JOURNAL OF DATA SCIENCE | Vol.2023:05
eISSN:2805-5160

MULTI-LABEL TEXT CLASSIFICATION FOR INDONESIAN LANGUAGE IT JOURNAL WITH K-NEAREST NEIGHBORS (KNN)

Redho Aidil Iqrom[1], Tri Basuki Kurniawan[2]

[1,2]Teknik Informatika, Pascasarjana, Universitas Bina Darma, Indonesia

Email: redhoaidiliqrom@gmail.com, tribasukikurniawan@binadarma.ac.id

Abstract

Classification is the process of finding a model or function that explains or distinguishes concepts or data classes, intending to estimate the category of an object whose label is unknown, and various types of classification, one of which is the classification of text documents. Document text classification based on label category is one of the mandatory components in the retrieval system to provide better and more accurate information. Based on existing research, only single-label Classification of text documents is carried out, and it is infrequent for multi-label Classification of IT journals, especially in the Indonesian language. Therefore, this research is aimed at multi-label text classification using the K-Nearest Neighbors (KNN) method, and the OnevsRest Classifier approach model, where the classification process will be determined by the closest k = n value in the category of documents that are similar and the multi-labels are in prediction with One vs. Rest Classifier. Training and testing are done with a dataset of 500 Indonesian IT journals. The test results are sufficient to give good results with an accuracy of 84% and a hamming loss of 0.076.

Keywords

Classification, k-nearest neighbours, one vs. rest classifier, single label, multi-label

Introduction

Data exchange at this time can be said to be fast and more advanced by keeping up with information technology updates that are constantly evolving. In other words, these advances make it easier for us to find the information data we need. Data is the result of direct observation of events or facts from phenomena in the real world that are equipped with specific values (Sri, 2018), and data can be in the form of text, audio, pictures, and videos. Obtain information data now; most people can find it quickly through various media, including social and digital media. We can find informative data on digital media through different web pages. Still, based on the Net Craft Web Server Survey in May 2008, the total number of active websites is 168 million compared to now, in 2022, the number of active sites is 1.1 billion (NETCRAFT, 2017).

With an increase in the number of web pages, there is a slight problem finding the correct information according to keywords or field themes, such as journal information data in the form of research abstracts, because there could be one keyword in one journal-title. Still, it contains

Submission: 18 June 2023; **Acceptance:** 17 August 2023

other keywords, usually called multi-labels, especially for final-year students looking for references for their research, where students often use one or more keywords (Tremblay, 2013). Therefore it is essential to process text categories in journals according to their classes. One way that can be used is by classifying journals and processing them using the Information Retrieval concept to provide better and more accurate information in the abstract search process in journals according to their contents.

Classification systematically groups several objects, ideas, books, or other objects into particular classes or groups based on the same characteristics (ISMANU, 2012). Based on the classification categories, there are single-label and multi-label forms. For single-label, many methods can be used to classify one of them as K-Nearest Neighbors (KNN). K-Nearest Neighbors is a classification method that classifies new data based on the distance of the new data to some nearest neighbours/data (Wati, 2023). The advantage is that this algorithm is very suitable for multi-label cases. Even KNN can be superior to other classifiers. Different methods should be used for multi-label, such as OnevsRest (Gaustautaite, 2022) and Binary Relevance (Zhang, 2018), which have special techniques to handle multi-label problem classification.

Methodology

1. Document Transformation

The document transformation process, also known as the preprocessing stage, must be carried out before entering the training & testing stage because document text is unstructured data.

Several processes are carried out at this representation stage, starting from tokenization, removing punctuation, stop words, stemming, and lemmatization, generating a term or phrase. Next, we enter the Feature Extraction process using TF-IDF, where in this process, each term is given a weight with the following formula (M. Ardiansyah, Nurjaya, 2022):

$$w_{t,d} = (1 + log_{10}tf_{t,d})log_{10}n/df_{t}, \quad (1)$$

n is the total number of documents in all documents, tft,d is the number of occurrences of terms in documents d, and dft is the number of documents that contain words in all documents.

