Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2025 Vol. 7 No. 2 Hal: 114-120 e-ISSN: 2714-9730

Machine Learning-Based News Classification: Comparison of KNN Accuracy with Hyperparameter Tuning

Muhamad Nur Gunawan[1]✉, Nuryasin[2], Syopiansyah Jaya Putra[3], Sarah Arhami[4]

[1,2,3,4]Syarif Hidayatullah Islamic State University, Jakarta, Indonesia

nur.gunawan@uinjkt.ac.id

Abstract

This study aims to develop an automatic news text classification system using the K-Nearest Neighbor (KNN) algorithm with a hyperparameter tuning approach. Manual classification by editors is considered inefficient, so an accurate and lightweight automated approach is needed. News datasets were obtained through web scraping of bbc.com sites with five main categories, namely business, technology, entertainment, science, and health. This research follows the CRISP-DM methodology which consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Feature representation is done using TF-IDF and preprocessing includes stopword removal as well as pattern-based noise cleaning. Two experimental scenarios were performed: first, using complete data without balancing; Second, using more balanced undersampling data. Hyperparameter tuning was performed with k-value variations from 1 to 50 and validated with 5-fold cross-validation. The results showed that the model with balanced data and a value of k=11 produced an accuracy, precision, recall, and F1-score of 95%. The system was also successfully implemented into a Flask-based web application that can be used by news editors for real-time text classification. This study emphasizes the importance of parameter optimization and preprocessing in text classification and shows that simple algorithms such as KNN remain competitive if supported by good data processing.

Keywords: *Text Classification, KNN Algorithm, Parameter Tuning, CRISP-DM, Digital News.*

1. Introduction

The rapid development of information technology in the last decade has created a major transformation in the way humans access, produce, and consume information, including in the form of digital news texts. Online media is now the main channel for disseminating information, replacing the dominance of conventional print media. In the midst of such a surge in digital news volume, the main challenge faced by media organizations is how to manage and classify news texts efficiently and accurately without relying on manual manpower that is time-consuming and prone to errors[1]. The need for an automated news classification system is becoming increasingly urgent, especially to support decision-making, the preparation of thematic archives, and the presentation of news based on readers' interests. Text mining as a statistical approach to text data plays an important role in managing the complexity of unstructured data generated by online news platforms[2]. Through the text classification process, the system can identify and group news into specific categories such as politics, technology, business, or health, making it easier to manage and utilize.

Machine learning algorithms are the dominant approach in building an adaptive and precise news text classification system. Among the various classification algorithms available, K-Nearest Neighbor (KNN) is known for its simplicity of implementation as well as its effectiveness in handling high-dimensional data such as text [3], [4]. The KNN approach works on the principle of spatial proximity between documents in vector space, where a new document is classified based on the majority of categories from its closest neighbor [5]. Although it is classified as non-parametric, the effectiveness of KNN is highly dependent on the selection of the k parameter, the representation of text features (TF-IDF, bag-of-words), and the preprocessing techniques used [6]. Therefore, the hyperparameter tuning process in KNN is a crucial aspect to optimize its performance in text classification [7].

However, in practice, many studies have not explicitly explored the impact of hyperparameter tuning on the performance of KNN classifications in the news text domain. Most previous studies have focused more on comparisons between algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree without paying special attention to the optimization of internal parameters within the KNN itself [8], [9]. In addition, the news classification approach also still faces various challenges such as data imbalance between news categories, the presence of noise in text, and high variability of language styles [10]. Therefore, further

Receipt: 23-01-2025 | Revision: 11-04-2025 | Publish: 17-07-2025 | doi: 10.60083/jidt.vi0.661

studies are needed that specifically test the impact of the hyperparameter tuning process on the accuracy and stability of the KNN model in classifying multicategory news texts.

