






Ensemble-based Methods for Multi-label Classification on Biomedical Question-Answer Data

Abid Famasya Abdillah¹⁾ , Cornelius Bagus Purnama Putra²⁾ , Apriantoni³⁾ ,
Safitri Juanita⁴⁾ , Diana Purwitasari^{5)*} 

¹⁾²⁾³⁾⁵⁾ Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia

Jl. Teknik Kimia, Keputih, Sukolilo, Surabaya

¹⁾abid.famasya@gmail.com, ²⁾baguspurnama98@gmail.com, ³⁾apriantoninadiwaryo@gmail.com, ⁵⁾diana@if.its.ac.id

⁴⁾ Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia

Gedung A, B, C dan AJ, Kampus ITS, Sukolilo, Surabaya

⁴⁾safitri.juanita@budiluhur.ac.id

⁴⁾ Department of Information System, Universitas Budi Luhur, Indonesia

Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan, 12260. DKI Jakarta

⁴⁾safitri.juanita@budiluhur.ac.id

Abstract

Background: Question-answer (QA) is a popular method to seek health-related information and biomedical data. Such questions can refer to more than one medical entity (multi-label) so determining the correct tags is not easy. The question classification (QC) mechanism in a QA system can narrow down the answers we are seeking.

Objective: This study develops a multi-label classification using the heterogeneous ensembles method to improve accuracy in biomedical data with long text dimensions.

Methods: We used the ensemble method with heterogeneous deep learning and machine learning for multi-label extended text classification. There are 15 various single models consisting of three deep learning (CNN, LSTM, and BERT) and four machine learning algorithms (SVM, kNN, Decision Tree, and Naïve Bayes) with various text representations (TF-IDF, Word2Vec, and FastText). We used the bagging approach with a hard voting mechanism for the decision-making.

Results: The result shows that deep learning is more powerful than machine learning as a single multi-label biomedical data classification method. Moreover, we found that top-three was the best number of base learners by combining the ensembles method. Heterogeneous-based ensembles with three learners resulted in an F1-score of 82.3%, which is better than the best single model by CNN with an F1-score of 80%.

Conclusion: A multi-label classification of biomedical QA using ensemble models is better than single models. The result shows that heterogeneous ensembles are more potent than homogeneous ensembles on biomedical QA data with long text dimensions.

Keywords: Biomedical Question Classification, Ensemble Method, Heterogeneous Ensembles, Multi-Label Classification, Question Answering

Article history: Received 20 December 2021, decision after peer review 4 February 2022, accepted 23 February 2021, available online 28 April 2022

I. INTRODUCTION

The increasing information accessible via the Internet affects users' habits to find and answer questions in various domains, including the biomedical domain [1]. Although medical knowledge becomes more accessible, users often have to filter various results of their queries to find the precise information they were looking for [2]. A question-answer (QA) system could narrow down the results of users' queries. QA systems aim to give users precise answers to natural language questions [3]. A QA system generally consists of three modules [4]. The first module is question processing, allowing users to ask questions dynamically. This stage produces a set of representations and data categorization based on the questions asked [5]. The second module is candidate retrieval, which collects the document

* Corresponding author

relevant to the question. The last module is the answer processing, which evaluates and returns the most appropriate answer. Overall, this mechanism needs a deep analysis of user questions to extract the relevant information.

In the biomedical domain, a QA system is expected to understand user questions and provide satisfactory answers that help a diagnosis [6]. There is a risk of misdiagnosis if the generated answers are not precise. However, because many questions are inputted daily, it is not easy to categorize domains and understand questions. Therefore, a QA decision support system is needed to mark questions quickly and precisely based on historical QA data. At the question processing stage, a QA system often has problems narrowing down the domains to find answers. The sub-process that handles this mechanism is question classification (QC)—an essential process to provide information in the following QA stages to get an accurate answer [5]. The problem becomes more complex if the question has a diagnosis that refers to multiple entities (multi-label). A QA system can be rendered ineffective [7]. Therefore, a classification method is needed to accelerate the process.

Past studies have successfully overcome the multi-label classification problem with an ensemble learning approach to obtain more generalizable results [8]. This is because the ensemble method considers the different inductive biases in different models [9]. Chen et al. [10] applied the ensemble method for multi-label text classification with Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) to obtain the global and local semantic information from the corpora. Onan et al. [11] used various ensemble learning algorithms with base learner algorithms and text representations to solve text classification problems. Similarly, Onan [12] used ensemble learning algorithms for text genre classification based on feature engineering and language function analysis. The result shows that the ensemble model can improve text classification performance.

