Multi-Label Classification for Doctor's Behavioral Pattern Matching During Online Medical Interview using Machine Learning

Safitri Juanita[1,5], Diana Purwitasari[2,4], I Ketut Eddy Purnama[1,3,4], Abid Famasya Abdillah[2] and Mauridhi Hery Purnomo[1,3,4*]

[1]Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, [2]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia [3]Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia [4]University Center of Excellence on Artificial Intelligence for Healthcare and Society, Indonesia [5]Department of Information Systems, Universitas Budi Luhur, Jakarta, Indonesia
*Corresponding author's Email: hery@ee.its.ac.id

Abstract: In recent years, many studies on medical texts have attracted the attention of researchers. Medical text studies have few multi-label data targets because it is challenging to understand dependencies between labels. Therefore, this study investigates a collection of medical texts by addressing complex problems in the behavioural pattern of Doctor's answer text in Online Health Consultation (OHC) by suggesting a pattern of six medical interview functions ranging from fostering doctor-patient relationships to treatment-related behaviours and responding to emotions. There are many proposed MLC methods to solve a multi-label problem. However, this study proposes an MLC model that can improve MLC accuracy, especially in multilingual medical datasets: English and Indonesian. This study proposes 16 MLC models using two feature extraction methods, compares all proposed models, and evaluates model performance using three perspectives. The results show that from 3 perspectives, the MLC model that consistently outperforms other models in the English dataset is a T-BR-RF model (TF/IDF, Binary Relevance, and Random Forest). In contrast, using the Indonesian dataset, the T-BR-AD Model (TF/IDF, Binary Relevance and Adaboost) outperforms other MLC models. The feature extraction method that helps optimize the performance of MLC models is TF-IDF compared to the Word2Vec method.

Keywords: Online Health Consultation, Multi-label Classification, Medical Interview Functions, Medical Text, Behavioral Pattern Matching.

1. Introduction

WHO confirmed that in 2020 the coronavirus pandemic had spread worldwide and currently confirms that there are more than 239 million cases [1]. This situation led some people who needed medical consultation to adopt new ways to get the virtual medical care that could limit direct contact between doctors and patients [during pandemics], such as using social media for telemedicine [2]. The telemedicine platform provides an online health consultation (OHC) for people to consult a doctor for free and receive higher traffic visits than before the Coronavirus outbreak. OHC helps people receive informational and emotional support to increase their willingness to share and seek health information [3].

Further, this study also found that a user in an OHC will only share or seek information related to their health [3]. On the other hand, based on research [4], people prefer to receive recommendations for their minor health problems with online health consultation rather than coming to a hospital or clinic, spending time waiting, and going home without proper consultation from a doctor's. In another study [5], users feel comfortable conducting online health consultations, mainly if a medical consultation provides practical information according to their needs. Based on previous studies' benefits of online health consultation for the community, it aims to produce helpful information for readers and health information seekers, especially for users who ask questions in OHC to get a proper and accurate diagnosis. For this reason, it requires adequate medical consultation.

Received: October 11th, 2022. Accepted: September 23rd, 2023
DOI: 10.15676/ijeei.2023.15.3.5

Medical consultation requires effective communication between patient and doctor's, and the patient's role is much more important in clinical encounters [6]. Although various obstacles sometimes disrupt the health consultation process with medical personnel, patients feel the benefits of consultation when reaching the core functions of consultation (for example, personal care, ongoing care, and more.) [6]. However, when conducting an effective medical consultation at OHC, the challenges faced are the limitations of patients in describing symptoms of the disease using text, and doctors also have limitations in providing answers to questions because the answers are in the form of text.

Therefore, doctors need communication skills; if the patient does not take an active role in the online health consultation session, the doctor's must collect information from the patient before making a diagnosis. Based on the exposure of previous research and the problems above, we suggest that the doctor's text answer at OHC follow the six functions of a medical interview [7]: building relationships, gathering information, providing information, making decisions, enabling illness and treatment-related behaviours, responding to emotions from the medical interview.

Medical text data is complex, and the amount of data has increased rapidly in recent years, requiring the application of artificial intelligence in the health sector; one currently overgrowing is online health consultation [8]. However, slightly artificial intelligence studies in medical texts have focused on multi-label data, such as [9], [10] using single-label data. Therefore, it is necessary to investigate further the multi-label classification method (MLC), which has good performance in predicting multi-label in medical texts—contrasting with recent research on medical texts using multi-label classification [11]–[14].

This study investigated the communication patterns of doctor response texts following a medical interview function [7] and aims to find a multi-label classification (MLC) model that can improve MLC performance, particularly in the classification of medical texts. Therefore, this paper proposed an MLC model for predicting the matching of doctor behavior patterns during medical interviews in Indonesian online health consultations. Furthermore, this study examines the applicability of the proposed MLC model using English health consultation data (the more widely used language). Consequently, integrating information including two languages enables our suggested model to produce a more comprehensive model, boost the model's adaptability and resilience, and provide significant insights into cross-language patterns in health consultations.

The contribution of this research is to propose the best-performing of MLC model for predicting Doctor's Behavioral Pattern Matching During Online Medical Interview using multilanguage datasets in English and Indonesian. The MLC approach uses two stages, problem transformation and adaptation algorithm, as shown in Figure 4. Because extracting informational features can significantly improve the performance of the classification model and reduce computational complexity [15]. Therefore, this study also compares two feature extraction methods: TF/IDF and Word2Vec, to find out which methods determine the proposed model to improve its performance.

We organize the remainder of the paper: Section 2 discusses multi-label classification and related work. In Section 3, we provide details of the Proposed Model Multi-Label Classification Method. Section 4 presents experimental results in testing and comparing several models and discusses their implications. Furthermore, in section 5, we conclude the experimental results.

2. Related Work

The multi-label Classification (MLC) method is part of the machine learning approach through supervised learning [16]. MLC is a classification in which an instance can be part of several labels simultaneously [17] and has been widely used in various topics [18], such as in images [19], the internet of things [20], action dependencies in the video [21] and many more. However, this paper investigated multi-label classification using a collection of Doctor's text answers which suggest contains six medical interview functions [7] and annotated by medical

experts; in other words, each Doctor's text answer has more than one medical interview function; this condition is named multi-label.

This section introduces previous research on problem-solving in text data using an MLC approach, especially in medical texts. Then explain the difference in the solution of the previous model compared to the proposed model. Various studies have used the MLC approach in text classification, including [22]–[26] and more. However, few studies still explicitly discuss multi-label using medical text data as follows; Research to predict gene function because one gene may have many functions with the MLC approach [11]. Experiments were carried out with the Weka library, using several methods, namely decision tree, j48, and methods from Mulam Library: Label Powerset, Binary Relevance, Random K-Label sets (Rakel), MLKNN (Multi-label K-Nearest Neighbors). All methods were compared, and it was found that the BR-DT Pru algorithm was superior to other algorithms. Another research uses biomedical documents which have a very extreme set of labels [12]. Biomedical documents have two forms: biomedical literature classification and clinical records. This study compares machine learning classification with deep learning to determine which method performs best in classifying biomedical text documents. The machine learning classification method uses Binary relevance in the problem transformation stage and SVM, logistic regression, random forest, and extra tree in the algorithm adaptation stage.

