DESIDOC Journal of Library & Information Technology, Vol. 42, No. 4, July 2022, pp. 218-226, DOI : 10.14429/djlit.42.4.17733

Automated Multi-Label Classification on Fertilizer-Themed Patent Documents in Indonesia

Aris Yaman[#,*], Bagus Sartono[$], Agus M. Soleh[$], Ariani Indrawati[#] and Yulia Aris Kartika[#]

[#]National Research, and Innovation Agency (BRIN), Indonesia
[$]Department Statistics and Data Science at IPB University, Indonesia
[*]E-mail: arisyaman@apps.ipb.ac.id

ABSTRACT

Patent literature research has a high scientific value for the industrial, commercial, legal, and policymaking communities. Therefore, patent analysis has become crucial. Patent topic classification is an important process in patent topic modeling analysis. However, the classification process is time-consuming and expensive, as it is usually carried out manually by an expert. Moreover, a patent document may be categorised in more than one category or label, further complicating the task. As the number of patent documents submitted increases, creating an automated patent classification system that yields accurate results becomes increasingly critical. Therefore, in this paper, we analyse the performance of two algorithms with regard to multi-label classification in patent documents: multi-label k-nearest neighbor (ML-KNN) and classifier chain k-nearest neighbor (CC-KNN), combined with latent Dirichlet allocation (LDA). These two methods have a considerable advantage in handling the continuously updated dataset; they also exhibit superior performance compared to other multi-label learning algorithms. This study also compares these two algorithms with the term frequency (TF)-weighting measure. The optimal value obtained is based on the following evaluation parameters: micro F1, accuracy, Hamming loss, and one error. The result shows that the ML-KNN method is better than the CC-KNN method and that the multi-label classification based on topics (patent LDA) is better than the TF-weighting technique.

Keywords: Topic modeling; Multi-label classification; Patent document; LDA; ML-KNN; CC-KNN

1. INTRODUCTION

Patent rights are a type of intellectual property rights (IPR), which are exclusive rights granted to innovators in the field of technology for a set period to carry out the innovation themselves or give permission to others for the same[1]. Patenting innovations has several advantages, including strengthening the market position and competitive advantage, increasing return on investment or profit, generating additional income from licensing, gaining access to new markets and technology through cross-licensing, reducing the risk of illegal imitators, enhancing the ability to raise funds and obtain grants, and boosting the public impression of a company[2]. Therefore, patent analysis has become crucial. Patent literature research can reveal important technical details and connections, explain business patterns, offer innovative industrial solutions, and help investors make important investment decisions[3–5]. Generally, patent-analysis experts are required to have a specific level of experience in a variety of research topics. Unfortunately, the rapid growth of patents in both quantity and quality has led to an increase in the workload of patent experts. Consequently, efficiency and consistency in analyzing patent documents have decreased[6]. One of the crucial processes in patent literature analysis is patent topic classification, in which patents covering similar topics or technological areas are grouped. Thus, developing an automated classification system for patent documents has become extremely important, as it will help both inventors and patent-analysis experts identify patents on similar topics[7].

However, developing an accurate automated patent document classification system is quite difficult for various reasons. First, the International Patent Classification (IPC) system is complicated, with a hierarchical structure and several labels[8-9]. Second, patent documents' complexity poses a concern; patent documents are complicated and typically contain extensive jargon or new technical terms resulting from technological advances[10]. Third, as knowledge and technology evolve over time, a patent documents may have several categories, and so we must simultaneously categorise a patent document into many labels, which is referred to as multi-label classification.

Unfortunately, the majority of the classification problems investigated in machine learning, especially in patent topic classification modeling, are single-label classification problems[11]. Multi-label classification differs from binary and multi-class classification in that it is more difficult to learn; in multi-label classification, one must classify an object into more than one label simultaneously[11-12]. There are at least two commonly methods to overcome difficulties in multi-label classification: the problem transformation method and

Received : 01 December 2021, Revised : 04 April 2022
Accepted : 05 May 2022, Online published : 19 July 2022

the adoption algorithm method. The problem transformation method solves multi-label problems by converting them to single-label problems. Meanwhile, the adoption algorithm solves the issue by applying algorithms to single-label instances that are relevant to multi-label problems[13]. Based on this, all single-label classification methods may be applied with the problem transformation approach. However, in multi-label classification, labels are sometimes correlated; consequently, using the problem transformation method with the binary relevance technique becomes problematic. To overcome problem issues, the classifier chain k-nearest neighbor (CC-KNN)[14] and the multi-label k-nearest neighbor (ML-KNN)[15] approaches attempt to handle the multi-label problem by focusing on the correlation between labels. Continuously updated patent documents with new entries comprise the primary reason for employing both methods in this study, as both methods have an advantage when handling several problems and deal well with simultaneous changes in the problem domain[16] as well as the correlation between labels. Furthermore, textual data processing is often challenging due to its unstructured nature. Therefore, the distribution of explanatory variable terms must be transformed into topic distribution. One of the most popular topic modeling methods is patent latent Dirichlet allocation (LDA), which can provide attribute meaning to the explanatory variables[17].

This paper aims to analyse the performance of ML-KNN and CC-KNN with LDA in classifying patent documents into multi-labels. It also compares the topic modeling-based approach and term frequency (TF)-weighting measure. The evaluation is based on the micro F1, accuracy, Hamming loss, and one error values.

2. RELATED WORKS

Various studies on modeling topics in patent documents have contributed significantly to developing knowledge in this field, especially in patent modeling. Until recently, they were modeled on patent documents developed in both directions. The first development involved modeling based on patent representation models[18–20]. Meanwhile, other developments were concerned with a patent technology classification model[6-7,21].

2.1 Patent Document Classification Modeling

Several studies have explored automatic patent classification systems. Wu et al. suggested a hybrid genetic-based support vector machine model (HGA-SVM) for automatic patent classification and tested the results with patents related to semiconductor equipment technology[6]. Chen and Chang used a three-phase model to classify WIPO patent data[7]. Lee and Hsiang tried to apply the BERT language model to classify patent data[21]; this study used the data in the “claim” section of patent documents as a corpus (a collection of several papers), which will be further analysed in patent classification.

