Optimizing the Classification of SNBP Student Candidate Files Using the K-Nearest Neighbor (KNN) Algorithm

Diah Ayu Larasati, Amil Ahmad Ilham and Ady Wahyudi Paundu
Department of Informatics, Hasanuddin University, Gowa, Indonesia

Keywords: SNBP, K-Nearest Neighbor (KNN), Certificate Classification, International

Abstract: In addition to academic achievement, non-academic achievement is one of the critical factors that can increase the chances of being accepted into the SNBP (Seleksi Nasional Berdasarkan Prestasi) pathway. There are several criteria for non-academic achievement certificates that are used as assessment criteria according to the portfolio from the Ministry of Education and Culture. This research aims to optimize the classification of each certificate according to its grouping in the portfolio to make it easier for reviewers to check files. The KNN algorithm's evaluation using the cross-validation method and evaluation metrics such as precision, recall, and F1-score shows that text preprocessing has a significant influence on model performance. The best experiment, experiment 3, which uses complete preprocessing including stopword removal, gives the highest accuracy of 0.722. The lowest performance was consistently found in the "Undefined" class, with an F1-score of only 0.39 in the 3rd experiment. These results show that the KNN method with TF-IDF vectorization and cosine similarity can classify proof of achievement well.

1 INTRODUCTION

SNPMB (Seleksi Nasional Penerimaan Mahasiswa Baru) is a selection system organized by the Indonesian government to attract outstanding students to continue their education at higher levels, such as State Universities and State Polytechnics. One of the selection channels in SNPMB is SNBP (Seleksi Nasional Berdasarkan Prestasi), which considers report cards and evidence of academic and non-academic achievements uploaded by students on the official SNPMB website.

So far, the selection committee has classified thousands of proofs of achievement manually, which takes a lot of time and effort. Therefore, an automated classification system is needed to help the selection process. Classification in the context of information systems is a technique of mapping data into certain predetermined classes. One of the most popular classification algorithms is K-Nearest Neighbor (KNN). This algorithm works by classifying new data based on its proximity to data that already has a known class, based on distance calculations.

The KNN algorithm has advantages due to its simplicity and effectiveness in handling small to medium datasets. However, the accuracy of KNN is greatly influenced by the selection of the K value and the quality of the data used. In this research, a text mining and OCR (Optical Character Recognition) approach is used to process raw data in the form of certificate files into text form. This research explicitly compares the KNN algorithm's accuracy against three different text preprocessing scenarios.

The main contribution of this research is a comprehensive analysis of the effect of text preprocessing on classification accuracy. In addition, this research provides practical benefits by offering a system that can accelerate and streamline the selection process, while also serving as a basis for the development of future achievement document classification methods.

2 METHODOLOGY

2.1 Optical Character Recognition (OCR)

Optical Character Recognition, or OCR, is a technology that converts text in image or picture file format into text that can be read and edited by

Larasati, D. A., Ilham, A. A. and Paundu, A. W.
Optimizing the Classification of SNBP Student Candidate Files Using the K-Nearest Neighbor (KNN) Algorithm.
DOI: 10.5220/0014276800004928

In *Proceedings of the 1st International Conference on Research and Innovations in Information and Engineering Technology (RITECH 2025)*, pages 231-237
ISBN: 978-989-758-784-9

computer applications (e.g., Notepad, and Microsoft Word) (Firdaus et al., 2021).

In image-type media, text content cannot be extracted directly because it is considered part of the visual image, so a special method is needed to extract text from an image, namely Optical Character Recognition or OCR (Nugraha, 2024). In the context of this research, OCR is a fundamental stage because the proof of achievement data uploaded by students is in the form of digital files that need to have their text information extracted. The OCR process involves several stages of image preprocessing, such as rotating to correct the document's orientation, cropping to focus on the area containing the text, and excluding non-certificate files to ensure the quality of the data to be processed.

OCR is frequently included as part of systems for handling documents, efforts to make things digital, and programs for entering data digitally. OCR can also be used to understand writing done by hand, but this use is not as common because handwriting can be very different from person to person. Older OCR setups relied on rules created by people, which meant people had to do a lot of work and the results were not very exact. Older types of machine learning models include Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbors (KNN) (Francis & Sangeetha, 2025).

