A Comparative Study of Multi-Label Classification for Document Labeling in Ethical Protocol Review

Rizka Wakhidatus Sholikah[1], Diana Purwitasari[2], Mohammad Zaenuddin Hamidi[3]
[1]Department of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia 60111
[2,3]Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia 60111
E-mail: [1]wakhidatus@its.ac.id, [2]diana@if.its.ac.id

Abstract

An ethical clearance document ensures that the research will protect the subject in accordance with existing ethical principles. The ethical clearance is issued by the Research Ethics Commission (KEP). KEP will conduct a review of the proposed ethical protocol based on the seven standards contained in a protocol. The review process is done manually by KEP. This process often creates bottlenecks in research due to the large number of protocols that must be reviewed, so that the process to get ethical clearance takes a long time. This can affect the setback in the schedule of the research process. Therefore, in this research, a comparative study was conducted on the problem of multi-label classification to automate the ethical protocol review process. Automation of the labeling process can increase the effectiveness of the review process because it can provide an overview to the reviewer regarding the label of a document before conducting a more in-depth review process. The experiment results show that the use of the traditional machine learning approach produces better performance than the deep learning approach. The machine learning method with the best results is Naïve Bayes+BoW with precision, recall, and F-score values of 0.76, 0.80, and 0.78, respectively.

Keywords: Ethical protocol, multi-label classification, automatic labeling.

1. INTRODUCTION

Research in healthcare that includes humans as subjects must have ethical clearance document. The ethical clearance document aims to anticipate research that contains physical, social, and economic risks to the research's subject. In addition, healthcare research must respect and protect research subjects' life, health, privacy, and dignity [1] [2]. Furthermore, research can be accepted if it can be proven by a valid scientific method. In Indonesia, the regulation for research in healthcare are contained in *Undang-Undang No. 36 Tahun 2009 tentang kesehatan* (Health Law No. 36/ 2009) [3] and *Peraturan Pemerintah No. 39 Tahun 1995* (Government Regulation No. 39/1995) [4].

The issuance of ethical clearance was carried out after the researcher submitted the ethical protocol to the Research Ethics Commission (KEP). KEP will conduct a review of the proposed ethical protocol. The review of the ethical protocol is based on the International Ethical Guidelines for Biomedical Research Involving Human Subjects by the Council for International Organizations of Medical Sciences (CIOMS) [5]. The ethical guidelines consist of 25 guidelines that must be reviewed on the proposed ethical protocol. In Indonesia, the 25 guidelines are simplified into seven standards: social value, scientific value, expenses and benefits, risk, inducements, privacy protection, and informed consent [1]. There is a checklist form that is used to manually check each standard in the ethical protocol. From the results of the quick review, a decision can be drawn whether a protocol is exempt, expedited, and full board [6] [7] [8].

Along with the development of research in healthcare, the demand for ethical clearance is also increasing. The process of reviewing ethical protocol by KEP is carried out manually for each document to ensure that each research complies with the guidelines and standards.

Consequently, when demand increases, the time required to issue ethical clearance also increases, which affects the postponement of the research schedule—considering that research in the health sector cannot be carried out without ethical clearance.

These problems can be overcome by boosting the effectiveness and efficiency of the review process. One of the solutions is utilizing computer technology to automate the process. The automation of the review process provides an overview to reviewers regarding the content of the proposed protocol before performing in depth review. Moreover, it also reduces time and subjectivity in the review process. Meanwhile, it should be noted that the automation process is not intended to replace the reviewer's role but as a tool to support the decision-making process.

The ethical protocol document is used as a data source to build the automation model. The document contains a research summary consisting of the framework, objectives, benefits, and results to be achieved in the research. The document is a text with an unstructured format. Unstructured text is quite challenging to be processed. Special approach in Natural Language Processing (NLP) is needed to process the ethical protocol. Several studies applied NLP approach in unstructured text to build models. Metadata such as abstract, title, and keywords are used to build a multi-label classification model for research articles [9]. Pradhan et al. [10] use abstracts and titles of scientific articles to automatically find the suitable reviewer by performing topic detection. In law, legal documents are utilized to find the risks of legal actions [11]. Another challenge in this research is the limited number of labeled datasets. Meanwhile, the ethical protocol document dataset is confidential data that is not easy to collect.