This study reports that the MLC approach with the Binary Relevance method and SVM with a linear kernel performs better than other machine learning algorithms. Another research uses data on patients with Psychotic Disorder Disease (PDD) [13]. This study uses data on PDD patients because a psychotic patient may have symptoms that lead to several PDD diseases, thus requiring the MLC method to assist in diagnosis. This study evaluates 15 MLC methods, some of them problem transformation (PT) using Binary Relevance (BR), Label Powerset (LP) with Homer, and Algorithm Adaptation (AA) using four algorithms, namely ML-KNN, PCT, ML-RF, ML-DT, Rakel. This study indicates that one of the two that has the best performance on the transformation problem is the powerset label. At the same time, Naive Bayes (NB) and Naive Bayes Tree (NBTree) consistently perform best on the PDD dataset.

Another study used a multi-label approach to identify various diseases in the patient's electronic medical record dataset. This study proposes machine learning for classification of diseases using the extracted data from electronic health records (EHR) and also used deep learning for training the model using deep neural networks, which this research claims can help prevent misdiagnosis [14]. Previous studies used machine learning approaches to effectively solve multi-label problems on medical text datasets, as machine learning models are often easier to understand and implement. Thus, it can be concluded that:

- Some studies above used two stages in solving multi-label problems: transformation problems and adaptation algorithms. In line with this research, we use both stages at the problem transformation stage using the Binary Relevance (BR) and Label Powerset (LP) methods.
- At the adaptation algorithm stage, we use four classification algorithms, namely Random Forest (RF), Adaboost (AD), KNN, and Multilayer Perceptron (MLP). Then combined, these two methods into eight proposed MLC models, and several previous studies have not discussed using this method.
- Several studies above compare various MLC models to determine which model performs best in the classification. In line with this study, we propose and investigate the best MLC method using several proposed MLC models to predict the pattern of medical interview functions in Doctors' text answers using multilingual datasets: Indonesian and English. The datasets were collected by web scraping from an online health consulting service website and labelled with the help of a team of medical experts with six medical interview functions.
- Previous studies measuring the performance of the MLC model mostly used an example-based perspective. This study conducted a more detailed investigation using three perspectives: example-based, label-based, and rating-based.

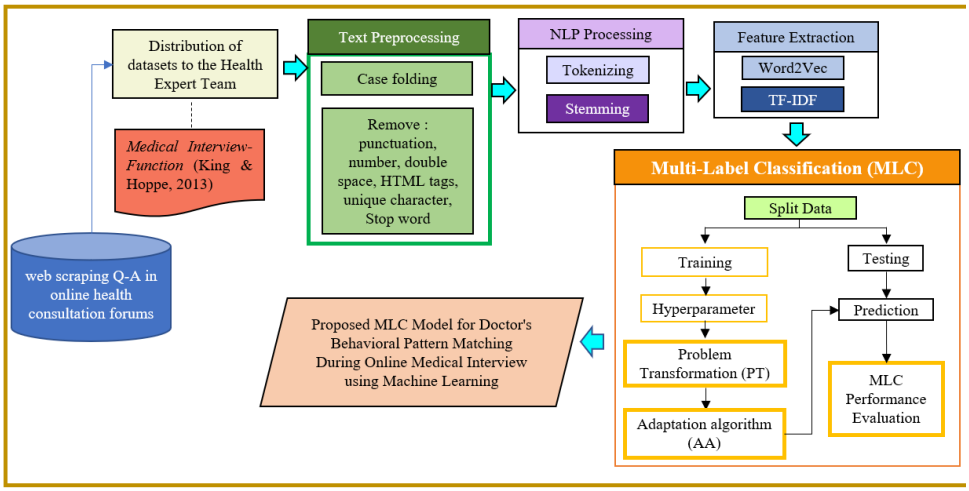

Figure 1. Proposed Framework Multi-Label Classification for Doctor's Behavioral Pattern Matching During Online Medical Interview using Machine Learning

In this study, we intentionally worked with a smaller dataset of 1000 datasets containing doctor's answer texts with six labels for several reasons: (1) The nature of our research required precise expert annotations, which are more feasible on a smaller scale; (2) Small datasets allow us to focus on the intricacies of label imbalance and multi-label classification, providing a deeper model performance analysis; (3) Working with a smaller dataset aligns with best practices, enabling thorough model evaluation and enhancing the robustness of our findings. Despite its size, our dataset is representative of real-world scenarios, and the insights gained are valuable for addressing similar challenges in larger datasets.

Our study proposed a machine learning approach for several reasons: it is more suitable as it can work reasonably with smaller datasets, is faster to train and deploy, is a more practical choice for applications that require real-time or near-real-time processing, and is suitable for researchers with limited computing resources. In addition, based on the research of [15], feature extraction helps classification performance, so in this study, we also investigate two feature extraction methods, TF-IDF and Word2Vec, to find out which method can help the performance of the proposed model. From the explanation above, the contribution of this study is proposing a multi-label classification model for medical texts, significantly to predict doctors' behavioral pattern matching during the online medical interview using a collection of Doctors' answer texts in various languages (Indonesian and English).

The proposed multi-label classification model combines problem transformation and adaptation algorithms with 8 MLC models, which then also compare two feature extraction methods, TF-IDF and Word2Vec so that a total of 16 combined models. We evaluate the proposed models using three perspectives with ten metric measurement methods.

3. Proposed Model Multi-Label Classification Method

This section introduces a proposed Multi-Label Classification (MLC) framework to predict the Pattern of Medical Interviews Function in Online Health Consultation Texts. The stages of the research process in the proposed framework are shown in Figure 1; we explain this section into four stages: collecting data at A, pre-processing data at B, proposing a model for MLC at C, and MLC Evaluation Measures at D.

A. Collecting Dataset

The consultation date range on the dataset is as follows: Alodokter from 8 December 2014 to 28 February 2021; Steadyhealth from 2 February 2005 to 12 April 2018. We provide the dataset to a team of medical experts, who then label each instance according to the six medical

F1: Good afternoon YJ. Thank you for the question *(Selamat siang YJ. Terima kasih atas pertanyaannya)* **[Fostering the relationship]**
F2 : Do you or in your family have a history of allergies? (*Apakah Anda atau dalam keluarga Anda memiliki riwayat alergi?)* **[Gathering information]**
F3 : When the throat is itchy, don't cough as hard as it will only increase the itching and pain. Even if you cough too often, it becomes even more inflamed and can cause injuries...(*Bahkan jika Anda batuk terlalu sering, akan semakin meradang dan dapat menyebabkan cedera...)* **[Providing information]**
F4 : Sore or itchy throat is usually caused by the flu or a virus or bacterial infection...(*Tenggorokan sakit atau gatal biasanya disebabkan oleh flu atau infeksi virus atau bakteri...*) **[Decision making]**
F5 : .. To fix this, drink lots of warm water, don't talk so that the throat feels better and use an over-the-counter cough syrup *(untuk mengatasinya, banyak minum air hangat, jangan bicara agar tenggorokan terasa lebih enak dan gunakan obat batuk yang dijual bebas)* **[Enabling disease and treatment-related behavior]**
F6 : ...Most will give side effects of drowsiness so that your sleep can be more comfortable at night. Greetings Health...*(Kebanyakan akan memberikan efek samping kantuk sehingga tidur Anda bisa lebih nyenyak di malam hari.....salam sehat)* **[Responding to emotions of the medical interview]**

Figure 2. An example of a dataset from the Online Health Consultation (OHC) website was then identified by a team of medical experts whether followed the six functions of a medical interview

interview functions (F) [7]: (F1) Building relationships, (F2) Gathering information, (F3) Providing information, (F4) decision making, (F5) enabling disease and treatment-related behaviours and (F6) responding to emotional, medical interviews. We display an example of a doctor's text answer used in the experiment and labelling by a team of medical experts in Figure 2 and the distribution of the instances number on each label as shown in Figure 3.

