INDONESIAN JOURNAL OF MULTIDISCIPLINARY SCIENCE

P-ISSN: 2808-5957
E-ISSN: 2808-6724
ijoms.internationaljournallabs.com

DATA MINING USING NEURAL NETWORK FOR RESEARCH TOPIC CLASSIFICATION BASED ON INSTITUTIONAL RESEARCH ROADMAP

Yudi Feriandi[1,2*], Desi Siti Suhartini[1], Budi Permana[1], Christina Juliane[1]

[1] *Master of Infromation System, Sekolah Tinggi Manajemen dan Informatika LIKMI Bandung, Bandung, West Java, Indonesia*

[2] *Faculty of Medicine, Universitas Islam Bandung, Bandung, Indonesia*

[*] *yudi.feriandi@unisba.ac.id*

ARTICLE INFO

Published: April 30th, 2023

Keywords: *classification, data mining, neural network, text mining, visualization*

ABSTRACT

In the last five years, lecturers and students at the Faculty of Medicine at the Islamic University of Bandung have produced hundreds of studies. Still, studies on the suitability of various studies with topics according to the institution's research roadmap have not been carried out. This research aimed to classify research documents based on 12 research roadmap topics. The data used in this study are the research titles of lecturers and students in 2015-2021, amounting to 1064 data. The document extraction process uses text mining, while the document grouping process is carried out using a supervised method with an artificial neural network algorithm. At the text mining stage, preprocessing procedures are carried out in case folding, tokenization, and filtering, followed by analysis through weighting using IDF and evaluating accuracy, precision, and recall. The Neural Network Algorithm can classify by level. The classification results using the neural network algorithm show that overall from the training data, the average precision is 74.7%, recall is 74.3%, and accuracy is 74.3%. Of the 12 research topics, the training dataset obtained high accuracy, precision, and recall values found on herbal medicine, Islamic insert, industrial health, tuberculosis, and vaccination. The four topics are in line with the five institution's leading research topics. The results of dataset testing analysis found that the research topics carried out by students and lecturers of Unisba School of Medicine were distributed among the 12 research topics on the institutional roadmap with the increased trends in institutional leading research topics.

INTRODUCTION

The Faculty of Medicine at the Islamic University of Bandung is one of the faculties under the auspices of the Islamic University of Bandung which was established in 2004. With a mission to become an independent, advanced, and leading medical faculty in ASIA by 2030. One of the obligations as well as efforts to realize this mission is the implementation of higher education obligatory missions (education-research-community services) in the field of research and community service managed by the Research and Community Service Unit (UPPM) of the Faculty. In the last five years, 238 lecturer studies and 821 final project studies have been carried out. However, the research track record portfolio and its suitability with the vision and mission of the institution have not been mapped due to the unavailability of a reliable information system for knowledge management in the field of research and only using Microsoft Excel and Google Drive as data storage facilities.

The Faculty Research Unit has not analyzed in real-time the suitability between lecturer and student research topics and the roadmap because there is no specific grouping and categorization. In addition, the similarity of the title and the repetition of research topics makes it difficult to trace. As a result, it can become a factor that hinders the continuity of research and achievement of outcomes in the form of Intellectual Property Rights, patents, and downstream research. Efforts should be made to identify patterns of variations in titles and research topics that have been carried out so that various inputs for strategic planning in the research field can be identified.

Data mining is a process of finding information automatically to find the benefits of a data set. Data Mining is closely related to knowledge discovery in database/KDD which is the process of converting raw data into useful information. The text mining process is part of data mining which includes several stages such as information retrieval, categorization, POS tagging, clustering, and others according to the "Knowledge Discovery in Databases" framework. Text mining identifies patterns in data that are correct, unique, useful, and understandable (Hassani et al., 2020). With text mining, KDD becomes an interactive and iterative process. The text mining process consists of selection, preprocessing, transformation, data mining, and interpretation/evaluation stages (Faisal et al., 2007; Pramadhani & Setiadi, 2014; Suntoro, 2019).

Text mining processes large amounts of textual data. The textual data used can be in the form of writing, documents, or text. Due to the need for large resources to process textual data, an initial processing stage or preprocessing of textual data is needed before the text mining process is carried out according to the algorithm to be applied (Findawati & Rosid, 2020; Yuliana et al., 2019).

