Aspect-based sentiment analysis on student reviews using the Indo-Bert base model.

Ahmad Jazuli[1*], *Widowati*[2], and *Retno* Kusumaningrum[3]

[1]*Doctoral Program in Information Systems, School of Postgraduate Studies, Diponegoro University, Semarang, 50275, Indonesia*.
[1]*Informatics Engineering, Faculty of Engineering, Universitas Muria Kudus, Kudus, 59332, Indonesia.*
[2]*Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, 50275, Indonesia.*
[3]*Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang, 50275, Indonesia.*

Abstract. This study aims to gain a deeper understanding of online student reviews regarding the learning process at a private university in Indonesia and to compare the effectiveness of several algorithms: Naive Bayes, K-NN, Decision Tree, and Indo-Bert. Traditional Sentiment Analysis methods can only analyze sentences as a whole, prompting this research to develop an Aspect-Based Sentiment Analysis (ABSA) approach, which includes aspect extraction and sentiment classification. However, ABSA has inconsistencies in aspect detection and sentiment classification. To address this, we propose the BERT method using the pre-trained Indo-Bert model, currently the best NLP model for the Indonesian language. This study also fine-tunes hyperparameters to optimize results. The dataset comprises 10,000 student reviews obtained from online questionnaires. Experimental results show that the aspect extraction model has an accuracy of 0.890 and an F1-Score of 0.897, while the sentiment classification model has an accuracy of 0.879 and an F1-Score of 0.882. These results demonstrate the effectiveness of the proposed method in identifying aspects and sentiments in student reviews and provide a comparison between the four algorithms.

Keywords: ABSA, Naïve Bayes, K-NN, Decision Tree, Indo-Bert

* Corresponding author: ahmadjazuli@students.undip.ac.id

1. Introduction

In today's digital era, online reviews have become an important and influential source of information, especially in education [1]. Online reviews from students can provide valuable insights into the quality and effectiveness of the learning process [2]. These reviews can cover various aspects, from the course material and teaching methods to the facilities provided [3]. However, analyzing these reviews is not an easy task. Traditional Sentiment Analysis methods can only analyze sentences and often fail to capture deeper nuances and context [4]. Therefore, it is important to develop more sophisticated and in-depth methods.

This research aims to gain a deeper understanding of online student reviews of the learning process at a private university in Indonesia [5]. We focus on developing an Aspect-Based Sentiment Analysis (ABSA) approach, which includes aspect extraction and sentiment classification [6]. With this approach, we hope to capture more information from student reviews and provide a more accurate and holistic picture of the learning process. Universities can use this information to make necessary improvements and adjustments [2]. In addition, the results of this research can also be used by other researchers interested in analyzing online reviews or other text data [7]. By understanding the strengths and weaknesses of the ABSA method, other researchers can further develop and improve this method [8].

Traditional Sentiment Analysis methods can only analyze sentences and often fail to capture deeper nuances and context [9]. Therefore, it is important to develop more sophisticated and in-depth methods. In this research, we compare the effectiveness of several popular algorithms in sentiment analysis, namely Naive Bayes, K-NN, Decision Tree, and Indo-Bert [10].

Naive Bayes, K-NN, and Decision Tree algorithms have long been used in machine learning and have proven effective in various applications. However, they have limitations in analyzing text data, especially in capturing context and nuances in reviews [11]. On the other hand, Indo-Bert is a BERT-based NLP model specifically trained for the Indonesian language, offering more advanced text analysis capabilities [12]. We hope to determine the best method for analyzing student reviews by comparing these four algorithms [13]. This information is important for universities wanting to understand and improve their learning processes and for researchers and practitioners interested in sentiment analysis and NLP. In addition, the results of this research can also be used by other researchers interested in analyzing online reviews or other text data. By understanding the strengths and weaknesses of each algorithm, other researchers can develop and improve their sentiment analysis methods [14].