The ethical clearance proposal begins with the researcher submitting an ethical protocol. Furthermore, each proposed protocol will be reviewed for compliance with the seven standards and labeled based on the suitable standard. The mapping of the seven standards is used to determine the decision of a protocol, whether it is approved or requires revision. Each document can have more than one standard, so the multi-label approach is suitable for this problem.

Several studies have been conducted for multi-label text classification. In a study conducted by Sari et al. [12], multi-label classification is solved by a deep learning approach. The method used is Long Short Term Memory (LSTM) with Word2Vec. The results show that LSTM and Word2Vec can outperform other state-of-the-art methods [12]. Research on patient notes data on multi-label classification was carried out using several algorithms, including Support Vector Machine (SVM), Convolutional Neural Network (CNN), Continuous Bag-of-Words (CBoW), and Bidirectional Gated Recurrent Unit with Hierarchical Attention Mechanism (HA-GRU) [13]. In this study, tokenization stages were also carried out on input text to improve the classifier's performance. Multi-label classification on text-social media is done to determine the topic of content or post. Hate speech can also be detected through classification on social media content. Omar et al. [14] use traditional machine learning methods such as Random Forest, Decision Tree, and Linear SVC. Multi-label classification with EUR LEX legal document data and MEDLINE medical abstract data was solved using a one-vs-rest binary relevance (BR) approach [15]. The problem of multi-label classification is solved by using the deep learning approach by Pal et al. [16]. Bidirectional LSTM (Bi-LSTM) and Graph Attention Network (GAT) are used to build an end-to-end architecture. Isnaini et al. [17] perform multi-label classification to predict topics from Indonesian news documents. This research uses Term Frequency-Inverse document frequency (TF-IDF) as a feature extractor, while K-Nearest Neighbor (KNN) is used as a classifier. Previous studies show that traditional machine learning and deep learning approaches have advantages depending on the data used.

Therefore, we conduct a comparative study in multi-label classification problems for ethical protocol documents. The multi-label classification method is proposed to perform labeling of ethical protocol documents automatically. The proposed strategy applies several multi-label classification methods, including traditional machine learning and current method of deep learning. We also perform several approach in NLP to process the unstructured text documents. In this study, data enrichment was carried out to overcome the problem of data limitations and improve the performance of the model.

2. RESEARCH METHOD

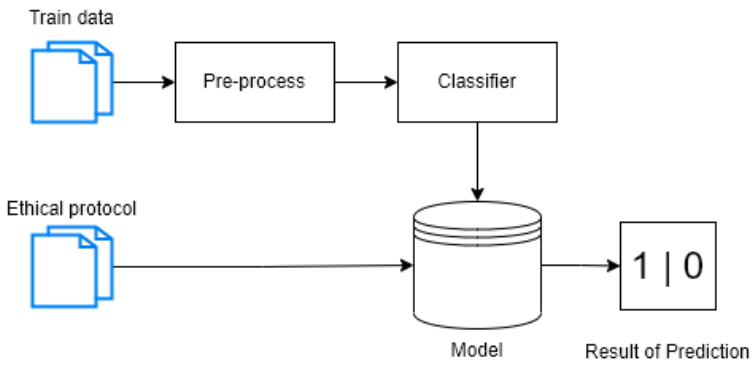


Figure 1. Step-by-step multi-label classification for ethical protocol review

In this study, multi-label classification was carried out through two main processes: pre-processing and classification, as shown in Figure 1. The detail of each process will be explained in the next sub-chapter.

2.1 Pre-process

Preprocess is the stage to prepare the dataset before entering into the classification method. In this study, several stages were carried out: case folding, tokenization, stopword removal, and vectorizer. Case folding is a process to change characters into lower form (lowercase). The output of the case folding process is tokenized to separate terms from a sentence or paragraph. The term in this context is a single word (not a phrase). The next step is stopword removal, which removes words with a high frequency and has no value when used as features. Examples of terms in Indonesian that are included in stop words are 'saya (I),' 'yang (which),' 'pada (on),' and so on. Table 1 shows an example of the pre-processing. In Table 1, the blue color shows the difference after case folding, while the red color shows the deleted word when performing stopword removal.

The next step is to vectorize the output of the stopword removal stage. Vectorization is a process to convert terms into vectors that computers can understand. In addition, the vectorizer also acts as a feature extraction before entering the classification algorithm. This study uses TF-IDF and BoW vectorizer in the traditional machine learning approach. The deep learning approach uses the embedding layer and GloVe [18] as vectorizers.

