Sentiment Analysis of Health Protocol Policy Using K-Nearest Neighbor and Cosine Similarity

N. Ritha[1*], N. Hayaty[1], T. Matulatan[1], A.Uperiati[2], M.Rathomi[1], M.Bettiza[1], F.Farasalsabila[1]

{nola.ritha@umrah.ac.id, nurul.hayaty@umrah.ac.id, tekad.matulatan@umrah.ac.id, alenaup7@gmail.com, radzi@umrah.ac.id, mbettiza@umrah.ac.id, fidid8@gmail.com}

[1] Informatics Department, Universitas Maritim Raja Ali Haji, Indonesia
[2] Software Engineering Department, State Polytechnic of Batam, Indonesia

Abstract. As an effort to handle COVID-19 transmission, the Indonesian government introduced various regulations such as Health Protocol Policy. Meanwhile, the policy has become a public discussion on social media, such as Twitter. To classify people's opinions on the social media, sentiment analysis technique is used. This method is widely utilized in natural language processing (NLP), data mining, and information retrieval. Therefore, this study aimed to analyze Twitter users' opinions toward health protocol policy. There were several approaches to conducting sentiment analysis, including the TF-IDF method for word weighting and the K-Nearest Neighbor algorithm for data classification. Also, the similarity between two documents or texts was measured using cosine similarity. The processed data were categorized into three, namely positive, negative, and neutral. The system classification results showed 43.94% positive sentiment, 9.06% negative, and 46.99% neutral assuming the value of K = 4.

Keywords: Sentiment Analysis, Health Protocol, K-NN, Cosine Similarity.

1 Introduction

COVID-19 is an epidemic that began in the Chinese city of Wuhan and has now spread around the world. The application of this Health Protocol is one of the most appropriate solutions for reducing the number of instances of COVID-19 viral infection, and it is still being implemented to all Indonesians. The Indonesian government's attempts to apply this Health Protocol have given birth to a plethora of perceptions, rumors, and other information that has entered the community, giving rise to both positives and negatives [1].

The government's recommendation to the public regarding the implementation of health protocol cannot be separated from social media users' thoughts and views that are particularly expressed through Twitter. One of the methods for categorizing types of thoughts or opinions from social media users is sentiment analysis [2].

Therefore, this study aims to determine Indonesian's emotions about the country's Health Protocol based on tweets or comments on Twitter using sentiment analysis. The dataset utilized was derived from Indonesian-language tweets' monitor by considering the keyword "Health Protocol". Subsequently, the crawled data are processed with the K-Nearest Neighbors approach in order to obtain sentiment analysis results on the Indonesian

ICSEDTI 2022, October 11-13, Tanjungpinang, Indonesia

DOI 10.4108/eai.11-10-2022.2326274

government's policy regarding the Health Protocol and determine whether public opinion about its implementation is negative or positive [3].

The K-Nearest Neighbors was used because it is effective in processing huge amounts of training data, resistant to noise, and has high consistency in producing accurate results. The model was utilized with word weights calculated by the TF-IDF approach [4]. After obtaining the results, the model was evaluated with a confusion matrix to determine its accuracy [5].

2 Research Methods

2.1 Related Research

There are several research journal literatures that analyze behavior using a data mining technique. Furthermore, this case also relates to the connection between data analysis and other machine learning variables [6]. Several methods which had been used, especially in analyzing sentiment behavior. Research by [7] conducted a sentiment analysis of online learning using K-NN algorithm and the highest accuracy results were obtained when K = 10 with an accuracy value of 84.65% with a precision of 87%, a recall of 86%, f measure 87% and an error rate of 0.12%. While research [8] conducted a sentiment analysis of movie reviews, the dataset gathered from to assess audience responses to the movies they watch in two groups: positive and negative. The text mining analysis in extracting information is obtained and categorised using Nave Bayes. Chi-Square will be used to measure response sentiments. In contrast to the research conducted by [9] comparing analytical techniques such as K-Nearest Neighbor, Naive Bayes, and Super Vector Machine to analyze sentiment behavior utilizing Twitter data sets. However, in another context was proposed by [10] the analysis show that the Super Vector Machine (SVM) performs better, reaching a value of 85% when compared to the Unsupervised Learning approach.