2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning method that can be used for classification and regression. It uses training data and classifies test data based on the distance to the training data (Ramya & Pinakas, 2014). The goal is to find the number of k nearest neighbours between the test and training data. Then the classification results are determined through the class labels that are the most in the range of the k nearest neighbours.

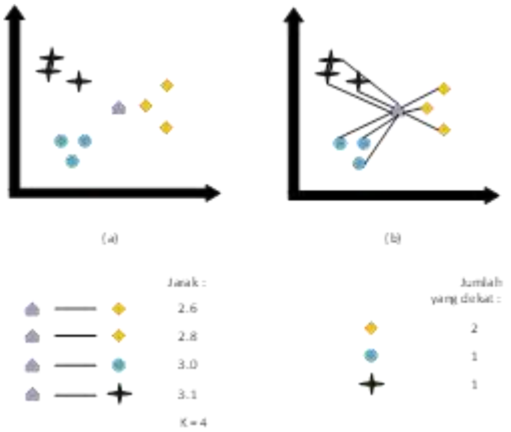

Figure 1. KNN Algorithm Illustration

Several kinds of distance calculations can be used in the process, namely; Euclidean, Manhattan, Minkowski, Mahalanobis

So the basic principles of KNN are (Nikhath, Subrahmanyam, & Vasavi, 2016) :

1. Determine the distance to several nearest neighbours.

The Euclidean distance calculation is one way to measure the distance between an object and its neighbours. Then how to find out the number of nearest neighbours? Here the method is used by determining the value of k at the beginning. This value of k is identical to the number of nearest neighbours.

2. Define Class

To determine the class can be known by considering the nearest neighbours. One often applied method is to sort and observe the results of as many as k pieces.

Results and Discussion

Testing this classification system measures the model's ability to predict the label on a document. The output is whether the document has another label category. For this reason, this test measures the difference in the mean score between the predicted and actual labels by measuring accuracy, precision, recall, f1-score and hamming loss. Two evaluation processes for the model can be carried out, first by the direct evaluation technique and second by the cross-validation technique (Pereira et al., 2018).

1. Evaluation of the KNN Predict Model

In Table 1, there is a classification report, one of the functions in the sci-kit-learn library, to generate reports on the performance of models that have been tested. The precision value is the ratio between the number of optimistic predictions and the number of positive predictions made by the model. The recall is the ratio between the number of correctly positive predictions and the number of positive data. F1-score is the harmonic mean of precision and recall, and support is the amount of data that belongs to each class in the data.

The results of testing the model directly with a value of k = 2 get sufficient accuracy with a mean score of 0.63 and a hamming loss of 0.073, where the smaller score of the hamming loss indicates the model is running quite well. Still, we can try to increase the accuracy score by There are several ways, one of which we can try is the cross-validation technique.

Table 1. Classification Report

K=2	Precision	Recall	F1-Score	Support
Data mining	0.90	0.86	0.88	21
Augmented reality	1.00	1.00	100	100
Big data	0.74	0.81	0.77	31
naïve bayes	1.00	1.00	1.00	100
Algoritma	0.70	0.70	0.70	10
Information systems	0.61	0.50	0.55	22

Accuracy Score: 0.63; Hamming Loss: 0.073

2. Cross Validation

The cross-validation technique is a performance evaluation technique of a model that aims to test the model's ability to generalize data that has never been seen before (Santos et al., 2018). Cross-validation works by dividing the data into several parts or "folding" and then training and testing the model in each part. Each iteration uses one part as validation (or testing) data, while the other is used as training (or training) data. The model can be tested by testing each part to determine its performance on different data.