B. Process Data

This section describes in detail several stages ranging from text pre-processing, NLP, and feature extraction.

- Text Preprocessing

The research data we collect is still raw and requires a cleaning stage. The following are the pre-processing stages used a library from python in this experiment [27]; (1) change all strings in attribute to lowercase using the lower method; (2) remove noise text; at this stage, we perform some removal actions using translate method such as: to remove punctuation and change it with space also delete sentences containing numbers and change it with space; using join and split method to remove double space, because the previous cleaning process caused many spaces between words, thus requiring a double-spaced cleaning; uses regex strings to remove sentences containing HTML tags, and also use this method to delete special characters that have not erased from the previous stage such as containing *, RP, or other characters from the dataset;

In this stage, we also perform stop word removal is the process of removing a word that is not needed. These unnecessary words are referred to in the stop word list [27], [28]. The most important characteristic to determine the stop word is usually the word that occurs most often, for example, conjunctions such as "why," "when," "until," and others. In our experiment, we used Satya's Indonesian stop word list for the Indonesian dataset. In contrast, for the English dataset, we use the English stop word from a python package named the nltk.corpus package ("stopword").

- NLP Processing

The next stage is NLP processing, tokenization, and stemming [27]. The tokenizing stage is separating each word in a doctor's text answer. An example of tokenizing in the following word

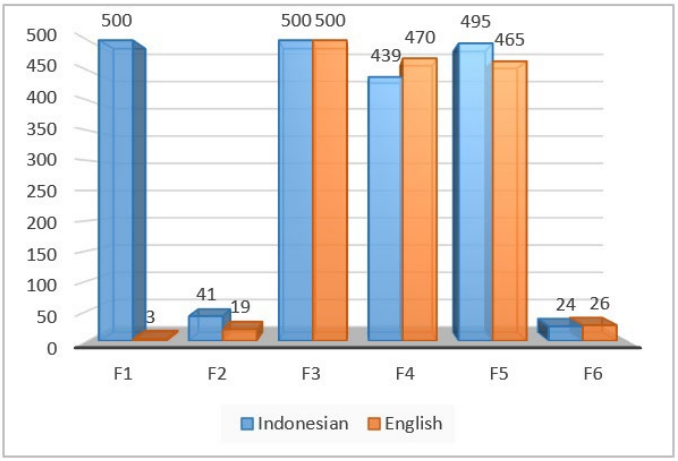

Figure 3. The Distribution of The Dataset Contains a Collection of Doctor's Text Answers on Each Label.

snippet is "sore throat is generally caused by..." then, in this step, generate three tokens "sore", "throat", and "generally." This stage also represents the dataset into several scores, such as the number of words generally obtained from the tokenization process. The next stage uses the Stemming algorithm, whose purpose is quite specific and straightforward, looking for the morphological roots of a word [28]. In the stemming stage, we use the Python Library name Sastrawi for the Indonesian dataset [29], while the English dataset uses a package in python named nltk.stem.porter.

- Feature Extraction

The next step is to perform feature extraction on the dataset to determine the importance of the word (term) in each doctor's text answer in the corpus. In this study, we compared the feature extraction process using Word2Vec and TF-IDF methods to find out which methods help improve MLC accuracy performance. The following is a detailed explanation of how the TF-IDF and Word2Vec models are processed;

- **TF-IDF** is the most well-known and used weighting method. This method performs word weighting by calculating the frequency value of the occurrence of a word in each document and counting the occurrence of a word in the total number of documents [30]. There are three stages of how TF-IDF feature extraction works [31]: **(1) Term Frequency (TF):** This component measures the frequency of a term (word) within a document. It indicates how often a term appears in a document and is calculated as the number of times a term appears in the document divided by the total number of terms in the document. The following equation:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad (1)$$

(2) Inverse Document Frequency (IDF): IDF measures the importance of a term across a collection of documents. It calculates the inverse of the fraction of documents that contain the term. Terms that are common across all documents receive a lower IDF score, while terms that are unique to a specific document receive a higher IDF score. The following equation:

$$IDF(t) = 1 + log\left(\frac{Total\ number\ of\ documents\ in\ the\ corpus}{Number\ of\ documents\ containing\ term\ t}\right) \quad (2)$$

(3) TF-IDF Score: The TF-IDF score for a term in a document combines the TF and IDF values. It indicates how important a term is in a specific document within the context of the entire corpus. The following equation:

$$TF - IDF(t,d) = TF\ (t,d)\ x\ IDF(t) \quad (3)$$

This study implements the TF-IDF feature extraction model using the Python package Scikit-TfidfVectorizer learn's function.

- **Word2Vec** is a particularly efficient predictive model for raw text with learning word embedding. This method combines CBOW and Skip-Gram to convert words into vectors [32], [33]. The following describes how Word2Vec performs feature extraction [34]; The Word2Vec model first constructs a vocabulary from training data, then determines and learns the vector representation of each word. Word2vec contains two training algorithms; (1) **Continuous Bag of Words (CBOW):** CBOW predicts the target word based on its surrounding context words. It takes a context window of words (e.g., "context_size" words to the left and right of the target word) as input and tries to predict the target word in the middle. (2) **Skip-gram:** Skip-gram, on the other hand, predicts the surrounding context words given the target word. It takes the target word as input and aims to predict the context words within the context window.

  In this study, the embedding of Word2Vec as a feature extraction technique enables the proposed multilabel classification model to recognize subtle semantic relationships in text data, thereby enhancing the model's ability to handle multi-label classification in the context of online medical consultations. This study implements the Word2Vec feature extraction model using the Gensim Python package.

C. Proposed Model for Optimization of Multi-label Classification (MLC) in Medical Texts (Hybrid-MLC)

Multi-label classification is a standard method for modelling objects with multiple meanings [35]. In this study, we divide the MLC approach into two stages: Problem Transformation (PT) and Adaptation Algorithm (AA) [23], [36]–[38], as shown in Figure 1 and Figure 4. Problem transformation is the first stage of multi-label classification that transforms the multi-label problem into a single-label, wherein this study used Binary Relevance (BR) and Label Powerset (LP). A detailed description of the proposed model for PT is as follows.

- Binary Relevance (BR) is a method for modifying the original data set with multiple labels into one label by dividing the original data set into label groups. Each instance can be in multiple label groups [39], [40]. Binary relevance has a hidden weakness of the inability to utilize label correlation to improve the generalization ability to learn systems [35].
- Label Powerset (LP) is a problem transformation method for transforming a multi-label problem into a multi-class problem using a single multi-class classifier trained with all the unique label combinations found in the training data [11], [26].

The next step is to classify the dataset using a single-label or conventional classification algorithm. This step is called the adaptation algorithm (AA). In the AA stage, to find the suitable algorithm for the proposed MLC model, based on related literature and conducting some preliminary experiments also recognizing the complexity of multilabel and imbalanced datasets, in this study we used Random Forest (RF), AdaBoost (AD), K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) as our primary algorithms after weighing their robustness, adaptability, and track record in similar settings. The consistent superiority demonstrated by these algorithms in our exhaustive evaluations validates our selection, confirming their suitability as the most effective instruments for addressing the challenges inherent to our dataset and research objectives.