The research abovementioned used ensemble learning with homogeneous ensembles. None has examined heterogeneous ensembles for combining dissimilar models. Therefore, it is necessary to analyze a multi-label text classification using heterogeneous ensembles in long text dimensions. This study proposes ensemble learning with heterogeneous deep learning and machine learning ensembles for multi-label text classification in long text dimensions. We used several algorithms to evaluate heterogeneous ensemble learning: four machine learning algorithms (SVM, kNN, Decision Tree, and Naïve Bayes) and three deep learning algorithms (CNN, LSTM, and Transformers) as base learners. We used a non-factoid type of question from Indonesian biomedical QA systems with various word representations in the long text dimension. This study contributes to ensemble methods in deep learning and machine learning approaches for a multi-label text classification of biomedical questions. This study is expected to increase the accuracy matrix of long text data dimensions. Additionally, we evaluated the heterogeneous models of the ensemble method with machine learning-based homogeneous models and deep learning-based homogeneous models.

II. RELATED WORKS

Previous studies have implemented multi-label text classification in the biomedical domain. Liu [13] proposed joint learning from Label Embedding and Label Correlation (LELC) based on the correlation of labels and multi-layer attention. Text features extraction is performed with Bidirectional Gated Recurrent Unit Network (Bi-GRU). This mechanism has a reasonably good evaluation of multi-label data classification results, with an average Hamming Loss value of 0.02 and an average Micro-F1 value of 80% from ten testing datasets. Ibrahim et al. [14] proposed a Generic Hybridized Shallow Neural Network (GHS-NET) for multi-label text classification. They used GHS-NET leverages and CNN to extract the most discriminative features and the BI-LSTM layer to acquire local features of the text accurately. The result showed that GHS-NET enhances recall and F1-score by 6% and 1% for hallmarks of the cancer dataset, 16% and 11% for the chemical exposure dataset, 6% and 8% for the Intensive Care dataset (MIMIC-III).

As for ensemble learning, Onan [15] conducted a comparative analysis of feature sets and classifiers to categorize web pages. They adopted four different feature selections—information gain, consistency, correlation, and chi-square-based feature selections—and four ensemble learning methods, i.e., Boosting, Bagging, Dagging, and Random Subspace. Each method used four different base learners, namely Naive Bayes, C4.5, FURIA, and kNN algorithm. The result showed that AdaBoost and Naive Bayes algorithms obtained the best average predictive performance at 88.1%. The result indicated that Random Subspace and Bagging ensemble methods and consistency-based and correlation-based feature-selection methods achieved higher accuracy rates.

Meanwhile, Xia et al. [16] proposed a stacked ensemble approach that could simultaneously exploit label correlations and ensemble members' weights learning. They developed weighted stacked ensembles to facilitate the classifiers selection and ensemble members' construction for the multi-label classification. Subsequently, they used paired label correlations to determine the weights of the ensemble members and an accelerated proximal gradient algorithm and block coordinate derivation technique to reach an optimal solution of the ensemble mechanism. The results showed good accuracy, with an average value above 80% for some tested datasets.

In a text classification using the deep learning approach, Zulqarnain et al. [5] used Word2Vec to represent words in a vector space. They evaluate the impact of using various Word2Vec pre-trained word embeddings on several deep learning approaches. The result shows that Word2Vec models can learn semantic and syntactic relation of words efficiently, improving the performance of classification models significantly. In another study, Noh [17] used FastText as text representation to obtain the vector representations of *out of vocabulary* (OOV) tokens in the biomedical domain. The result shows that FastText performed the best in the come cases as biomedical word embeddings.