Modern OCR technology uses deep learning approaches to improve character recognition accuracy, especially on documents with varying image quality. In this research, the EasyOCR library was chosen for its ability to handle various fonts and layouts of certificate documents accurately. The PDF to image conversion process is performed using the PyMuPDF library, which maintains the visual quality of the original document.

2.2 Text Mining

Text mining is extracting useful information and knowledge from extensive, unstructured textual data (Pertiwi, 2022). Text mining is a well-known method used in areas like computer science, information science, mathematics, and management to find useful information from large amounts of data (Kumar et al., 2021). In the context of document classification, text mining consists of three main stages: preprocessing, data mining, and postprocessing. The preprocessing stage includes data selection, text cleaning, and feature extraction, to transform documents into representations that machine learning algorithms can process. Text mining can use standard dictionary ways, ways that let computers learn, and advanced computer learning ways. The goal of text mining is to figure out what the text is trying to say (Lai & Chen, 2023).

In this research, text mining is vital in converting OCR-generated text into numerical features that the KNN algorithm can process. This process involves various preprocessing techniques such as case folding, punctuation removal, space normalization, and stop words removal. Each preprocessing technique has a different impact on the quality of the resulting features and ultimately affects the classification accuracy. Figure 1 presents the design of the system workflow that outlines the main processes involved in this research.

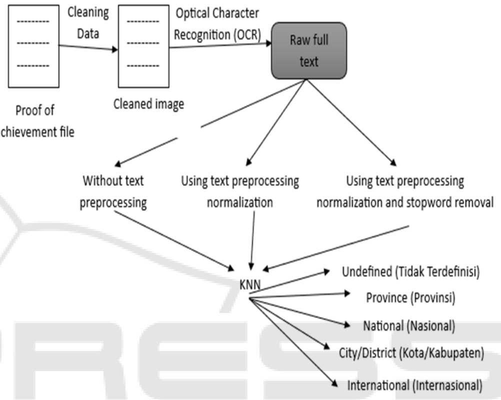

Cleaning Data
Optical Character Recognition (OCR)
Raw full text
Proof of achievement file
Cleaned image
Without text preprocessing
Using text preprocessing normalization
Using text preprocessing normalization and stopword removal
KNN
Undefined (Tidak Terdefinisi)
Province (Provinsi)
National (Nasional)
City/District (Kota/Kabupaten)
International (Internasional)

Figure 1: The design of the system workflow developed.

2.3 K-Nearest Neighbor

The K-NN method works by sorting items into groups by looking at the training information that is most like the item being checked (Danny et al., 2024). The K-Nearest Neighbor algorithm is a supervised learning method that performs classification based on closeness or similarity (Cholil et al., 2021). This algorithm works by finding the K nearest neighbors of the test data to the training data and determining the class based on the majority of the neighboring classes. KNN has lazy learning characteristics because it does not build an explicit model during the training phase but instead stores all the training data and performs computations during the prediction process.

$$cos\ (P,Q)\ =\frac{PxQ}{|P|\ x\ |Q|}=\frac{\sum_{i=1}^{n}Freq(\omega_i|P)\ x\ Freq(\omega_i|Q)}{\sqrt{\sum_{i=1}^{n}Freq(\omega_i|P)^2}\ x\ \sqrt{\sum_{i=1}^{n}Freq(\omega_i|Q)^2}} \quad (1)$$

This research uses a similarity measurement method using cosine similarity (1), which is suitable for text vector data resulting from TF-IDF. The most commonly used distance measurement is Euclidean distance. In this research, similarity measurement uses cosine similarity, which is calculated by the following equation 1.

P and Q are different document vectors, and the components of the two vectors are the frequencies of certain words in the document. Cosine similarity was chosen because of its effectiveness in measuring the similarity of text documents represented in the form of TF-IDF vectors.

2.4 Term Weighting

Term weighting is a method of weighting words in documents used to convert text into numerical representations in text mining algorithms. This weighting aims to provide a value that reflects the importance or relevance of a word in the document and the document collection as a whole.