2. Methods

In this study, we utilized Google Colab with Python version 3.10.11 [15]. Fine-tuning was performed using Torch version 2.0.1+cu118 and Transformers version 4.29.2 [16]. We used the Tesla T4 NVIDIA-SMI 525.85.12 GPU and CUDA version 12.0 . The flow of this research is shown in Figure 1.

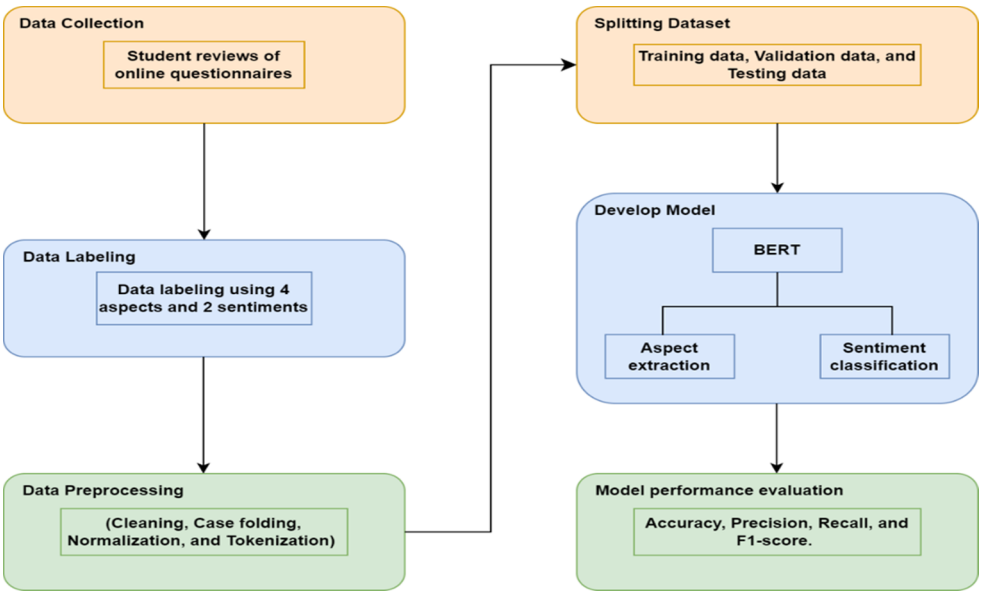


Figure 1. System architecture

2.1. Data Collection. This study uses 10,000 student review data obtained from online questionnaires on the academic portal of a private university in Indonesia [17]. The dataset statistics can be seen in Table 1.

Table 1. The statistic of the dataset.

Aspect	Sentiment	Total
Lecturer	Positive	1.250
	Negative	1.250
Curriculum	Positive	1.250
	Negative	1.250
Infrastructure	Positive	1.250
	Negative	1.250
Service	Positive	1.250
	Negative	1.250

2.2. Data Labeling. The data was labeled with 4 aspects (lecturer, curriculum, infrastructure, and service) [18]. We labeled sentiments with 2 classes (positive and negative) using values as markers. Each aspect in the review with positive sentiment was given a value of 2, while negative sentiment was given a value of 1 and 0 for each aspect that did not contain sentiment [19]. Labeling with negative sentiment was given to each review that contained sentences of criticism, disappointment, and dissatisfaction for each aspect [16]. Whereas positive sentiment was given to reviews that contained sentences without any sense of disappointment, such as quite satisfied, and suggestions to maintain each aspect. An example of student review data labeling is seen in Table 2.

Table 2. An example of data labeling

Reviews	Lecturer	Curriculum	Infrastructure	Service
MK sesuai perkembangan dosen sering TELAT pegawai sekretariat KURANG RESPONSIP****	1	2	0	1
JAM DINDING SERING MATI PENYAMPAIAN DOSEN MUDAH DIPAHAMI!!!!!	2	0	1	0
Kurikulim perlu diupdate, dosen boseni pelayanan prodi baik	0	1	0	2
Lift sering mati tapi dosennya tegas mk yang diberikan sesuai dengan pekerjaan dan pegawai pada ramah	2	2	1	2
Ruangan nyaman namun dosen sering kosong	1	0	2	0

2.3. Data Preprocessing. The obtained dataset is unstructured. Therefore, we applied data preprocessing, including cleaning used to clean data from symbols, emojis, punctuation, and others [20]. To convert uppercase letters to lowercase, we use case folding [11]. Normalization converts non-standard words into standard ones according to Indonesian language guidelines. Tokenization is used to break sentences into words according to the language rules [21].