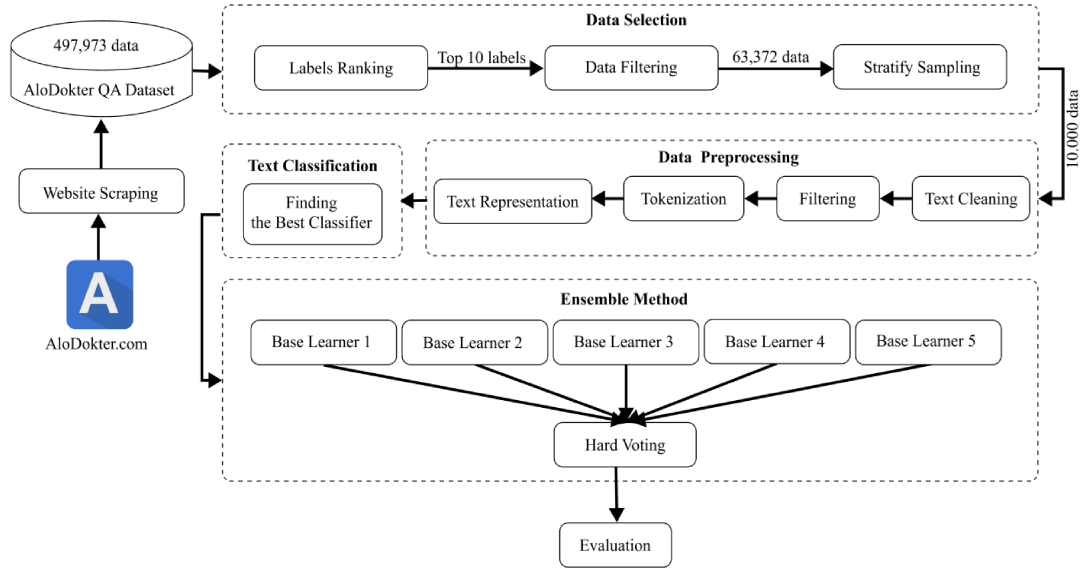


Fig. 1. A mechanism for multi-label classification with ensemble approach

III. METHODS

This section describes the methodological aspects of the proposed method. We focus on the implementation of ensemble methods to improve multi-label classification performance. An overview of the methodology is presented in Fig. 1.

A. Data Preparation

There are two main steps in data preparation. First, this study uses QA data collected from the Indonesian language-based health consultation forum on alodokter.com. BeautifulSoup and Scrapy libraries supported the data collection process from 2015 to 2021. This mechanism generated 497,973 data containing questions, answers, disease topics, and related metadata, such as question date and answer date. This study does not use problem-topic filtering in the data collection, so the results consisted of 1,214 topics. The topic from each question was used as a data label, so manual labeling was not required. For instance, a question asked, “I was dizzy and got diarrhea last night...”, has diarrhea and dizziness as a topic, so this question was automatically labeled as diarrhea and dizzy by default.

The next step was data selection, which started with an exploratory data analysis to find and remove duplicate questions and answers. This process resulted in 263,451 QA data. Then, we ranked the distinct labels from the whole dataset. A label ranking aimed to filter most frequently consulted diseases and remove labels outlier. This process was performed by calculating label frequency, with the top 10 labels selected as the filtering criteria. This mechanism resulted in 63,372 rows of data. Finally, we conducted a stratified data sampling to obtain 10,000 data for training and testing.

B. Data Pre-processing

The data pre-processing aimed to clean any irrelevant and noisy text to improve the model performance. For example, the question on the QA systems contains informal words. Therefore, this study's training dan testing data only use the answer data because the language is formal. As such, the data were of good quality and consistent for the next modeling process. The data processing was carried out in two stages. The first was text processing. This process used different preprocessing mechanisms in the machine learning-based and deep learning-based algorithms. In the machine learning-based algorithm, the data pre-processing was conventional: changing case folding, removing

punctuations, symbols, numeric, non-alphabet characters, and double-spaced characters. The data was then filtered by removing stopwords and tokenized to separate each word. Meanwhile, the deep learning-based algorithm does not need this. Previous research [14] reports that deep learning learns feature representation itself. Therefore, any characters could help deep learning learn better in the modeling processing. The pre-processing mechanism in deep learning is only double-spaced characters removal.

The second stage was to perform text representation to transform word tokenization into vector representations as to the input of machine learning algorithms. Various word representation models were used as a comparison to get the best text representation. This study used three-word representations: Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and FastText. TF-IDF is a mechanism to determine the word frequency in a document. It aims to obtain a series of dominant words following a modeling context [18]. This feature representation obtains feature extraction results adaptive to problems in a specific domain, such as medical data. We used TF-IDF to compare the word embedding approach as the baseline. Then, Word2Vec could represent words in a vector space. This process is used to efficiently learn the text semantic and syntactic relation to improve classification models' performance [5]. Finally, we used FastText to test how it deals with OOV problems, where words are present in testing data but not learned during the training phase [17]. These three feature representations are used to obtain an effective model with a good performance for multi-label data classification.

Every text representation was combined with machine learning-based in the classification process. We used 300 dimensions for word embedding. Then, we used the Gensim library to implement Word2Vec and FastText. These results were relevant data representations for the following modeling stage.

C. Classification

This study used two classification scenarios: single and ensemble. The single method evaluates the performance of each model individually, while the ensemble method combines multiple models to perform better than a single model [8]. In the single method, this study used three machine learning algorithms: Decision Tree, Support Vector Machine (SVM), kNN, and Naïve Bayes. The Decision Tree algorithm aims to handle imbalanced data. This algorithm uses hierarchical categorization to classify given data so that they can be classified correctly by their given properties [19]. SVM is effective in high-dimensional spaces that fit the feature representations. Meanwhile, kNN is a simple model, but it works well in arbitrarily complicated decision boundaries, and Naïve Bayes is a good baseline model.

The following classification scenario is an ensemble method with three deep learning algorithms: CNN, LSTM, and Transformers. CNN recognizes local patterns in a sequence by processing multiple words simultaneously and is suitable for text processing tasks [5]. LSTM is an advancement of RNN architecture to store long-term memory dependencies. LSTM also handles vanishing gradient problems, becoming a baseline in many sequential problems [5]. Transformers is the latest model that works well in many NLP-related tasks, such as text classification, sequence labeling, and multi-label classification [20]. Transformers use an attention mechanism that takes a whole chunk of sequential data to learn its feature representation. Finally, Bidirectional Encoder Representations from Transformers (BERT) is Transformers trained to a large number of texts and has been reported to reach state-of-the-art performance in various tasks [21].