Table 1. Example of pre-process

Document	*Virus corona telah menyebar dari China ke ratusan negara lainnya dalam kurun waktu kurang dari 2 bulan. Di masa pandemi seperti ini, sistem kesehatan yang kuat dan metode penanganan oleh pemerintah menjadi salah satu faktor yang dapat meminimalisasi penyebaran. Akan tetapi, tidak semua negara memiliki pondasi yang kuat pada sistem kesehatannya. Banyak negara yang mengalami krisis tenaga kesehatan dikarenakan tingginya pasien aktif COVID-19 yang harus dirawat di Rumah Sakit.*
Case folding	*virus corona telah menyebar dari china ke ratusan negara lainnya dalam kurun waktu kurang dari 2 bulan. di masa pandemi seperti ini, sistem kesehatan yang kuat dan metode penanganan oleh pemerintah menjadi salah satu faktor yang dapat meminimalisasi penyebaran. akan tetapi, tidak semua negara memiliki pondasi yang kuat pada sistem kesehatannya. banyak negara yang mengalami krisis tenaga kesehatan dikarenakan tingginya pasien aktif covid-19 yang harus dirawat di rumah sakit.*
Tokenization	*virus corona telah menyebar dari china ke ratusan negara lainnya dalam kurun waktu kurang dari 2 bulan. di masa pandemi seperti ini, sistem kesehatan yang kuat dan metode penanganan oleh pemerintah menjadi salah satu faktor yang dapat meminimalisasi penyebaran. akan tetapi, tidak semua negara*

	memiliki pondasi yang kuat pada sistem kesehatannya. banyak negara yang mengalami krisis tenaga kesehatan dikarenakan tingginya pasien aktif covid-19 yang harus dirawat di rumah sakit.
Stopword removal	*virus corona menyebar china ratusan negara kurun waktu kurang 2 bulan. masa pandemi, sistem kesehatan kuat metode penanganan pemerintah menjadi salah satu faktor dapat meminimalisasi penyebaran. tetapi, tidak semua negara memiliki pondasi kuat sistem kesehatannya. banyak negara mengalami krisis tenaga kesehatan dikarenakan tingginya pasien aktif covid-19 harus dirawat rumah sakit.*

2.2 Classifier

This research uses two approaches to solving multi-label classification problems: the traditional machine learning approach and the deep learning approach. The SVM, Random Forest (RF), KNN, and Naive Bayes (NB) methods are used in the traditional machine learning approach. In the deep learning approach, the basic Deep Neural Network (DNN), Convolutional Neural Network (CNN), Long Short Term Memory (LSTM), and Bi-Directional LSTM (Bi-LSTM) methods are used.

SVM

SVM is a classification method that utilizes a hyperplane as a decision boundary between two classes. The decision boundary maximizes the distance between classes based on the support vector. Support vector is the outermost data object that is closest to the hyperplane. An illustration of SVM can be seen in Figure 2. However, because of the behavior of SVM, SVM cannot be directly used for multi-label cases. The reason is that the decision boundary in SVM is a hyperplane that is only suitable for binary classification. In this study, the Scikit-learn MultiOutputClassifier library was used to overcome this problem.

NB

Naive Bayes is a classification method that uses Bayes' theorem by assuming independence between features. The Naive Bayes method is easy to create and implement and is suitable for large data. In the Naive Bayes classification, the Bayes theorem calculates the posterior probability $P(c|x)$. Given c as a class and x as a feature, the calculation of the posterior probability $P(c|x)$ using Bayes' theorem can be calculate by the Equation (1).

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad , \tag{1}$$

where $P(c)$ is the prior probability of class, $P(x|c)$ is the likelihood between feature and class, and $P(x)$ is the probability of feature x.

RF

Random forest is an ensemble technique that consists of a collection of several decision trees [19]. RF creates final decisions based on majority voting or averaging predictions. The majority voting makes decisions by choosing the most class from the results of each decision tree. For example, if an RF has five decision trees, three decision trees predict an instance belongs into class A, and two predict class B. The final result of RF is class A.

K-NN

KNN assumes that instances belonging to the same class are in adjacent positions in the vector space. So that the data points of a similar instance are adjacent to each other, based on these assumptions, the K data points that have the highest similarity or the smallest distance to the target data point belong to the same class as the target data point. In KNN, the initial data point (target data point) is taken randomly as many as the classes. Then calculate the distance between the initial data point to other data points. The last step is to take the closest K data points

as class members.

