International Journal of Advances in Data and Information Systems
Vol. 3, No. 1, April 2022, pp. 30~37
ISSN: 2721-3056, DOI: 10.25008/ijadis.v3i1.1233

Sentiment Analysis of Kampus Mengajar 2 Toward the Implementation of Merdeka Belajar Kampus Merdeka Using Naïve Bayes and Euclidean Distence Methods

Abdul Rozaq[1], Yessi Yunitasari[2], Kelik Sussolaikah*[3], Eka Resty Novieta Sari[4]
[1,2,3,4]Department of Informatics Engineering, Universitas PGRI Madiun, Indonesia

Article Info

Article history:

Received Jan 19, 2022
Revised Mar 10, 2022
Accepted Apr17, 2022

Keywords:

Sentiment Analysis
Naïve Bayes
K-NN
MBKM

ABSTRACT

The Ministry of Education and Culture initiated the *Merdeka Belajar Kampus Merdeka* (MBKM) program. Several programs in *Merdeka Belajar Kampus Merdeka* (MBKM) Program include industrial internships, independent projects, student exchanges, community service projects, humanitarian programs, and so on. *Kampus Mengajar* 2 is one of the programs had been running. The program received various responses from the public, which were expressed on social media. The Supervisor at *kampus mengajar* 2 was also active in providing various comments on *kampus mengajar* 2 telegram groups in the form of good, bad, and neutral comments. These comments have the potential to generate a growing sentiment among the general public and academics. Based on these issues, the researcher analyzed the kampus mengajar 2 sentiments toward the implementation of *Merdeka Belajar Kampus Merdeka* program with the data source being comments on the supervisors' telegram group. The data obtained from the telegram group is classified as good, bad, or neutral using the Naive Bayes method and K-Nearest Neighbors on up to 591 data points. The data is then divided into two parts: training data and testing data. Testing data can account for up to 20 percent of total data, with the remaining 80 percent serving as training data. The accuracy results on sentiment analysis show that the Naive Bayes method outperforms the KNN method, with 99.30 percent for Naive Bayes and 97.20 percent for K-Nearest Neighbors.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Kelik Sussolaikah
Department of Informatic Engineering,
PGRI Madiun University,
Auri Street, Madiun, East Java, Indonesia
Email: kelik@unipma.ac.id

1. INTRODUCTION

A student is an agent of change who must be able to adapt and use technology to develop themselves. The Ministry of Education and Culture launched *Kampus Merdeka* program in the terms of giving students the freedom to learn things outside of their study program, so it will enhance student competence in responding to modern era needs as well as significant social, cultural, and technological changes [1][2][3]. *Merdeka Belajar Kampus Merdeka* (MBKM) program is a policy

that allows students to enhance their skills based on their talents and interests so that the students are ready to join the working world [4].

Kampus Mengajar is a component of *Merdeka Belajar Kampus Merdeka* (MBKM) program designed to enhance student competence. [5] This program is really helpful in enhancing the quality of learning at the elementary level by involving students in participating and contributing for one semester. While on duty, students face a variety of challenges that will test their leadership, creativity, innovation, problem solving, communication, and teamwork skills [6].

The implementation of the *Merdeka Belajar Kampus Merdeka* (MBKM) program will be successful if the university is willing and dare to change its mindset toward, an adaptive and flexible curriculum. The achievement of the *Merdeka Belajar Kampus Merdeka* (MBKM) program is accompanied by pros and cons in the academic world [7][8]. Academics and ordinary citizens frequently criticize government policies on social media [9]. Inspiration, complaining, expressing opinions, commenting on products, and expressing satisfaction or disappointment with something can be found on social media. Twitter is a popular social media platform for expressing ideas, thoughts, and feelings about government policy [10] [11].

The presence of various public opinions and comments on the *Merdeka Belajar Kampus Merdeka* policy on social media, particularly on Twitter, has become a community sentiment [12]. Not only Twitter but also Telegram is frequently used to convey various opinions, including satisfaction with the implementation of the MBKM program, specifically *Kampus Mengajar* 2.

According to a study [13] titled A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm, the comparison of the results of the proposed algorithm implementation and five other classification algorithms on 10 datasets selected from the UCI repository revealed a significant improvement in the algorithm's classification.

The previous study [14] The data set used in the study includes 7262 legitimate examples and 3793 phishing attacks with 30 features. Cross-validation is used to train and test classifiers ten times, and the experiment is repeated 20 times. The average value is taken into consideration for evaluation. The results show that SCAK-NN has the highest accuracy, F-Measure, and True Positive Rate, as well as the lowest mean absolute error and false positive rate. Future work will include testing the classifier with real-world examples and extracting the most relevant features to improve phishing attack prediction.