Data mining operations by their nature are divided into two, namely prediction (prediction-driven) and discovery (discovery-driven). Thus, text mining can also be used for both operations. One of the uses of text mining is for classifying data. The main purpose of text mining with the classification method is to group several data into certain classifications. Classification is grouping documents based on labeled training data. The difference with clustering is that in classification, the class/category is predetermined, while clustering is not. Various algorithms can be used to group a number of data such as Expectation maximization (EM), Naive Bayes classifier, Support vector machines (SVM), artificial neural networks, K-nearest neighbor algorithms, and so on (Davis et al., 2007; Faisal et al., 2007; Findawati & Rosid, 2020; Goyal & Vohra, 2012; Yuliana et al., 2019).

An Artificial Neural network (ANN) is a complex non-linear model built from components that individually behave like a regression model. Neural networks are machine learning models that mimic learning aspects of past experiences to predict the future. In general, there are two classes of machine learning techniques, namely supervised and unsupervised. In the supervised method, the model is created based on the training dataset. In this method, the categories have been determined in the training dataset through manual tagging with one or more labels. The classification algorithm is then trained with the dataset which means it can predict the new document category given post-training. The level of classification accuracy will be largely determined by the classification algorithm or strategy used (Davis et al., 2007; Goyal & Vohra, 2012; Purwati et al., 2020; Yuliana et al., 2019).

Previous research conducted by Muslihudin and Zahrotun using the shared nearest neighbor algorithm for lecturer research allows the integration of text mining algorithms into research information systems to map research trends. In this study, the accuracy, precision, and recall of the algorithm used were not described (Muslihudin & Zahrotun, 2017). Text mining research uses the neural network classification conducted by Yuliana et al. on web-based public complaint text data resulting in an accuracy of 43% (Yuliana et al., 2019). Research comparing the use of three text mining algorithms by Wibowo et al. (2022) shows that the Neural Network classification model has the highest accuracy value (94.1%) while the lowest accuracy value is the Naïve Bayes model (79%), but this research was conducted on a small number of datasets.

Even though the evaluation results were quite good, research on the classification of lecturer and student research using the neural network algorithm and its relation to KDD and its application to organizational strategic aspects is still limited. Therefore, this study aims to classify research by lecturers and students at the Faculty of Medicine, Islamic University of Bandung using the Artificial Neural Network algorithm to evaluate the suitability of the selection of research topics with the research roadmap and leading research institutions.

METHOD

In this study, the Orange version 3.34.1 application was used with the Artificial Neural Network algorithm through the following research stages:

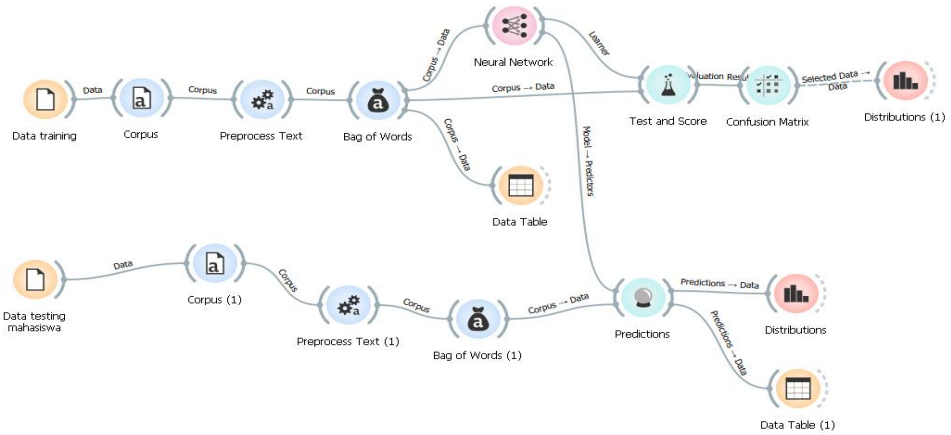

Figure 1. Stages of Research Using the Orange Application

Before dataset analysis is carried out, initial processing of text data is carried out (Figure 2) which aims to remove parts or text that are not needed so that optimal data quality is obtained for the next stage.