2.2 Patent Representation Modeling (Patent Topic Modeling)

Classification of patents using automated technology is effective in various technological innovation activities, including checking, detecting, and reducing the possibility of patent infringement. Previous studies have focused predominantly on classifying patents rather than on patent representation[6-7,21]. Document representation is essential because it determines the characteristics of a patent; it also addresses the contents of a patent and is represented effectively in a structured form[17].

Research on patent document representation models (topical modeling in patents) has been carried out before. Chen, *et al.* introduced the patent LDA model for analyzing root topics in a patent document[18]; this study claimed that the patent technique of LDA can provide a better indicator of the perplexity value than the conventional LDA method. Hu, *et al.* introduced the hierarchical feature extraction model (HFEM) method[19]. They compared it with three other models: the single neural network model (CNN), long–short-term memory (LSTM), and BiLSTM. The study results showed that the introduced model had a better performance in terms of precision and recall when modeling topics in a patent document. Kim et al. used a Word2vec-based latent semantic analysis (W2V-LSA) model for modeling patent topics[20]. This method can be

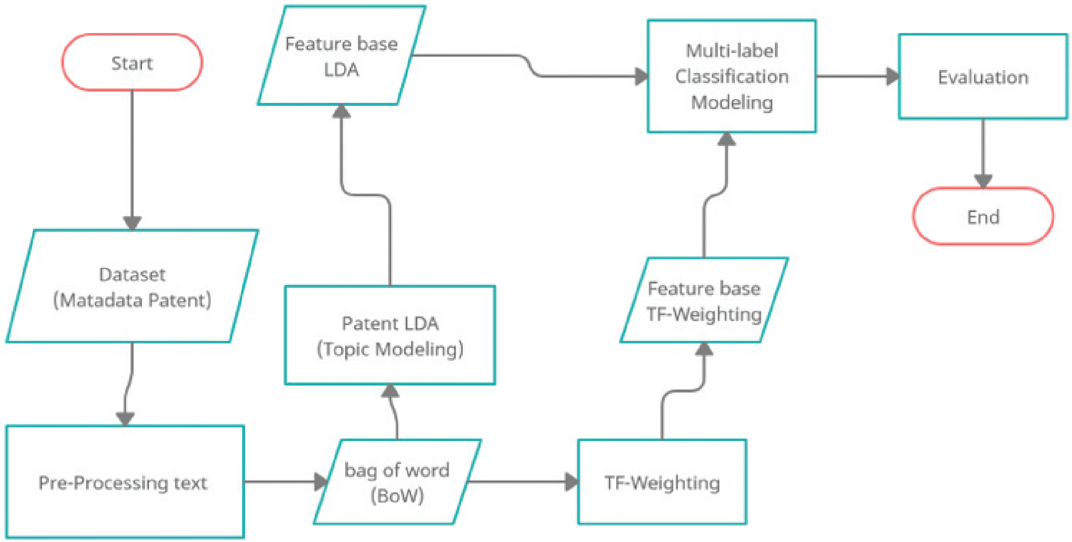

Start
Feature base LDA
Multi-label Classification Modeling
Evaluation
End
Dataset (Matadata Patent)
Patent LDA (Topic Modeling)
Feature base TF-Weighting
Pre-Processing text
bag of word (BoW)
TF-Weighting

Figure 1. Research methodology.

Table 1. Sample of raw data

Title	Abstract	Categories technology
PEMBUATAN INHIBITOR NITRRIFIKASI BERBAHAN LATEKS-CHITOSAN PADA PRODUK PUPUK	Dalam invensi ini inhibitor nitrifikasi berbahan lateks-chitosan digunakan untuk melapisi produk pupuk yang berfungsi sebagai penapis pelepasan kandungan pupuk NPK, sehingga dalam pelepasan kandungan pupuk lebih terkendali dan lebih efisien dibandingkan dengan pupuk tanpa penapis. Inhibitor nitrifikasi berbahan lateks-chitosan bersifat biodegradable atau ramah Iingkungan. Tahapan proses pembuatan inhibitor nitrifikasi berbahan lateks-chitosan antara lain meliputi; formulasi yaitu pembuatan inhibitor nitrifikasi berbahan lateks-chitosan dilakukan dengan mencampurkan lateks cairan dan chitosan cairan yang pada saat kering terbentuk komposit dengan perbandingan lateks-chitosan bervariasi dari 20:80 sampai dengan 80:20 (contoh ; 30:70, 40:60, 50:50, 60:40, dan seterusnya), tahapan pelapisan pada permukaan granul pupuk dengan metode spraying, dan tahapan pengeringan dengan hembusan udara panas membentuk struktur penapis menjangrkau seluruh permukaan pupuk NPK.	B01 C05
KONVERSI GAS ASAM KE PUPUK BERBASIS SULFAT ATAU FOSFAT	Suatu metode dijelaskan untuk membuat pupuk berbasis sulfat dan fosfat dari hidrogen sulfida. Metode tersebut meliputi mengumpan suatu aliran yang mengandung suatu hidrogen sulfida dan udara dengan volum yang besar ke tungku, di mana aliran tersebut dibakar untuk menghasilkan suatu aliran gas yang kaya akan sulfur dioksida. Aliran gas yang kaya akan sulfur dioksida tersebut kemudian diumpan ke suatu reaktor untuk menghasilkan suatu aliran asam sulfat dan suatu aliran limbah yang mengandung karbon dioksida, nitrogen, oksigen, pengotor-pengotor sisa dan sejumlah sulfur dioksida sisa yang tidak bereaksi. Aliran asam sulfat tersebut akhirnya dikonversi menjadi suatu pupuk berbasis sulfat atau fosfat.	C05
PROSES PEMBUATAN PUPUK ORGANIK GRANUL DARI LIMBAH AGAR DAN PUPUK ORGANIK YANG DIHASILKAN DARI PROSES TERSEBUT	Invensi ini berhubungan dengan proses pembuatan pupuk organik granul dari limbah agar dan pupuk organik yang dihasilkan dari proses tersebut dengan : • Mengeringkan limbah agar terlebih dahulu sampai kadar air 5-6%; • Menepungkan limbah agar ysng telah kering tersebut • Dengan alat penepung; • Mengayak dolomit dan kaptan sebagai bahan tambah; • Mencampurkan 50% berat tepung limbah agar dengan 7% berat dolomit dan 7% berat kaptan sebelum dimasukkan ke dalam granulator; • Memasukkan campuran (d) ke dalam granulator; • Menyalakan granulator dengan kecepatan 40 rpm, • Menyemprotkan 36% berat air ke dalam campuran bahan di dalam granulator sampai terbentuk butiran-butiran granul; • Mengeluarkan butiran granul yang sudah jadi dari granulator; • Mengayak granul tersebut dengan ayakan berukuran 2 dan 4 mesh untuk mendapatkan ukuran granul yang seragam; dan; • Mengeringkan butiran granul yang sudah diayak tersebut sampai kadar airnya 2-3%.	CO5