The following is a detailed explanation of Classification algorithm or a conventional classification algorithm in the AA stage using four classical classifier methods:

- Random Forest (RF) [41][42]: is one of the methods in the Decision Tree, an ensemble of unpruned classification or regression trees. RF is a powerful tool capable of delivering performance because reduces overfitting, less sensitive to outliers and noisy data than some other algorithms, and can provide a measure of feature importance.
- Adaboost (AD): This algorithm combines weak classification functions to form a more robust classifier [42].
- KNN is a lazy learning method widely used in data mining, especially when datasets have little or no prior knowledge of data distribution [43].
- Multilayer Perceptron (MLP) is a method that repeatedly adjusts the weights and thresholds to minimize the difference between the target output and the resulting output [44], also type of artificial neural network that can be used for deep learning tasks and to learn complex hierarchical representations of data, can handle various types of data

In this study, the two stages are combined and create eight proposed models of the MLC model. This study also compares the two-feature extraction (FE) methods: TF-IDF (T) and Word2Vec (W), to determine which FE method helps the model get the best performance. So, in the experimental stage, the eight models are combined with the FE method, with 16 models, as shown in Figure 4.

The abbreviations in Figure 4 contain T-BR-RF=TF IDF-Binary Relevance-Random Forest), T-BR-AD=TFIDF-Binary Relevance-Adaboost, T-BR-KNN=TFIDF-Binary Relevance-KNN, T-BR-MLP=TFIDF-Binary Relevance-Multilayer Perceptron, T-LP-RF=TFIDF-Label Powerset-Random Forest, T-LP-AD=TF IDF-Label Powerset-Adaboost, T-LP-KNN=TF IDF-Label Powerset-KNN, T-LP-MLP=TF IDF-Label Powerset-Multilayer Perceptron, W-BR-RF=Word2Vec-Binary Relevance-Random Forest, W-BR-AD=Word2Vec-Binary Relevance-Adaboost, W-BR-KNN=Word2Vec-Binary Relevance-KNN, W-BR-MLP=Word2Vec-Binary Relevance-Multilayer Perceptron, W-LP-RF=Word2Vec-Label Powerset-Random Forest, W-LP-AD=Word2Vec - Powerset Label – Adaboost, W-LP-KNN=Word2Vec - Label Powerset – KNN, W-LP-MLP=Word2Vec-Label Powerset-Multilayer Perceptron.

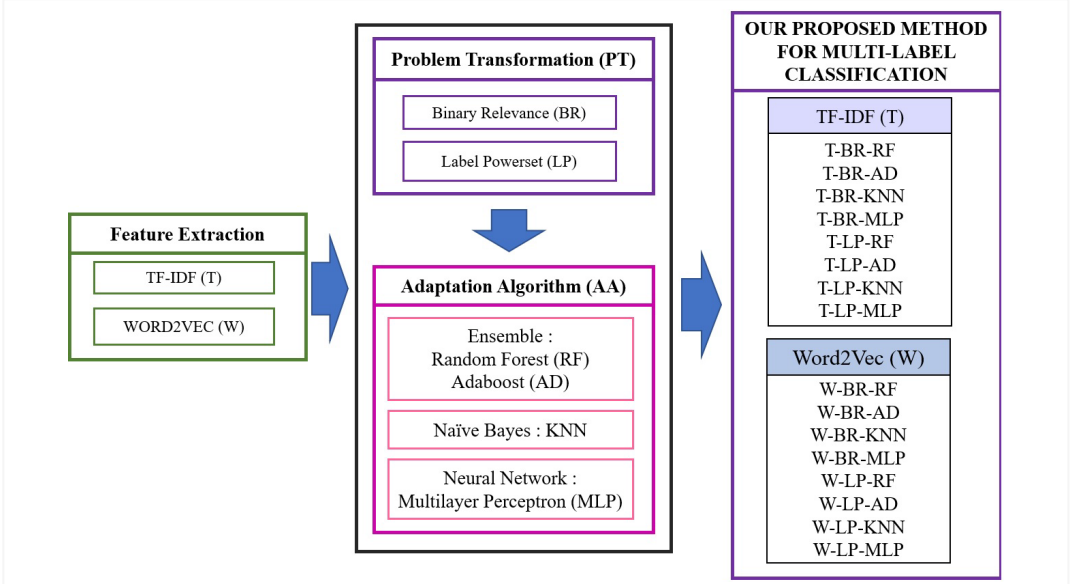

Figure 4. The Proposed Multi-Label Classification Model for Doctor's Behavioral Pattern Matching During Online Medical Interview

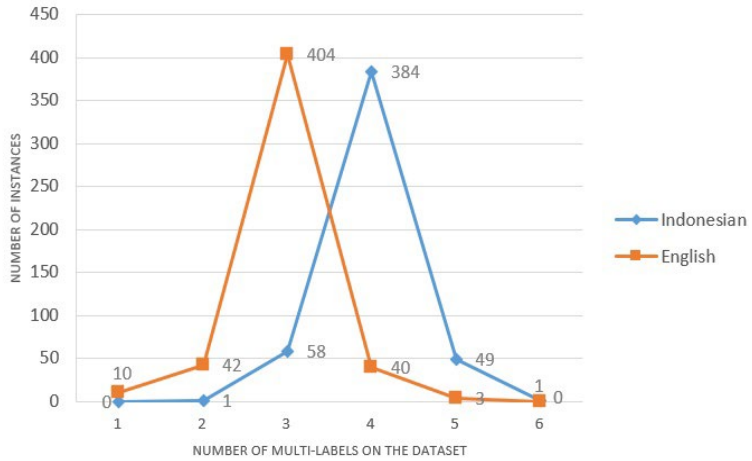

Figure 5. Multi-label visualization data based on two datasets containing collections of doctors' answers text from online health consultations

We train the proposed model using datasets containing a collection of doctor's text answers from OHC in Indonesian and English using the hyperparameter technique to determine the best parameters for each model's performance accuracy. Hyperparameter is a search strategy that evaluates all candidates with a few resources and repeatedly uses more resources to select the best candidate.

(2) Multi-Label Classification Evaluation Measures

In traditional supervised classification, performance evaluation measures are usually by accuracy or F-measure. However, different from MLC, performance in multi-label is more complicated as each instance can be associated with multiple labels simultaneously [17]. Therefore, in this study, we conduct several steps to evaluate the performance of the MLC method from three perspectives: example-based, label-based, and ranking-based. The following is an explanation from each perspective [45].

- Example-based perspective measures the classification performance for each sample based on the average difference between the actual and predicted label sets across all samples [39]. At this stage, we use multiple measurement metrics, such as the most well-known multi-label measurement, to predict errors in MLC named Hamming Loss (HL), precision, recall and F1-Score.
- The label-based perspective measures the classification performance for each label individually and reports the average performance across all labels [45]. The MLC model's evaluation needs to be measured using this perspective because the classification performance measurement will be carried out on all labels and then averaged [45]. This study uses a macro average because the macro average calculates the metric independently for each label and then takes the average (treating all classes equally) [40]. The following is a detailed explanation of macro averaging approach with three metrics; Precision Macro (P_M), Recall Macro (R_M), and F1-Score Macro ($F1_M$) [40], [46]; Precision Macro (P_M) is an average precision over the data (d) class variables (precision per label) as in the following formula:

  $$P_M = \sum_{i=1}^{d} \frac{TP_i}{TP_i + FP_i} \quad (1)$$

  Recall Macro (R_M) over the variables in the data (d) class (recall per label) as in the following equation:

  $$R_M = \sum_{i=1}^{d} \frac{TP_i}{TP_i + FN_i} \quad (2)$$

  F1-Score Macro ($F1_M$) is Relationships between data's positive labels and those assigned by a classifier on a per-class basis. The following formula:

$$F1_M = \frac{(\beta^2+1)\ X\ P_M\ X\ R_M}{\beta^2\ P_M\ + R_M} \quad (3)$$

where TP is True Positive, FP (False Positive), and FN (False Negative).