D. Ensemble Method

The ensemble method is a set of learning models with individual predictions (single models) combined, so that component models resolve each other's weaknesses. The ensemble method has three approaches: Bagging, Boosting, and Stacking [9]. This research uses a Bagging method with a hard voting mechanism based on multiple classification models. Bagging is an ensemble method determined based on the average decision tree for regression cases or voting for classification cases.

Most ensemble method research uses the Bagging approach with homogeneous models, such as deep learning-based or decision-making-based, as a base learner. Past research rarely used heterogeneous models as a base learner. In this research, our model uses a heterogeneous model. We evaluated the variation in base learner numbers: three, five, seven, and nine. For decision making, we use Predict then Combine (PTC) in each model to better the performance of the ensemble method [22], which means each model yields its predicted label outputs, and then it is combined in an ensemble setting.

Additionally, we compare the heterogeneous models of the ensemble method with homogeneous models of the ensemble method. We used two homogeneous models of the ensemble method: machine learning-based and deep learning-based. The machine learning-based homogeneous models used a single model's three best base learners to overcome the limitation of the voting approach, which must be an odd number. Meanwhile, the deep learning-based homogeneous models used three deep learning methods (CNN, LSTM, and BERT).

E. Evaluation

In the evaluation stage, the effectiveness of both model classification scenarios was quantified by calculating the result of feature representation using different word insertions. Then, multi-label classification was run using a single model for each word embedding representation. Then, combinations of ensemble models were analyzed and compared to find the best composition. This study used four conventional evaluation metrics to compare the classification performance: accuracy, precision, recall, and Micro-F1 [13]. These metrics measure the de-facto classification problems, especially in single-label classification problems.

Because we carried out multi-label text classification, this study used two metrics to measure the model performance: Label Ranking Average Precision (LRAP) and Hamming Loss. LRAP is a metric for multi-label classification problems to assign better ranks to the topics associated with each repository [23]. The LRAP calculates the average of each ground truth label with an accurate ratio against the total labels with a lower score. It can be seen in (1).

$$LRAP(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{1}{\|y_i\|_0} \sum_{j:y_{ij}} \frac{|L_{ij}|}{rank_{ij}} \quad (1)$$

where variable y defines an actual label of a sample, while variable \hat{y} represents a predicted label of samples. Then, $n_{samples}$ is number of samples, L_{ij} is the predicted probability of labels compared to its rank where $L_{ij} = \{k: y_{ik} = 1, \hat{y}_{ik} \geq \hat{y}_{ij}\}$. The $rank$ is defined as $rank_{ij} = |\{k: \hat{y}_{ik} \geq \hat{y}_{ij}\}|$ to compute the cardinality of the set from the number of elements in a set. Finally $\|\cdot\|_0$ is the non-zero elements in a vector.

As example in our multi-label data, the first datum has actual labels [0, 1, 1, 1] with probability of predicted scores of [0.6, 0.1, 0.9, 0.7, 0.3]. In the second datum, actual labels are [1, 0, 1, 1] with predicted [0.7, 0.2, 0.6, 0.9]. First, LRAP calculates the sum of correct and incorrect predicted labels, then calculates the average scores of each predicted label with a weighted average strategy. The average score is obtained with regard to its class occurrence.

Meanwhile, Hamming Loss calculates XOR value of the incorrectly predicted labels and averages them across the dataset [23]. The calculation of this metric can be seen in (2).

$$Hamming Loss = \frac{1}{nl} \sum_{i=1}^N \sum_{j=1}^L [\hat{y}_j^{(i)} \neq y_j^{(i)}] \quad (2)$$

where variable n is the number of training samples, while $y_j^{(i)}$ is the true labels for the i training samples and the j class. Then, variable $\hat{y}_j^{(i)}$ is the predicted labels for the i training samples and j class. To calculate Hamming loss, the actual labels are [0, 1, 1, 1] and the predicted label are [1, 0, 1, 1]. The first stage of Hamming loss is calculated by XOR value or misclassification between actual labels with predicted labels. Then the sum of incorrectly predicted labels is divided by the number of samples. In other terms, a smaller Hamming loss score indicates better model performance.

IV. RESULTS

A. Data Preparation

At this stage, we obtained the top ten labels from the data selection results: menstruation, pregnancy, drugs, stomach acid, babies, headaches, bumps, skin, allergies, contraceptives. Then, we used these labels to obtain the sampling data. We used 10,000 sampling data from the top ten labels. The distribution of the top ten labels on our selected data is illustrated in Fig. 2. It shows that maternal and child health was the most frequently asked topic, followed by skin diseases and minor ailments such as stomach acid and headaches.