In this research, the distance calculation used is Minkowski distance. Minkowski distance used to calculate the distance between two data points in high dimensional space. Let *X* is the first data point that has n dimension (x_1, x_2, x_3, …, x_n) and *Y* is the second data point that also has *n* dimension (y_1, y_2, y_3, …, y_n). Then, the distance can be calculated by following (2).

$$M(X,Y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \ , \quad (2)$$

where *p* is the order of the data point.

DNN

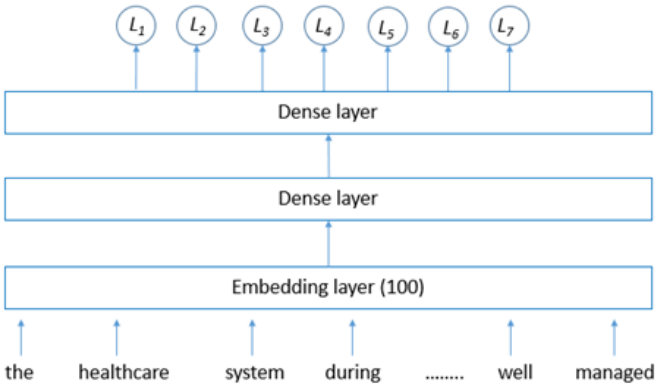
DNN is a neural network architecture that has more than one hidden layer. The architecture of DNN used can be seen in Figure 3. Each layer in the DNN is a sequence layer in which the output of layer *i* becomes the input of layer *i*+1. Input from DNN is abstract from an ethics protocol. Furthermore, the term from the abstract will enter the embedding layer to be changed into its vector representation. The output of the embedding layer becomes the input to the dense layer, which functions to extract essential features in the training process. Furthermore, the output layer uses seven neurons based on the number of labels in this study and the sigmoid as an activation function. In the training process, binary cross-entropy is used as the loss function and Adam as the optimizer.

CNN

CNN is one of the architectures of the Artificial Neural Network (ANN) in which there are convolutional and pooling layers. The convolutional layer consists of filters or kernels to identify local predictors. Usually, a convolutional layer is applied to the embedding matrix in text processing. This study used a filter with a length of 3 and stride 1. A convolutional layer is followed by a max-pooling layer to produce a feature map consisting of the most preminent features. Max pooling takes the maximum value of a patch for each feature map. In addition, max-pooling aims to reduce the spatial size of the vector representation, thereby reducing parameters and computations. The next layer is a non-linear fully connected layer with a dropout of 0.5. The output layer consists of seven neurons with a sigmoid as the activation function. The architecture of CNN for text classification can be seen in Figure 4.

LSTM

LSTM is the development of a Recurrent neural network (RNN). RNN is a feed-forward network architecture that has a recurrent hidden state. RNN can remember the previous information for use in the current process. However, RNN has a problem where the information from long-term dependencies will be lost due to the vanishing gradient. LSTM can overcome this problem by replacing the self-connected hidden layer with a memory block. In the memory block, there is a mechanism that controls when to learn new information and when to forget old information. The illustration of the LSTM unit can be seen in Figure 5.