One frequently used cross-validation technique is k-fold cross-validation (Jung & Hu, 2015), where the data is divided into k parts of the same size. Then, k iterations are carried out where one part becomes testing data in each iteration, and the other becomes training data. Model performance is calculated as the average of the performance in each iteration. This technique helps to reduce the possibility of overfitting and more accurately estimate model performance on data that has never been seen before.

Table 2. Classification Report Cross Validation

K=2 | cv=6	Precision	Recall	F1-Score	Support
Data mining	0.76	0.76	0.76	68
Augmented reality	1.00	1.00	100	500
Big data	0.64	0.60	0.62	137
naïve bayes	1.00	1.00	1.00	500
Algoritma	0.88	0.76	0.82	59
Information systems	0.64	0.51	0.56	91

Accuracy Score: 0.64; Hamming Loss: 0.075

By using the cross-validation, an accuracy score of 0.64 is obtained and a hamming loss of 0.075, where there is a slight increase in terms of accuracy without cross-validation. The average is 0.63. The evaluation uses the values k = 2 and cv = 6 as parameters in model testing. These values are the best values among the automated processes designed to obtain k and k-fold values suitable for use in the learning model process. Furthermore, it can be seen below the plot for finding the best k-fold value.

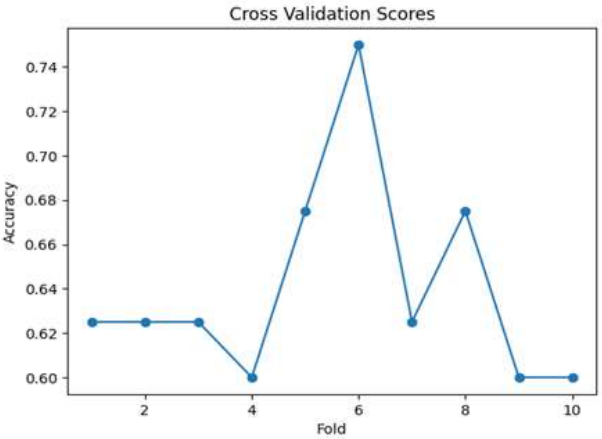

Figure 2. Search K-fold Validation Scores

The best value score from Figure 2 shows the number 6 with the highest accuracy score among the other numbers, equal to 0.78, wherein determining the k-fold is iterated on each parameter number 1-10. The score is obtained in each iteration starting from 0.63, 0.60, 0.66, 0.78, 0.63, 0.66, 0.60, 0.60.

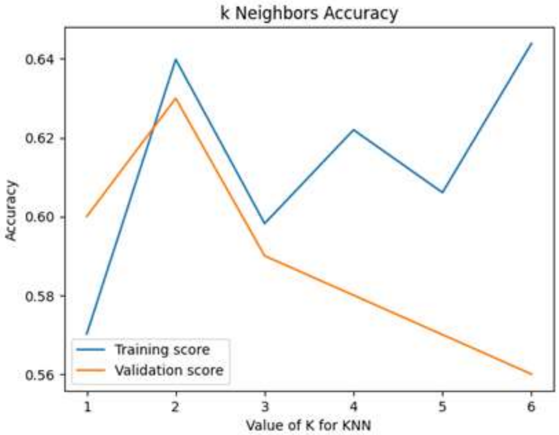

Figure 3. Search for value k for KNN

From the plot in Figure 3, we can see that six numbers have been determined as candidates to fill in the value of k, which will later be used in the model learning process. From these results, we can conclude that k=2 is the best value because the accuracy training and validation scores are not much different and relatively high.

4. Discussion

The model performance is expected when it arrives at the training & testing process, and efforts to improve it must always be top performers. Still, the model evaluation results using the test method with and without cross-validation can be seen in Figure 4.