Long text can generally be weighted using the usual term weighting method such as TF-IDF (Ni'mah & Syuhada, 2022). The term weighting process consists of several main components. First, Term Frequency (TF) calculates the frequency of occurrence of a word in a particular document, where the more often the word appears, the greater its weight. Second, Inverse Document Frequency (IDF) measures the uniqueness of the word in the entire document collection by calculating the logarithm of the ratio of total documents to the number of documents containing the word and combining these two components results in TF-IDF weight, which gives high weight to words that appear frequently in a particular document but rarely appear in the overall document collection (Sholehhudin et al., 2018).

In this research, the term weighting implementation uses the TF-IDF scheme with ngram_range parameters of 1-2 to capture the context of unigrams and bigrams, and max_features of 1000 to limit the feature dimension and avoid overfitting. Bigrams are essential to capture meaningful phrases in the context of certificates of achievement, such as “first place” or “national level”, which have different meanings when separated into individual words. TF-IDF performs effectively when the same words are used again (Gani et al., 2022)

2.5 Text Classification

Text classification is a supervised learning process that aims to find a model to distinguish data classes or concepts and predict the class of unidentified objects. Unlike unsupervised clustering, text classification requires training data that has been labeled to build a prediction model (Iqbal Mubarok et al., 2024).

Text classification specifically focuses on categorizing and organizing text data to facilitate easier management and analysis. These techniques leverage natural language processing, machine learning, and data mining to extract meaningful patterns and categorize text data (Taha et al., 2024). In the context of this research, text classification is used to categorize proof of achievement based on its level of achievement: International, National, Provincial, Regency / City, and Undefined.

The success of text classification depends on the quality of preprocessing and feature representation used. Improper preprocessing can produce noise that interferes with the learning process, while a good feature representation can improve the algorithm's ability to distinguish between classes.

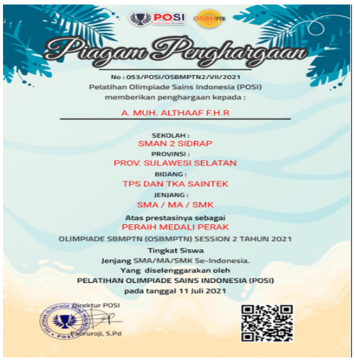
3 RESULTS

3.1 Initial Data Collection and Preparation

The data processing begins by collecting proof of achievement in PDF/PNG/JPG format from 2022, 2023, and 2024 with 4,338, 6,578, and 11,472 files, respectively. This research focuses on the PDF format to homogenize the data processing process. Each PDF file page is converted into an image in JPG/PNG format using the PyMuPDF library with sufficient resolution to maintain the quality of the text to be extracted.

The image preprocessing stage is done manually to ensure data quality. Rotating was done to correct the orientation of tilted or upside-down documents, while cropping focused on areas that contained important information. After data cleaning, 3,335 files for 2022, 2,846 for 2023, and 4,571 for 2024 were obtained.

Each image file is then named according to the convention of SNBP registration year—ID—page number of the original PDF. This systematic naming facilitates tracking and debugging during system development. As illustrated in Figure 2, examples of the cleaned certificate images are presented.



POSI OSBMPTN

Piagam Penghargaan

No : 053/POSI/OSBMPTN2/VII/2021
Pelatihan Olimpiade Sains Indonesia (POSI)
memberikan penghargaan kepada :

A. MUH. ALTHAAF F.H.R

SEKOLAH :
SMAN 2 SIDRAP
PROVINSI :
PROV. SULAWESI SELATAN
BIDANG :
TPS DAN TKA SAINTEK
JENJANG :
SMA / MA / SMK

Atas prestasinya sebagai
PERAIH MEDALI PERAK
OLIMPIADE SBMPTN (OSBMPTN) SESSION 2 TAHUN 2021
Tingkat Siswa
Jenjang SMA/MA/SMK Se-Indonesia.
Yang diselenggarakan oleh
PELATIHAN OLIMPIADE SAINS INDONESIA (POSI)
pada tanggal 11 Juli 2021

Direktur POSI
Fahruroji, S.Pd

Figure 2: Examples of cleaned certificate images.

