Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment

Agus Pamuji

Abstract— **Each user who interacts with the Internet and information technology can provide feedback on the specific application. One of the applications which were used as social media is youtube. Various comments are given so that it becomes a challenge for the organizers. In this study, the concept of data mining was needed through the K-Nearest Neighbor method as a tool for classifying in addition to investigating comments that have the potential to be sentimental. There are three factors that are observed as input data, namely Services Quality, Information Quality, and Responsibility when the dataset is collected from social media applications. The initial phase of analysis is to extract the datasets from youtube then carried out pre-processing to the analysis phase. As the result shows that the method which was proposed is able to describe the accuracy rate of up to 88% through the confusion matrix technique. Therefore, the performance of K-Nearest Neighbor has provided a classification with positive and negative sentiment analysis classes**

Keywords: Sentiment Analysis, Social Media, K-Nearest Neighbor, Data Mining, Confusion Matrix

Abstrak— **Setiap pengguna yang berinteraksi dengan teknologi internet dan informasi dapat memberikan umpan balik terhadap aplikasi tertentu. Salah satu aplikasi yang digunakan seperti media sosial yaitu youtube. Berbagai macam komentar yang diberikan sehingga menjadi tantangan kepada pengelola. Dalam penelitian ini, konsep data mining diperlukan melalui metode K-Nearest Neighbour sebagai alat bantu untuk mengklasifikasi disamping mengivstigasi komentar – komentar yang berpotensi sentimen. Ada tiga faktor yang diobservasi sebagai data input yaitu Services Quality, Information Quality, and Responsibility ketika dataset dikumpulkan dari aplikasi media sosial. Tahap awal analisa dengan mengekstrak dataset dari youtube kemudian dilakukan pre-processing sampai pada tahap analisis. Hasil akhir menunjukan bahwa metode yang diusulkan mampu mendeskripsikan tingkat akurasi mencapai 88% melalui teknik confusion matrix. Dengan demikian, kinerja K-Nearest Neighbour telah memberikan klasifikasi dengan kelas analisis sentimen positif dan negatif**

***Kata Kunci*: Analisis Sentimen, Media Sosial, K-Nearest Neighbour, Data Mining, Confusion Matrix.**

I. INTRODUCTION

Technology users unwittingly continue to increase, especially in information technology interactions [1]. Social media is a medium which is used by users to provide comments or suggestions that have the potential to have positive or negative values. There are a variety of behaviors that are owned each user when interacting with a technology. One of them is the attitude of sentiment which has as a positive and negative reaction. Sentiment can be considered as an exaggerated attitude contained in the services, products, services, religion, politics and other topics. [2]. Someone who is uses the media and have a sentiment is usually was found on the text-based comments and and continuously increasing. Therefore, comment recipients will have face challenges regarding managing a wide variety of comments including sentimental comments. Research that discusses the study of sentiment analysis has been carried out by many researchers, academics and also practitioners in the field of computing or behavioral analysis. In addition, this study gives a lot of special attention due to the reasons for the feedback to the managers and owners of the message data in the form of many comments with various types. In an effort to accelerate the comment management process, it is necessary was applied on the sentiment analysis to investigate whether the comments were generally positive or negative [3]. In this context, a K-Nearest Neighbor (KNN) method would be applied and adoption for classifying comments [4]. Data which was contained the many comment would be explored in the Indonesian language content available on social media applications such as Youtube. We would have explored the review of various comments in a sentence which were means sentiment come from each user [5]. The purpose of analyzing the sentiment behavior of social media users was to increase the accuracy of the classification. The performance of the K-Nearest Neighbor method would be carried out using a Confusion Matrix and two classes.

Data mining can be conceptualized as the process of extracting knowledge from large amounts of data [6]. The application of data mining methods is growing rapidly in the last two decades because it is triggered by the increasing amount of data generated and its methods tend to be more efficient than traditional analytical methods which have been applied for a long time [7] [8]. One application of data mining in the field of information systems, and especially behavior is in processing data such as available comments on social media for example are Facebook, Twitter, YouTube and so on [9].