In addition, the study [15] The Naïve Bayes has proven to be a tractable and efficient method for classification in multivariate analysis. However, features are typically correlated, which contradicts the Naive Bayes assumption of conditional independence and may impair the method's performance. In this study, the proposed sparse Naive Bayes achieves competitive results in terms of accuracy, sparsity, and running times for balanced datasets when compared to well-referenced feature selection approaches. In the case of unbalanced (or different importance) classes in a dataset, a better compromise between classification rates for the different classes is achieved.

The study [16] is titled Towards Applying Internet of Things and Machine Learning For The Risk Prediction of COVID-19 In Pandemic Situation Using Naive Bayes Classifier For Improving Accuracy. The goal is to use Random Forest and Naive Bayes classifiers to predict from data collected using sensors. Groups of people are identified, and the disease's impact can be reduced for the population group with the highest population. Random Forest has a 97 percent accuracy rate, while Naive Bayes has a 99 percent accuracy rate.

According to [17] sentiment Classification of Roman-Urdu Opinions Using Naive Bayesian, Decision Tree, and KNN Classification Techniques. The majority of sentiment mining research has been conducted in English. There is currently little research being done on sentiment classification in other languages such as Arabic, Italian, Urdu, and Hindi. Three classification models are used in this study for text classification using the Waikato Environment for Knowledge Analysis (WEKA). Opinions in Roman-Urdu and English are taken from a blog. These extracted opinions are documented in text files to create a training dataset with 150 positive and 150 negative opinions as labeled examples. The testing data set is fed into three different models, and the results in each case are analyzed. The results show that Naive Bayes outperforms Decision Tree and KNN in terms of accuracy, precision, recall, and F-measure.

Sentiment Analysis Research [18] on the titled Sentiment Analysis of Social Media Twitter With Case of Anti LGBT Campaign In Indonesia Using Naïve Bayes, Decision Tree, And Random Forest Algorithm reviews that the sentiment analysis obtained in this study shows that Twitter users in Indonesia provide more neutral comments. In this study, an accuracy of 86.43 percent was obtained from testing data using the Naive Bayes Algorithm in RapidMiner tools, where the accuracy is higher than the other algorithms, Decision Tree and Random Forest which is 82.91 percent.

The study [19] on Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm. According to the findings of the study, the Modified Naive Bayes classification can classify academic performances with research activities better than the decision tree and ordinary Nave Bayes classifications, with 96.15 percent and 94.23 percent, respectively. Nonetheless, this study's findings are appropriate for preliminary classification at the university research office level before being submitted to the respective national educational authority for further review.

Based on the background and previous research, it is hoped that the sentiment analysis research of the kampus mengajar 2 towards the implementation of *Merdeka Belajar Kampus Merdeka* (MBKM) will be able to classify the sentiments that occur with a more effective and accurate method without having to do it manually so that assisting policymakers in evaluating programs have been implemented properly.

2. RESEARCH METHOD

2.1 Research Stages

Data crawling, preprocessing, feature extraction, and classification are the stages of sentiment analysis research. For 5 months of *kampus mengajar* 2, the data was compiled from various comments on the supervisor's telegram group. The next step after crawling the data was data labeling. The data is classified into three categories: positive, negative, and neutral.

The following stage is preprocessing. Preprocessing is stage in which we clean the data before extracting its features. Preprocessing can help to avoid data interference, imperfect data, and inconsistent data. The following preprocessing was used in this study:

1. The technique of transforming letters to lowercase is known as case folding.
2. Tokenization is the division of a remark text into tokens
3. Stopword removal is the process of removing words from the parsed result that aren't' relevant.'

2.2 Feature Extraction

Feature extraction procedure in responses is Term Frequency and TF-IDF (Term Frequency-Inverse Document Frequency). This stage includes the following steps:

1. Determine the Term Frequency
   Term Frequency is the frequency with which a term appears in a corpus.
2. Determine the TF-IDF (Term Frequency-Inverse Document Frequency)
   The Inverse document frequency is calculated first in order to calculate the TF-IDF (idf). After determining the idf, the TF-IDF weight value is calculated by multiplying the term frequency value by the inverse document frequency value for each term.

2.3 Classification of Comments

Classification is the process of categorizing new comments with unknown sentiment values [20][19]. Stages of classification using Naive Bayes and K-NN.