Extracting text from images is done using the EasyOCR library configured for Indonesian and English, considering that many certificates of achievement use both languages. The OCR results in the form of text are then stored as full text for each year of registration.

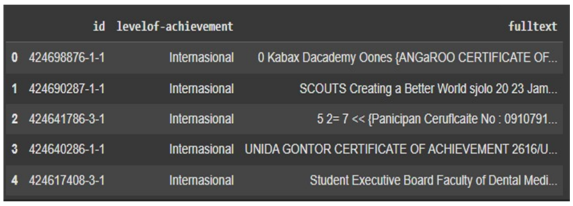
3.2 Text Preprocessing

Pre-processing includes case folding, data cleaning, tokenization, and stopword removal to prepare raw data in a format suitable for analysis, modeling, and machine learning tasks (Lewu et al., 2024).

Data Pre-processing is the first step in machine learning in which the data gets transformed/encoded so that it can be brought in such a state that now the machine can quickly go through or parse that data (Maharana et al., 2022).

The achievement level labeling process is carried out based on SNBP guidelines that categorize achievements into five classes, namely International, National, Provincial, Regency/City, and Undefined. Labeling is done manually by involving domain experts to ensure the accuracy of categorization. Data distribution shows the dominance of national achievement levels, while international achievement levels are in the minority. To address the imbalance in the data, stratified sampling was conducted using the amount of international data as a reference, resulting in a balanced dataset with 205 national data points, 109 district/city data points, 108 provincial data points, 102 international data points, and 102 undefined data points.

The implementation of three preprocessing scenarios was carried out in stages to analyze the impact of each technique on classification accuracy. The first scenario without preprocessing directly uses raw text from OCR, which is then vectorized using TF-IDF. Figure 3 provides an example of the OCR-generated full text, where each recognized segment is annotated with the corresponding achievement level label. The second scenario applies text cleaning, which includes case folding to convert all characters into lowercase letters, punctuation removal to remove irrelevant punctuation marks, number removal to remove numbers that do not provide essential information, and whitespace normalization to normalize excess spaces. The third scenario adds stopword removal to the text cleaning process. The removed stopwords come from a combination of the Sastrawi and NLTK libraries.



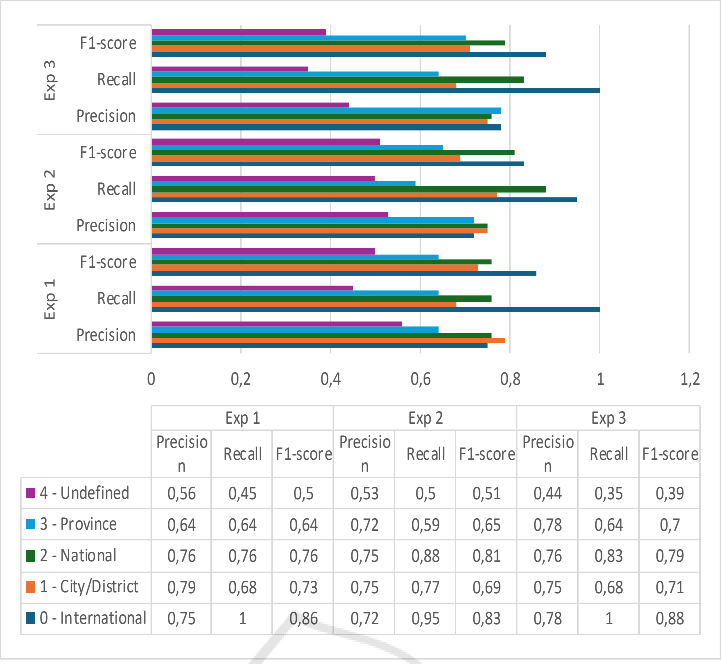
	id	levelof-achievement	fulltext
0	424698876-1-1	Internasional	0 Kabax Dacademy Oones {ANGaROO CERTIFICATE OF...
1	424690287-1-1	Internasional	SCOUTS Creating a Better World sjolo 20 23 Jam...
2	424641786-3-1	Internasional	5 2= 7 << {Panicipan Ceruflcaite No : 0910791...
3	424640286-1-1	Internasional	UNIDA GONTOR CERTIFICATE OF ACHIEVEMENT 2616/U...
4	424617408-3-1	Internasional	Student Executive Board Faculty of Dental Medi...