2.2 Methods

Text mining, also known as text analytics, is a branch of study that combines linguistics with computer science, as well as statistical and machine learning approaches. During the analysis, text processing is always performed through the following stages, namely Tokenization, Stop Word Removal, Stemming, and Lemmatization [11]. In this study, Twitter crawling was employed and the data were collected from November 2021 to January 2022.

Crawling was employed in retrieving data from tweets using the Twitter API's capabilities. This method involves collecting tweets with the keyword 'health protocol' (*protokol kesehatan* in Bahasa). It is important to note that the API usage is the legal access granted to users by Twitter [12].

Prior to performing the classification stage, the K-Nearest Neighbor method was used. The first step is to clean up the data and turn it into tokens. This is known as pre processing. The token acquired from the pre processing results will be utilized in the following phase, which is the computation of word weighting using TF-IDF. The resulting weight will then be used to compute the closeness distance between documents using the K-Nearest Neighbor method. The K-nearest neighbor (KNN) method is a non-parametric, instance-based supervised learning technique [13]. It is primarily used to assign an unknown data record to the class of the closest existing examples. This assignment is determined in classification tasks

based on the distances between this unknown record and all of the available instances. After computing all of the distances, the K-Nearest examples are chosen. A new record is predicted by the class majority among the K-Nearest Neighbor. To measure between two documents or texts using the cosine similarity and validate the performance of KNN using confussion matrix [14].

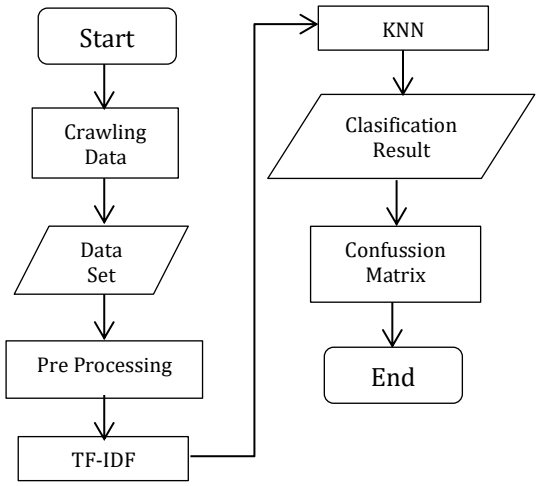

Figure 1 *Overview the process*

3 Results and Discussion

3.1 Data Preprocessing

In this study, The dataset used will be divided into 2 types of data which is training data and test data. Training Data is used to train the algorithm in finding the appropriate model. While Test Data is new data that does not yet have a class so a classification process is needed to determine the appropriate class. The data obtained includes the class (label) from the polarity of sentiment. There are 3 classes defined, namely positive, negative and neutral. Before carrying out the classification stage using the K-Nearest Neighbor algorithm, a pre-processing process is needed so that later each word has a weight from the TF-IDF results, then it can be processed to the next stage. The following is the result after preprocessing the data.

Table 1. Data Pre-processing Results

Doc	Before	After	Sentiment
D1	teman', 'rencana', 'libur', 'kebijakan', 'pemerintah', 'protokol', 'kesehatan', 'libur'	teman', 'rencana', 'libur', 'bijak', 'pemerintah', 'protokol', 'sehat', 'libur'	Positive
D2	himbauan','masyarakat', 'mematuhi', 'protokol', 'kesehatan', 'disiplin', 'mencegah', 'penyebaran', 'covid'	imbau','masyarakat', 'patuh', 'protokol', 'sehat', 'disiplin', 'cegah', 'sebar', 'covid'	Neutral
D3	patuhi', 'protokol', 'kesehatan'	patuh', 'protokol', 'sehat'	Neutral
⋮	⋮	⋮	⋮
D15	'pembinanya', 'patuhi', 'protokol', 'kesehatan'	pembina', 'patuh', 'protokol', 'sehat'	Positive

3.2 Cosine Distance

The Cosine method mainly used to find the similarity between two objects basend on cosine degree between two objects. The formula to calculate the cosine distance is

$$D(X,Y) = 1 - Cosine\ \theta \tag{1}$$

Where θ is the degree from one objects to other objects with corresponding to origin plane. The result range from 0 to 1 where 1 means two objects were not similar and 0 means two objects are identical.