1. Naïve Bayes
   The Naive Bayes method is widely used in data classification. In this method, for example, if a new comment is entered, the probability of the comment being in a positive or negative class is calculated using the results of the training process [21][22]. The Bayes theorem equation, which states that:

$$P(A|B) = \frac{P(B \mid A)\, P(A)}{P(B)} \quad (1)$$

   Description :

   P(A|B) = The possibility of an event A if it is known that B

   P(B|A) = The possibility of an event B if it is known that A

   P(A) = Possibility of event A Peluang kejadian A

   P (B) = Possibility of event B

2. K-Nearest Neighbor

   The K-Nearest Neighbor (KNN) algorithm categorized objects regarding the learning data closest to the object. KNN includes a supervised learning algorithm in which the class with the most appearances becomes the class that results from the classification [23][24].

   Metric distances, such as Euclidean distance, are used to define proximity. Equation 2 can be used to calculate Euclidean distance:

$$Dxy = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

   Description:

   x_i: attribute x for i

   y_i : attribute y for i

   n : number of attributes

Performance Evaluation classifier used for evaluation is the confusion matrix. The confusion matrix is an important tool in the visualization method used on machine learning machines that typically fall into two or more categories. The visualization stage aims to make the data produced by classification easier to read. Figure 1 shows the entire sequence of research stages.

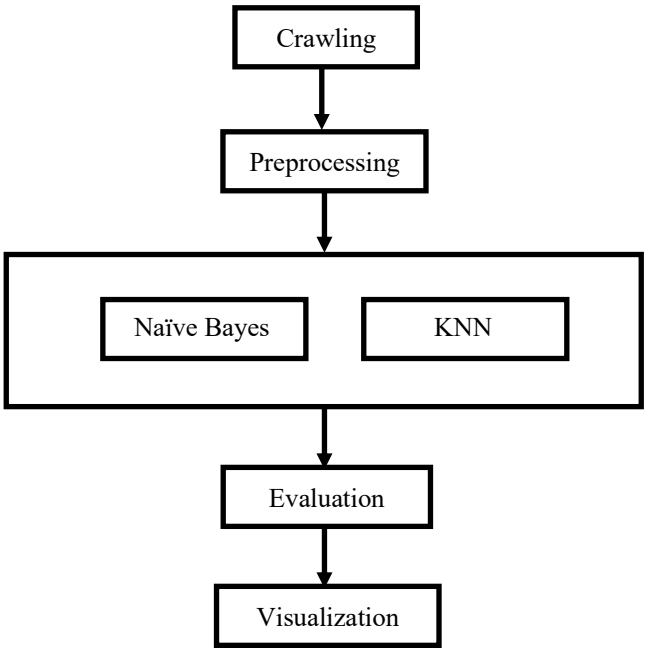

Figure 1. Research Stages

3. RESULTS AND DISCUSSION

Sentiment analysis on the implementation of *Kampus Mengajar* 2 based on data from the supervisor's telegram group conversation; various comments were classified as positive, negative, or neutral. The Naive Bayes and K-Nearest Neighbors methods are used in the sentiment analysis process. The two methods were chosen because they were used to compare the data accuracy of one method to another. The Naive Bayes method is a supervised learning algorithm that can make accurate predictions. The Naive Bayes method has the advantage of not requiring a large amount of training data to determine the estimated parameters needed in the classification process, which makes the classification process more effective and efficient [25][26]. While the K-Nearest Neighbor method is easier to implement, experiments with this method show that it can provide good performance for independent data (which does not have word dependence) [27].

The two methods are used in this sentiment analysis so that they can be used as benchmarks for the accuracy of the results of the classification of sentiments that occur on *kampus mengajar* 2 as many as 591 data for 5 months. The data is categorized into two: training data and testing data. Prior to further processing, each comment data must be labeled with positive, negative, and neutral labels. Table 1 shows the data labeling procedure

Table 1. Data Labeling

Data	Label
terima kasih untuk pengelola KM2 kami sudah menerima honor DPL	Positif
Sosialisasi Program MSIB Batch 2 Merdeka Belajar	Netral
Saya ga bisa login ulang, username dan atau password salah	Negatif

The data is cleaned after entering the preprocessing process so that there is no more inconsistent data. Some of the processes that have been passed include changing uppercase letters to lowercase letters, dividing a collection of text in the form of comments into tokens, and removing

words that are irrelevant to what is required. The appropriate data is divided into two parts, with training data accounting for 20% of the total data and testing data accounting for the remaining 80%. Table 2 shows an example of how the preprocessing process is used.