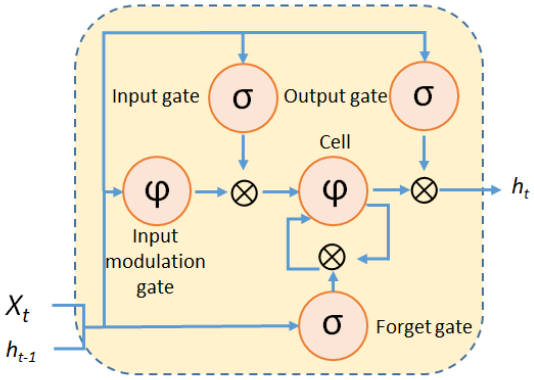


Figure 4. The illustration of LSTM cell

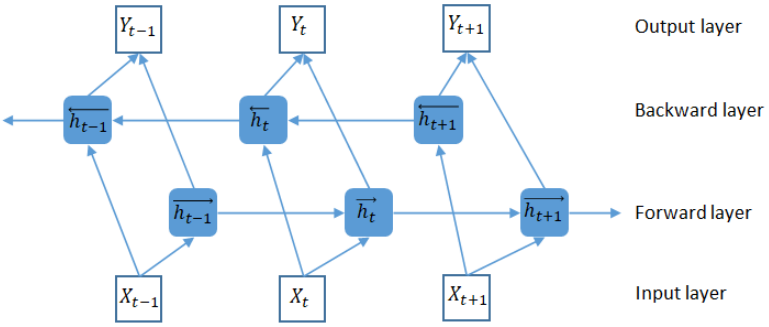


Figure 5. Overview of the Bi-LSTM architecture

Figure 2. The DNN Architecture

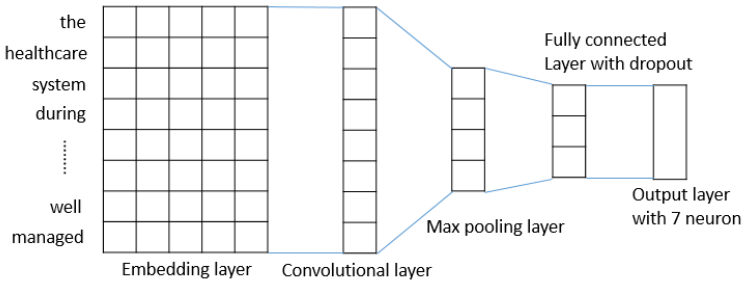


Figure 3. The DNN architecture

Given X_t as the input of the LSTM at time t, the hidden state h_t can be calculate by following (3) - (8).

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (3)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (4)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (5)$$
$$g_t = \varphi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (6)$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \quad (7)$$
$$h_t = o_t \otimes \varphi(c_t) \quad , \quad (8)$$

where i, f, c, o represents input gate, forget gate, cell activation, and output gate, respectively. The W refers to self-updated weight of the hidden layer and b is the bias. The $\sigma(.)$ and $\varphi(.)$ are sigmoid and *tanh* activation function, respectively. The $\otimes$ operator refers to element-wise multiplication. In LSTM, all the gate values and hidden layer outputs have values between [0,1].

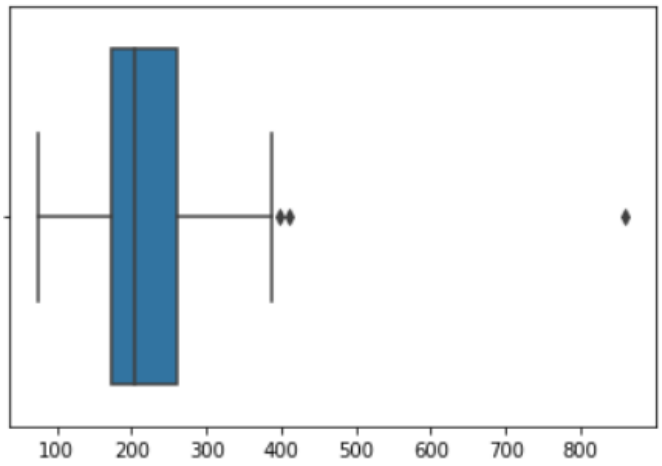


Figure 6. The length distribution of ethical document collections

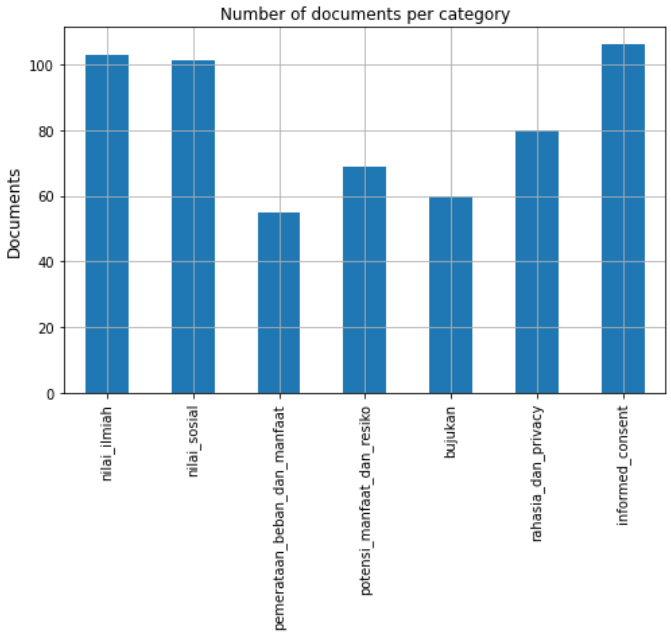


Figure 7. The label distribution

Bi-LSTM

Bi-LSTM (bi-directional LSTM) is an extension of the vanilla LSTM. LSTM has an architecture of forward sequence where information from time t-1 will affect the process at time t. The idea of Bi-LSTM is to create sequences from two directions: forward and backward. Furthermore, the forward and backward outputs must be combined before being parsed into the next layer. Combinations can be done with '*sum*' (the output is added), '*mul*' (the output is multiplied), '*concat*' (the output is concatenated, the number will be doubled), and '*ave*' (the average of the outputs). Figure 6 shows an overview of the Bi-LSTM.