an alternative for modeling blockchain technology patents, allowing future trends and research in blockchain technology to be obtained and further investigated.

3. METHODOLOGY

Steps in this research include 5 stages, are: Retrieving dataset, pre-processing text, features extraction, classification modeling, and evaluation. Steps of patent documents classification in this study are illustrated in Fig. 1.

3.1 Dataset

The data used in this study are metadata, particularly the titles and abstracts from a collection of patent documents in Indonesia from 1945 to 2017, retrieved from the Indonesia Intellectual Property Database (https://pdki-indonesia.dgip.go.id/) with the keyword "fertiliser". Sample data can be seen in Table 1.

3.2 Text Data Preprocessing

Prior to the categorisation process, it is necessary to process the patent documents' text into a structured form. The text pre-processing phase involves a few steps: case folding, tokenizing, stopword removal, and stemming. Several algorithms have been developed for pre-processing text, especially Indonesian; these are the Nazief and Adriani algorithm[22], Yusof and Sembok algorithm[23], Arifin and Setiani algorithm[24], and Vega Ahmad algorithm[24]. The Nazief and Adriani algorithm is better than the other three algorithms[24]. Therefore, in this study, the Nazief and Adriani algorithm is utilised in the preprocessing stage, with the NLTK package ("Indonesian" stopword) and Sastrawi for Stemming the word into root word.

3.3 Topic Modeling: Latent Dirichlet Allocation

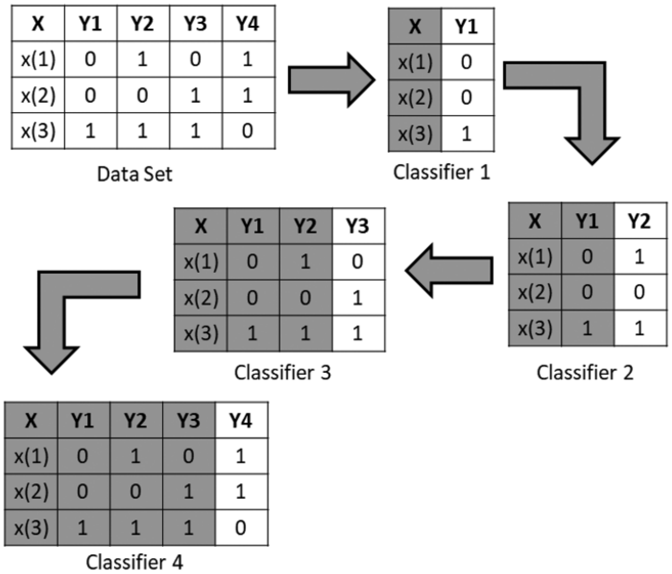
The Latent Dirichlet Allocation (LDA) algorithm provides an output list of the weighted topics for each document[25]. LDA is one of the most popular topic modeling methods[26]. A topic consists of a particular set of words that constitute the topic, and one document may consist of several topics, each with a distinct probability. Topic modeling aims to find a word or group of words in a document that represents a topic. Accordingly, topic modeling is defined as a series of algorithms that can find the central theme of a large, unstructured text[27]. This stage produces the "bag of the topic (BoT)" output of each document, which is used in LDA topic modeling. There are four distinct dimensions for determining the level of topic coherence that can be combined: segmentation, probability calculation, confirmation measure, and aggregation function[28]. To interpret the topics generated by LDA and facilitate the comprehension of their meanings, the participation of experts is required[29]. The saliency parameter can determine the most optimal word order, making interpretation easier for experts[30]. The relevance size can also help in determining the words or terms that will allow experts to interpret the issue more easily[31].

3.4 Term Frequency

As opposed to LDA, TF involves term weighting to assess the value of each term to the document. It represents each document as a vector of word frequencies[32]. In this study, the Term Frequency (TF) method is compared with the BoT method in multi-label classification.

3.5 Multi-label Classification

The multi-label classification methods applied in this study are ML-KNN and CC-KNN. ML-KNN is included in the lazy learning algorithm that adopts the conventional KNN algorithm to identify multi-label classification cases. ML-KNN will first identify multi-label membership in the training data and then use an algorithm adopted using the maximum a posteriori (MAP) principle to determine the labels of a patent. This MAP principle is applied to the Bayes probability, requiring the prior and posterior probability values obtained from the training data[15,33]. CC-KNN has the same learning algorithm foundation as ML-KNN. The fundamental difference between the two lies in how each method places the correlation between labels. ML-KNN considers the correlation between labels by changing the algorithm on conventional KNN, whereas CC-KNN considers the correlation between labels by conducting the KNN procedure sequentially (Fig. 2). Both CC-KNN and ML-KNN use the problem transformation approach. Transformation problems can be integrated easily with various algorithms on a single label[34]. The sequencing method for performing a single classification is a notable aspect of the CC-KNN process. In a single classification technique, different ordering can result in different projected label performance[35]; the classification order used in this paper is based on the number of labels on the response variable.