The number of label distribution can be seen in Table 1, where most data have single labels, followed by two and three labels. Since we used stratified sampling, these results also represent actual data found in the real-world scenario. The one-question-answer data is usually marked with one label, which means that the diagnosis is straightforward. Meanwhile, multiple labels may result in more complex scenarios with multiple diagnoses—in line with the findings that online health consultation tends to handle simple diseases [24]. Menstruation is a label often discussed in multi-label data. There are 538 data containing menstruation labels. Moreover, this label often appears together with the pregnancy and contraception labels. This result indicates that the three topics are often asked simultaneously in this data sample.

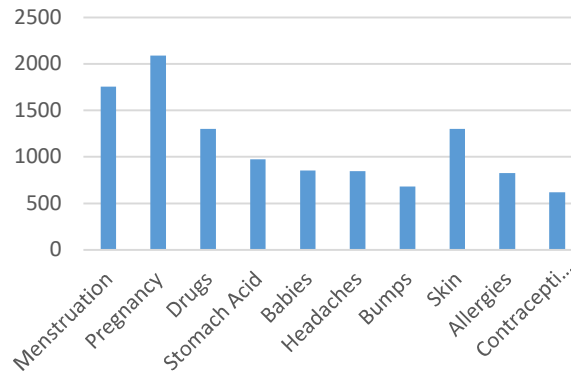


Fig. 2 Distribution of the top ten label

TABLE 1
 THE NUMBER OF LABEL DISTRIBUTION

Number of Labels Distribution	Total
1	8796
2	1164
3	40

B. Classification Results

In the first stage, we conducted a single model classification to find the best learners to solve multi-label classification. We used machine-learning-based and deep-learning-based approaches and four classification algorithms for the machine learning method (SVM, kNN, Decision Tree, and Naive Bayes) and three text representations (TF-IDF, Word2Vec, and FastText). These models and text representations from machine learning are combined into 12 models. Additionally, we used three algorithms in the deep learning method: CNN, LSTM, and BERT. We used a pre-trained model that implements the word embeddings layer as text representation. Finally, all of these base learners were determined based on the best performance of the single model. In technical classification, we split the data into training and testing data with a ratio of 90%:10%. In sum, we conducted multiple experiments to evaluate the best performant model among all scenarios.

TABLE 2
 THE RESULT OF SINGLE MODEL BASED ON MACHINE LEARNING APPROACH

Method	Text Representation	Precision	Recall	Micro-F1	Accuracy	LRAP	Hamming Loss
SVM	TF-IDF	0.85	0.73	79%	68.2%	0.774	0.0414
kNN		0.80	0.68	74%	62.1%	0.720	0.0544
Decision Tree		0.74	0.73	73%	64.7%	0.747	0.0647
Naïve Bayes		0.72	0.71	73%	51.1%	0.619	0.0659
SVM	Word2Vec	0.87	0.70	77%	64.6%	0.736	0.0455
kNN		0.82	0.69	75%	63.6%	0.716	0.0514
Decision Tree		0.46	0.48	47%	38.2%	0.504	0.1207
Naïve Bayes		0.64	0.77	70%	46.5%	0.650	0.0730
SVM	FastText	0.88	0.70	78%	66.1%	0.749	0.0437
kNN		0.80	0.66	73%	60.4%	0.704	0.0556
Decision Tree		0.46	0.49	48%	37.4%	0.504	0.1201
Naïve Bayes		0.60	0.77	67%	39.7%	0.606	0.0837
CNN	Word embeddings	0.84	0.77	80%	68.5%	0.781	0.0420
LSTM		0.82	0.71	76%	68.9%	0.757	0.0534
BERT		0.81	0.76	79%	70.5%	0.797	0.0444

Based on Table 2, in the machine-learning-based, the SVM algorithm with TF-IDF outperformed other algorithms. The highest F1-score of this approach (79%) is better than Word2Vec (77%) and FastText (78%). However, from its lowest measure, Word2Vec (47%) and FastText (48%) are substantially worse than TF-IDF (73%). In terms of algorithms, SVM performs best compared to other algorithms. The F1-score evaluation shows that SVM is consistently achieved the highest F1-scores in all text representations, followed by kNN and Naïve Bayes. Meanwhile, the least performant algorithm is the Decision Tree. Interestingly, this phenomenon is caused by Word2Vec and FastText as text representations. It indicates that both text representations do not help Decision Tree obtain good and

convergent performance results. On the other hand, this result also shows that the use of text representation based on word frequency could affect the performance of a single model better than the use of text representation based on word weights. In this case, TF-IDF yields better performance results than other text representations such as, such as Word2Vec and FastText, in a machine learning approach.