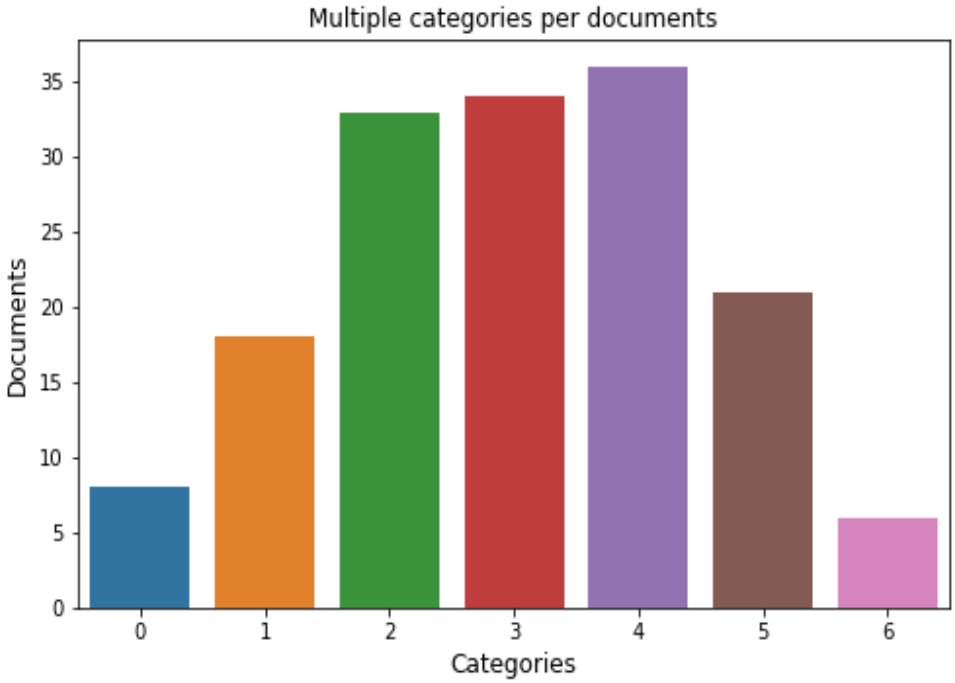


Figure 8. Multi-label documents distribution

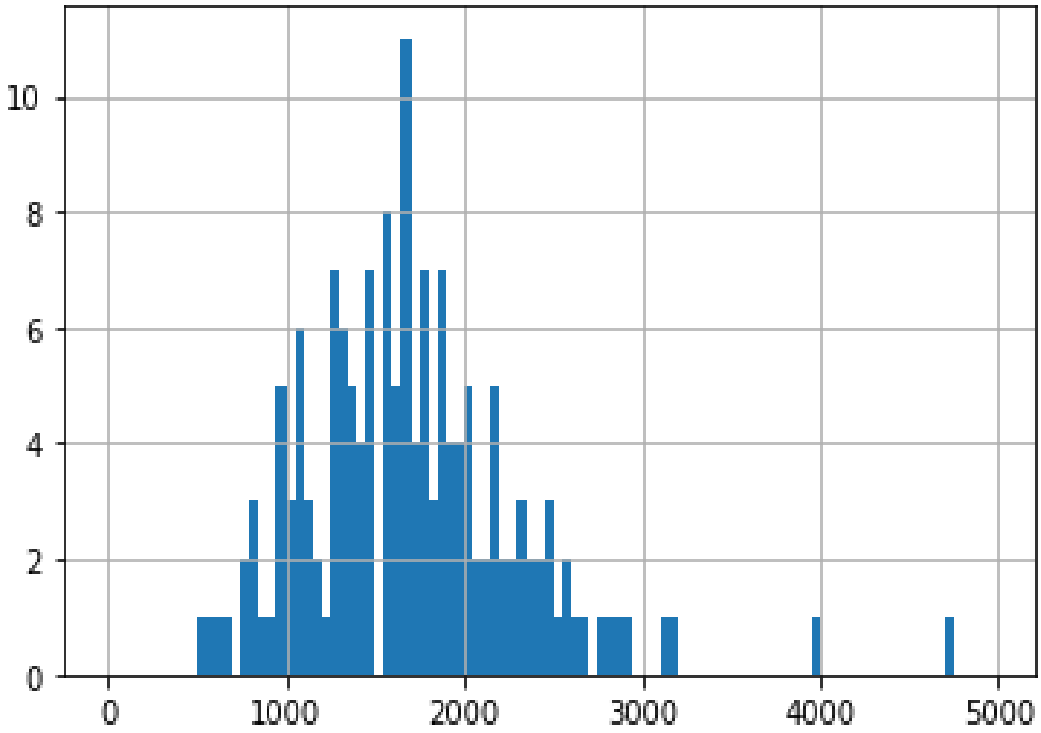


Figure 9. The length distribution of additional dataset

3. RESULT AND ANALYSIS

a. Data

In this study, the data used is the 22 ethical protocol documents from KEP. The length distribution of ethical protocol documents can be seen in Figure 7. The shortest document consists of 80 tokens, and the longest is 400 tokens, while the average document has a length of 180 to 260 tokens.

The number of ethical protocol collections is very limited to produce a robust model. However, the ethical protocol document is confidential, so it is not easy to obtain. Therefore, we create a scenario in the experimental process by obtaining additional data to increase the model. We hypothesized that the data from a similar domain could potentially increase the model performance. The additional data are collected from theses on medicine, nursing, and public health. The total of this additional data is 200 documents. Figure 8 shows the distribution of documents for each standard: *nilai ilmiah* (scientific value), *nilai sosial* (social value), *pemerataan beban dan manfaat* (distribution of expenses and benefits), *potensi manfaat dan risiko* (potential risk), *bujukan* (inducements), *rahasia dan privacy* (privacy protection), and informed consent.

The distribution of the number of labels for each document can be seen in Figure 9. document collection shows a normal distribution where the dominating document has labels between 3 to 5. From Figure 9, 0 means a document has only one label, 1 means a document has two labels, and so on. Meanwhile, the length distribution of the document collection is shown in Figure 10.

b. Results

This study was conducted to compare traditional machine learning methods with deep learning methods to solve multi-label text classification problems. There are three types of scenarios related to the data enrichment. The first scenario is training without additional data. The second scenario uses only additional data in the training process. The last scenario is to use a combination of additional data with ethical protocol data to conduct the training process.

