testing analysis of public sentiment towards the implementation of booster vaccines in Indonesia using the KNN implies that with a value of K = 9 and 10-fold cross validation for testing the validity of the data obtained, the values for accuracy, recall and precision are 82.60%, 82, 60%, and 83.89% respectively.

3. Based on the results of this accuracy, the system created to analyze public sentiment regarding the implementation of booster vaccines in Indonesia using the KNN algorithm can run well as expected.

References

[1] R. Nuraini, "Kasus Covid-19 Pertama, Masyarakat Jangan Panik," 2020. https://indonesia.go.id/narasi/indonesia-dalam-angka/ekonomi/kasus-covid-19-pertama-masyarakat-jangan-panik (accessed Mar. 11, 2022).

[2] C.-19 Hotline, "Tok! Vaksin Booster gratis untuk Seluruh Masyarakat Indonesia," 2022. https://covid19.go.id/artikel/2022/01/11/tok-vaksin-booster-gratis-untuk-seluruh-masyarakat-indonesia (accessed Mar. 11, 2022).

[3] A. Tanggu Mara, E. Sediyono, and H. Purnomo, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Metode Pembelajaran Dalam Jaringan (DARING) Di Universitas Kristen Wira Wacana Sumba," *Jointer - J. Informatics Eng.*, vol. 2, no. 01, pp. 24–31, 2021, doi: 10.53682/jointer.v2i01.30.

[4] L. R. Dharmawan, I. Arwani, and D. E. Ratnawati, "Analisis Sentimen pada Sosial Media Twitter Terhadap Layanan Sistem Informasi Akademik Mahasiswa Universitas Brawijaya dengan Metode K- Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, pp. 959–965, 2020, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/7099

[5] S. Rohwinasakti, B. Irawan, and C. Setianingsih, "Sentiment Analysis on Online Transportation Service Products Using K-Nearest Neighbor Method," *Proc. Int. Conf. Comput. Information, Telecommun. Syst. CITS 2021*, vol. 7, no. 3, pp. 9312–9321, 2021, doi: 10.1109/CITS52676.2021.9618301.

[6] S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso, and R. Nooraeni, "Data mining dengan R konsep serta implementasi," *Jakarta: InMedia*, 2018.

[7] M. Andreotta *et al.*, "Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis," *Behav. Res. Methods*, vol. 51, no. 4, pp. 1766–1781, 2019.

[8] T. Pawar, P. Kalra, and D. Mehrotra, "Analysis of Sentiments for Sports data using RapidMiner," in *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2018, pp. 625–628.

[9] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[10] Exportcomments, "Export Social Media Comments," 2022. https://exportcomments.com/ (accessed Jul. 29, 2022).

[11] M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," *Int. J. Control Theory Appl*, vol. 9, no. 27, pp. 255–260, 2016.

[12] B. Zaman and E. Winarko, "Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 5, no. 2, 2011.

[13] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving text preprocessing for student complaint document classification using sastrawi," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 874, no. 1, p. 12017.

[14] O. R. Hartono, "Indonesian Stoplist," *kaggle*, 2016. https://www.kaggle.com/datasets/oswinrh/indonesian-stoplist (accessed Jul. 01, 2022).

[15] P. Pascasarjana, M. Ilmu, S. Tinggi, M. Informatika, D. A. N. Komputer, and N. Mandiri, "Analisis Sentimen Review Kosmetik Pada Website Femaledaily Menggunakan Metode Naive Bayes Dan Support Vector Machine Berbasis Particle Swarm Optimization," 2019.

[16] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," *J. Electr. Comput. Eng.*, 2015.

[17] A. Mishra and S. Vishwakarma, "Analysis of tf-idf model and its variant for document retrieval," in *2015 international conference on computational intelligence and communication networks (cicn)*, 2015, pp. 772–776.

[18] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science & Business Media, 2011.

[19] A. Bode, “K-nearest neighbor dengan feature selection menggunakan backward elimination untuk prediksi harga komoditi kopi arabika,” *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017.