This research is designed to answer this need through the application of the KNN algorithm combined with the k-parameter tuning process in an experimental space based on news datasets from the BBC. This dataset was collected through a web scraping process and includes five main categories: business, technology, health, entertainment, and science, which represent a variety of online news content and writing styles [11]. The method used refers to the CRISP-DM (Cross Industry Standard Process for Data Mining) framework which consists of six systematic stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [12]. The model was developed using Google Colab and implemented in a Flask-based web application to ensure the application of the model directly in the user's environment [13], [14]. This study specifically contributes to evaluating how parameter tuning affects classification results based on accuracy, precision, recall, and F1-score metrics.

The uniqueness of this research lies in two things: first, an experimental approach that compares the classification results between the standard KNN model and the KNN model as a result of hyperparameter tuning; second, the application of the CRISP-DM framework which is rarely used explicitly in news classification studies, even though this framework has been proven effective in the practice of the data mining industry [15]. The results of this study are expected not only to enrich theoretical understanding of the performance of the KNN algorithm in the text domain, but also to make a practical contribution to the development of a machine learning-based news classification system that can be used by media editors and information application developers. Thus, this research has a contribution value both scientifically and applicatively.

A literature review shows that although the KNN algorithm has been widely used in document classification, parameter tuning approaches are rarely carried out systematically, especially in the context of online news classification [16], [17]. Most studies still rely on arbitrary k-value selection without a cross-validation process or grid search. In fact, in many cases, errors in determining the k-value can have a significant impact on the accuracy of the model's predictions [18]. In addition, the issue of data imbalance, i.e. the uneven distribution of labels between categories, is also a common problem in the classification of news texts, which often leads to predictive bias against the majority class [19]. This study tries to overcome this issue through undersampling techniques in the dominant news categories and noise removal based on sentence position patterns.

In the realm of text preprocessing, this study adopts a classic but still effective approach that includes tokenization, stopword removal, letter normalization, and transformation into TF-IDF representations. This technique has been shown to improve the accuracy of text classification models in various studies [20], [21]. The vector representation generated from the TF-IDF process allows the KNN model to more accurately calculate the distances between documents in multidimensional space, which is the basis for classification. Euclidean distance calculations in the Vector Space Model (VSM) were chosen as a metric for similarity between documents, as they are also widely used in the Information Retrieval literature [22], [23].

In general, this study aims to: (1) develop an automatic news classification system based on the KNN algorithm, (2) evaluate the performance of the model before and after hyperparameter tuning, and (3) apply the CRISP-DM framework in the online news text classification process. With this approach, this study is expected to be able to fill gaps in the literature related to machine learning-based news classification, especially those related to parameter optimization in non-parametric algorithms such as KNN. This study will also highlight the extent to which classification results can be improved through simple but systematic tuning of k-values, as well as their impact on key evaluation metrics.

Based on the above description, the main research questions can be formulated as follows: To what extent can k-parameter tuning improve the accuracy of news text classification using the K-Nearest Neighbor algorithm compared to the no-tuning model? This question is the main focus of the research and will be answered through controlled experiments with data that has been systematically prepared. This research not only contributes to answering theoretical questions about KNN performance, but also presents a replication approach that can be adapted by researchers or other practitioners who want to develop a news classification system based on machine learning.

2. Research Methods

This study uses an experimental quantitative approach that aims to test the performance of the K-Nearest Neighbor (KNN) algorithm in automatically classifying news texts, especially by applying the hyperparameter tuning process to the value of the k-parameter data preparation, modeling, evaluation, and deployment. The dataset used came from the results of web scraping of international news sites bbc.com conducted in the period from May to June 2020. The dataset consists of five news categories, namely business, technology, entertainment, science, and health, with a total of 3685 raw data. After data cleaning, duplication filtering, and noise removal based on sentence patterns at the end of the text (e.g., email addresses, journalists' social media, or calls to respond), the amount of data was filtered to 813 lines. The distribution of the data includes 290 business