2.4. Data Splitting. The next step is to divide the data into training, validation, and testing data. This study used a data-splitting percentage of 80% as the training set and 20% as the validation set. For the test set, we used 10% of the validation data.

2.5. Develop a model. We used the BERT method to develop the ABSA model [22]. BERT Tokenizer is used by applying [SEP] separator, [CLS] classification, [PAD] padding, and [UNK] unknown [23]. Next, we apply encoding plus and attention mask [24]. This research uses the Indo-Bert pre-trained model because Indo-Bert is one of Indonesia's most comprehensive NLP dataset models [25]. This study applies to fine-tuning using the following hyper-parameters: batch size (16, 32), learning rate (2e-5), epoch (15, 30), and dropout (0.1, 0.3, 0.5). We conducted trials for each epoch on each batch size to get the best model.

2.6. Model Performance Evaluation. In this study, model performance evaluation uses a confusion matrix (accuracy, precision, recall, and f1-score) [26]. Accuracy is a matrix to determine how often the model can correctly classify items [27].

3. Results and Analysis

3.1 Naïve Bayes Classifier Model

After obtaining the training and testing datasets, the next step is to perform feature extraction and apply the Naïve Bayes Classifier algorithm. This study uses the Naïve Bayes algorithm to classify reviews into three sentiment categories: positive, negative, and neutral. The testing using the training and testing data resulted in an accuracy of 89.40%. Figure 2 illustrates the process in RapidMiner for the Naïve Bayes Classifier algorithm.

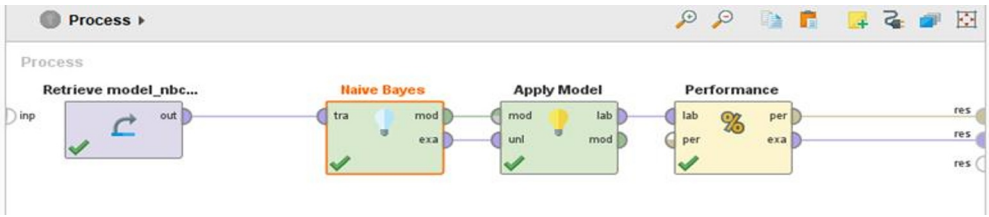


Figure 2. Naïve Bayes Classifier Algorithm Model

From the testing results obtained using RapidMiner, class precision and class recall were derived. Precision represents the ratio of relevant documents to the total retrieved documents, while recall represents the ratio of relevant documents retrieved to the total relevant documents. The Naïve Bayes Classifier algorithm with an accuracy of 89.40%. The precision for positive prediction is 100.00%, negative prediction is 86.23%, and neutral prediction is 86.55%. Additionally, there is a class recall for true positive of 86.13%, true negative of 86.08%, and true neutral of 100.00%. According to previous research, Naïve Bayes Classifier is considered one of the best algorithms for classification on traditional sentiment analysis.

3.2. K-Nearest Neighbors (K-NN) Algorithm Model.

The K-Nearest Neighbors (K-NN) algorithm was also tested in this study. Figure 3 shows the K-NN algorithm model.