The K-Nearest Neighbor concept, a classification algorithm is a method which is classifies objects based on the learning data that is closest to the object. On the other hand, the K-Nearest Neighbor (K-NN) Algorithm is a classification method for a set of data based on previously classified data learning [10]. KNN methods and algorithms are included in supervised learning, where the results of the new query instance are classified based on the majority of the distance proximity of the categories in K-NN [11].

Near or far neighbors are usually calculated based on Euclidean Distance [12], or it can also use the distance

[1] *Dosen Bimbinaan dan Konselina Islam. IAIN Svekh Nuriati Cirebon, Jalan Periuanaan by Pass Sunyaragi Kota Cirebon Jawa Barat.45132 (e-mail: jurnal.agus.pamuji@gmail.com)*

formula to another, as described in the article of Vector Space Model and Measurement Distance. Proximity can be thought of as the inverse of distance, or inversely proportional to distance. The smaller the distance between two instances, the greater the "closeness" between the two instances. Thus, the K Nearest Neighbors of an instance x are defined as the K instances that have the smallest distance (nearest, closest) to x.

Learning on the data is depicted in a multidimensional space with each dimension representing each feature/feature of the data. The new data classification was conducted by looking for the labels of the k nearest neighbors. The label that appears the most becomes the new data label. If k = 1, the new data is labeled with the nearest neighbor label. The distance commonly used is the Euclidean distance.

In the learning phase, this algorithm to store feature vector and classification of learning data. In the classification phase, the same features are calculated for the test data (whose classification is unknown). The distance from this new vector to all learning data vectors is calculated and the closest K number is taken. The new point is labeled based on the most labels of the points [13].

Related to the study of sentiment analysis, there are aspects which are of concern such as positive and negative responses. Responses would appear when a faced with certain situations. For example, various of comments on social media such as blogs and twitter have the potential to contain informal text where each user can understand what the message is conveying. However, the system can't understand it.

II. LITERATURE REVIEW

A. Related Work

There are many research journal literatures which are to examine behavior with a data mining approach. Furthermore, this case cannot be separated from the relationship between the data analysis and various methods in the context of machine learning. In addition, several methods which had been used, especially in analyzing sentiment behavior, had been disclosed. Sentiment behavior was analyzed, such as Suchita V Wawre et al (2013), whose conducted their research with objects and data sets in the form of Twitter users when providing film reviews. [14]. The methods that used are Naïve Bayes and Super Vector Machine. In These two methods were compared with the result that Naïve Bayes is superior to the Super Vector Machine. However, in another context was proposed by Bhavita (2017), the results of the analysis show that are the Super Vector Machine (SVM) is better and even reaches a value of 85% when compared to the Unsupervised Learning technique [15]. Likewise, the dataset taken was derived from the results of a review of customer attitudes towards the product.

Pamungkas (2021) and also proposed analyzing sentiment behavior using data sets from Twitter by comparing analytical techniques such as K-Nearest Neighbor, Naïve Bayes, and Super Vector Machine. Thus, the results show that SVM is better, although SVM is still was considered relevant technique, sentiment analysis research had been carried out by some researchers using Naïve Bayes with the same results. As proposed by Sari (2019) when analyzing customer behavior provides an overview of online stores. The researcher uses three classes as a classification, namely positive, negative and neutral [16]. In addition, the study added an emotional icon conversion feature. The final result shows that Naïve Bayes effectively has 96.44% before using the additional feature, namely the conversion of emotional icons which changes to 98%. In the same year as proposed by Fauzi, they still use social media such as Instagram [17]. The research material used the Naïve Bayes method and adds features, namely tokenize comments and Positive Transform Cases. The result of the last review was that some of the methods used by some researchers use additional features with the aim of being able to find out changes in the effectiveness of each test result. The using of the K – Nearest Neighbor method in the field of machine learning. In fact, the KNN method was considered as an effective method in analyzing behavior, especially sentiment analysis [18].