news, 180 entertainment, 151 technology, 121 science, and 71 health, showing the inequality of class distribution. To overcome the data imbalance, a random undersampling technique was carried out by taking 100 data each per category except for the health category which retained 71 data due to population limitations.
The pre-processing process of text is carried out with a standard approach in text mining, namely: conversion of letters to lowercase format, removal of non-alphabetic characters, tokenization, stopword removal using the nltk library, and normalization of sentence structure. After that, the transformation into a numerical representation is carried out using the TF-IDF vectorization scheme to extract the features of the key words in each document. This representation is used as a basis for calculating the distances between documents in high-dimensional vector spaces using the Euclidean distance metric. Further, the processed data is divided into two experimental schemes: first, experimenting with all filtered data without equalizing the number of classes, and second, experimenting with the data resulting from undersampling to make the number of categories more balanced. The modeling was carried out using the KNN algorithm implemented with the scikit-learn library in the Google Colaboratory environment. In the first stage, modeling is carried out without parameter tuning. Next, a hyperparameter tuning process was carried out with variations in k-values from 1 to 50 to find the optimal parameters based on an average score of 5 cross-validations.
The model evaluation was carried out using four main metrics: accuracy, precision, recall, and F1-score. In addition, a confusion matrix is used to observe the distribution of classifications between categories. The best model of the tuning results is determined based on the highest accuracy value of all the k-values tested. Based on the tuning results, the best k-value for the first experiment (all data) is 29, and for the second experiment (balanced data) is 11. All evaluations were carried out by stratified sampling to maintain the proportionality of the class distribution during training and model testing. Finally, the KNN model that has been obtained is implemented into a web-based system using the Flask framework. The web interface is designed to allow users to enter news text and is automatically classified into the five available categories. The application is then deployed using a PaaS (Platform as a Service) platform with integration through Git to facilitate continuous development and public use. All stages and procedures in this study are carried out systematically so that they can be replicated by other researchers with similar data contexts and computational environments.

3. Results and Discussion

This study produced two news text classification models based on the K-Nearest Neighbor (KNN) algorithm with two experimental scenarios, namely the first experiment using the entire filtered dataset (813 data), and the second experiment using the undersampling dataset (500 data). The dataset was obtained through a web scraping process from the BBC website with five main categories: business, technology, entertainment, science, and health. The initial distribution of data shows that there is a disparity between labels, namely: business (290), entertainment (180), technology (151), science (121), and health (71). To overcome this, undersampling is carried out so that each category has 100 data each, except health.
In the preprocessing stage, the text is cleaned from noise such as email information, journalists' social media, and feedback sentences, which generally appear in the last two to three sentences of the news. The next preprocessing process includes converting letters to lowercases, removing non-alphabetic characters, tokenization, stopword removal (using NLTK), and transforming text into vectors using TF-IDF.
The first model (using all data) and the second model (using balanced sample data) were each tested through a hyperparameter tuning process with a k-value from 1 to 50. The tuning results showed that the optimal value for the first model was k=29, and for the second model was k=11. The entire experiment used 5-fold cross-validation.

Here is a summary of the metrics from the model evaluation (Table 1):

Table 1. Confusion Metrics

Model Scenario	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Complete Dataset	92	92	92	92
Sample Dataset	95	95	95	95

In the second model (balanced dataset), the confusion matrix shows a very precise classification of almost all categories, as follows (Table 2):

Table 2. Classification

	Pred B	Pred E	Pred H	Pred S	Pred T
Actual B	18	1	0	1	1
Actual E	0	24	0	0	0
Actual H	0	0	21	0	0
Actual S	0	0	0	31	0
Actual T	2	0	0	1	18

The categories with the most stable predictions are science and health, while technology and business show minor prediction errors between the two. This is confirmed from the following data citations:

"In the business category, some texts contain technology terms such as 'AI', 'Google', which give rise to misclassifications into technology labels."

"The health label remains stable even though the dataset is small, because it has specific terms such as 'covid patient' that are not ambiguous."