Figure 3: Example of OCR Output (fulltext) with labeled achievement levels.

3.3 Model Evaluation

Model evaluation is performed using a stratified train-test split with a ratio of 80:20 to ensure a balanced class distribution in the training and testing data. The K=3 value in the KNN algorithm was selected based on the results of preliminary experiments that showed optimal performance at this value. A small K value was chosen to avoid bias towards the majority class in the dataset.

The KNN model is built using the cosine distance metric, which is more suitable for text-based data or sparse vectors such as the results of the TF-IDF process. After the model is trained, the test data will be predicted to calculate the evaluation metric. A 3-fold cross-validation process is performed using the cross_val_score function to obtain more robust performance estimates. Each fold maintains a stratified class distribution to ensure the validity of the evaluation. The evaluation metrics calculated include accuracy and a classification report that includes precision, recall, and F1-score for each class as shown in Figure 4. These evaluation results provide a comprehensive overview of the model's performance in general and for each class. To provide a clearer performance overview, Table 1 presents a detailed comparison of the evaluation metrics derived from the three experiments.



	Exp 1			Exp 2			Exp 3		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
■ 4 - Undefined	0,56	0,45	0,5	0,53	0,5	0,51	0,44	0,35	0,39
■ 3 - Province	0,64	0,64	0,64	0,72	0,59	0,65	0,78	0,64	0,7
■ 2 - National	0,76	0,76	0,76	0,75	0,88	0,81	0,76	0,83	0,79
■ 1 - City/District	0,79	0,68	0,73	0,75	0,77	0,69	0,75	0,68	0,71
■ 0 - International	0,75	1	0,86	0,72	0,95	0,83	0,78	1	0,88

Figure 4: Comparison chart of precision, recall, and accuracy of the three experiments.

Table 1: Detailed comparison of evaluation metrics from the three experiments.

Experiment	Preprocessing	Accuracy	Class	Precision	Recall	F1-score	Support
Exp 1	Without Cleaning	0.714	0 (International)	0.75	1	0.86	21
			1 (City/District)	0.79	0.68	0.73	22
			2 (National)	0.76	0.76	0.76	41
			3 (Proviice)	0.64	0.64	0.64	22
			4 (Undefined)	0.56	0.45	0.5	20
Exp 2	OCR Cleaning + Casefolding + Punct	0.706	0 (International)	0.72	0.95	0.83	21
			1 (City/District)	0.75	0.77	0.69	22
			2 (National)	0.75	0.88	0.81	41
			3 (Province)	0.72	0.59	0.65	22
			4 (Undefined)	0.53	0.5	0.51	20
Exp 3	OCR Cleaning + Casefolding + Punct + Stopword	0.722	0 (International)	0.78	1	0.88	21
			1 (City/District)	0.75	0.68	0.71	22
			2 (National)	0.76	0.83	0.79	41
			3 (Province)	0.78	0.64	0.7	22
			4 (Undefined)	0.44	0.35	0.39	20

4 CONCLUSIONS

This research concludes by creating an automatic classification system for SNBP proof of achievement selection using the K-Nearest Neighbor (KNN) algorithm. The KNN algorithm's evaluation using the cross-validation method and evaluation metrics such as precision, recall, and F1-score shows that text preprocessing has a significant influence on model performance.

The best experiment, experiment 3, which uses complete preprocessing including stopword removal, gives the highest accuracy of 0.722. The model in this experiment also performed best on the international class, with an F1-score of 0.88 and recall of 1.00, meaning that all International data in the test set was correctly classified. The National class performed consistently high in all experiments, with F1-score between 0.76 and 0.81, indicating that documents in this category have features easily recognized by the model. The lowest performance was consistently found in the "Undefined" class, with an F1-score of only 0.39 in the 3rd experiment. This indicates that documents in this category have text structures that are inconsistent or similar to other classes, making it challenging to distinguish automatically. In general, adding preprocessing stages such as stopword removal improves the quality of feature representation. It positively impacts classification results, both overall (accuracy) and per class, except for the undefined class.