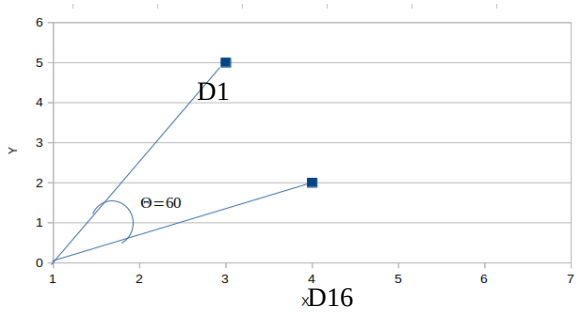

Figure 2 *Counting similarity between two documents using Cosine distance*

3.3 TF-IDF Word Weighting

Calculating how often the words appear in the document in relation with other documents, using the TF method [15]

$$TF\ (t,d) = \frac{count(t)}{count\ (d)} \tag{2}$$

Where t is number of occurrence for each word, and d is the number of documents. The greater the result means the word is appear many times in all documents, and the lower the result means the word is rarely being used.
IDF method is used to know which documents is refered to related to specific word,

$$\text{IDF}\,(t, \text{D}) = \log \frac{number\ of\ documents}{number\ of\ documents\ containing\ t} \tag{3}$$

Where t is the word that used as search filtered.The higher IDF number means the word is not commonly used or just a specific documents that use the word. The lower the IDF number means the word is commonly used.

Each words that has relation with corpus data that will determined the attitude or sentiment from the word categories. For that process we need to calculate the TF-IDF

$$\text{TF} - \text{IDF}\,(t, d, \text{D}) = TF(t, d) * IDF(t, D) \tag{4}$$

Where t is number of word appearances in all documents, d is number of documents that contains the word, and D is total number of all documents.

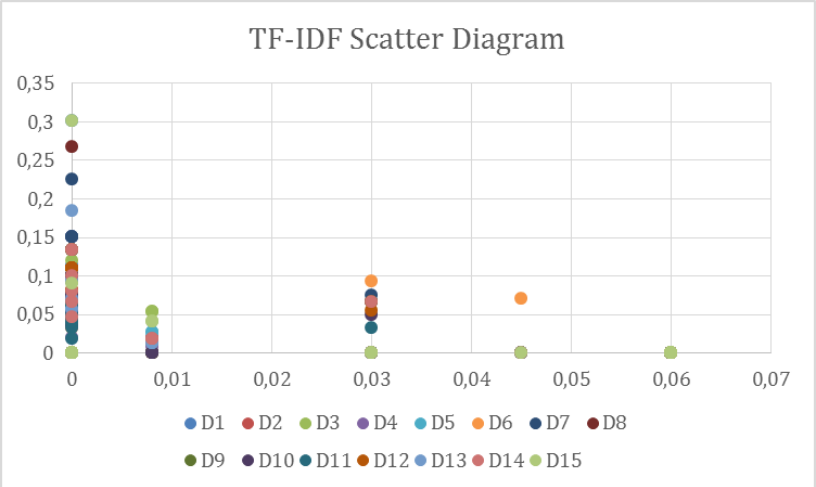

Figure 3 *Scatter Diagram of Normalized TF-IDF comparing similarity with D16*

Table 2. TF-IDF results

Query	**Term Frequency-Inverse Document Frequency (TF-IDF)**						
	D1	**D2**	**D3**	**D4**	**D5**	**…**	**D16**
sampai	0	0	0	0	0	…	0.06
lawan	0	0	0	0	0	…	0.06
korona	0	0	0	0	0	…	0.06
kendali	0	0	0	0	0	…	0.06
pandemi	0	0	0	0	0	…	0.06
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
pembina	0	0	0	0	0	…	0