The KNN model was then implemented as a web-based application using the Flask framework. The system provides a text input box and a classification button, allowing users (editors) to instantly classify new news texts. The following is a description of the system interface:

Input Box: An area to enter news text

"Classification" button: The classification process begins

Output: The category label of the classification result is displayed automatically

During the application testing process, the system can classify news text in just a few seconds. Here is an example of the actual classification results based on the input of the new test text:

Text input: "Covid-19 cases are rising again in rural areas with limited access to hospitals."

Classification result: health

Text input: "Apple and Google invest in AI chips for next-gen smartphones."

Classification result: technology

However, there are still a few misclassifications such as:

Text input: "Google Play Music expands feature to podcast creators."

Classification result: business

It should be included in the entertainment category.

The developed application supports real-time classification and can be used as a news editor tool to save time and improve editorial work efficiency. The model is proven to be stable in texts that have clear and consistent lexical characteristics between categories.

The results of the study show that the K-Nearest Neighbor (KNN) algorithm is able to provide high performance in the classification of news texts, especially after the hyperparameter tuning process is carried out. These findings are in line with previous studies that stated that KNN is one of the simple but effective algorithms in dealing with the problem of classification of high-dimensional texts [4], [5]. In this study, the tuned KNN model produced an accuracy, precision, recall, and F1-score value of 95% in the balanced dataset. This figure is higher than the accuracy of the untuned model (92%) in the complete dataset, reinforcing the finding that optimal k-value selection greatly affects the performance of the nearest neighbor-based classification model.

The superior performance of the second model also reinforces the results of research that emphasizes the importance of data balancing in text classification [9]. The unbalanced distribution of classes in the initial dataset with categories such as business (290 data) and health (71 data) led to bias in predictions to the majority class, as also reported by Dadgar et al. (2016) in SVM's study of news datasets. Using the undersampling technique, the researchers managed to stabilize the distribution between labels and improve the accuracy of the model's predictions. This supports the argument that for algorithms such as KNN, the proportional distribution of classes is highly influential in the process of determining the majority of labels from the nearest neighbors.

The application of hyperparameter tuning in this study makes an important methodological contribution. So far, many studies have used KNN directly without exploring the optimal value of the k parameter [16]. In fact, the selection of a k value that is too small can cause the model to be too sensitive to noise (overfitting), while too large a k value can reduce sensitivity to minority patterns (underfitting) [7]. In this study, the tuning process was carried out by testing k-values from 1 to 50 and validated through 5-fold cross-validation, resulting in an optimal value of k=11 for a balanced dataset and k=29 for a complete dataset. This suggests that there is no single universally applicable k-value, but rather depends on the distribution and complexity of the data.

Theoretically, these results also confirm the advantages of TF-IDF-based feature representation in aiding the distance-based classification process [3]. With TF-IDF, words that are too common in the document are given a lower weight, while class-specific words are given a higher weight. This is very important in distinguishing news texts that have similar language structures but different content contexts, such as "Apple develops new AI chips" (technology) and "Google releases quarterly financial statements" (business). A study by [19] also mentioned

that the use of TF-IDF in text classification can improve accuracy by up to 15% compared to regular binary representations, especially in short documents such as news.

These findings have strong practical significance in the context of the media industry. As conveyed by [24], digital media today faces challenges in quickly sorting information to avoid information overload. An automated classification system like the one developed in this study allows editors to classify articles more efficiently without reading the entire content of the news. The web-based application built allows for live text input and real-time classification, supporting a digital newsroom working model that demands speed and accuracy. With this automatic classification, editorial workflows become more standardized and automated.

High performance on specific labels, such as health and science, also provides additional insights. Although the amount of data for the health category is relatively small, the model still provides stable classification results. This is allegedly due to the existence of specific and unambiguous keywords in the category, such as "covid", "patient", and "vaccine", which lexically do not appear in other categories. In contrast, categories such as technology and business have a lot of overlapping keywords, such as "Google" or "Apple," which can lead to ambiguity. This phenomenon is also observed by [17] in the emotional classification of Indonesian texts using KNN, where ambiguous words are the main source of prediction errors.