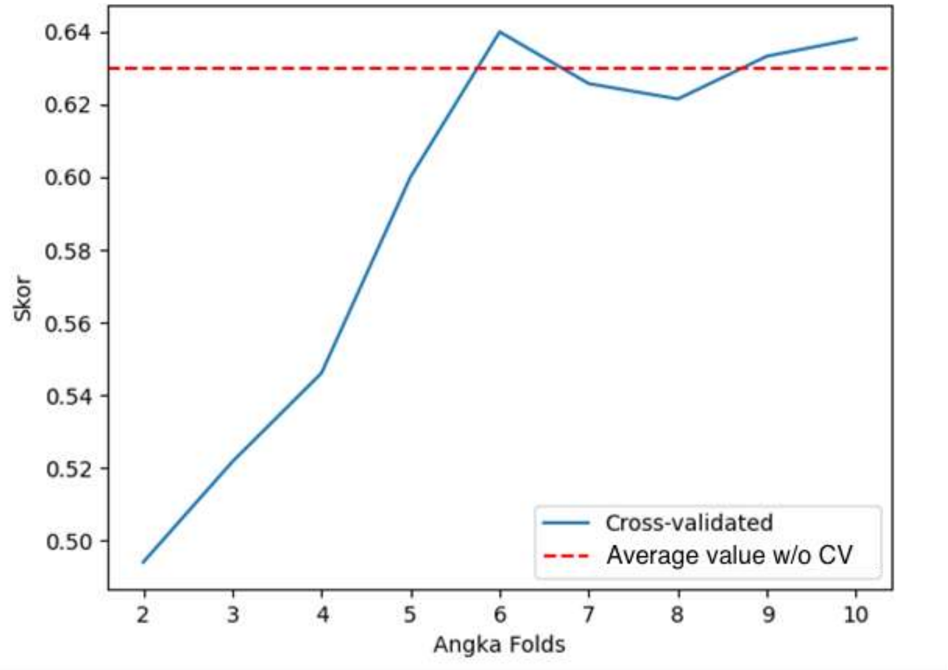

Figure 4. Model Performance Comparison

The results are not too significant. The difference's just that there is a slight difference in the accuracy score. The mean score is marked with a red dotted line of 0.63, and cross-validation is in the form of a straight blue line with several variations of the accuracy score, starting from 0.48, 0.52, 0.55, 0.58, 0.64 … 0.63. So, the k-folds value is set to six because the resulting accuracy is 0.64. The difference is 0.01 with accuracy without cross-validation, which should be expected to improve model performance which is better and more accurate in a training and testing process on data. However, this does not mean that the model that has been trained cannot classify documents in the class of predetermined labels. The results of the classification test can be seen in Table 3.

Table 3. Classification Test Results

Title	Recommend Labels	Accuracy	Precision	Recall	F1-Score
Algoritma Naive Bayes To Find Estimated Student Study Time	"naive bayes", "big data", "augmented reality"	0.500000	0.666667	0.666667	0.666667
Big Data Analysis of Covid-19 Spread With Business Intelligence (Bi)	"big data", "augmented	0.666667	1.000.000	1.000.000	1.000.000

	reality", "data mining"				
Big Data and Its Utilization in Libraries	'Libraries' big data", "augmented reality", "data mining’'	0.666667	1.000.000	1.000.000	1.000.000
Implementation of the Naive Bayes Algorithm in Determining Credit	"naive bayes", "big data", "information systems'systems'	0.333333	0.333333	0.333333	0.333333
Implementation of augmented reality as a learning medium for the introduction of mathematical arithmetic operations in kindergarten or preschool Permata Bunda Langsa	"augmented reality", "algoritma", "information systems'systems'	0.833333	1.000.000	1.000.000	1.000.000
Implementation of augmented reality as an introduction media for the Sukabumi Polytechnic computer study program using the marker-based tracking method on brochures	"algoritma", "augmented reality", "information systems'systems'	0.666667	1.000.000	1.000.000	1.000.000
Implementation of the Naive Bayes Classification Method to Predict Chili Quality	"naive bayes", "augmented reality", "algoritma"	0.500000	0.000000	0.000000	0.000000
IoT Security With Deep Learning and Big Data Technology	"augmented reality", "data mining", "big data"	0.166667	0.000000	0.000000	0.000000
Build a mobile application based on augmented reality as a teaching aid in selecting clothes	"algoritma", "naive bayes", "augmented reality"	0.666667	1.000.000	1.000.000	1.000.000
The Naive Bayes Method for Determining Bidikmisi Scholarship Recipients at Mulawarman University	"naive bayes", "augmented reality", "information systems'systems'	0.666667	0.500000	0.500000	0.500000