III. RESEARCH METHODOLOGY

This paper contains the field of data mining studies involving machine learning techniques. Thus, as for the method used with the concept of quantitative research [19]. In investigating there are the existence of comments that have the potential to be positive and negative, this method will explain the stages in processing with a data mining approach. The stages of the data mining process include preparing a dataset with the meaning of collecting data. Related to this, the conceptual would used private dataset derived from social media like youtube [20]. The topic of sentiment analysis was conducted on the public policy as medium which used by government. The second stage, using the method, was implementing the K-Nearest Neighbor method to be able to classify training data. The next stage was understanding the model and the appropriate knowledge as outputs after entering the analysis process using the KNN method [21].

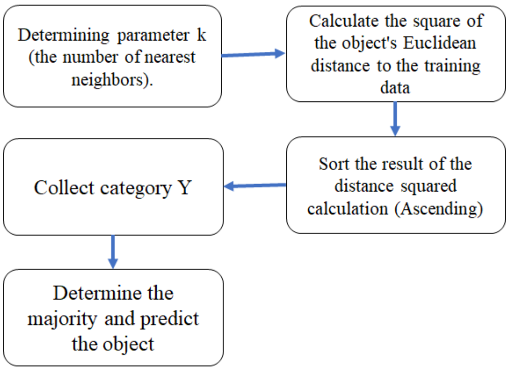
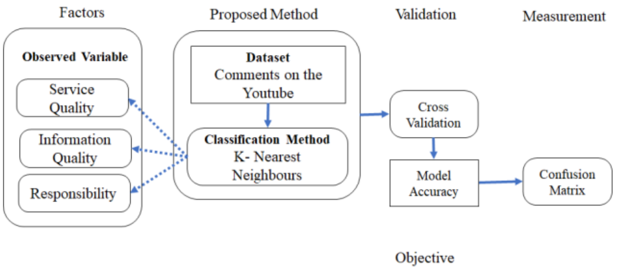

Determining parameter k (the number of nearest neighbors).
Calculate the square of the object's Euclidean distance to the training data
Sort the result of the distance squared calculation (Ascending)
Collect category Y
Determine the majority and predict the object

Figure 1. K-NN Analisa Analysis Steps

When the classification using the K-NN method has several stages of the analysis process. First, determine the parameter K [22]. The K parameter in KNN is used as a determination of the number of neighbors was identified. After the nearest neighbor was known, and it can do to calculate the distance of the object using the Eucledian technique. The next stage, sorting of the results of calculation of the distance squared, then displaying the data categories that were considered as Y with the diagram above [23].



G Figure 2. Framework for Analysis of the K-NN . Method

Because of the picture above as a diagram of the K-NN analysis framework. There are three factors which were used as inputs, namely Services Quality, Information Quality, and Responsibility. While the datasets were collected through private datasets through the YouTube as social media application. The K-NN method can be applied when the dataset is ready to go through the pre-processing process. The next step is to validate through cross validation and measurement of model performance using the Confusion Matrix.

The last stage is to evaluate the performance of the model that provides an overview of the level of accuracy or error. The KNN method would be tested on the model using the Confusion Matrix. The following is a research method carried out with the following image.

A. K-Nearest Neighbour

K-Nearest Neighbor as one of method for machine learning had been knowing a method for classifying objects based on datasets which were used as learning data [24]. Its implementation was expected to use the closest distance or similarity to the object. Due to K-Nearest Neighbor as a technique in machine learning, this method can store to each feature vector and the classification of learning data. In the classification phase, the same features were calculated for the test data (whose classification is unknown). The distance from this new vector to the learning data vector was calculated. Next, the closest K number will be taken. The newly classified points are predicted to be included in the most classification of these points. Each point near or far will be calculated using Euclidean Distance techniques. While the text-based classification, the determination of Euclidean Distance menggukan cosine similarity. The greater the value of Euclidean Distance the more remote degree of similarity between the test data and data learning. If the value of Euclidean Distance getting smaller the more closely the degree of similarity between the test data with the data of learning. However, the text classification, the greater the cosine similarity value the more closely the degree of similarity between the test data and learning data, and vice versa if the cosine similarity of its value the smaller the more remote degree of similarity between the test data and learning data. The following is the Euclidean Distance equation [25].