- The ranking-based perspective compares the label ranking predictions generated by the classification, averaging the results of overall samples [13]. This experiment used multiple measurement metrics, such as Coverage Error (CE), to evaluate the MLC Method from a Ranking-Based perspective. This metric goal is to predict all labels correctly, so it is necessary to include all labels with a score greater than or equal to the actual label. Label ranking average Precision (LRAP) is the average of each ground truth label associated with each sample, the ratio of the accurate label to the total label with the lowest score. Label ranking Error (LRE) expresses the number of times irrelevant labels are ranked higher than relevant labels [45].

4. Result and Discussion

This section presents the experimental results using two datasets containing doctor's answer text in Indonesian and English. We compare the performance of our proposed multi-label classification (MLC) models and suggest the best model of all the proposed models. In addition, this study also compared two feature extractions to determine which method helps our proposed MLC model get better accuracy performance.

The experiment in this study uses the python programming language with the Scikit-Learn package. We Explain this section into three sections, Data visualization of multi-label quantities A, Experimental result B, and Discussion of experimental result discovery C.

A. Data visualization for the distribution of multi-label quantities

This section displays the results of tagging or labelling by a team of medical experts on both datasets: Indonesian and English, using six medical interview functions. The results in Figure 5 show the distribution of the number of multi-label instances. The result showed that most of the two datasets had 3-4 labels on each instance. In contrast, the Indonesian dataset had four labels on 384 instances. In contrast, the English dataset has three labels on 404 instances. No instance in the Indonesian dataset has only one label, but the corpus has at least 2 to 6 labels. In contrast, no instance has six labels in the English data set, but the corpus has at least 1 to 5 labels.

B. Experimental results of our proposed MLC Method

The experiment conducted only focused on hyperparameter tuning with HalvingGridSearch and utilized a deliberate 80/20 test split. This focused approach allowed us to explore how model parameter tuning impacts performance, especially on our complex dataset. Although other splitting ratios exist, this study selected consistency to emphasize the significance of hyperparameter optimization. Future research could explore various splits, but our study lays a solid foundation for understanding the power of hyperparameter tuning in improving model adaptability to challenging datasets. This study chose HalvingGridSearch because it effectively explores the hyperparameter space while avoiding overfitting, a common issue in machine learning. By carefully optimizing parameters, HavingGridSearch strikes a balance between model complexity and generalization ability, ensuring that the proposed MLC model can effectively adapt to challenging datasets such as the dataset used in this study.

This study also compares TF-IDF (T) and Word2Vec, two Feature Extraction (FE) techniques (W). Determine which FE method can improve the accuracy of the proposed MLC models. Thus, in this study, we present experimental results by combining three stages using FE-PT-AA: Feature Extraction (FE), Problem Transformation (PT), and Adaptation Algorithm (AA). For example, the T-BR-RF model stands for TF-IDF - Binary Relevance - Random Forest. Other proposed combinations, or the proposed MLC Model in this study, are shown in Figure 4.

For each proposed MLC model, Table 1 shown the optimal model parameters generated by the hyperparameter procedure using HalvingGridSearch. The first column contains the names of the 16 proposed MLC models, whereas columns 2 through 6 contain the best parameter (P) values for each proposed MLC model, along with the order of parameters from two datasets.The

first column contains the names of the 16 proposed MLC models, whereas columns 2 through 6 contain the best parameter (P) values for each proposed MLC model, along with the order of parameters from two datasets: (1) Indonesian (ID) and (2) English (EN). For example, column parameters 1 to 5 in the T-BR-RF model have the same parameter values in both datasets. In contrast, the W-LP-RF model has two different parameters in column P-1, where the Indonesian dataset uses max_depth = 3, while the English dataset uses max_depth = None.

Table 1. The Best Parameter Values for Each Proposed Model

FE-PT-AA	**P-1 (ID; EN)**	**P-2 (ID; EN)**	**P-3 (ID; EN)**	**P-4 (ID; EN)**	**P-5 (ID; EN)**
T-BR-RF	max_depth=none	Min_samples_split =5; Min_samples_split =10	n_estimators= 48		
W-BR-RF	max_depth=none	Min_samples_split =5; Min_samples_split =10	n_estimators= 48		
T-BR-AD	algorithm= ’SAMME’	learning_rate = 0.5	n_estimators= 48		
W-BR-AD	algorithm= ’SAMME’	learning_rate = 0.1	n_estimators= 48		
T-BR-KNN; W-BR-KNN	algorithm=’kd_tree’ ; algorithm: 'auto'	n_neighbors=48	weights= ’distance’		
T-BR-MLP	activation= ‘logistic’	hidden_layer_sizes = 189	learning_rate= ’invscaling’ ; learning_rate: 'constant'	solver: 'adam'	
W-BR-MLP	activation = ’tanh’	hidden_layer_sizes = 189	learning_rate= 'constant'	solver:'sgd' ;solver:'adam'	
T-LP-RF	Max_depth=3; max_depth: None	min_samples_split =5; min_samples_split: 10	n_estimators = 48		
W-LP-RF	max_depth=3; max_depth: None	min_samples_split= 10	n_estimators = 48		
T-LP-AD	algorithm= ’SAMME.R’	learning_rate = 0.1	n_estimators = 48		
W-LP-AD	algorithm= ’SAMME’	learning_rate = 0.1	n_estimators = 48		
T-LP-KNN; W-LP-KNN	algorithm=’kd_tree’; algorithm:'ball_tree'	n_neighbors = 48	weights = ’distance’		
T-LP-MLP	activation=’logistic’	hidden_layer_sizes =189	learning_rate= ‘constant’	max_iter =100	solver= ’adam’
W-LP-MLP	activation=’tanh’	hidden_layer_sizes =189	learning_rate= ’constant’	max_iter =100	solver= ’adam’

- Example-Based

The performance evaluation results on the proposed MLC model with an example-based perspective using eight proposed MLC models also by comparison of two feature extraction

methods. Performance measurement is Displayed in Table 2 using HL, P, R, and F1. HL is better for lower values, but higher values are better for P, R, and F1 measurements.

Table 2. Metrics Evaluation Results using Example-Based Perspective on the proposed MLC model for Doctor's Behavioral Pattern Matching During Online Medical Interviews

EF-PT-AA	Indonesia				English			
	HL	P	R	F1	HL	P	R	F1
T-BR-RF	0.037	0.963	0.985	0.971	**0.022**	**0.976**	**0.987**	**0.978**
T-BR-AD	**0.035**	0.963	**0.987**	**0.973**	0.023	**0.976**	0.985	0.977
T-BR-KNN	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
T-BR-MLP	**0.035**	**0.965**	0.985	0.972	**0.022**	**0.976**	0.986	0.977
T-LP-RF	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
T-LP-AD	0.040	0.958	0.985	0.969	0.025	0.973	0.983	0.973
T-LP-KNN	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
T-LP-MLP	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
W-BR-RF	0.037	0.963	0.985	0.971	0.025	**0.976**	0.981	0.974
W-BR-AD	0.037	0.963	0.985	0.971	0.023	**0.976**	0.984	0.976
W-BR-KNN	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
W-BR-MLP	0.037	0.963	0.985	0.971	0.027	0.974	0.978	0.972
W-LP-RF	0.037	0.963	0.985	0.971	0.023	**0.976**	0.984	0.976
W-LP-AD	0.037	0.963	0.985	0.971	0.025	0.973	0.983	0.974
W-LP-KNN	0.037	0.963	0.985	0.971	0.023	**0.976**	0.983	0.976
W-LP-MLP	0.037	0.963	0.985	0.971	0.025	0.973	0.983	0.974

Based on the results shown in Table 2 shows that the best performance with an example-based perspective on the Indonesian dataset for HL is in the models: T-BR-AD, TBR-MLP with a value of 0.035, P metric in T-BR-MLP with a value of 0.965, metric R on the T-BR-AD model with a value of 0.987, and F1 on the T-BR-AD model with a value of 0.973.