These results show that the KNN method with TF-IDF vectorization and cosine similarity can classify proof of achievement well. This approach can speed up the SNBP selection process automatically and efficiently and reduce the burden of manual classification by the committee. For future research, further preprocessing steps such as stemming or lemmatization should be added to make the text feature representation cleaner and more informative. It is also recommended to explore other classification algorithms such as Support Vector Machine (SVM), Naive Bayes, or Random Forest, which have different approaches in handling text data, especially on OCR extracted data that tends to be inconsistent.

Image preprocessing in this study was performed manually (e.g., rotation and cropping) due to efficiency considerations and time constraints. This approach was chosen so that the research could focus on testing the performance of the classification algorithm, while automation of preprocessing could be pursued in future studies.

DISCLAMER

AI tools were used to help with this writing so that the grammar would be improved.

REFERENCES

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118–127. https://doi.org/10.31294/ijcit.v6i2.10438

Danny, M., Muhidin, A., & Jamal, A. (2024). Application of the K-Nearest Neighbor Machine Learning Algorithm to Preduct Sales of Best-Selling Products. *Brilliance: Research of Artificial Intelligence*, *4*(1), 255–264. https://doi.org/10.47709/brilliance.v4i1.4063

Firdaus, A., Syamsu Kurnia, M., Shafera, T., Firdaus, W. I., Teknik, J., Politeknik, K., & Sriwijaya -Palembang, N. (2021). Implementasi Optical Character Recognition (OCR) Pada Masa Pandemi Covid-19 *1. *Jurnal JUPITER*, *13*(2), 188–194.

Francis, S. A., & Sangeetha, M. (2025). A comparison study on optical character recognition models in mathematical equations and in any language. *Results in Control and Optimization*, *18*, 100532. https://doi.org/10.1016/j.rico.2025.100532

Gani, M. O., Ayyasamy, R. K., Alhashmi, S. M., Sangodiah, A., & Fui, Y. T. (2022). ETFPOS-IDF: A Novel Term Weighting Scheme for Examination Question Classification Based on Bloom's Taxonomy. *IEEE Access*, *10*(December), 132777–132785. https://doi.org/10.1109/ACCESS.2022.3230592

Iqbal Mubarok, M., Purwantoro, P., & Carudin, C. (2024). Penerapan Algoritma K-Nearest Neighbor (Knn) Dalam Klasifikasi Penilaian Jawaban Ujisan Esai. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(5), 3446–3452. https://doi.org/10.36040/jati.v7i5.7676

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(1), 100008. https://doi.org/10.1016/j.jjimei.2021.100008

Lai, Y. W., & Chen, M. Y. (2023). Review of Survey Research in Fuzzy Approach for Text Mining. *IEEE Access*, *11*(February), 39635–39649. https://doi.org/10.1109/ACCESS.2023.3268165

Lewu, R. Y., Kusrini, K., & Yaqin, A. (2024). Comparing text classification algorithms with n-grams for mediation prediction. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *18*(2). https://doi.org/10.22146/ijccs.93929

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Ni'mah, A. T., & Syuhada, F. (2022). Term Weighting Based Indexing Class and Indexing Short Document for Indonesian Thesis Title Classification. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *6*(2), 167–175. https://doi.org/10.29303/jcosine.v6i2.471

Nugraha, K. A. (2024). Penerapan Optical Character Recognition untuk Pengenalan Variasi Teks pada Media Presentasi Pembelajaran 69. *Jurnal Buana Informatika*, 69–78.

Pertiwi, L. (2022). Penerapan Algoritma Text Mining, Steaming Dan Texrank Dalam Peringkasan Bahasa

Inggris. *BIMASATI(Bulletin of Multi-Disciplinary Science and Applied Technology)*, *1*(3), 100–104.

Sholehhudin, M., Fauzi Ali, M., & Adinugroho, S. (2018). *Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)*. *2*(11), 5518–5524.

Taha, K., Yoo, P. D., Yeun, C., Homouz, D., & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review*,