X	Y1	Y2	Y3	Y4
x(1)	0	1	0	1
x(2)	0	0	1	1
x(3)	1	1	1	0

Data Set

X	Y1
x(1)	0
x(2)	0
x(3)	1

Classifier 1

X	Y1	Y2
x(1)	0	1
x(2)	0	0
x(3)	1	1

Classifier 2

X	Y1	Y2	Y3
x(1)	0	1	0
x(2)	0	0	1
x(3)	1	1	1

Classifier 3

X	Y1	Y2	Y3	Y4
x(1)	0	1	0	1
x(2)	0	0	1	1
x(3)	1	1	1	0

Classifier 4

Figure 2. Classifier chain method illustration 34.

3.6 Evaluation

In this study, four evaluation parameters are used: micro F1, accuracy, Hamming loss, and one error. Parameter accuracy is the proportion of correctly predicted labels to the total number of labels in the predicted label set and the truth label of an instance[33]. Micro F1 can be effectively used in the case of multi-label classification, considering that, in this study, some labels appear very rarely compared to others[36]. One error evaluates the number of times the top-ranking label is not in the proper label set of the instance (label in the test data)[37]. Hamming loss estimates the number of times the label–instance pair (in the test data) is misclassified[38]. The smaller the values of the Hamming loss and one error parameters, the better the model built. Furthermore, the higher the values of the accuracy and micro F1 parameters, the better the model built.

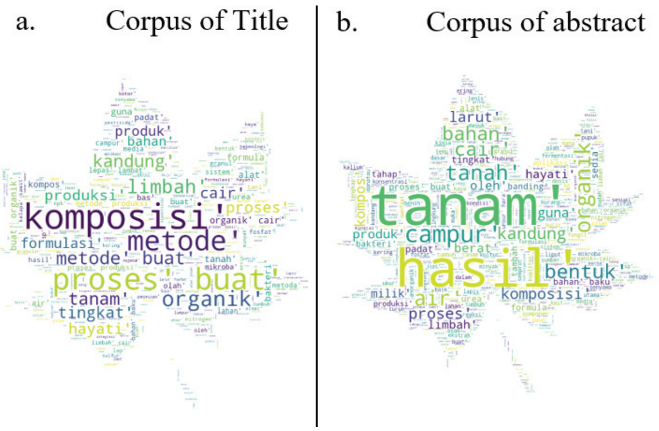


Figure 3. Distribution of words.

(Source: Primary Data Processed, 2021)

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

This study's description of word distribution uses the word cloud technique, Fig. 3 shows the distribution of the words contained in the titles and abstracts, respectively, of the patent corpus.

Figure 3a can be interpreted as follows: most of the documents (especially the patent titles) discuss fertiliser technology inventions related to composition, manufacturing methods and processes, and biological organic fertilisers and their formulations. According to Fig. 3b, most of the patent documents (in the abstracts) discuss fertiliser technology inventions, the impact on plants, mixed processes for making fertilisers, organic fertilisers (such as compost), premium content, and the superior form of fertiliser (liquid). However, these descriptions are too general and do not describe the topic of each document. Therefore, a topic modeling approach is needed for assessing possible topics in a corpus.

4.2 Topic Coherence and Classification Evaluation

Based on the comparison charts of the number of topics against the coherence measure and the number of topics against the micro F1 measure (Fig. 4), the best coherence (largest value) is obtained on 18 topics, whereas the best number of topics based on the classification evaluation measure is obtained at 23 points (largest value of micro F1). The best number of topics obtained based on the coherence measure is not in line with the best number of topics obtained based on the classification evaluation measure (micro F1).

This study's ultimate goal is to achieve the best multi-label classification modeling; hence, determining the best number of topics should be based on the best classification evaluation value. Therefore, in this study, 23 topics were used as the basis for further modeling.

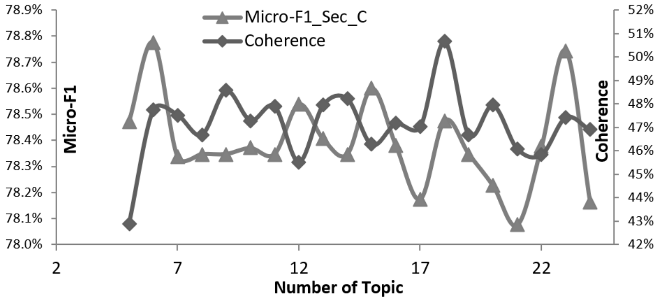


Figure 4. Topic Coherence and Micro-F1. (Source: Primary Data Processed, 2021)

4.3 Modeling Topics by Patent LDA

Topic modeling with patent LDA can produce output in the form of a collection of words for each topic. The set of words can be elaborated on, and their relationship can be translated as a specific topic. Table 2 provides information regarding the ten words representing each topic; it is presented for ease of interpretation[31]. In essence, the words extracted from the 23 selected topics by limiting them to 10 words are presented in Table 2. This word extraction is intended to facilitate the interpretation of each topic. For instance, when considering Topic 2, we can see a collection of words representing it, namely "process," "create," "biological," "organic," "bas," "formulation," "formula," "liquid," "leaves," and "chemistry." Thus, we can know that Topic 2 concerns "the process of making bio-organic fertilisers." Similarly, Topics 3 to 23 can be further elaborated (not discussed further in this study).

4.4 Multi-Label Classification Result

Figure 5 shows a comparison between the CC-KNN and ML-KNN methods, with the LDA approach. The ML-KNN model provides an optimum value of the model evaluation parameters that is better compared to the CC-KNN model. The optimum values of micro F1 and accuracy are obtained when k (the number of nearest neighbors) equals 6 (83.11 %, 81.58 %). Furthermore, Hamming loss and one error reach their optimum when the values are 0.114 and 0.184, respectively. In both cases, k equals 6.