$$P1\ dan\ P2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (1)$$

B. Consine Similarity

Another performance measurement on K-Nearest Neighbours (KNN) used the Cosine Similarity method, which is a method to calculate the similarity between two pieces of information [26]. In addition, the Consine Similarity can do compare the similarities between the information in the document [27]. Determination of the suitability of the information with the query was considered as measurement (similarity measure) between the vector of the document (D) to the query vector (Q) [28]. The more similar a document vector is to a query vector, the document can be seen as more compatible with the query [29]. By using Cosine Similarity it will be represented using the following equation:

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

C. Confusion matrix

The Confusion Matrix is a matrix of size M x N which is used to evaluate the performance of the classification model, where N represents the number of target classes. [30]. The matrix compares the actual target values with those predicted by the machine learning model [31]. This provides a holistic explanation of how well the built classification model performs and the types of errors. In this study, we will use two classes (positive and negative) was called the Confusion Matrix for Binnary Classification [32].

TABLE I
CONFUSION MATRIX CRITERIA

Valid Classifications	Classifications action	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Based on the table above, the values of (TN), (FP), (FN), and (TP) will be obtained for accuracy, precision and recall values. On the value of accuracy testified how accurate the model can classify the data correctly [33]. In addition, the accuracy value was a comparison between data which was confirmed to be classified as correct and the entire data.

The accuracy value can be obtained by Equations. Furthermore, the precision value provides information on the number of data that has a positive category that is classified correctly divided by the total data that is classified as positive. Finally, the precision value can be generated by Equations. Meanwhile, the recall value shows how many percent of the data are positive categories that the model has confirmed correctly that the recall value is obtained by Equation [34].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \qquad (3)$$

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (4)$$

$$Recall = \frac{TP}{FN+TP} * 100\% \quad (5)$$

IV. RESULTS AND DISCUSSION

A. Datasets

The process in the first data mining was to collect data which can be used as analysis material [35]. There is a factor attribute or parameter will be a class or label. There are two datasets used, namely private datasets and public datasets. Private datasets are datasets obtained from organizations, agencies, companies that will be the object of research. Public datasets are data sets obtained from public repositories that have been agreed apart from being known by data mining researchers [36]. Therefore, the dataset which were taken is included in the category of private dataset. The youtube application contains both positive and negative comments related to public policies so that they can be taken for analysis.

D.

B. Data Pre-Processing

Furthermore, the dataset that has been taken from the youtube application, then the data enters the pre-processing phase as a condition for classification. Each sentence will be broken down into words as a tokenizing stage followed by a stopword removal process. Each word is verified by equating word forms, eliminating invalid ones, and reducing vocabulary frequency. The rest, every word will be changed to lowercase, excluding numbers which is called Case-folding. The last step, basic words that have meaning to be analyzed using the stemming algorithm [38].

C. C lassification of sentiment comments

Analysis of comments containing sentimental meaning, negation will be applied. Next, count the number of positive and negative sentiments in the data to be tested. If there is an indication that positive sentiment exceeds the number of negative sentiments, the test data is positive sentiment. However, on the contrary, the test data is a negative sentiment.

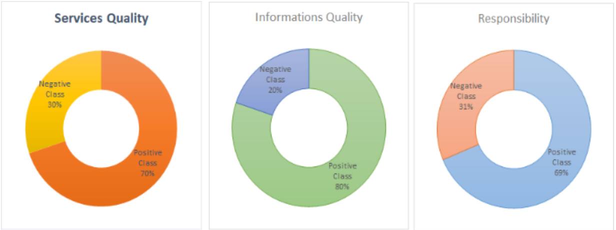


Figure 3. K-NN . Classification Results

The classification stage is carried out using the K-Nearest Neighbor Algorithm through three stages, namely the training stage, the testing stage and the performance evaluation. There are several stages in the K Nearest Neighbor algorithm, namely, determining the k parameter (number of closest neighbors), Calculating the square of the object's Euclidean distance to the given training data. Sort the results in ascending order (in order from high to low value), and Collect Y category (Classification of nearest neighbors based on k value). Comment data will be processed using K-Nearest Neighbor. The data presented are as follows. In the table below, service quality is represented by X, information quality is represented by Y, and responsibility is denoted by z.