3.4 K-Nearest Neighbors Classification

After completing the word weighting process, the next step was the implementation of the K-Nearest Neighbors algorithm. In this phase, the value of K was determined assuming it is equivalent to five (K = 5). The next stage was the calculation of the test data proximity to the training data using the cosine similarity distance method. From the word weights obtained, the scalar multiplication between the test and the overall training data were calculated and added. Furthermore, the length of each document, including test data, was determined by squaring the weight of each word in the text, adding the squared values, and calculating the root. Table 3 shows the calculation results.

Table 3. Document Length Calculation Results

D1	D2	D3	D4	D5	...	D15
0	0	0	0	0	...	0.0036
0	0	0	0	0	...	0.0036
0	0	0	0	0	...	0.0036
0	0	0	0	0	...	0.0036
0	0	0	0	0	...	0.0036
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.1625	**0.069583**	**0.020232**	**0.070546**	**0.046509**	**...**	**0.055253**
0.4031	**0.2638**	**0.1422**	**0.2656**	**0.2157**	**...**	**0.2351**

After the document length was obtained, the correlation between the test data and the entire training data was estimated using cosine similarity. From the outcome, the distance values from the smallest to the largest were sorted. This process is commonly known as the document ranking process. Table 4 shows the overall results of the calculated distance.

Table 4. Cosine Similarity results

No	Document	Distance	Rank
1	D16,D4	0.00231	1
2	D16,D13	0.00275	2
3	D16,D1	0.00338	3
4	D16,D5	0.00852	4
⋮	⋮	⋮	⋮
15	D16,D6	0.09904	15

Since the value of K is assumed to be five (K = 5), about five documents ranking from 5 and above were selected in order to observe the sentiments.

Table 5. Document Ranking Results

No	Document	Distance	Rank	Sentiment
1	D16,D4	0.00231	1	Neutral
2	D16,D13	0.00275	2	Neutral
3	D16,D1	0.00338	3	Positive
4	D16,D5	0.00852	4	Neutral
5	D16,D15	0.00873	5	Negative

From the table above, it was observed that D16 (test data) is Neutral. This is because positive sentiment is more dominant than others.

3.5 Validation

Accuracy testing is needed to determine the level of effectiveness of the classification. Testing accuracy in this study using a confusion matrix. Accuracy testing is carried out to test the validity of the accuracy value. The test results based on the K value were carried out 9 times with the overall results can be seen in table (6).

Table 6. Results of K Value Accuracy

K Value	Accuracy
2	57%
3	60%
4	62%
5	56%
6	57%
7	56%
8	60%
9	57%
10	58%

The results of the highest level of accuracy based on the system obtained by 62% obtained with a value of K = 4 with the following calculations.

Table 7. Confusion Matrix Table

Actual	Prediction		
	Positive	Negative	Neutral
Positive	33	2	10
Negative	13	0	6
Neutral	7	0	29