Table 2. Words/Phrases (Terms) distribution for each topic

	term 1	*term 2*	*term 3*	*term 4*	*term 5*	*term 6*	*term 7*	*term 8*	*term 9*	*term 10*
Topik 1	Organic	liquid	base	Urine	Feed	Formula	fermentation	method	biopesticides	Full
Topik 2	process	for	biological	Organic	bass	Formulation	Formula	liquid	leaf	chemical
Topik 3	method	product	production	Baku.	ingredient	slag	steel	process	Latex	Pupu.
Topik 4	Bacteria	Bacillus	late	consortium	ash	Paddy	fast	biological	experience	reaction
Topik 5	Tani	Planting	seed	oil	stand	dirty	Lapis	wash	spread	battery
Topik 6	composition	application	control	free	agriculture	Fertilizer	Chlorida	anti	free	for
Topik 7	nitrogen	Lebih slow	mix	mold	microbial	substance	potassium	Porpor	Vermikompos	PGPR.
Topik 8	Mikoriza	active	land	SENAH	ingredient	fungi	fungus	Arbuscular	slow	biological
Topik 9	waste	liquid	Change	factory	complete	rich	skin	Ph.	installation	type
Topik 10	water	Cultivation	method	Sea	benefit	low	in accordance	food	Entomopathog	production
Topik 11	Bladder	Sulfur.	composition	block	garden	for	sour	Floating	Nitrification	zinc
Topik 12	congested	spray	grass	Granular	grow	tool	Electrical	capsule	Bibit	liquid
Topik 13	results	side	type	compound	residue	pesticide	Azolla	termite	process	product
Topik 14	compost	mud	Inhibitor	precision	waste	rotten	is lost	Hara	ingredient	expired
Topik 15	Methods	Bio.	technology	micro	salute	Bioorganic	Granule	biological	industry	production
Topik 16	Exercise	integration	mineral	down	Unit	gas	scale	separate	color	Bio.
Topik 17	Phosphor	compound	vegetable	NPK	Fill	reactor	extraction	free	Full	post
Topik 18	Urea	Lap.	charcoal	Kendara	animal	sand	blood	automatic	work	active
Topik 19	machine	tool	system	seed	Transplanter	House	Basmi.	conversion	smell	sesbania
Topik 20	Bule	phosphate	Ammonium	Granule	owned by	polymer	push	shape	nitrate	rock
Topik 21	Palmoil	combination	media	Hydrocarbon	Trichoderma.	process	Cyanobacteria	Biomass	Spiriru	Lapuk
Topik 22	Paddy	magnesium	condition	husk	salt	powder	peanut	semi-	Humum	land
Topik 23	eat	bioplastics	Salur.	continuous	system	open	applicator	plot	Cacah	Designation

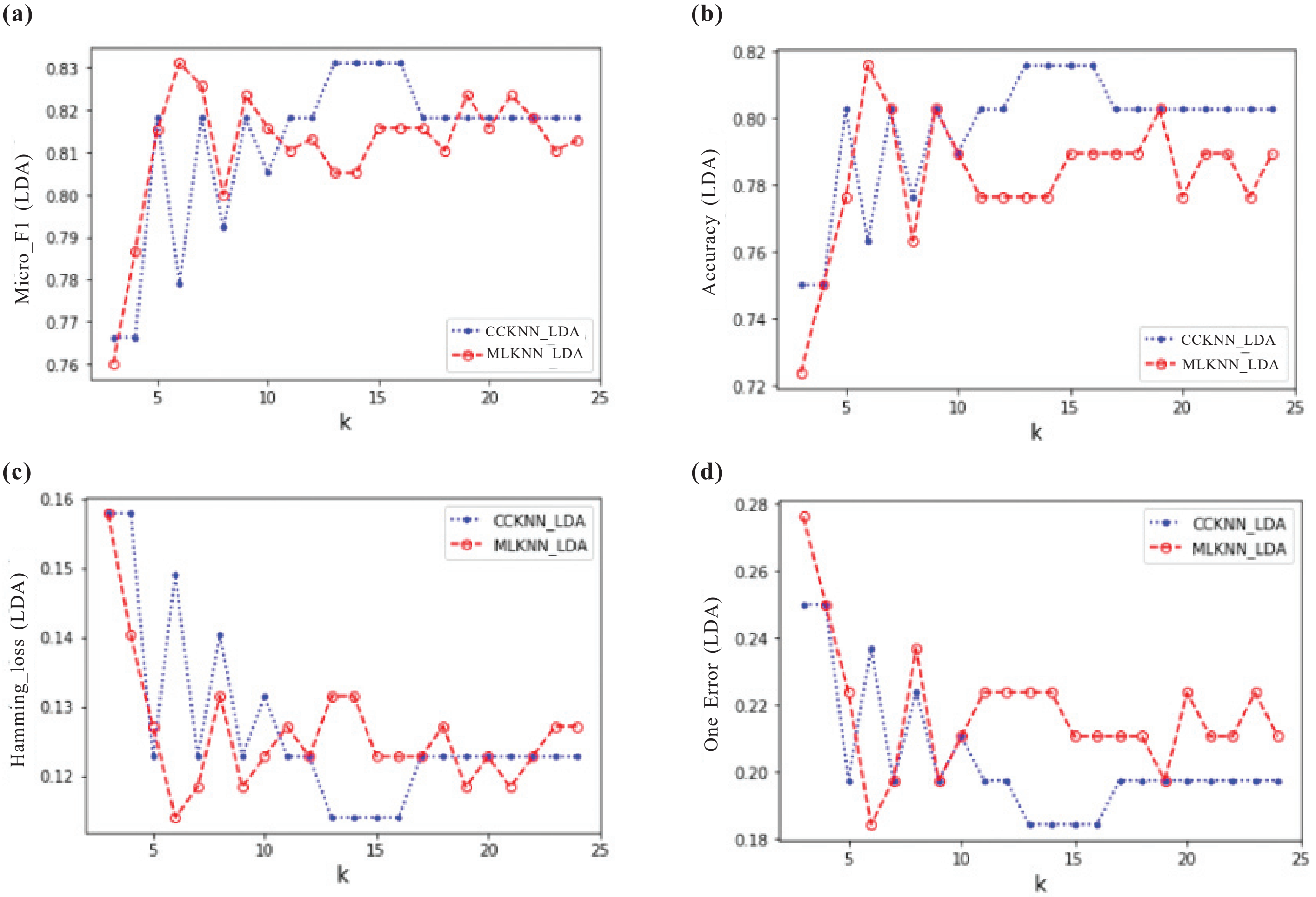


Figure 5. CC-KNN vs ML-KNN (Topic Modeling Approaches)

Comparison of CC-KNN and ML-KNN methods with topic explanatory variables, part a) comparison is seen from the evaluation parameters of Micro-F1, b) evaluation with accuracy value, c) evaluation on Hamming-loss value, and d) evaluation on One-error value.