In the first scenario, the 22 documents owned are separated between training and testing data with a ratio of 60:40. The first scenario tests the model's performance with extremely limited training data. Several machine learning methods are used: SVM, RF, NB, and KNN. In the machine learning approach, BoW and TFIDF methods are applied to extract features from text documents. In the deep learning approach, we used DNN, CNN, LSTM, and Bi-LSTM as comparison. As a feature extractor, the Glove method on LSTM is used, while in DNN, CNN, Bi-LSTM, we added the embedding layer. The result of the first scenario can be seen in Table 2. In the first scenario, it can be seen that the SVM+BoW method obtains the highest result with a value of 0.9. The NB+TFIDF return the highest score in F1-score with 0.78. For recall, the highest value is 0.87 using the CNN method.

In the second scenario, the data from the thesis document is used as training data. Meanwhile, we used the same test data as in the first scenario for data testing. The second scenario intends to determine the effectiveness of using other documents in the same field to transfer knowledge in the learning process. The test results can be seen in Table 3. The experiment results show that in the second scenario, the highest precision, recall, and F-score values are using the NB with TFIDF method with 0.70 each.

In the third scenario, additional data and ethical protocol data are combined as training data. While the data testing still uses the same data as the data in scenarios one and two. The third scenario has a goal to measure the effect of adding additional data (different documents with similar topics) to the original training data (ethical protocol). The results of the trials in Table 4 show that the highest precision and F-score values come from the KNN+TFIDF method with 0.74 and 0.72, respectively. For recall, the highest result was obtained from LSTM with 0.74.