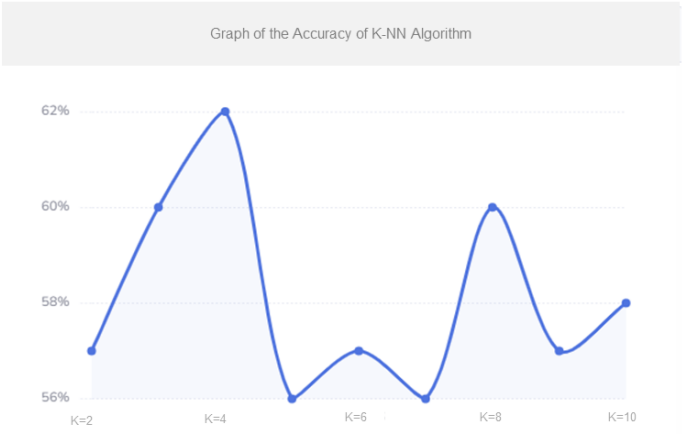

Figure 4 *Visualization of Overall K Accuracy Test*

4 Conclusions

Based on the findings of the analysis and discussion, it is possible to conclude that the data for this research effectively applied the K-Nearest Neighbor algorithm for sentiment classification on Twitter tweets including the terms 'health protocol' (*protokol kesehatan* in Bahasa). To conduct research, the preprocessing procedure of filtering data and word weighting calculations using TF-IDF was successfully carried out using 1600 data, 1500 tweet data grouping for training data and 100 tweets for test data. The maximum accuracy test result, 62%, is obtained by utilizing the K value = 4. According to the system categorization findings, there is 46.999% positive emotion, 9.063% negative sentiment, and 43.938% neutral sentiment.

Acknowledgments

This research was fully funded by the LP3M PUP program for the 2022 budget year, Universitas Maritim Raja Ali Haji

References

[1] Hustinawaty., Dwiputra, R.A.A., and Rumambi, T. 2019. Public Sentiment Analysis Of Pasar Lama Tangerang Using K-Nearest Neighbor Method And Programming Language R. *Jurnal Ilmiah Informatika Komputer. Vol.24. No. 2.*

[2] Liu, B. 2015. Sentiment analysis : Mining Opinions, Sentiments, and Emotions. USA: Cambridge University Press. Cambridge: USA

[3] Pratama, A.Y. Umaidah, Y. dan Voutama, A. 2021. Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square [Kasus

Omnibus Law Cipta Kerja]. *Jurnal Sains Komputer & Informatika [J-SAKTI]*. Vol. 5. No.2.

[4] Samsir. 2021. Analisis Sentimen Pembelajaran Daring pada Twitter di Masa Pandemi COVID-19 menggunakan Metode Naive Bayes. *Jurnal Media Informatika Budidarma*.Vol.5. No.1. 157-163.

[5] Williams, L.A. 2017. Pushing the Envelope of Sentiment Analysis Beyond Words and Polarities (doctoral dissertation). School of Computer Science & Informatics. Cardiff University

[6] Pamuji, A., 2021, Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment, Jurnal Informatika dan Sistem Informasi (JUISI), Vol. 07, No. 01.

[7] Isnain, A.R., Supriyanto, J., and Kharisma, M.P. 2021. Implementation of K-Nearest Neighbor (K-NN) AlgorithmFor Public Sentiment Analysis of Online Learning. Indonesian Journal of Computing and Cybernetics Systems (IJCCS), Vol. 15, No. 2.

[8] Amrullah, A.Z., Anas, A.S., and Hidayat, M.A.J., 2020, Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier dengan Seleksi Fitur Chi Square, Jurnal BITe, Vol. 2, No. 1, 40-44.

[9] Pamungkas, F.S., and Kharisudin, I., 2021 Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter, Prosiding Seminar Nasional Matematika , Vol. 4

[10] B. B. K, A. P. Rodrigues and N. N. Chiplunkar, "Comparative Study of Machine Learning Techniques in Sentimental Analysis," in International Conference on Inventive Communication and Computational Technologies, -, 2017.

[11] Ignatow, G. and Mihalcea, R. 2018. An Introduction to Text Mining: Research Design, Data Collection, and Analysis. SAGE Publications. USA

[12] Saputra, P.Y. 2017. Implementasi Teknik Crawling Untuk Pengumpulan Data dari Media Sosial Twitter. Jurnal Dinamika Dotcom. Vol.8. No.2.

[13] Taunk, K., De, S., and Verma, S., 2019, A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, International Conference on Intelligent Computing and Control Systems (ICCS).

[14] Suwanda, R., Syahputra, Z. And Zamzami, E. M., 2019, Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K, Journal of Physics: Conference Series.1566.

[15] Lane, H. Howard, C. dan Hapke, H.M. 2019. Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python. Manning Publication.Co. Shelter Island