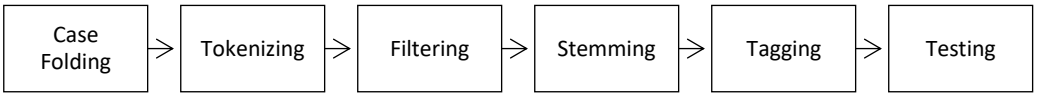

Figure 2. Preprocessing Flow in Research

1) Folding Case: Change the letters in the document to lowercase and eliminate all punctuation marks other than letters. Only letters of the alphabet are retained, while other characters are eliminated and treated as delimiters.
2) Tokenizing: Tokenization is an attempt to decompose text data into smaller structures (tokens) such as words and phrases.
3) Filtering (Delete Stopwords): Elimination of punctuation marks and removal of Indonesian language stopwords and stopwords related to research design can confuse the text mining process such as the words "influence", "relationship", "description" and so on as well as words indicating the location of the research. This is based on initial training which is confused by words related to design and research location which lowers the level of accuracy, precision, and recall of the training dataset. This process aims to eliminate words that are not useful or have no effect on the mining process.
4) Stemming and Tagging: Tagging is the search for the initial form/each past word or stemming word. Because Indonesian does not have a standard and past form like English, this stage is not carried out.
5) Analyzing (Vector Representation For Text): The analyzing stage is the stage of determining how far the connectedness is between the words in the existing document. After the tokenization and filtering process, the next step is feature weighting. The TF/IDF algorithm is known to be quite efficient and accurate for calculating the weight of word terminology in information retrieval. The mathematical formula used to calculate the weight (W) of each document against keywords is:

$$W\ a, t = tf\ dt * IDF$$

With:

W = document weight - n

d = document

t = keyword

tf = terms frequency, accquired from: $\frac{amount\ of\ word\ (t)\ in\ a\ document}{total\ of\ all\ words\ in\ a\ document}$

IDF = inverse document frequency, from the formula: $\log_2 d/df$

To evaluate the performance of the classification model, the Confusion Matrix method is used which is based on the predictive accuracy of a model. To get the value of predictive accuracy, it is necessary to calculate the amount of data that is correctly predicted and the amount of data that is predicted incorrectly. The results of these calculations are tabulated into the Confusion Matrix table which divides the classification results into the following categories:

1) True Positive (TP): classification results related to the correct category.
2) False Positive (FP): classification results related to the wrong category.
3) True Negative (TN): the classification results are not related to the correct category.
4) False Negative (FN): the classification result is not related to the wrong category

The confusion matrix data is then used for performance evaluation to evaluate classifier performance and at least consists of recall, precision, and accuracy. A recall is a positive data set that is correctly classified as positive data. Precision is a data set classified as true positive. Accuracy is the accuracy of data classification determined from training data with the following equation:

1) Precision = TP/ (TP + FP)
2) Recall = TP/ (TP + FN)
3) Accuracy = (TN+TP) / (TN + FN + TP + FP)

RESULT AND DISCUSSION

In this study, primary qualitative data were obtained from the document master book of research registers for lecturers and students. The data used is in the form of lecturer and student research title text. The data that will be used as samples in this study are 487 titles of training data from 12 research roadmap topics of the Faculty of Medicine Unisba from 2019 to 2021 and training datasets of 577 research data from 2015 to 2018. It is not possible to conduct training with the same number of datasets on all -12 research topics due to variations in existing research titles.

Table 3. Sample Dataset Based on Research Topics

No	Research Topics	Dataset Amount
1	Aging	12
2	Herbal medicine	91
3	Islamic insert	27
4	Cancer	21
5	Nutrition	20
6	Industrial Health	82
7	Communicable disease	36
8	Non-communicable disease	118
9	Reproduction and development	31
10	Stunting	14
11	Tuberculosis	27
12	Vaccine and immunization	8
	Total	487

The learning rate value is determined by testing various Learning Rate values. In this study, experiments were carried out with learning rates of 0.001, 0.002, 0.003, 0.01, 0.02, and 0.03. The optimal LR parameter is obtained at 0.001. The model with the highest accuracy value from the results of training and testing is obtained with the following parameters:

1) Hidden layers: 50
2) Alpha: 0.001
3) Max iterations: 1000

4) Replicable training: True

Furthermore, the model was tested using the confusion matrix table obtained from the neural network model itself for 487 data records with the results depicted in Table 4.

Table 4. Confusion Matrix

Research Topics	A	HM	II	C	N	IH	CD	NCD	RD	S	T	VI	∑
A	6	1	0	0	0	1	0	3	1	0	0	0	12
HM	0	86	0	0	0	0	1	4	0	0	0	0	91
II	1	2	19	1	0	1	0	3	0	0	0	0	27
C	0	1	0	11	1	0	0	8	0	0	0	0	21
N	0	2	0	0	7	2	0	8	0	1	0	0	20
IH	0	1	0	0	1	73	0	7	0	0	0	0	82
CD	0	2	1	0	0	2	9	15	3	1	3	0	36
NCD	0	4	1	1	2	8	3	97	1	0	1	0	118
RD	0	0	1	0	0	2	1	11	15	1	0	0	31
S	0	0	0	0	0	0	2	3	1	8	0	0	14
T	0	0	0	0	0	0	0	3	0	0	24	0	27
VI	0	0	0	0	0	0	0	0	1	0	0	7	8
∑	7	99	22	13	11	89	16	162	22	11	28	7	487