However, this study also has limitations that need to be observed. First, the dataset used was limited to just five categories and sourced from a single English-language news portal. This limits the generalization of results to other language contexts or different types of media. Second, only one algorithm is used, namely KNN, with no direct comparison to other algorithms such as SVM, Naïve Bayes, or Decision Tree. In fact, some previous studies have shown that SVM can provide competitive accuracy results in text classification [8]. Third, the classification system that was built is still limited to manual text input, does not yet support file uploads or batch classification in bulk.

The main contribution of this research lies in three aspects. First, methodologically, this study introduces the practice of systematic hyperparameter tuning on the KNN algorithm for the classification of news texts, an approach that has not been widely adopted in the literature before. Second, from a technical point of view, this model is practically implemented in the form of a ready-to-use web application, not just an experimental model. Third, this study contributes a new understanding of how feature representation and data balancing can optimize the performance of distance-based classification models.

The implications of these findings cover two main domains. In the academic field, this research can be a reference for the development of further studies in the domain of text classification, especially those that use a distance-based learning approach. In practical terms, applications from this system can be extended to local English-language media, community news portals, or content recommendation systems. In addition, the approach used can also be applied in other domains such as spam email classification, consumer opinion, or social media analysis.

As a suggestion for further research, it is recommended that similar research be conducted on news data in Indonesian by expanding the coverage of categories, such as politics, economics, sports, and culture. Future research could also integrate other algorithms as a comparator, or use ensemble learning methods to improve accuracy. In addition, the development of classification systems that support inputs in the form of document files such as PDF or DOCX, as well as mass classification on a large scale, will make the system more relevant in the context of professional journalistic content production.

Overall, the study successfully shows that with proper preprocessing, controlled data balancing, and systematic parameter tuning, a simple algorithm such as KNN can provide excellent news text classification results. These results also prove that the effectiveness of a model does not always depend on the complexity of the algorithm architecture, but rather on how the data is prepared and the parameters are optimized. Therefore, this approach can be an efficient alternative for news organizations that require a lightweight yet high-precision classification system.

4. Conclusion

This study proves that the K-Nearest Neighbor (KNN) algorithm is able to automatically classify news text with excellent performance, especially after systematic hyperparameter tuning . Through two experimental scenarios using a complete dataset and an undersampling dataset it was found that the model with balanced data and optimal parameters (k=11) produced 95% accuracy, precision, recall, and F1-score. This shows that the effectiveness of the classification model is determined not only by the selection of the algorithm, but also by the data processing strategy and parameter optimization.

The use of the CRISP-DM method in the framework of this research provides a structured workflow, starting from business understanding, data understanding, data preparation, modeling, evaluation, to deployment. The application of comprehensive text preprocessing techniques , feature representation with TF-IDF, and balancing data between categories proved to be the main supporting factors for the success of the model. In addition, the

final implementation in the form of a web-based application shows that this system is not only theoretically relevant but also applicable in a digital editorial environment.

The contributions of this research include strengthening the methodological foundation in the use of KNN for text classification, affirming the importance of hyperparameter tuning, as well as empirical evidence that simple algorithms can compete in accuracy if supported by the right data processing process. This research also opens up new space for the development of an automatic text classification system that is lightweight, easy to implement, and able to adapt to various types of text data.

However, this study has limitations, including using only one language (English), limited to five news categories, and testing only one machine learning algorithm. Therefore, for further development, it is recommended to experiment with larger and multilingual datasets. Future research can also explore a combination of algorithms or ensemble learning approaches, as well as build a text classification system capable of processing documents in file formats such as PDF or Word and supporting batch classification.

Thus, this research provides a strong foundation for advanced studies in the field of machine learning-based news text classification and contributes to the development of information systems that are efficient, precise, and relevant to the needs of today's digital media industry.