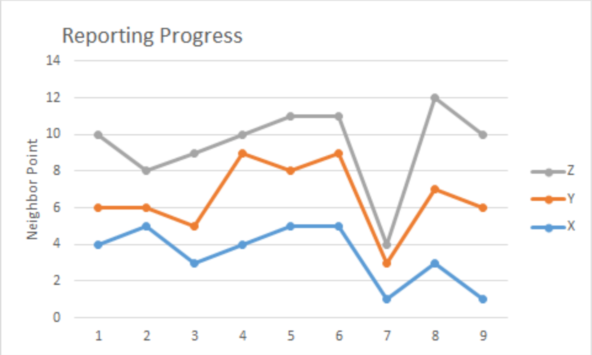


Figure 4. K-NN . Sentiment Analysis

The data analyzed were 357 comments containing sentiment, while the validation used 10-fold cross validation to test the model with the aim of getting accurate results.

D. Model Evaluation

The following is a table of experimental results with changes in the value of K [39]. There are 18 iterations indicated by the value of each measurement criteria.

TABLE II
K-NN ANALYSIS MODEL PERFORMANCE

K	Accuracy	Presisi	Recall
1	74	83	68
2	70	88	69
3	60	70	72
4	65	88	82
5	78	80	68
6	76	72	73
7	60	83	69
8	62	86	86
9	68	81	66
10	81	72	69
11	85	83	60
12	83	83	65
13	72	62	75
14	69	76	63
15	68	69	79
16	74	75	86
17	* 88	83	78
18	80	87	77

Based on the experimental table for changes in K, the table above shows the highest accuracy value is at K = 17 with an accuracy value of 88%. Next in the evaluation stage using Confusion.

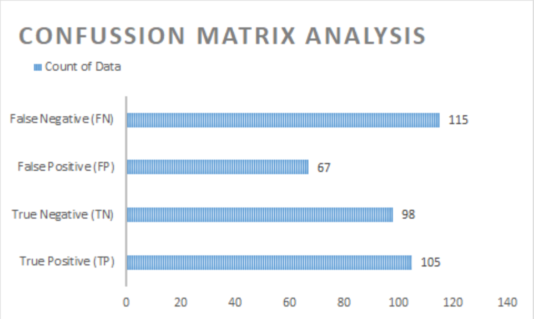


Figure 5. Confusion Matrix Test Results

Based on the table, the prediction results show the greatest accuracy value at K = 17. The accuracy value reaches 88%, precision is 83%, and recall is 78%. The K-Nearest Neighbor method has good performance in machine learning but requires a lot of time to process the model and analyze the training data. Methods in machine learning can be collaborated with other selected features. Feature selection can improve learning performance. In this study, we did not have feature selection and it was proven from the experimental results that it was quite good. Future research, machine learning can be collaborated or added feature selection.

Data mining algorithms can be categorized as simple data mining techniques [40] [6]. However, the main drawback is when large data is found and contains noise [41].

Discussion about machine learning, the k-Nearest Neighbors (kNN) method is the most popular and simplest algorithm in machine learning classifier. Along with its popularity, kNN is widely used to classify data in the fields of science and engineering as well as economics and business. K-NN was first introduced by T. Cover and P. Hart in 1967 where this algorithm classifies sample classes based on their closest neighbor classes. K-NN is often referred to as a lazy learner, because kNN learns and classifies data without building a model. Unlike model-based classification algorithms, the kNN classifier only needs to remember all the training data in memory.

V. CONCLUSION

The K-NN method is considered effective and has good performance even without feature selection. The table above shows good performance by achieving an accuracy exceeding 50%. The success of K-NN without feature selection is supported by a strict pre-processing process, but it takes a lot of time and effort in terms of technical data. Sentiment analysis on public policy was successfully classified with two positive and negative classes on all comments received on YouTube social media. The test results get the highest accuracy score of 88%, precision 83%, and recall 78% The final result, the decision is that public policies on social media get positive sentiment.