(Source: Primary Data Processed, 2021)

Figure 6 shows a comparison between the LDA model and TF-weighting technique. On the left (parts a and b), the figure shows that the optimum values (largest) of micro F1 and accuracy are obtained when k is 6 (83.11 % and 81.58 %). Furthermore, Hamming loss and one error (parts c and d) reach their optimum when the values are 0.114 and 0.184, respectively. The optimum values of these four parameters are obtained when using the ML-KNN method with a topic approach. Hence, the ML-KNN topic approach is the best method. This method is compared with the CC-KNN method with the TF approach (the right side of Fig. 6). The optimum values of micro F1 and accuracy are obtained when k equals 6 (83.11 %, 81.58 %). Hamming loss and one error reach their optimum when the values are 0.114 and 0.184, respectively. It can be seen that the optimum values in Fig. 5 are also the optimum values in Fig. 6. Therefore, the LDA model provides a better model.

As per the results of this study, ML-KNN method is better than CC-KNN method for estimating a patent document's multi-label classification, especially for Indonesian patents with a fertiliser theme. This finding has practical implications regarding the relationship between technology labels. In addition, it is known that the number of the best closest neighbors is six. Estimating the technological classification of a patent document may depend on the six documents that have the closest similarity distance. Another interesting finding is that the topic approach is the best for modeling the classification of patent document technology. If understood further, this may facilitate an understanding of the importance of giving meaning to a document (topic representation) before assessing the technology categorisation of a patent document.

5. CONCLUSION

The ML-KNN and CC-KNN methods are quite suited to handling multi-label classification issues, as indicated by the accuracy values over 80 per cent; the ML-KNN approach provides better results than the CC-KNN one, with the best number of closest neighbors being six. Estimating the technological classification of a patent document may depend on the six documents that have the closest similarity distance.

There is no indication of a relationship between the interpretability of the topic and the performance of multi-label classification modeling. Based on the evaluation values of the best multi-label classification parameters, there are 23 topics in

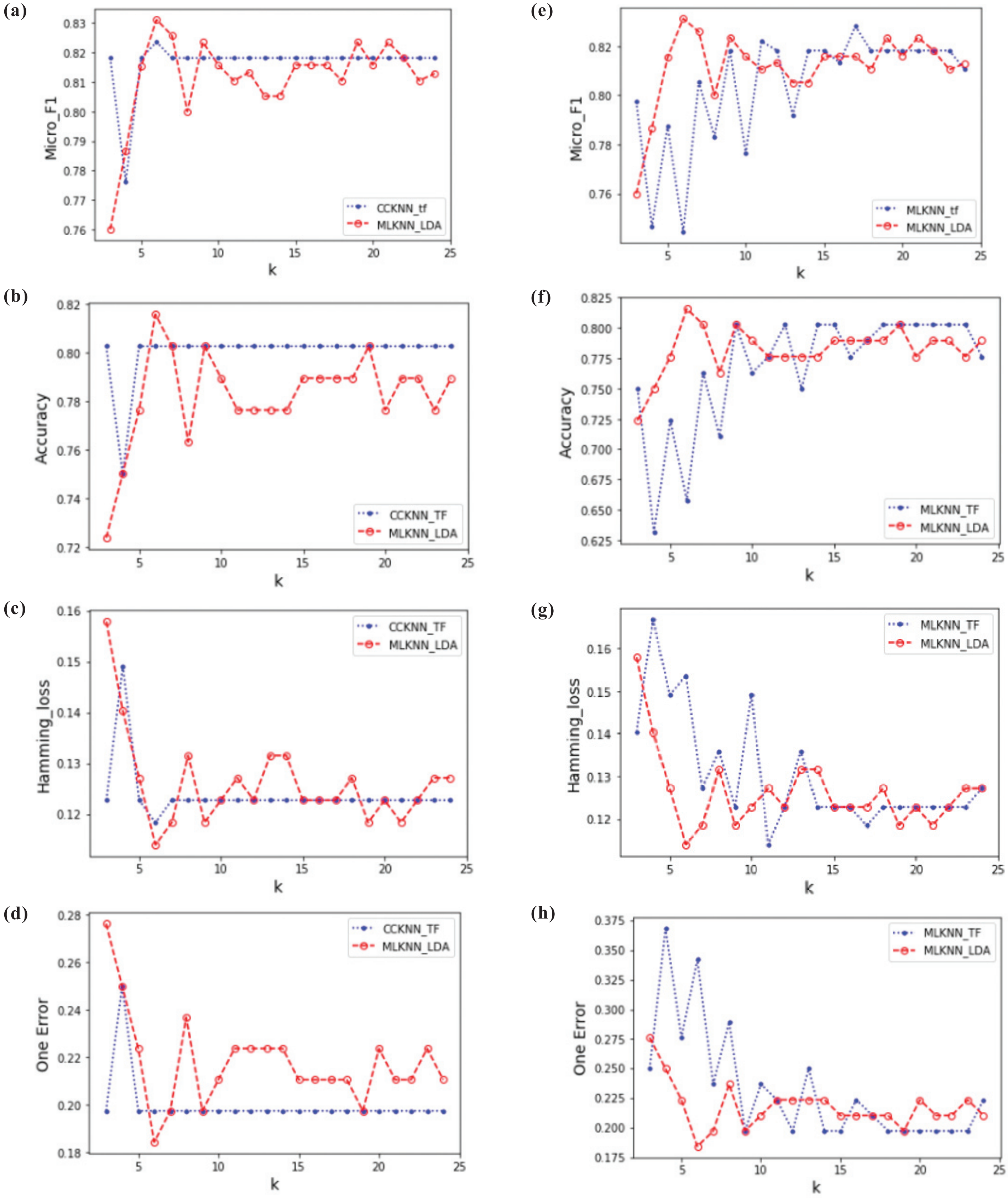


Figure 6. Comparison between LDA and TF Approach.