Table 3 shows the results of testing on 10 document data that have been prepared to test the model that has been trained. From these results, it can be seen that the label predictions displayed are pretty good at predicting multi-labels in a document, with each accuracy showed starting from 0.16, 0.33, 0.50, 0.66, …. 0.83, although the prediction results of the other labels are bit unsustainable due to the limited labels.

Conclusion

According to the study's findings, it can be said that OnevsRest Classifier technique is used in addition to K-Nearest Neighbours (KNN) for the classification of multi-label text in IT journals, primarily Indonesian. The iterations, which involved trying each of the established numerical parameters, revealed that two was the ideal k value. With the best k-folds at cv=6, the evaluation of model performance test results using cross-validation yields a mean score of 0.64. The findings and an accuracy score of 0.83 show that this method effectively predicts multi-labels, even though some labels are wrong.

References

Gostautaite, D & Sakalauskas, L (2022). Multi-label classification and Explanation Methods for Student's Learning Style Prediction and Interpretation. Applied Science 12(11).

Hakim, Rahmad (2019).Aplikasi Prediksi Kelulusan Mahasiswa berbasis K-Nearest Neighbor (KNN). Jurnal Teknologi Informasi dan Multimedia, Vol.1, No.1

ISMANU, I. (2012). Pengadaan Bahan Pustaka Untuk Perpustakaan Sekolah. *Universitas Negeri Malang*, 1–10.

Jung, Y., & Hu, J. (2015). A K-fold Averaging Cross-validation Procedure. *Journal of Nonparametric Statistics*, *27*, 1–13. https://doi.org/10.1080/10485252.2015.1010532

Ling Zhang, Min & Kun Li, Yu & Xin Geng (2018). Binary relevance for multi-label learning: an overview. Frontier of Computer Science 12.

M. Ardiansyah, Nurjaya, and M. I. R. (2022). *Data Mining Dan Implementasinya Untuk Klasifikasi Loyalitas Pelanggan*. Tangerang Selatan: Pascal Books.

NETCRAFT. (2017). How many active sites are there? Retrieved January 20, 2023, from Net Craft website: http://www.netcraft.com/active-sites/

Nikhath, A. K., Subrahmanyam, K., & Vasavi, R. (2016). Building a K-Nearest Neighbor Classifier for Text Categorization. *International Journal of Computer Science and Information Technologies*, *7*(1), 254–256.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. C. (2018). Correlation analysis of performance measures for multi-label classification. *Information Processing & Management*, *54*(3), 359–369. https://doi.org/https://doi.org/10.1016/j.ipm.2018.01.002

Ramya, M., & Pinakas, J. A. (2014). Different Type of Feature Selection for Text Classification. *International Journal of Computer Trends and Technology*, *10*(2), 102–107. https://doi.org/10.14445/22312803/ijctt-v10p118

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier]. *IEEE Computational Intelligence Magazine*, *13*(4), 59–76. https://doi.org/10.1109/MCI.2018.2866730

Sri, A. (2018). Pengantar Konsep Informasi, Data, dan Pengetahuan. *Modul Pembelajaran*, (1), 11–18.

Tremblay, K. (2013). Oecd assessment of higher education learning outcomes (ahelo): Rationale, challenges and initial insights from the feasibility study. *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges*, *1*, 113–126. https://doi.org/10.1007/978-94-6091-867-4