Tabel 2. Application of Data Preprocessing

Data Samples Before Preprocessing	Preprocessed data
Karena takutnya ada batasan waktu sehingga belum mengisi, jadi nilai akhir tidak bisa maksimal	Takut batas waktu hingga belum isi nilai akhir tidak maksimal

The preprocessed data then used to extract features. The feature extraction process begins by calculating the Term Frequency and then proceeds to calculate the TF-IDF (Term Frequency-Inverse Document Frequency). The Inverse document frequency is calculated first in order to calculate the TF-IDF (idf). After determining the idf, the TF-IDF weight value is calculated by multiplying the term frequency value by the inverse document frequency value for each term. When the feature extraction process is finished, move on to the comment classification process. The classification of comments employs two methods: Naive Bayes and KNN, both of which will be processed to produce the most accurate results. The classification of comments related to *kampus mengajar* 2 results in an accuracy of 99.30 percent for Naive Bayes and 97.20 percent for K-Nearest Neighbors.

Table 3 shows the accuracy, precision, recall, and F-measurement classifications for *kampus mengajar* 2 data set.

Table 3. Accuracy, precision, recall, and F-measure classification

Method	Accuracy	Precision	Recall	F-Measure
Naïve Bayes	99.30 %	96.77 %	91.67 %	94.15%
KNN	97.20 %	96.50 %	94.28 %	95.37%

The naive Bayes approach outperforms the method and KNN in *kampus mengajar* 2 data set in terms of accuracy, precision, and F-measure values, as seen in the table above. Meanwhile, the KNN technique outperforms the naive Bayes method in terms of recall. Based on these findings, it is clear that the Naive Bayes method outperforms the K-Nearest Neighbors method in classifying comments.

4. CONCLUSION

Sentiment analysis on various comments related to the *kampus mengajar* 2 as part of *Merdeka Belajar Kampus Merdeka* (MBKM) program is based on a telegram group conversation of supervisor *kampus mengajar* 2. The data obtained from the crawling results are classified as positive, negative, or neutral using the Nave Bayes method and K-Nearest Neighbors on a total of 591 data points, which are categorized into two: training data and testing data. The training data is 80 percent of total data, with the remaining 20 percent serving as testing data. With an accuracy result of 99.30 percent for Naive Bayes and 97.20 percent for K-Nearest Neighbors, it can be concluded that using of the Naive Bayes method for processing sentiment analysis can better classify comments. Meanwhile, when tested on crawling data related to the implementation of *kampus mengajar* 2, which is component of *Merdeka Belajar Kampus Merdeka* (MBKM) program, the Nearest Neighbors method has a slightly lower level of accuracy.

REFERENCES

[1] E. Purike, "Political Communications of The Ministry of Education and Culture about 'Merdeka Belajar, Kampus Merdeka (Independent Learning, Independent Campus)' Policy: Effective?," *EduLine J. Educ. Learn. Innov.*, vol. 1, no. 1, pp. 1–8, 2021, doi: 10.35877/454ri.eduline361.

[2] K. Krishnapatria, "Merdeka Belajar-Kampus Merdeka (MBKM) curriculum in English studies program: Challenges and opportunities," *ELT Focus*, vol. 4, no. 1, pp. 12–19, 2021, doi: 10.35706/eltinfc.v4i1.5276.

[3] S. Andari, W. Windasari, A. Chandra Setiawan, and A. Rifqi, "Student Exchange Program of Merdeka Belajar-Kampus Merdeka (MBKM) in Covid-19 Pandemic," *JPP (Jurnal Pendidik. dan Pembelajaran)*, vol. 28, no. 1, pp. 30–37, 2021, doi: 10.17977/um047v28i12021p030.

[4] A. A. Sihombing, S. Anugrahsari, N. Parlina, and Y. S. Kusumastuti, "Merdeka Belajar in an Online Learning during The Covid-19 Outbreak: Concept and Implementation," *Asian J. Univ. Educ.*, vol. 17, no. 4, pp. 35–48, 2021, doi: 10.24191/ajue.v17i4.16207.

[5] C. M. Yudhawasthi and L. Christiani, "Challenges of Higher Educational Documentary Institutions in Supporting Merdeka Belajar Kampus Merdeka Program," *Khizanah al-Hikmah J. Ilmu Perpustakaan, Informasi, dan Kearsipan*, vol. 9, no. 2, p. 193, 2022, doi: 10.24252/kah.v9cf2.