While the best performance on the English dataset for model measurement with HL on two models, namely T-BR-RF and T-BR-MLP, with a value of 0.022. while for the P measurement, most of the models showed a high P value, namely 0.976, except for four models, T-LP-AD, W-LP-AD, and W-LP-MLP, with a value of 0.973 and W-BR-MLP, with value 0.974. while the measurement using R and F1 metrics, which got the best performance, was the T-BR-RF model with an R-value of 0.987 and an F1 value of 0.978.

The measurement results in the example-based perspective demonstrate that the T-BR-AD model significantly outperformed the Indonesian dataset by having the highest scores on HL (0.035), R (0.987), and F1 (0.973). In contrast, the T-BR-RF model significantly outperformed the other models in the English dataset, achieving the highest scores on the measurements of HL (0.022), P (0.976), R (0.987), and F1 (0.978). The proposed MLC model uses the TF-IDF (T) feature extraction method to outperform other models in both datasets. Thus, it can be concluded that TF-IDF (T) contributes to the superiority of the proposed MLC model over other models by enhancing classification performance more than the Word2VEC (W) method.

- Label-based

Evaluation metrics based on a label-based perspective use a macro averaging approach with three metrics, namely Precision Macro (P_M), Recall Macro (R_M), and F1-Score Macro ($F1_M$), which appear in Table 3. The macro average assumes “equal weight” for the label and each instance. Based on the label-based perspective, Table 3 shows that neither model significantly outperformed the other models in both data sets in a label-based perspective using macro-average.

Based on the results shown in Table 3 shows, the model that outperforms other models using the Indonesian dataset is the T-BR-AD model with metric values of P_M (0.700), R_M (0.722), and

$F1_M$ (0.711). In comparison, the model that has superior performance to other models in the English dataset is the T-BR-MLP model with metric values of P_M (0.768), R_M (0.691), $F1_M$ (0.722).

Table 3. Metrics Evaluation Results using Label-Based Perspective with macro-averaging on the proposed MLC model for Doctor's Behavioral Pattern Matching During Online Medical Interviews

EF-PT-AA	Indonesia			English		
	P_M	**R_M**	**$F1_M$**	**P_M**	**R_M**	**$F1_M$**
T-BR-RF	0.642	0.667	0.653	**0.768**	0.637	0.673
T-BR-AD	**0.700**	**0.722**	**0.711**	0.573	0.542	0.550
T-BR-KNN	0.642	0.667	0.653	**0.768**	0.635	0.672
T-BR-MLP	0.643	0.667	0.654	**0.768**	**0.691**	**0.722**
T-LP-RF	0.642	0.667	0.653	**0.768**	0.635	0.672
T-LP-AD	0.642	0.667	0.653	0.487	0.500	0.493
T-LP-KNN	0.642	0.667	0.653	**0.768**	0.635	0.672
T-LP-MLP	0.642	0.667	0.653	**0.768**	0.635	0.672
W-BR-RF	0.642	0.667	0.653	0.601	0.580	0.588
W-BR-AD	0.642	0.667	0.653	0.740	0.595	0.633
W-BR-KNN	0.642	0.667	0.653	**0.768**	0.635	0.672
W-BR-MLP	0.642	0.667	0.653	0.655	0.538	0.559
W-LP-RF	0.642	0.667	0.653	**0.768**	0.635	0.671
W-LP-AD	0.642	0.667	0.653	0.487	0.500	0.493
W-LP-KNN	0.642	0.667	0.653	**0.768**	0.635	0.672
W-LP-MLP	0.642	0.667	0.653	0.487	0.500	0.493

- Ranking-based

The evaluation metrics are based on a Ranking-based perspective using three measurement metrics; Namely Label Ranking Average Precision (LRAP), Coverage Error (CE), and Label Ranking Error (LRE). A lower value is better for CE and LRE, but for LRAP, a higher value is better. The results in Table 4 show that from a ranking-based perspective, no model outperforms other models in the two datasets.

Based on the results shown in Table 4 shows, the model that outperforms other models using the Indonesian dataset is the T-BR-MLP model with metric values of LRAP (0.963), CE (4.130), and LRE (0.063). In comparison, the model that has superior performance to other models in the English dataset is the T-BR-RF model with metric values of LRAP (0.971), CE (3.180), and LRE (0.034).

Table 4. Metrics Evaluation Results using Ranking-Based Perspective on the proposed MLC model for Doctor's Behavioral Pattern Matching During Online Medical Interviews

EF-PT-AA	INDONESIA			ENGLISH		
	LRAP	CE	LRE	LRAP	CE	LRE
T-BR-RF	0.961	4.140	0.067	**0.971**	**3.180**	**0.034**
T-BR-AD	0.962	**4.130**	0.067	**0.971**	3.190	0.036
T-BR-KNN	0.961	4.140	0.067	0.969	3.210	0.038
T-BR-MLP	**0.963**	**4.130**	**0.063**	0.970	3.190	0.035
T-LP-RF	0.961	4.140	0.067	0.969	3.210	0.038
T-LP-AD	0.957	4.160	0.077	0.968	3.210	0.038
T-LP-KNN	0.961	4.140	0.067	0.969	3.210	0.038
T-LP-MLP	0.961	4.140	0.067	0.969	3.210	0.038
W-BR-RF	0.961	4.140	0.067	0.968	3.230	0.040
W-BR-AD	0.961	4.140	0.067	0.970	3.200	0.037

EF-PT-AA	INDONESIA			ENGLISH		
	LRAP	CE	LRE	LRAP	CE	LRE
W-BR-KNN	0.961	4.140	0.067	0.969	3.210	0.038
W-BR-MLP	0.961	4.140	0.067	0.966	3.250	0.043
W-LP-RF	0.961	4.140	0.067	0.970	**3.180**	0.037
W-LP-AD	0.961	4.140	0.067	0.968	3.210	0.038
W-LP-KNN	0.961	4.140	0.067	0.969	3.210	0.038
W-LP-MLP	0.961	4.140	0.067	0.968	3.210	0.038

C. *Discussion*

Our research investigates the complexities of multi-label classification, in which each data instance is assigned multiple labels. This multi-label characteristic introduces a fundamental difficulty: the variable distribution of instances across individual labels. Some labels are significantly more common than others. For example, Labels 1, 3, 4, and 5 have a significantly greater number of occurrences than Labels 2 and 6. The variance in instance numbers across labels profoundly affects model performance. Traditional machine learning algorithms may exhibit a bias toward the majority class labels, leading to suboptimal performance on minority class labels. We actively addressed this challenge through a two-stage modeling approach. In the Problem Transformation (PT) stage, we explored both Binary Relevance and Powerset Label strategies, adapting our approach to the specific characteristics of our data. These strategies allow the models to better account for label imbalance and variation in instance numbers.