REFERENCES

[1] K. C. Loudon and J. P. Laudon, Management Information Systems, New Jersey: Pearson, 2016.

[2] R. P. Fitrianti, A. Kurniawati and D. Agusten, "Implementasi Algoritma K - Nearest Neighbor Terhadap Analisis Sentimen Review Restoran Dengan Teks Bahasa Indonesia," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, Yogyakarta, 2019.

[3] C. Catal, "A Sentiment Classification Model Based On Multiple Classifiers," *Applied Soft Computing,* Vols. -, no. -, pp. 1 - 19, 2016.

[4] S. Ernawati and R. Wati, "Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel," *Khatulistiwa Informatika,* vol. 6, no. 1, pp. 64 - 69, 2018.

[5] A. A. L. Cunha, M. C. Costa and M. .. A. C. Pacheco, "Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks," in *Lecture Note in Computer Science*, -, Springer, 2019, pp. 561 - 570.

[6] A. A and S. B., Data Analysis and Data Mining: An Introduction, Oxford: Oxford University Press, 2012.

[7] R. Sullivan, Introduction to Data Mining for the Life Sciences, New Delhi: Humana Press, 2012.

[8] L. Cao, "Data Science: A Comprehensive Overview," *ACM Computinkg Survey,* vol. 50, no. 3, pp. 1 - 42, 2017.

[9] R. Novendri, A. S. Callista, D. N. Pratama and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *BULLETIN OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING,* vol. 1, no. 1, pp. 1 - 15, 2020.

[10] R. S. King, Cluster Analysis and Data Mining. An Introduction, Singapore: Mercury, 2015.

[11] M. Nikhitha and M. A. Jabbar, "K Nearest Neighbor Based Model for Intrusion Detection System," *International Journal of Recent Technology and Engineering (IJRTE),* vol. 8, no. 2, pp. 2258 - 2262, 2019.

[12] M. Bordoloi and S. K. Biswas, "Machine Learning based Sentiment Analysis using Graph Based Approach," in *ICCCNT,* Kanpur, India, 2019.

[13] P. Attewell and D. Monaghan, Data Mining for the Social Sciences: An Introduction, California: University of California Press, 2015.

[14] S. V. Wawre and S. N. Deshmukh, "Sentiment Classification using Machine Learning Techniques," *International Journal of Science and Research (IJSR),* vol. 5, no. 4, pp. 819 - 821, 2016.

[15] 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.

[16] A. P. Natasuwarna, "Analisis Sentimen Keputusan Pemindahan Ibukota Negara Menggunakan Klasifikasi Naive Bayes," in *Seminar Nasional Sistem Informasi dan Teknik Informatika*, 2019.

[17] D. Iskandar and Y. K. Suprapto, "Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C 4.5 Dan Naïve Bayes," *NERO,* vol. 2, no. 1, pp. 37 - 43, 2015.

[18] H. W. Y. R. M. Jelodar, "A NLP framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on youtube comments," *Multimed Tools Application,* vol. 80, no. -, p. 4155 – 4181, 2021.

[19] V. Bairagi and M. V. Munot, Research Methodology : A Practical and Scientific Approach, New York: CRC Press, 2019.

[20] G. Ignatow and R. Mihalcea, An Introduction to Text Mining Research Design Data Collection and

Analysis, Singapore: SAGE, 2018.

[21] A. Campbell, Data Visualization Guide: Clear Introduction to Data Mining, Analysis, and Visualization, New York: CRC Press, 2021.

[22] S. Sahara, R. A. Permana and Hariyanto, "Particle Swarm Optimization pada Analisa Review Software Antivirus Menggunakan Metode K-Nearest Neighbors," *Informatic for Educators and professional,* vol. 4, no. 2, pp. 123 - 132, 2020.