TABLE 3
 THE RESULT OF HETEROGENEOUS ENSEMBLES

Method	Base Learner	Precision	Recall	Micro-F1	Accuracy	LRAP	Hamming Loss
Machine Learning-based Homogeneous Ensembles	SVM-TF IDF, SVM-Word2Vec and SVM-FastText	0.88	0.72	79.2%	67.1%	0.757	0.0421
Deep Learning-based Homogeneous Ensembles	CNN, BERT, and LSTM	0.87	0.77	81.7%	71.5%	0.802	0.0384
Heterogeneous Ensembles	CNN, BERT and SVM- TF-IDF	0.87	0.78	82.3%	71.8%	0.805	0.0375

TABLE 4
 THE RESULT OF ENSEMBLE MODEL

Number of base learners	Base Learner	Precision	Recall	Micro-F1	Accuracy	LRAP	Hamming Loss
3	CNN, BERT, and SVM-TF IDF	0.87	0.78	82.3%	71.8%	0.805	0.0375
5	CNN, BERT, SVM-TF IDF, SVM-Word2Vec and SVM-FastText	0.88	0.74	80.8%	69.2%	0.778	0.0395
7	CNN, BERT, SVM-TF IDF, SVM-Word2Vec, SVM-FastText, LSTM, and kNN-Word2Vec	0.89	0.74	81%	69.7%	0.781	0.0389
9	CNN, BERT, SVM-TF IDF, SVM-Word2Vec, SVM-FastText, LSTM, kNN-Word2Vec, kNN-TF IDF, and Decision Tree-TF IDF	0.89	0.74	81.1%	69.5%	0.781	0.0386

In the deep learning approach, we found that deep learning methods had a good performance, with an average micro F1-score value above 75%. Overall, the deep learning approach obtains the top three of a single model. This result shows that the deep learning approach has more optimal results than the standard machine learning results. Specifically, CNN reached a higher score with a micro F1 score of 80% than LSTM, which obtained an F1 score of 76%. On the other hand, BERT obtains the highest accuracy scores among deep learning methods and all single-model methods with 79.7%. Then, only BERT reaches an F1 score of 79%, making CNN the preferred method for single labels. The micro F1-score is the balance between precision and recall score. Hence this metric is the most important in the multi-label scenario [13]. Moreover, the single model classification results show that misclassification cases often occur in skin labels with bumps or allergies labels and pregnancy labels with menstruation labels. It happens because of the possibility of using similar words in the questions. The results of the single model classification can be seen in Table 2.

C. Ensemble Model

After conducting a single model experiment, we carried out the ensemble model experiments using a hard voting scenario. There are three scenarios: machine-learning-based homogeneous ensemble, deep-learning-based homogeneous ensemble, and heterogeneous ensemble. Different numbers of base learners from the top three to the top nine were also studied for the heterogeneous ensemble. The result shows that a heterogeneous ensemble works best at combining three base learners (CNN, BERT, and SVM-TFIDF), compared to five, seven, and nine base learners. The precision scores of three learners are not better (87%) than nine learners (89%), seven learners (89%), and five learners (88%). However, due to the high recall value, it could achieve good overall F1-scores. With three base learners, the best F1-score is 82.3%, followed by nine base learners with 81.1%. In other metrics such as accuracy, LRAP score, and Hamming loss, we found that these three learners are also the most optimal. Based on the LRAP and Hamming loss score, the more base learners used, the less effective the results. Therefore, increasing the number of base learners in the model may not be a feasible option [25]. The results of this classification process can be seen in Table 3.

For other learning strategies in Table 4, heterogeneous-based ensembles with three learners (CNN, BERT, SVM with TF-IDF) obtain the best result with an F1 score of 82.3%, followed by deep learning and machine learning ensembles. It also can be seen that the combination of machine learning methods into deep learning yields a 0.6%

higher F1-score compared to homogenous deep learning. In contrast, homogenous ensemble yields insignificantly lower F1-score at 1.7% for machine learning and 0.6% for deep learning. Machine learning achieves the best result (88%). Nevertheless, the recall score of the heterogeneous ensemble generated the highest score at 78%. This finding is consistent with LRAP and Hamming Loss score that heterogeneous ensemble is better overall [23]. Using different models and decision fusion perspectives, the heterogeneous ensemble is more diverse and shows better generalization performance [25]. As in the single model classification, in the ensemble model, misclassification cases also occurred in many skin labels with bumps or allergies labels.

V. DISCUSSION

The biomedical QA system has become an integral part of health information sourcing for many Internet users. Therefore, a reliable QA system is needed to recognize questions. They tend to have more than one medical entity (multi-label) [7]. This study focuses on multi-label question classification using an ensemble method. The result shows that the ensemble method generates the highest Micro-F1 score, which means that this method effectively classifies questions in biomedical QA systems. Compared to the single models, the ensemble model could reach a 2.3% improvement in Micro-F1 score than CNN, considered the best single model. This is in line with the previous studies [9], [26], showing that ensemble models perform better than single models. However, previous studies mostly used a homogeneous ensemble, although a combination with a heterogeneous ensemble is possible. Our findings highlight that heterogeneous ensembles perform better than homogeneous ensembles. Combining multiple learners with dissimilar models could improve multi-label text classification on biomedical QA data.