Additionally, in the Adaptation Algorithm (AA) stage, we harnessed the strengths of four distinct machine learning algorithms—Random Forest, K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and AdaBoost. This strategic selection enabled our models to excel in scenarios where label distribution varies significantly. Each algorithm brought unique capabilities to the ensemble, enhancing our models' adaptability. In summary, the variance in instance numbers across labels is a central consideration in our research.

It has a significant effect on the performance of multi-label classification models. Our comprehensive modeling approach strategically handles label imbalance and label distribution complications by including Problem Transformation and Adaptation Algorithm stages, along with the selection of flexible algorithms. These components, along with our complex evaluation viewpoint incorporating example-based, label-based, and rating-based metrics, contribute to a comprehensive understanding of the performance of multi-label classification models in the setting of varying label instance numbers. This study provides useful information and indicates how our research may be applied to actual scenarios in the broader field of machine learning.

5. Conclusion

Using Machine Learning, we investigate the multi-label classification (MLC) model in this study to predict Doctor's Behavioral Pattern Matching During Online Medical interviews. This study uses the doctor's answer text in Indonesian and English from online health consultations. This study uses the MLC approach by proposing eight models, which in the experiment are a combination of 2 stages, namely problem transformation (PT) and Adaptation Algorithm (AA). The PT stage uses BR and LP, while in the AA stage, four classical classification algorithms are RF, AD, KNN, and MLP. This study also uses two feature extraction (FE) methods to determine which can help the eight models' performance. The combination of 8 MLC models and 2 FE methods makes a combination of 16 models. Using the hyperparameter method, all these models

were tested, compared, and evaluated using three approaches: example-based, label-based, and ranking based.

According to the experimental results, the proposed model shows the best performance, the T-BR-AD (TF/IDF, Binary Relevance, and Adaboost) model, specifically for Indonesian language datasets based on the evaluation of metrics from two perspectives, which are example-based and label-based. Meanwhile, The T-BR-RF (TF/IDF, Binary Relevance, and Random Forest) model, developed specifically for the English dataset, evaluates metrics from two perspectives: example-based and rank-based.

Furthermore, the results show that the TF-IDF feature extraction method supports the performance of the proposed MLC model more effectively than the Word2Vec feature extraction method. For future research, we intend to add more datasets (large amounts of data), process the datasets using deep learning approaches, and investigate imbalanced datasets.

6. Acknowledgements

This work received support from the Ministry of Research and Technology or the National Research and Innovation Agency of the Republic of Indonesia in 2022 and partially received financial support from the University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS) Institut Teknologi Sepuluh Nopember, Indonesia. The authors also thank Rizaldy Ramadhan and Bellinda Zalzabillah Tazkira, last year's medical students at Airlangga University, for annotating the dataset.

7. References

[1] WHO, "WHO Coronavirus Disease (COVID-19) Dashboard," *covid19.who.int*, 2021. https://covid19.who.int/ (accessed Jan. 03, 2021).

[2] Y. Li and K. Zhang, "Using social media for telemedicine during the COVID-19 epidemic," *Am. J. Emerg. Med.*, vol. 395, pp. 1016–1017, 2020, doi: 10.1016/j.ajem.2020.08.007.

[3] T. Mirzaei and P. Esmaeilzadeh, "Engagement in Online Health Communities: Channel Expansion and Social Exchanges," *Inf. Manag.*, vol. 58, no. 1, p. 103404, 2020, doi: 10.1016/j.im.2020.103404.

[4] S. Maeen and S. Zykov, "Towards social network - Integrated E-Health: Identify user attitudes," in *Information Technology and Quantitative Management (ITQM)*, 2015, vol. 55, pp. 1174–1182, doi: 10.1016/j.procs.2015.07.091.

[5] J. Huh *et al.*, "Personas in online health communities," *J. Biomed. Inform.*, vol. 63, pp. 212–225, 2016, doi: 10.1016/j.jbi.2016.08.019.

[6] M. A. Mazzi, M. Rimondini, E. van der Zee, W. Boerma, C. Zimmermann, and J. Bensing, "Which patient and doctor behaviours make a medical consultation more effective from a patient point of view. Results from a European multicentre study in 31 countries," *Patient Educ. Couns.*, vol. 101, no. 10, pp. 1795–1803, 2018, doi: 10.1016/j.pec.2018.05.019.

[7] A. King and R. B. Hoppe, "'Best Practice' for Patient-Centered Communication: A Narrative Review," *J. Grad. Med. Educ.*, vol. 5, no. 3, pp. 385–393, 2013, doi: 10.4300/jgme-d-13-00072.1.

[8] T. Davenport and R. Kalakota, "The Potential for Artificial Intelligence in Healthcare," *Futur. Heal.*, vol. 6, no. 2, pp. 94–98, 2019, doi: 10.2139/ssrn.3525037.

[9] C. VanDam, S. Kanthawala, W. Pratt, J. Chai, and J. Huh, "Detecting clinically related content in online patient posts," *J. Biomed. Inform.*, vol. 75, no. July, pp. 96–106, 2017, doi: 10.1016/j.jbi.2017.09.015.

[10] M. Frias *et al.*, "Classification Accuracy of Hepatitis C Virus Infection Outcome: Data Mining Approach," *J. Med. Internet Res.*, vol. 23, no. 2, pp. 1–17, 2021, doi: 10.2196/18766.

[11] E. A. Tanaka, S. R. Nozawa, A. A. Macedo, and J. A. Baranauskas, "A multi-label approach using binary relevance and decision trees applied to functional genomics," *J. Biomed. Inform.*, vol. 54, pp. 85–95, 2015, doi: 10.1016/j.jbi.2014.12.011.