A: Aging
HM: Herbal Medicine
II: Islamic Insert
C: Cancer
N: Nutrition
T: Tuberculosis
IH: Industrial Health
CD: Communicable Disease
NCD: Non-communicable Disease
RD: Reproduction and Development
S: Stunting
VI: Vaccine and Immunization

The results of calculating accuracy, precision, and recall, the results are shown in Table 5 below:

Table 5. Results of the Performance Evaluation of the Text Mining Algorithm

Research Topics	Accuracy	Precision	Recall
Average of the Whole Class	74,3%	74,7%	74,3%
Aging	98,6%	85,7%	50,0%
Herbal medicine*	96,3%	86,9%	94,5%
Islamic insert	97,7%	86,4%	70,4%
Cancer	98%	84,6%	52,4%
Nutrition	96,5%	63,6%	35,0%
Industrial health*	94,9%	82,0%	89,0%
Communicable disease*	93,0%	56,2%	25,0%

Non-communicable disease *	82,3%	59,9%	82,2%
Reproduction and development	95,3%	68,2%	48,4%
Stunting	98,2%	72,7%	57,1%
Tuberculosis *	98,6%	85,7%	88,9%
Vaccine and immunization*	99,8%	100,0%	87,5%

Note: * the values for accuracy, precision, and recall are above the average for all classes

The classification results using the neural network algorithm show that overall from the training data, the average Precision is 74.7%, Recall is 74.3%, and the level of accuracy is 74.3%. The highest precision and accuracy values are in the vaccine research topic dataset while the highest recall value is in the herbal medicine topic dataset. The lowest precision and accuracy values are found in the non-communicable disease dataset, while the lowest recall values are in the infectious disease topic dataset. Datasets with high values of accuracy, precision, and recall are on the topics of herbal medicine, Islamic inserts, industrial health, tuberculosis, and vaccinations. The results of the training dataset visualization can be seen in Figure 3.

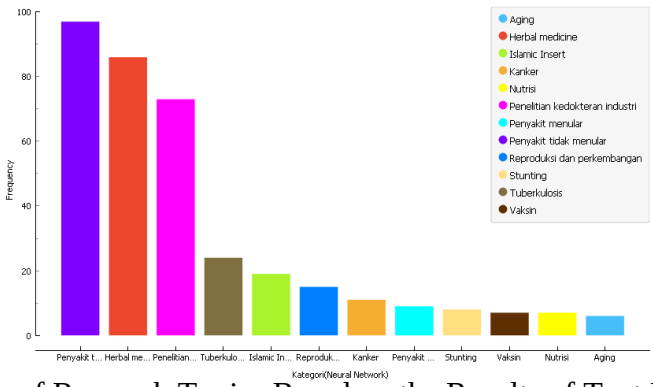

Figure 3. Distribution of Research Topics Based on the Results of Text Mining in the 2019 to 2021 Research Training Dataset

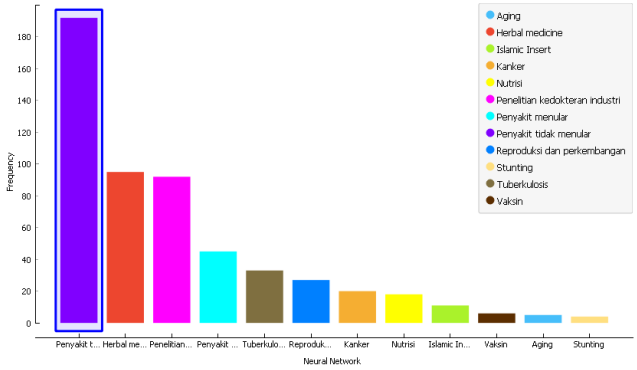

Figure 4. Distribution of Research Topics Based on the Results of Text Mining on Research Testing Datasets From 2015 to 2018

The total data for research titles in 4 (four) years starting from 2015 to 2018 amounted to 577 records, which can be visualized based on research topics based on the classification of text mining results which can be shown in Figure 4. From this visualization, it is illustrated that the pattern of text mining results has shown a pattern that is almost the same as the visualization of the training dataset. The use of two different periodizations between the training dataset and the testing dataset emphasizes that this text-mining process can be carried out continuously to observe patterns as the purpose of text mining is to look for patterns in a large amount of big data (Davis et al., 2007; Goyal & Vohra, 2012; Menasalvas & Wasilewska, 2016; Suntoro, 2019).