The study also highlights that a high number of base learners does not guarantee a model's performance. Our experiment shows that three base learners work best than five, seven, and nine. The precision scores increase as the number of base learners increases. However, the recall value also significantly decreases. The false-negative rate is higher following a high number of base learners. Meanwhile, TF-IDF could consistently outperform Word2Vec and FastText, but this is because both are trained in the general domain, so they do not fit the medical domain. On the contrary, since TF-IDF is built from our training dataset, it could better learn medical word representation. However, TF-IDF could not handle the OOV problem, which Word2Vec and FastText can solve. Finally, this result also indicates that Word2Vec and FastText alone are suboptimal to be used in a specific domain.

This study has several limitations. First, our data consists of only 10,000 samples. Moreover, the number of label distribution is imbalanced with single labeled as the most dominant. Future research can benefit from more data and more balance labels distributions. Second, we did not use various text representations in the deep learning approach. We only used the word embedding layer as the text representation. Therefore, we suggest using other text representations to add the number of base learners. Third, our ensemble method is simple because it is only based on hard voting. We suggest future research using different ensemble approaches, such as Stacked ensemble and Boosting ensemble.

VI. CONCLUSIONS

We conducted experiments to find the best model for the multi-label text classification in biomedical QA systems. Our findings show that the ensemble models perform better than single models on biomedical QA data with long text dimensions. It can help the biomedical QA systems narrow down the question domains to get an accurate answer. A heterogeneous ensemble generates the best ensemble model with an F1-score of 82.3%, which is a 2.3% improvement from the best single model. In the single model, the best machine learning is by TF-IDF and SVM model, which outperforms the combination of Word2Vec and FastText. As a deep learning model, CNN produces a better performance than the machine learning model. It makes CNN the best single learner and heterogeneous model with three base learners as the best model overall.

Future research in dataset and model development can also be planned. Balance labels in a dataset are a crucial part of any model development. However, many multi-label datasets can improve this area so that extensive feature engineering can be incorporated into the framework. Moreover, finding all possible combinations of parameters and hyperparameters of models is a non-trivial task due to high computing power. Therefore, we need an optimal method to solve this task. The optimization method can yield better results in classifying biomedical QA data with long text dimensions.

Author Contributions: *Abid Famasya Abdillah*: conceptualization, data collection, methodology, writing – review & editing. *Cornelius Bagus Purnama Putra*: conceptualization, methodology, writing – original draft. *Apriantoni*: conceptualization, methodology, writing – original draft. *Safitri Juanita*: data curation, resources. *Diana Purwitasari*: supervision, resources, formal analysis.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