[12] J. Du, Q. Chen, Y. Peng, Y. Xiang, C. Tao, and Z. Lu, "ML-Net: Multi-label classification of biomedical texts with deep neural networks," *J. Am. Med. Informatics Assoc.*, vol. 26, no. 11, pp. 1279–1285, 2019, doi: 10.1093/jamia/ocz085.
[13] S. O. Folorunso, S. G. Fashoto, J. Olaomi, and O. Y. Fashoto, "A multi-label learning model for psychotic diseases in Nigeria," *Informatics Med. Unlocked*, vol. 19, pp. 1–11, 2020, doi: 10.1016/j.imu.2020.100326.
[14] S. Khan and J. A. Shamsi, "Health Quest: A generalized clinical decision support system with multi-label classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 1, pp. 45–53, 2021, doi: 10.1016/j.jksuci.2018.11.003.
[15] R. Chaib, N. Azizi, N. Zemmal, D. Schwab, and S. B. Belhaouari, "Improved Multi-label Medical Text Classification Using Features Cooperation," in *Innovative System for Intelligent Health Informatics*, vol. 72, Springer, 2021, pp. 61–71.
[16] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021, doi: https://doi.org/10.1613/jair.1.12228.
[17] Z.-H. Min-Ling, "Review on multilabel classification models," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014.
[18] W. Liu, H. Wang, X. Shen, and I. Tsang, "The Emerging Trends of Multi-Label Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 8, pp. 1–21, 2021, doi: 10.1109/TPAMI.2021.3119334.
[19] J. Y. Park, Y. Hwang, D. Lee, and J. H. Kim, "MarsNet: Multi-Label Classification Network for Images of Various Sizes," *IEEE Access*, vol. 8, pp. 21832–21846, 2020, doi: 10.1109/ACCESS.2020.2969217.
[20] M. Jethanandani, A. Sharma, T. Perumal, and J.-R. Chang, "Multi-label classification based ensemble learning for human activity recognition in smart home," *Internet of Things*, vol. 12, p. 100324, 2020, doi: 10.1016/j.iot.2020.100324.
[21] P. Tirupattur, K. Duarte, Y. S. Rawat, and M. Shah, "Modeling Multi-Label Action Dependencies for Temporal Action Localization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1460–1470, doi: 10.1109/cvpr46437.2021.00151.
[22] S. M. Liu and J. H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1083–1093, 2015, doi: 10.1016/j.eswa.2014.08.036.
[23] M. Pushpa and S. Karpagavalli, "Multi-label Classification: Problem Transformation methods in Tamil Phoneme classification," in *7th International Conference on Advances in Computing & CommunicationsProcedia Computer Science*, 2017, vol. 115, pp. 572–579, doi: 10.1016/j.procs.2017.09.116.
[24] D. Rahmawati and M. L. Khodra, "Automatic multilabel classification for Indonesian news articles," in *International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2015, pp. 1–6, doi: 10.1109/ICAICTA.2015.7335382.
[25] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings ofthe Third Workshop on Abusive Language Online*, 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.
[26] S. Bi, P. Shi, Y. Du, B. Jin, and L. Yu, "Mining Knowledge within Categories in Global and Local Fashion for Multi-Label Text Classification," 2020, doi: 10.1109/IJCNN48605.2020.9207136.
[27] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *J. King Saud Univ. - Comput. Inf. Sci.*, 2020, doi: 10.1016/j.jksuci.2020.05.006.
[28] C. Luque, J. M. Luna, M. Luque, and S. Ventura, "An advanced review on text mining in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, pp. 1–16, 2019, doi: 10.1002/widm.1302.
[29] M. Adriani, J. Asian, B. Nazief, S. M. . Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A Confix-Stripping Approach," in *ACM Transactions on Asian Language*

Information Processing, 2007, vol. 6, no. 4, pp. 307–314, doi: 10.1145/1316457.1316459.
[30] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.
[31] M. Mohammedid and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, vol. 15, no. 3, pp. 1–21, 2020, doi: 10.1371/journal.pone.0230442.
[32] F. K. Khattak, S. Jeblee, C. Pou-prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of Word Embeddings For Clinical Text," *J. Biomed. Informatics X*, vol. 4, pp. 1–18, 2019, [Online]. Available: https://doi.org/10.1016/j.yjbinx.2019.100057.
[33] J. Santoso, E. I. Setiawan, C. N. Purwanto, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory," *Expert Syst. Appl.*, vol. 176, pp. 1–11, 2021, doi: 10.1016/j.eswa.2021.114856.
[34] M. A. Fauzi, "Word2Vec model for sentiment analysis of product reviews in Indonesian language," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, p. 525, 2019, doi: 10.11591/ijece.v9i1.pp525-530.
[35] M. L. Zhang, Y. K. Li, X. Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview," *Front. Comput. Sci.*, vol. 12, no. 2, pp. 191–202, 2018, doi: 10.1007/s11704-017-7031-7.
[36] R. Gupta and A. Kumar, "Indian sign language recognition using wearable sensors and multi-label classification," *Comput. Electr. Eng.*, vol. 90, pp. 1–13, 2021, doi: 10.1016/j.compeleceng.2020.106898.
[37] Z. Abdallah, A. El-Zaart, and M. Oueidat, "Comparison of multilabel problem transformation methods for text mining," in *5th International Conference on Digital Information and Communication Technology and Its Applications (DICTAP)*, 2015, pp. 115–118, doi: 10.1109/DICTAP.2015.7113182.
[38] R. Alazaidah and F. K. Ahmad, "Trending Challenges in Multi Label Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 127–131, 2016, doi: 10.14569/IJACSA.2016.071017.
[39] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, and A. Bahamonde, "Binary relevance efficacy for multilabel classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303–313, 2012, doi: 10.1007/s13748-012-0030-x.
[40] A. Y. Taha and S. Tiun, "Binary relevance (BR) method classifier of multi-label classification for arabic text," *J. Theor. Appl. Inf. Technol.*, vol. 84, no. 3, pp. 414–422, 2016.
[41] A. Fahmi, D. Purwitasari, S. Sumpeno, and M. H. Purnomo, "Performance evaluation of classifiers for predicting infection cases of dengue virus based on clinical diagnosis criteria," in *IES 2020 - International Electronics Symposium: The Role of Autonomous and Intelligent Systems for Human Life and Comfort*, 2020, pp. 456–462, doi: 10.1109/IES50839.2020.9231728.
[42] A. J. Wyner, M. Olson, J. Bleich, and D. Mease, "Explaining the Success of Adaboost and Random Forests as Interpolating Classifiers," *J. Mach. Learn. Res.*, vol. 18, pp. 1–33, 2017.
[43] H. Liu, X. Wu, and S. Zhang, "Neighbor selection for multilabel classification," *Neurocomputing*, vol. 182, pp. 187–196, 2016, doi: 10.1016/j.neucom.2015.12.035.
[44] J. Singh and R. Banerjee, "A study on Single and Multi-layer Perceptron Neural Network," in *3rd International Conference on Computing Methodologies and Communication, (ICCMC)*, 2019, pp. 35–40, doi: 10.1109/ICCMC.2019.8819775.
[45] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Inf. Process. Manag.*, vol. 54, no. 3, pp. 359–369, 2018, doi: 10.1016/j.ipm.2018.01.002.
[46] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

Safitri Juanita received a bachelor's degree in computer science from Budi Luhur University in 2004, a Master of Information Technology in 2009 from the University of Indonesia, and is currently a doctoral student at the Department of Electrical Engineering, Sepuluh Nopember Institute of Technology (ITS), Indonesia. Research Interest currently focus on text mining, such as Information Extraction, Natural Language Processing, and Information Retrieval, especially in medical texts.

Diana Purwitasari received a bachelor's degree in computer science from the Sepuluh Nopember Institute of Technology Surabaya, Indonesia, in 2001, a Master's degree in 2009 from Saga University, Japan, and a Doctoral degree in 2020 at the Department of Electrical Engineering, Sepuluh Nopember Institute of Technology (ITS), Indonesia. Her Research Interests focus on web mining, information retrival, social network analysis, computational intelligence.

I Ketut Eddy Purnama received a bachelor's degree in electrical engineering from Sepuluh Nopember Institute of Technology (ITS), Indonesia, a Master's degree in informatics engineering from the Institute Technology Bandung, and a Doctorate in biomedical engineering at the University of Groningen, the Netherlands. His research focuses on medical image analysis, microscopic image analysis, computer vision and images.

Abid Famasya Abdillah received a bachelor's degree in computer science from Politeknik Elektronika Negeri Surabaya Surabaya, Indonesia, in 2017, a master degree in 2022 from Department of Electrical Engineering, Sepuluh Nopember Institute of Technology (ITS), Indonesia. His research focuses on Data Intelligence, Health Tech.

Mauridhi Hery Purnomo received a bachelor's degree from the Sepuluh Nopember Institute of Technology (ITS), Indonesia, while a master's degree and doctorate from Osaka City University, Japan. He received a professor in 2003 from ITS, Indonesia, Expert in Robotics & Intelligent Systems. His research focuses on artificial intelligence and deep learning, robotic and signal processing, power systems, Power Systems, biomedical engineering.