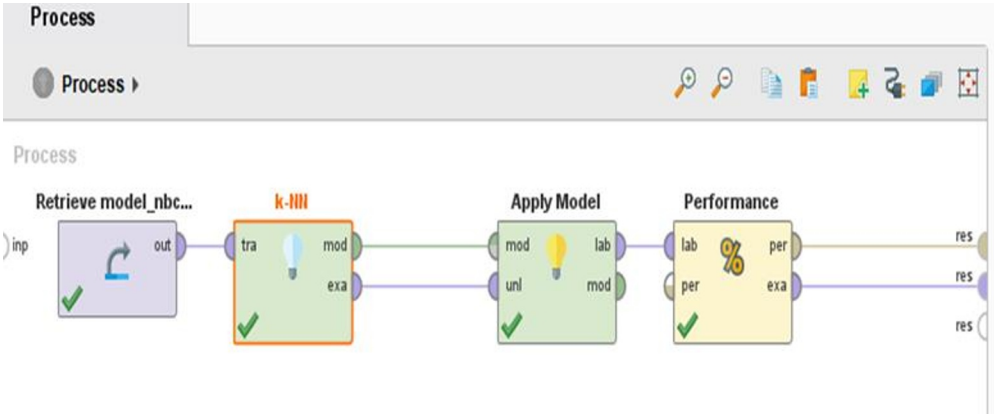


Figure 3. K-Nearest Neighbors (K-NN) Algorithm Model

The testing process of the K-NN algorithm resulted in an accuracy of 77.01%, as shown in Figure 4.

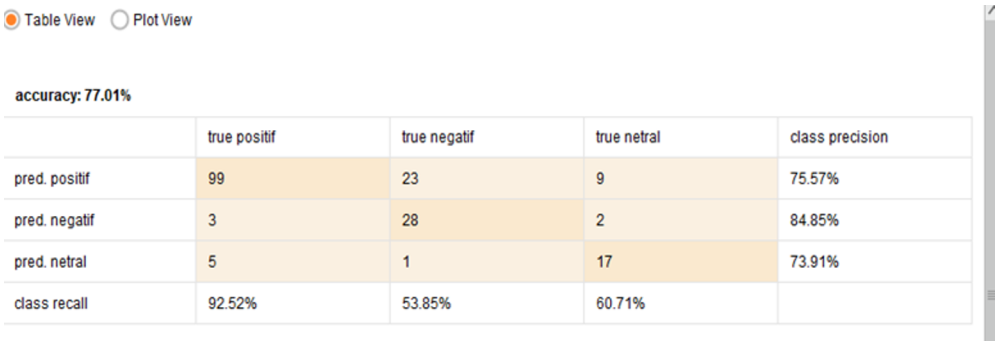


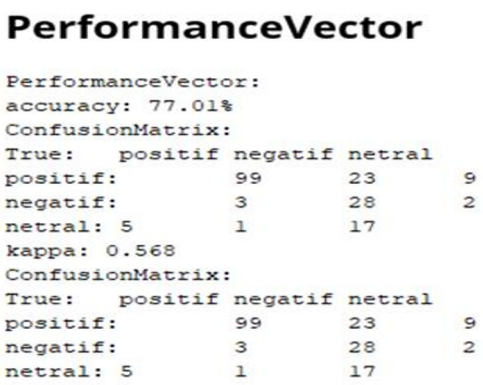
Table View Plot View

accuracy: 77.01%

	true positif	true negatif	true netral	class precision
pred. positif	99	23	9	75.57%
pred. negatif	3	28	2	84.85%
pred. netral	5	1	17	73.91%
class recall	92.52%	53.85%	60.71%	

Figure 4. Testing Results in RapidMiner

The design process in Figure 4 indicates the testing results of the K-NN algorithm with an accuracy of 77.01%. The precision for positive prediction is 75.57%, for negative prediction is 84.85%, and for neutral prediction is 73.91%. Additionally, there is a class recall for true positive of 92.52%, true negative of 53.85%, and true neutral of 60.71%.



PerformanceVector

```
PerformanceVector:
accuracy: 77.01%
ConfusionMatrix:
True:   positif negatif netral
positif:        99      23      9
negatif:        3       28      2
netral: 5       1       17
kappa: 0.568
ConfusionMatrix:
True:   positif negatif netral
positif:        99      23      9
negatif:        3       28      2
netral: 5       1       17
```

Figure 5. illustrates the description of the performance vector of the K-NN algorithm with a kappa value of 0.568.