- [1] S. Liu, H. Wang, B. Gao, dan Z. Deng, "Doctors' Provision of Online Health Consultation Service and Patient Review Valence: Evidence from a Quasi-Experiment," *Inf. Manag.*, no. March 2019, hal. 103360, 2020, doi: 10.1016/j.im.2020.103360.
- [2] A. Ben Abacha dan P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Inf. Process. Manag.*, vol. 51, no. 5, hal. 570–594, 2015, doi: 10.1016/j.ipm.2015.04.006.
- [3] E. Dimitrakis, K. Sgontzos, dan Y. Tzitzikas, "A survey on question answering systems over linked data and documents," *J. Intell. Inf. Syst.*, vol. 55, no. 2, hal. 233–259, 2020, doi: 10.1007/s10844-019-00584-7.
- [4] M. A. Calijome Soares dan F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 6, hal. 635–646, 2020, doi: 10.1016/j.jksuci.2018.08.005.
- [5] M. Zulqarnain, R. Ghazali, M. G. Ghouse, N. A. Husaini, A. K. Z. Alsaedi, dan W. Sharif, "A comparative analysis on question classification task based on deep learning approaches," *PeerJ Comput. Sci.*, vol. 7, hal. 1–27, 2021, doi: 10.7717/PEERJ-CS.570.
- [6] N. Chen, X. Su, T. Liu, Q. Hao, dan M. Wei, "A benchmark dataset and case study for Chinese medical question intent classification," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. Suppl 3, hal. 1–7, 2020, doi: 10.1186/s12911-020-1122-3.
- [7] M. Wasim, M. N. Asim, M. U. Ghani Khan, dan W. Mahmood, "Multi-label biomedical question classification for lexical answer type prediction," *J. Biomed. Inform.*, vol. 93, no. March, hal. 103143, 2019, doi: 10.1016/j.jbi.2019.103143.
- [8] M. P. Sesmero, J. A. Iglesias, E. Magán, A. Ledezma, dan A. Sanchis, "Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles," *Appl. Soft Comput.*, vol. 111, hal. 107689, 2021, doi: 10.1016/j.asoc.2021.107689.
- [9] J. Kazmaier dan J. H. van Vuuren, "The power of ensemble learning in sentiment analysis," *Expert Syst. Appl.*, vol. 187, no. June 2021, 2022, doi: 10.1016/j.eswa.2021.115819.
- [10] G. Chen, D. Ye, Z. Xing, J. Chen, dan E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, hal. 2377–2383, 2017, doi: 10.1109/IJCNN.2017.7966144.
- [11] A. Onan, S. Korukoğlu, dan H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Syst. Appl.*, vol. 57, hal. 232–247, 2016, doi: 10.1016/j.eswa.2016.03.045.
- [12] A. Onan, "An ensemble scheme based on language function analysis and feature engineering for text genre classification," *J. Inf. Sci.*, vol. 44, no. 1, hal. 28–47, 2018, doi: 10.1177/0165551516677911.
- [13] H. Liu, G. Chen, P. Li, P. Zhao, dan X. Wu, "Multi-label text classification via joint learning from label embedding and label correlation," *Neurocomputing*, vol. 460, hal. 385–398, 2021, doi: 10.1016/j.neucom.2021.07.031.
- [14] M. A. Ibrahim, M. U. Ghani Khan, F. Mehmood, M. N. Asim, dan W. Mahmood, "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification," *J. Biomed. Inform.*, vol. 116, no. November 2020, hal. 103699, 2021, doi: 10.1016/j.jbi.2021.103699.
- [15] A. Onan, "Classifier and feature set ensembles for web page classification," *J. Inf. Sci.*, vol. 42, no. 2, hal. 150–165, 2016, doi: 10.1177/0165551515591724.
- [16] Y. Xia, K. Chen, dan Y. Yang, "Multi-label classification with weighted classifier selection and stacked ensemble," *Inf. Sci. (Nij.)*, vol. 557, hal. 421–442, 2021, doi: 10.1016/j.ins.2020.06.017.
- [17] J. Noh dan R. Kavuluru, "Improved biomedical word embeddings in the transformer era," *J. Biomed. Inform.*, vol. 120, hal. 103867, Agu 2021, doi: 10.1016/j.jbi.2021.103867.
- [18] M. Mohammedid dan N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, vol. 15, no. 3, hal. 1–21, 2020, doi: 10.1371/journal.pone.0230442.
- [19] K. Pliakos dan C. Vens, "Mining features for biomedical data using clustering tree ensembles," *J. Biomed. Inform.*, vol. 85, hal. 40–48, Sep 2018, doi: 10.1016/j.jbi.2018.07.012.
- [20] Z. Shaheen, G. Wohlgenannt, dan E. Filtz, "Large Scale Legal Text Classification Using Transformer Models," Okt 2020, [Daring]. Tersedia pada: <http://arxiv.org/abs/2010.12871>.
- [21] R. Wang, R. Ridley, X. Su, W. Qu, dan X. Dai, "A novel reasoning mechanism for multi-label text classification," *Inf. Process. Manag.*, vol. 58, no. 2, hal. 102441, Mar 2021, doi: 10.1016/j.ipm.2020.102441.
- [22] V. L. Nguyen, E. Hüllermeier, M. Rapp, E. Loza Mencía, dan J. Fürnkranz, "On Aggregation in Ensembles of Multilabel Classifiers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12323 LNAI, hal. 533–547, 2020, doi: 10.1007/978-3-030-61527-7_35.
- [23] M. Izadi, A. Heydarnoori, dan G. Gousios, "Topic recommendation for software repositories using multi-label classification algorithms," *Empir. Softw. Eng.*, vol. 26, no. 5, hal. 1–33, 2021, doi: 10.1007/s10664-021-09976-2.
- [24] F. Zhou *et al.*, "Online Clinical Consultation as a Utility Tool for Managing Medical Crisis During a Pandemic: Retrospective Analysis on the Characteristics of Online Clinical Consultations During the COVID-19 Pandemic," *J. Prim. Care Community Heal.*, vol. 11, 2020, doi: 10.1177/2150132720975517.
- [25] M. A. Ganaie, M. Hu, M. Tanveer*, dan P. N. Suganthan*, "Ensemble deep learning: A review," 2021, [Daring]. Tersedia pada: <http://arxiv.org/abs/2104.02395>.
- [26] Z. Sun, C. Wang, Y. Zhao, dan C. Yan, "Multi-Label ECG Signal Classification Based on Ensemble Classifier," *IEEE Access*, vol. 8, hal. 117986–117996, 2020, doi: 10.1109/ACCESS.2020.3004908.

Publisher's Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.