Figure 6 on the left (a, b, c, and d) is a comparison between ML-KNN method with topic modeling and CC-KNN with TF-weighting technique. Figure 6 on the right (e, f, g, and h) is comparison between ML-KNN method with topic modeling and ML-KNN with TF-weighting technique

(Source: Primary Data Processed, 2021)

developing fertiliser technology. When viewed on the basis of the four evaluation parameters used in this study, multi-label classification modeling with the patent LDA (topic modeling) gives better results than the TF-weighting technique. Another interesting finding is that the representation modeling of patent documents can increase the validity when determining the classification of patent documents.

Future research should explore modeling with non-lazy learning methods adapted for multi-label classification purposes. If possible, regarding data sources, general Indonesian patent data should be used to ensure the resulting conclusions are not casuistic. A specific use of vocabulary is recommended for future research.

REFERENCES

1. President of the Republic of Indonesia. Law No. 13 of 2016:Patent. https://www.hukumonline.com/pusatdata/detail/lt57dfa79196118/undang-undang-nomor-13-tahun-2016/document (2016).
   (Accessed on 12 November 2020)
2. Artz, K.W.; Norman, P.M.; Hatfield, D.E. & Cardinal, L.B. A longitudinal study of the impact of R&D, patents, and product innovation on firm performance. *J. Prod. Innov. Manag.,* 2010, **27**, 725–740
   doi: 10.1111/j.1540-5885.2010.00747.x
3. Hongshu, C.; Guangquan, Z.; Donghua, Z. & Jie, L. Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014. *Technol. Forecast. Soc. Change.,* 2017, **119**, 39–52.
   doi: 10.1016/j.techfore.2017.03.009
4. Kim, G. & Bae, J. A novel approach to forecast promising technology through patent analysis. *Technol. Forecast. Soc. Change.,* 2017, **117**, 228–237.
   doi: 10.1016/j.techfore.2016.11.023
5. Yu, X. & Zhang, B. Obtaining advantages from technology revolution: A patent roadmap for competition analysis and strategy planning. *Technol. Forecast. Soc. Change.,* 2019, **145**, 273–283.
   doi: 10.1016/j.techfore.2017.10.008
6. Wu, C.H.; Ken, Y. & Huang, T. Patent classification system using a new hybrid genetic algorithm support vector machine. *Appl. Soft Comput. J.,* 2010, **10**, 1164–1177.
   doi: 10.1016/j.asoc.2009.11.033
7. Chen, Y.L. & Chang, Y.C. A three-phase method for patent classification. *Inf. Process. Manage.,* 2012, **48**, 1017–1030.
   doi: 10.1016/j.ipm.2011.11.001
8. WIPO. Guide to the International Patent Classification. *WIPO (World Intellect. Prop. Organ.* (2018). https://www.wipo.int/publications/en/series/index.jsp?id=183 (Accessed on 10 July 2019).
9. Fall, C.J.; Törcsvári, A.; Fiévet, P. & Karetka, G. Automated categorization of German-language patent documents. *Expert Syst. Appl.,* 2004, **26**, 269–277.
   doi: 10.1016/S0957-4174(03)00141-6
10. Cao, S. Speed of patent protection, rate of technical knowledge obsolescence and optimal patent startegy: Evidence from innovation patented in the US, China and several other countries. *McKinsey Q.,* 2014, **2**, 1–22. https://are.berkeley.edu/sites/default/files/job-andidates/paper/SiweiCao_JMP121014.pdf.
    (Accessed on 5 November 2020).
11. Almeida, A.M.G.; Cerri, R.; Paraiso, E.C.; Mantovani, R.G. & Junior, S.B. Applying multi-label techniques in emotion identification of short texts. *Neurocomputing,* 2018, **320**, 35–46.
    doi: 10.1016/j.neucom.2018.08.053
12. Elujide, I.; Fashoto, S.G.; Fashoto, B.; Mbunge, E.; Folorunso, S.O. & Olamijuwon, J.O. Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Informatics. Med. Unlocked*, 2021, **23**, 100545.
    doi: 10.1016/j.imu.2021.100545
13. Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehouse. Min.,* 2009, **3**, 1–13.
    doi: 10.4018/jdwm.2007070101
14. Spolaôr, N.; Cherman, E.A.; Monard, M.C. & Lee, H.D. A comparison of multi-label feature selection methods using the problem transformation approach. *Electron. Notes Theor. Comput. Sci.,* 2013, **292**, 135–151.
    doi: 10.1016/j.entcs.2013.02.010
15. Zhang, M.L. & Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.,* 2007, **40**, 2038–2048.
    doi: 10.1016/j.patcog.2006.12.019
16. Webb, G.I. Lazy Learning BT - Encyclopedia of Machine Learning. in (eds. Sammut, C. & Webb, G. I.) 571–572 (Springer US, 2010).
    doi: 10.1007/978-0-387-30164-8_443.
17. Yun, J. & Geum, Y. Automated classification of patents: A topic modeling approach. *Comput. Ind. Eng.,* 2020, **147**.
    doi: 10.1016/j.cie.2020.106636
18. Liang, C.; Weijiao, S.; Guancan, Y.; Jing, Z. & Xiaoping, L. A topic model integrating patent classification information for patent analysis. *Wuhan Daxue Xuebao (Xinxi Kexue Ban)/Geomatics Inf. Sci. Wuhan Univ.,* 2016, **41**, 123–126. https://www.researchgate.net/publication/309530756_A_topic_model_integrating_patent_classification_information_for_patent_analysis.
    (Accessed on 23 October 2020).
19. Hu, J.; Li, S.; Hu, J. & Yang, G. A hierarchical feature extraction model for multi-label mechanical patent classification. *Sustain.,* 2018, **10**, 219.
    doi: 10.3390/su10010219
20. Suhyeon, K.; Haecheong, P. & Junghye, L. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Syst. Appl.,* 2020, **152**.
    doi: 10.1016/j.eswa.2020.113401
21. Lee, J.S. & Hsiang, J. Patent classification by fine-tuning BERT language model. *World Pat. Inf.,* 2020, **61**, 101965.
    doi: 10.1016/j.wpi.2020.101965
22. Adriani, M.; Asian, J.; Nazief, B.; Williams, H.E. & Tahaghoghi, S.M.M. Stemming Indonesian : A confix-