Figure 5 illustrates the description format of the table presenting the analysis results of the K-Nearest Neighbors (K-NN) algorithm. It also includes a kappa value of 0.568 in the performance vector or performance vector, which lists the performance criteria values of the tested algorithm.

3.3. Decision Tree Algorithm Model

The third algorithm tested is the Decision Tree algorithm. Figure 6 shows the design process of the Decision Tree algorithm using RapidMiner.

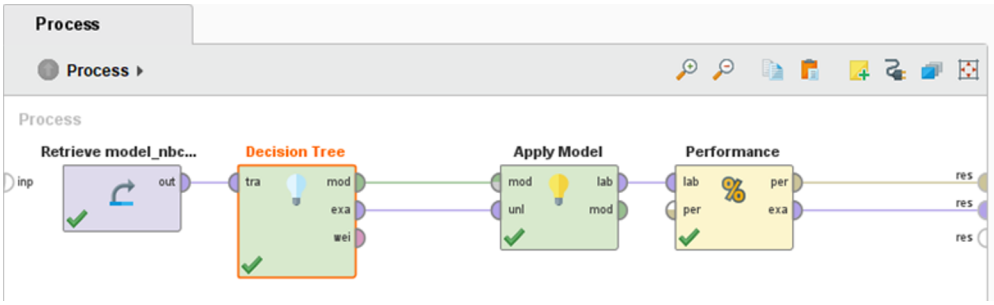


Figure 6. Decision Tree Algorithm Model

The testing process of the Decision Tree algorithm resulted in an accuracy of 62.03%, as shown in Figure 7.

Table View Plot View

accuracy: 62.03%

	true positif	true negatif	true netral	class precision
pred. positif	107	52	19	60.11%
pred. negatif	0	0	0	0.00%
pred. netral	0	0	9	100.00%
class recall	100.00%	0.00%	32.14%	

Figure 7. Testing Results in RapidMiner

The design process in Figure 7 indicates the testing results of the Decision Tree algorithm with an accuracy of 62.03%. The precision for positive prediction is 60.11%, negative prediction is 0.00%, and neutral prediction is 100.00%. There is also a class recall for a true positive of 100.00%, a true negative of 0.00%, and a true neutral of 32.14%.

PerformanceVector

```
PerformanceVector:
accuracy: 62.03%
ConfusionMatrix:
True:    positif negatif netral
positif:          107      52       19
negatif:          0        0        0
netral: 0         0        9
kappa: 0.153
ConfusionMatrix:
True:    positif negatif netral
positif:          107      52       19
negatif:          0        0        0
netral: 0         0        9
```

Figure 8. illustrates the description of the performance vector of the Decision Tree algorithm, with a kappa value of 0.153.

Figure 8 depicts the format of the table description presenting the analysis results of the Decision Tree algorithm. It also includes a kappa value of 0.153 in the performance vector or performance vector, which lists the performance criteria values of the tested algorithm.

3.4. Indo-Bert base Algorithm Model

The experimental results are divided into two tasks: aspect extraction and sentiment classification. Each task uses the same hyperparameters. The results of the aspect extraction experiment are shown in Table 3.

TABLE 3. The results of aspect extraction

Epoch	Batch size	Dropout	Accuracy	Precision	Recall	F1 score
15	16	0,1	0,859	0,850	0,853	0,851
15	16	0,3	0,878	0,874	0,870	0,872
15	16	0,5	0,846	0,849	0,843	0,846
15	32	0,1	0,869	0,862	0,864	0,863
15	32	0,3	0,851	0,849	0,852	0,850
15	32	0,5	0,837	0,830	0,835	0,832
30	16	0,1	0,842	0,849	0,844	0,846
30	16	0,3	0,883	0,889	0,885	0,887
30	16	0,5	0,857	0,851	0,855	0,853
30	32	0,1	**0,890**	**0,896**	**0,898**	**0,897**
30	32	0,3	0,876	0,871	0,879	0,875
30	32	0,5	0,860	0,864	0,868	0,866