In addition, an analysis of the implementation of leading research can also be carried out. Based on the research roadmap from two periods of the four-year roadmap according to the periodization of the institution's strategic plan, it can be seen that herbal medicine, industrial health, Islamic insertion, tuberculosis, vaccines, and stunting have increased. The strategic value of implementing continuous text mining can increase the accuracy of decision-making. This is because the data discovery process in knowledge management is carried out automatically and can even be integrated into a decision support system (Abu-Oda & El-Halees, 2015; Sudhana, 2022; Targowski, 2005).

CONCLUSION

The results of research conducted from the initial stage to the testing stage using Data Mining using the neural network algorithm show that overall from the training data, the average Precision is 74.7%, Recall is 74.3%, and the accuracy level is 74.3 %. The highest precision and accuracy values are found in the vaccine research topic dataset while the highest recall values are in the herbal medicine topic dataset. The lowest precision and accuracy values are found in the non-communicable disease dataset, while the lowest recall values are in the infectious disease topic dataset. There is an increasing trend in the proportion of research on five topics of institutional excellence, namely herbal medicine, industrial health, Islamic inserts, tuberculosis, and vaccines. The use of data mining for prediction and visualization can assist in evaluating the suitability of the research roadmap and provide recommendations for better research implementation policies at the Faculty of Medicine, Islamic University of Bandung. The use of data mining in organizational business processes will help make quick and accurate decisions but again depends on the accuracy of selecting methods, models, and various parameters used in the data mining process.

REFERENCE

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education : University student dropout case study. *International Journal of Data Mining & Knowledge Management Process, 5*(1), 15–27. https://doi.org/10.5121/ijdkp.2015.5102

Davis, C. M., Hardin, J. M., Bohannon, T., & Oglesby, J. (2007). *Data mining methods and applications* (K. D. Lawrence, S. Kudyba, & R. K. Klimberg, Eds.). Auerbach Publications. https://doi.org/10.1201/b15783

Faisal, M. R., Kartini, D., Arrahimi, A. R., & Oglesby, J. (2007). *Belajar data science: Text mining untuk pemula*. Scripta Cendekia.

Findawati, Y., & Rosid, M. A. (2020). *Buku ajar text mining* (R. Dijaya, Ed.). UMSIDA Press.

Goyal, M., & Vohra, R. (2012). Applications of data mining in higher education. *International Journal of Computer Science Issues (IJCSI)*, *9*(2).

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1). https://doi.org/10.3390/bdcc4010001

Menasalvas, E., & Wasilewska, A. (2016). Data mining as generalization: A formal model. In *Foundations and Novel Approaches in Data Mining* (pp. 99–126). Springer-Verlag. https://doi.org/10.1007/11539827_6

Muslihudin, M., & Zahrotun, L. (2017). Perancangan text mining pengelompokkan penelitian dosen menggunakan metode Shared Nearest Neighbor dengan Euclidean Similarity. *Prosiding SNATIF*, 849–855.

Pramadhani, A. E., & Setiadi, T. (2014). Penerapan data mining untuk klasifikasi prediksi penyakit ISPA (Infeksi Saluran Pernapasan Akut) dengan algoritma Decision Tree (ID3). *Jurnal Sarjana Teknik Informatika*, *2*(1).

Purwati, N., Nurlistiani, R., & Devinsen, O. (2020). Data mining dengan algoritma Neural Network dan visualisasi data untuk prediksi kelulusan mahasiswa. *Jurnal Informatika*, *20*(2), 156–163. https://doi.org/10.30873/ji.v20i2.2273

Sudhana, S. (2022). Implementasi konsep DIKW pada penggunaaan Microsoft Dynamics 365 CRM terhadap kinerja perusahaan (Studi kasus pada LKP LCS di Surabaya). *Jurnal Ilmu Siber*, *1*(3).

Suntoro, J. (2019). *DATA MINING: Algoritma dan implementasi dengan pemrograman PHP*. Elex Media Komputindo.

Targowski, A. (2005). From data to wisdom. *Dialogue and Universalism*, *15*(5), 55–71. https://doi.org/10.5840/du2005155/629

Wibowo, A. P., Saifudin, A., & Darmawan, A. S. (2022). Naïve Bayes, Neural Network dan K-Nearest Neighbor untuk klasifikasi topik tugas akhir. *Smart Comp: Jurnalnya Orang Pintar Komputer*, *11*(4).

Yuliana, D., Purwanto, & Supriyanto, C. (2019). Klasifikasi teks pengaduan masyarakat dengan menggunakan algoritma Neural Network. *Jurnal KomtekInfo*, *5*(3), 92–116. https://doi.org/10.35134/komtekinfo.v5i3.35