stripping approach. *Conf. Res. Pract. Inf. Technol. Ser.,* 2005, **38**, 307–314.
doi: 10.1145/1316457.1316459

23. Ahmad, F.; Yusoff, M. & Sembok, T.M.T. Experiments with a stemming algorithm for Malay words. *J. Am. Soc. Inf. Sci.,* 1996, **47**, 909–918.
24. Asian, J.; Williams, H.E. & Tahaghoghi, S.M.M. Stemming Indonesian. *Conf. Res. Pract. Inf. Technol. Ser.*, 2005 **38**, 307–314.
doi: 10.1145/1316457.1316459
25. Campbell, J.C.; Hindle, A. & Stroulia, E. Latent dirichlet allocation: Extracting topics from software engineering data. *Art Sci. Anal. Softw. Data,* 2015, 139–159.
doi:10.1016/B978-0-12-411519-4.00006-9.
26. Vayansky, I. & Kumar, S.A.P. A review of topic modeling methods. *Inf. Syst.,* 2020, **94**.
doi: 10.1016/j.is.2020.101582
27. Blei, D.; Carin, L. & Dunson, D. Probabilistic topic models. *Commun. Acm.,* 2012, **27**, 55–65.
doi: 10.1109/MSP.2010.938079
28. Röder, M.; Both, A. & Hinneburg, A. exploring the space of topic coherence measures. WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min., 2015, 399–408.
doi:10.1145/2684822.2685324.
29. Momeni, A. & Rost, K. Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. *Technol. Forecast. Soc. Change,* 2016, **104**, 16–29.
doi: 10.1016/j.techfore.2015.12.003
30. Chuang, J.; Manning, C.D. & Heer, J. Termite: Visualization techniques for assessing textual topic models *In* Proceedings of the Workshop on Advanced Visual Interfaces AVI 74–77, 2012.
doi:10.1145/2254556.2254572.
31. Sievert, C. & Shirley, K. LDAvis: A method for visualizing and interpreting topics. *In* Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014, 63–70.
doi:10.3115/v1/w14-3110.
32. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval._7896.pdf., 1988.
doi: 10.1016/0306-4573(88)90021-0
33. Cherman, E.A.; Spolaôr, N.; Valverde-Rebaza, J. & Monard, M.C. Lazy multi-label learning algorithms based on mutuality strategies. *J. Intell. Robot. Syst. Theory Appl.*, 2015, **80**, 261–276.
doi: 10.1007/s10846-014-0144-4
34. Read, J.; Pfahringer, B.; Holmes, G. & Frank, E. Classifier chains for multi-label classification. *Mach. Learn.,* 2011, **85**, 333–359.
doi: 10.1007/978-3-642-04174-7_17
35. Gustafsson, R. & Gustafsson, R. Ordering classifier chains using filter model feature selection techniques, 2017, 8–15. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1118469.
(Accessed on 11 August 2021).
36. Manning, C.D.; Raghavan, P. & Schütze, H. Introduction to modern information retrieval (2nd edition). Library Review vol. 53 (Cambridge University Press, 2004).
doi: 10.1108/00242530410565256
37. Younes, Z. & Abdallah, F. Multi-label classification algorithm derived from K-Nearest neighbor rule with label dependencies. *In* 2008 16th European Signal Processing Conference (IEEE, 2015). https://ieeexplore.ieee.org/document/7080359. (Accessed on 22 January 2021).
38. Díez, J.; Luaces, O.; del Coz, J.J. & Bahamonde, A. Optimizing different loss functions in multilabel classifications. *Prog. Artif. Intell.,* 2015, **3**, 107–118.
doi: 10.1007/s13748-014-0060-7

CONTRIBUTORS

Mr Aris Yaman received his Master's degree in Statistics and Data Science from the Bogor Agricultural University, Indonesia. He is currently a researcher at National Research, and Innovation Agency (BRIN), Indonesia. His research interests include: Machine learning, data mining, and data science.
He contributed in conceptualising the present study, collection of related literature, methodology, and data analysis.

Dr Bagus Sartono obtained PhD in Applied Economics (Statistics) from University of Antwerp, Belgian. He is currently Lecturer at Department Statistics and Data Science at IPB University, Indonesia. His research interests are: Data analysis, machine learning and data science.
He contributed to conceptualizing the present study, methodology, and data analysis adviser.

Dr Agus M. Soleh received his PhD in Statistics from Bogor Agriculture University, Indonesia. He currently works at the Department of Statistics, IPB University. Agus does research in Statistics.
He contributed to computational methods, and evaluation models for this study.

Ms Ariani Indrawati received his Master's degree in Informatics Engineering from Bina Nusantara University, Indonesia in 2016. She is currently a researcher at National Research, and Innovation Agency (BRIN), Indonesia. Her research interests include: Machine learning, data mining, and data science.
She contributed in conceptualising the present study and writing this article.

Mrs Yulia Aris Kartika received Master's degree in Computer Science from the Universitas Indonesia, Indonesia. She is currently a researcher at National Research, and Innovation Agency (BRIN), Indonesia. Her research interests include: Machine learning, data mining, data science and biodiversity informatics.
She contributed in conceptualising the present study, editing and writing this article.