Based on Table 3, it is known that the results of the aspect extraction experiment utilizing fine-tuning with hyper-parameters epoch (30), batch size (32), and dropout (0.3) yielded the best results with an accuracy value of 0.890 and F1-Score of 0.897. It turns out that a batch size of 32 can produce better accuracy compared to a batch size of 16. In reality, using a larger batch size requires more time [24]. The larger the epoch value, the longer the training time required [25]. The results of the aspect extraction experiment show that a smaller dropout value is superior. In this study, the Adam optimizer and the dropout value were adjusted according to the optimizer used. In addition, a learning rate of 2e-5 was used in the aspect extraction and sentiment classification stages because it can address issues with BERT

3.5. Final Results of Algorithm Comparison

After obtaining the results from the four algorithms, Naïve Bayes Classifier, K-Nearest Neighbors (K-NN), Decision Tree, and Indo-Bert, the final step is to compare them. Based on the comparison, the highest accuracy was achieved using the Naïve Bayes Classifier algorithm with an accuracy of 89.40%; however, in terms of Precision and Recall, Indo BERT outperforms the other algorithms, which is due to the smaller amount of data provided. Table 4 presents each algorithm's comparison results of accuracy, precision, and recall.

Algoritma	Accuracy	Precision	Recall
Naïve Bayes Classifier	89.40 %	87.59 %	89.73 %
K-NN	77.01 %	78.11 %	69.02 %
Decision Tree	62.03 %	53.37 %	44.06 %
IndoBert	**89.00 %**	**89.60 %**	**89.80 %**

Table 4. Comparison of the Three Algorithms

4. Conclusion.

This research successfully developed an effective Aspect-Based Sentiment Analysis (ABSA) model using the pre-trained IndoBERT model, currently the best NLP model for the Indonesian language. Through the fine-tuning of hyperparameters, we achieved the best aspect extraction model with an accuracy value of 0.890 and an F1-Score of 0.897. This model outperforms previous advanced models in the Indonesian language dataset domain. In addition, the comparison between several algorithms (Naive Bayes, K-NN, Decision Tree, and IndoBERT) shows the superiority of IndoBERT in sentiment analysis. Universities can use the results of this research to gain a deeper understanding of student reviews and make necessary improvements. For future research, we plan to improve ABSA results by trying the BERT knowledge graph approach, which is expected to provide a deeper understanding of sentiment in student reviews.

The acknowledgments: This work is supported by Post Graduate Research Grant with contract number **345-24/UN7.D2/PP/IV/2023**. This research was supported by the Laboratory of Computer Modelling, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia.

References

1. Jazuli A, Latubessy A, Nindyasari R. Arsitektur Web Service Di Lembaga Pendidikan Ma'Arif Demak. Indones J Technol Informatics Sci. 2nd ed. 2021;2(2):67–70.
2. Karaoglan Yilmaz FG, Yilmaz R. Learning Analytics Intervention Improves Students' Engagement in Online Learning. Technol Knowl Learn. 2021;(0123456789).
3. Kastrati Z, Dalipi F, Imran AS, Nuci KP, Wani MA. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. Appl Sci. 2021;11(9).
4. Salazar C, Aguilar J, Monsalve-Pulido J, Montoya E. Affective recommender systems in the educational field. A systematic literature review. Comput Sci Rev [Internet]. 2021;40:100377. Available from: https://doi.org/10.1016/j.cosrev.2021.100377
5. Liu Y, Soroka A, Han L, Jian J, Tang M. Cloud-based big data analytics for customer insight-driven design innovation in SMEs. Int J Inf Manage [Internet]. 2020;51(November 2019):102034. Available from: https://doi.org/10.1016/j.ijinfomgt.2019.11.002
6. Mehbodniya A, Rao MV, David LG, Joe Nige KG, Vennam P. Online product sentiment analysis using random evolutionary whale optimization algorithm and deep belief network. Pattern Recognit Lett. 2022;159:1–8.
7. Žitnik S, Blagus N, Bajec M. Target-level sentiment analysis for news articles. Knowledge-Based Syst. 2022;249:108939.
8. Yan H, Dai J, Ji T, Qiu X, Zhang Z. A Unified Generative Framework for Aspect-based Sentiment Analysis. 2021;2416–29.
9. Abdi A, Hasan S, Shamsuddin SM, Idris N, Piran J. A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion. Knowledge-Based Syst [Internet]. 2021;213:106658. Available from: https://doi.org/10.1016/j.knosys.2020.106658
10. Ahmad Jazuli TKR. Analisis Sentimen Terhadap Xiaomi Indonesia Menggunakan Naïve Bayes Method. Indones J Technol Informatics Sci [Internet]. 3rd ed. 2021;3(1):21–9. Available from: https://jurnal.umk.ac.id/index.php/ijtis/article/view/7514/pdf
11. Abella A, Araya León M, Marco-Almagro L, Clèries Garcia L. Perception evaluation kit: a case study with materials and learning styles. Int J Technol Des Educ.