[23] Martha, V. C. M, D. S. Naga and P. T. D. Rompas, "Perbandingan Pengklasifikasi k-Nearest Neighbor dan Neighbor-Weighted k-Nearest Neighbor Pada Sistem Analisis Sentimen dengan Data Microblog," *Jurnal Sains dan Teknologi,* vol. 1, no. 1, pp. 81 - 90, 2018.

[24] T. Denœux, O. Kanjanatarakul and S. Sriboonchitta, "A new evidential K-nearest neighbor rule based on contextual discounting with partially supervised learning," *International Journal of Approximate Reasoning,* vol. 113, no. -, pp. 287-302, 2019.

[25] J. Gou, H. Ma, . W. Ou, S. Zeng and Y. Rao, "A generalized mean distance-based k-nearest neighbor classifier," *Expert Systems with Applications,* vol. 115, no. 2, pp. 356-372, 2019.

[26] J. Yen, "Generalized Ordered Weighted Simplified Neutrosophic Cosine Similarity Measure for Multiple Attribute Group Decision Making," *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* , vol. 14, no. 1, pp. 1 - 12, 2020.

[27] X.-S. Yang, Introduction to Algorithms for Data Mining and Machine Learning, Hongkong: Academic Press, 2019.

[28] T. Thongtan and T. Phienthrakul, "Sentiment Classification using Document Embeddings trained with Cosine Similarity," in *Annual Meeting of the Association for Computational Linguistics: Student Research Workshop,* Florence - Italy, 2019.

[29] K. Park, J. S. Hong and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Applied Artificial Intelegence,* vol. 34, no. 5, pp. 396 - 411, 2020.

[30] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," in *International Conference on Machine Vision and Information Technology (CMVIT)* , Guangzhou - China, 2019.

[31] A.-J. Gallego, J.-. Zaragoza, J. J. .Valero-Mas and J. R. Juan, "Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation," *Pattern Recognition,* vol. 74, no. 1, pp. 531 - 543, 2018.

[32] J. Xu, Y. Zhang and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Information Science,* vol. 507, no. -, pp. 772 - 794, 2020.

[33] W. W, "Weakly Supervised Learning by a Confusion Matrix of Contexts. In: U. L., Lauw H. (eds) Trends and Applications in Knowledge Discovery and Data Mining," in *Lecture Notes in Computer Science,* Cham, Springer, 2019, pp. 59 - 64.

[34] A. Leque, A. Carrasco, A. Martin and A. d. L. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition,* vol. 91, no. -, pp. 216 - 231, 2019.

[35] K. Jamsa, Introduction to Data Mining and Analytics, Boston: Jones & Bartlett Learning LLC, 2021.

[36] U. Ojha and S. Goel, "A Study On Prediction of Breast Cancers Recurence Using Data Mining Techniques," in *International Conference on Cloud Computing, Data Science & Engineering,* -, 2017.

[37] A. Z. Saiz, C. Q. González, L. H. Gil and D. M. Ruiz, An Introduction to Data Analysis in R: Hands-on Coding, Data Mining, Visualization and Statistics from Scratch, New Jersey USA: Springer International Publishing, 2020.

[38] N. V. G. Raju and K. P. Lakshmi, K. G. P, "Prediction of chronic kidney disease (CKD) using Data Science," in *International Conference on Intelligent Computing and Control Systems (ICCS),*, Madurai - India, 2019.

[39] L. Kuang, H. Yan, Y. Zhu, S. Tu and . X. , "Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor, Journal of Intelligent Transportation Systems," *Journal of Intelligent Transportation Systems* , vol. 23, no. 2, pp. 161-174, 2019.

[40] W. L, . X. Q., L. H and Cao, "Multi-criteria decision making method based on improved cosine similarity measure with interval neutrosophic sets," *International Journal of Intelligent Computing and Cybernetics,* vol. 12, no. 3, pp. 414-423, 2019.

[41] K. Mittal, G. Aggarwal and P. Mahajan , "Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy," *International Journal Informatic and Technology,* vol. 11, no. -, pp. 535 - 540, 2019.