[6] I. Lhutfi and R. Mardiani, "Merdeka Belajar - Kampus Merdeka Policy: How Does It Affect the Sustainability on Accounting Education in Indonesia?," *Din. Pendidik.*, vol. 15, no. 2, pp. 243–253, 2020, doi: 10.15294/dp.v15i2.26071.

[7] D. Kodrat, "Industrial Mindset of Education in Merdeka Belajar Kampus Merdeka (MBKM) Policy," *J. Kaji. Perad. Islam*, vol. 4, no. 1, pp. 9–14, 2021, doi: 10.47076/jkpis.v4i1.60.

[8] I. H. Batubara *et al.*, "Bibliometric Mapping on the Research 'Merdeka Belajar' Using Vosviewer," *J. Pendidik. Progresif*, vol. 12, no. 2, pp. 477–486, 2022, doi: 10.23960/jpp.v12.i2.202207.

[9] B. K. Prahani *et al.*, "The Concept of 'Kampus Merdeka' in Accordance with Freire's Critical Pedagogy," *Stud. Philos. Sci. Educ.*, vol. 1, no. 1, pp. 21–37, 2020, doi: 10.46627/sipose.v1i1.8.

[10] D. Sunitha, R. K. Patra, N. V. Babu, A. Suresh, and S. C. Gupta, "Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries," *Pattern Recognit. Lett.*, vol. 158, pp. 164–170, 2022, doi: 10.1016/j.patrec.2022.04.027.

[11] Y. Zhang, K. Chen, Y. Weng, Z. Chen, J. Zhang, and R. Hubbard, "An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US," *Expert Syst. Appl.*, vol. 198, no. May 2021, p. 116882, 2022, doi: 10.1016/j.eswa.2022.116882.

[12] W. He and G. Xu, "Social media analytics: unveiling the value, impact and implications of social media analytics for the management and use of online information," *Online Inf. Rev.*, vol. 40, no. 1, p. OIR-12-2015-0393, Feb. 2016, doi: 10.1108/OIR-12-2015-0393.

[13] M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm," *Int. J. Comput. Eng. Inf. Technol.*, vol. 8, no. 6, pp. 90–95, 2016, [Online]. Available: www.ijceit.org

[14] R. S. Moorthy and P. Pabitha, "Optimal Detection of Phising Attack using SCA based K-NN," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1716–1725, 2020, doi: 10.1016/j.procs.2020.04.184.

[15] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.

[16] N. Deepa, J. Sathya Priya, and T. Devi, "Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy," *Mater. Today Proc.*, no. xxxx, 2022, doi: 10.1016/j.matpr.2022.03.345.

[17] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016, doi: 10.1016/j.jksuci.2015.11.003.

[18] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.

[19] S. Farhana, "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm," *Procedia Comput. Sci.*, vol. 194, pp. 224–228, 2021, doi: 10.1016/j.procs.2021.10.077.

[20] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[21] E. M. M. van der Heide, R. F. Veerkamp, M. L. van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, "Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle," *J. Dairy Sci.*, vol. 102, no. 10, pp. 9409–9421, 2019, doi: 10.3168/jds.2019-16295.

[22] D. van Herwerden, J. W. O'Brien, P. M. Choi, K. V. Thomas, P. J. Schoenmakers, and S. Samanipour, "Naive Bayes classification model for isotopologue detection in LC-HRMS data," *Chemom. Intell. Lab. Syst.*, vol. 223, no. November 2021, p. 104515, 2022, doi: 10.1016/j.chemolab.2022.104515.

[23] A. Jhamtani, R. Mehta, and S. Singh, "Size of wallet estimation: Application of K-nearest neighbour and quantile regression," *IIMB Manag. Rev.*, vol. 33, no. 3, pp. 184–190, 2021, doi: 10.1016/j.iimb.2021.09.001.

[24] Y. Amonkar, D. J. Farnham, and U. Lall, "A k-nearest neighbor space-time simulator with applications to large-scale wind and solar power modeling," *Patterns*, vol. 3, no. 3, p. 100454, 2022, doi: 10.1016/j.patter.2022.100454.

[25] T. Olsson, M. Ericsson, and A. Wingkvist, "To automatically map source code entities to architectural modules with Naive Bayes," *J. Syst. Softw.*, vol. 183, p. 111095, 2022, doi: 10.1016/j.jss.2021.111095.

[26] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.

[27] A. Islam, S. B. Belhaouari, A. U. Rehman, and H. Bensmail, "K Nearest Neighbor OveRsampling approach: An open source python package for data augmentation," *Softw. Impacts*, vol. 12, no. February, p. 100272, 2022, doi: 10.1016/j.simpa.2022.100272.