2021;(0123456789).
12. Nurdin A, Anggo Seno Aji B, Bustamin A, Abidin Z. Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks. J Tekno Kompak. 2020;14(2):74.
13. Rajaguru H, Sannasi Chakravarthy SR. Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. Vol. 20, Asian Pacific Journal of Cancer Prevention. 2019. p. 3777–81.
14. Idrus. Evaluasi Dalam Proses Pembelajaran. Eval Dalam Proses Pembelajaran. 2019;9(2):920–35.
15. KULKARNI SM, SUNDARI G. Comparative analysis of performance of deep cnn based framework for brain mri classification using transfer learning. J Eng Sci Technol. 2021;16(4):2901–17.
16. Agüero-Torales MM, Abreu Salas JI, López-Herrera AG. Deep learning and multilingual sentiment analysis on social media data: An overview. Appl Soft Comput. 2021;107.
17. Süzen N, Gorban AN, Levesley J, Mirkes EM. Automatic short answer grading and feedback using text mining methods. Procedia Comput Sci [Internet]. 2020;169(2019):726–43. Available from: https://doi.org/10.1016/j.procs.2020.02.171
18. Mohamad Beigi O, Moattar MH. Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification. Knowledge-Based Syst [Internet]. 2020;213(xxxx):106423. Available from: https://doi.org/10.1016/j.knosys.2020.106423
19. Ribeiro D, Matos LM, Moreira G, Pilastri A, Cortez P. Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection. Computers. 2022;11(4):1–15.
20. Skarpathiotaki CG, Psannis KE. Cross-Industry Process Standardization for Text Analytics. Big Data Res. 2022;27(2).
21. Qi B, Costin A, Jia M. A framework with efficient extraction and analysis of Twitter data for evaluating public opinions on transportation services. Travel Behav Soc [Internet]. 2020;21(December 2019):10–23. Available from: https://doi.org/10.1016/j.tbs.2020.05.005
22. Liu MZ, Zhou FY, Chen K, Zhao Y. Co-attention networks based on aspect and context for aspect-level sentiment analysis. Knowledge-Based Syst. 2021;217.
23. Kognisi PK, Risiko P, Jenis DAN, Bidori F, Puspitowati LI dan I, Wijaya IGB, et al. No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. Ind High Educ [Internet]. 2021;3(1):1689–99. Available from: http://journal.unilak.ac.id/index.php/JIEB/article/view/3845%0Ahttp://dspace.uc.ac.id/handle/123456789/1288
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21:1–67.
25. Fudholi DH, Nayoan RAN, Hidayatullah AF, Arianto DB. a Hybrid Cnn-Bilstm Model for Drug Named Entity Recognition. J Eng Sci Technol. 2022;17(1):730–44.
26. Da'u A, Salim N, Rabiu I, Osman A. Weighted aspect-based opinion mining using deep learning for recommender system. Expert Syst Appl. 2020;140.
27. Corso MP, Perez FL, Stefenon SF, Yow KC, Ovejero RG, Leithardt VRQ. Classification of contaminated insulators using k-nearest neighbors based on computer vision. Computers. 2021;10(9):1–18.