Acknowledgements

This research is a grant from the Center for Research and Publication (PUSLITPEN) of Universitas Islam Negeri Syarif Hidayatullah Jakarta in 2022

References

[1] C.-H. Chan, A. Sun, and E.-P. Lim, "Automated Online News Classification with Personalization BT - 4th Int. Conference Available: [Online]. Asian Digit. Libr.," 2001, pp. 1–10. [Online]. Available: http://ncsi-net.ncsi.iisc.ernet.in/gsdl/collect/icco/index/assoc/HASH01de.dir/doc

[2] V. Korde, "Text Classification and Classifiers: A Survey," *Int. J. Artif. Intell. Appl.*, vol. 3, no. 2, pp. 85–99, 2012, doi: 10.5121/ijaia.2012.3208.

[3] R. Jindal, R. Malhotra, and A. Jain, "Techniques for text classification: Literature review and current trends," *Webology*, vol. 12, no. 2, pp. 1–28, 2015.

[4] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 61–70, 2014, doi: 10.14257/ijdta.2014.7.1.06.

[5] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[6] 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.

[7] M. DEI, "Hyperparameter Tuning Explained Tuning Phases, Tuning Methods, Bayesian Optimization, and Sample Code!" 2019. [Online]. Available: https://towardsdatascience.com/hyperparameter-tuning-explained-d0ebb2bald35

[8] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification BT - Proc. 2nd IEEE Int. Conf. Eng. Technol. ICETECH 2016," 2016, pp. 112–116. doi: 10.1109/ICETECH.2016.7569223.

[9] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, pp. 221–232, 2017, doi: 10.22364/bjmc.2017.5.2.05.

[10] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[11] M. A. Fauzi, A. Z. Arifin, S. C. Gosaria, and I. S. Prabowo, "Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 2, no. 3, pp. 401–408, 2016, doi: 10.11591/ijeecs.v2.i2.pages.

[12] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects." 2014. [Online]. Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[13] Google, "Colaboratory - Google." 2020. [Online]. Available: https://research.google.com/colaboratory/faq.html

[14] F. (web framework) Wikipedia, "Flask (web framework) Wikipedia." 2020. [Online]. Available: https://en.wikipedia.org/wiki/Flask_(web_framework)

[15] F. Martinez-Plumed, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 4347, no. c, p. 1, 2019, doi: 10.1109/tkde.2019.2962680.

[16] J. Rajshree, S. B. Gaur, C. K. R., and M. Amit, "Text Classification using KNN with different Features Selection Methods Abstra," *Int. J. Res. Publ. Vol. 8-Issue. 1, July 2018 Text*, 2018.

[17] Arifin, "Classification of Emotions in Indonesian TextsUsing K-NN Method," *Int. J. Inf. Electron. Eng.*, vol. 2, no. 6, 2012, doi: 10.7763/ijiee.2012.v2.237.

[18] M. Sanjay, "Why and how to Cross Validate a Model." 2018. [Online]. Available: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

[19] X. Fang and J. Zhan, "Sentiment analysis using product review data," *J. Big Data*, vol. 2, no. 1, 2015, doi:

10.1186/s40537-015-0015-2.
[20] K. L. Sumathy and M. Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues An Overview," *Int. J. Comput. Appl.*, vol. 80, no. 4, pp. 29–32, 2013, doi: 10.5120/13851-1685.
[21] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.
[22] Informatikologi, "Algoritma K-Nearest Neighbor (K-NN) INFORMATIKALOGI." 2017. [Online]. Available: https://informatikalogi.com/algoritma-k-nn-k-nearest-neighbor/#1
[23] Informatikologi, "Vector Space Model (VSM) dan Pengukuran Jarak pada Information Retrieval (IR) INFORMATIKALOGI." 2016. [Online]. Available: https://informatikalogi.com/vector-space-model-pengukuran-jarak/#1
[24] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. K. Nielsen, "Digital News Report 2019," pp. 70–72, 2019, doi: 10.2139/ssrn.2619576.