Table 2. The results of the first scenario

Methods	Feature extractor	Precision	Recall	F1-Score
Machine Learning				
SVM	TFIDF	0.90	0.20	0.32
	BoW	**0.92**	0.24	0.38
RF	TFIDF	0.79	0.57	0.66
	BoW	0.74	0.50	0.60
NB	TFIF	0.74	0.61	0.67
	BoW	0.76	0.80	**0.78**
KNN	TFIDF	0.77	0.78	0.77
	BoW	0.76	0.48	0.59
Deep learning				
DNN	Embedding layer	0.74	0.63	0.68
CNN	Embedding layer	0.67	**0.87**	0.75
LSTM	Embedding layer	0.72	0.67	0.69
	Glove	0.80	0.52	0.63
Bi-LSTM	Embedding layer	0.69	0.76	0.72

Table 3. The results of the second scenario

Methods	Feature extractor	Precision	Recall	F1-Score
Machine Learning				
SVM	TFIDF	0.67	0.63	0.65
	BoW	0.67	0.63	0.65
RF	TFIDF	0.66	0.63	0.64
	BoW	0.67	0.65	0.66
NB	TFIDF	**0.70**	**0.70**	**0.70**
	BoW	0.67	0.61	0.64
KNN	TFIDF	0.68	0.61	0.64
	BoW	0.65	0.67	0.66
Deep learning				
DNN	Embedding layer	0.65	0.61	0.63
CNN	Embedding layer	0.64	0.61	0.62
LSTM	Embedding layer	0.65	0.67	0.66
	Glove	0.64	0.65	0.65
Bi-LSTM	Embedding layer	0.62	0.67	0.66

Table 4. The results of the third scenario

Methods	Feature extractor	Precision	Recall	F1-Score
Machine Learning				
SVM	TFIDF	0.68	0.65	0.67
	BoW	0.67	0.63	0.65
RF	TFIDF	0.64	0.63	0.64
	BoW	0.66	0.63	0.64

NB	TFIF	0.70	0.70	0.70
	BoW	0.67	0.61	0.64
KNN	TFIDF	**0.74**	0.70	**0.72**
	BoW	0.67	0.65	0.66
Deep learning				
DNN	Embedding layer	0.67	0.67	0.67
CNN	Embedding layer	0.67	0.65	0.66
LSTM	Embedding layer	0.64	**0.74**	0.69
	Glove	0.67	0.63	0.65
Bi-LSTM	Embedding layer	0.67	0.61	0.64

c. Analysis

The first scenario results show that the use of very small amounts of data is still able to produce a model that has high precision compared to the use of data enrichment. However, the recall value is relatively low. This indicates that the resulting model can only retrieve a few relevant documents, but most of them have the correct label. The test results in scenario one also show that the deep learning method does not increase performance. Deep learning produces good performance when trained with large amounts of data, so the results are not satisfactory in this scenario.

The second scenario shows that it is possible to use data from another domain (thesis document) as training data. However, the precision values for all methods have decreased compared to when using training data from the ethical protocol alone, even though the amount is small. The decrease in precision values occurs in machine learning and deep learning methods. That decrease may be due to the different characteristics between the ethical protocol document and the thesis document used as training data. The different characteristic leads to the resulting model not being able to adequately accommodate the features contained in the ethical protocol so that the quality of the model declines. On the other hand, there was a significant increase in recall values from several machine learning methods such as SVM, RF, and NB+TFIDF. The test results show that when using additional data sets for the training process, they are able to balance precision and recall values.

The third scenario combines additional data (thesis documents) and original data (ethical protocol) in the training process. The results show that the use of combined data produces better results in almost all methods than thesis data alone. However, the resulting precision is smaller than the first scenario. From scenarios one, two, and three, it can be concluded that adding data from other domains with the same topic is possible even though the precision results are not as good as when using ethical protocol data alone.

4. CONCLUSION

Automatic labeling of documents in the ethical protocol review can increase effectiveness and efficiency. Automatic labeling can be done using various methods for multi-label classification, both machine learning, and deep learning approaches. The experiment results show that the use of deep learning methods on very small ethical protocol data does not have a positive effect compared to traditional machine learning methods. The method that produces the best performance is Naive Bayes with precision, recall, and F-score 0.76, 0.80, and 0.78, respectively. Meanwhile, data enrichment from other domains with similar topics to ethical protocols is possible. However, the precision obtained is not as good as when only using ethical protocol data during the training process. In future work, the research will explore the transfer learning techniques in deep learning to improve the model's performance. The analysis regarding the effect on the effectiveness of the review will be carried out in future studies.

REFERENCES

[1] A. F. Abdillah, M. Z. Hamidi, R. N. E. Anggraeni and R. Sarno, "Comparative Study of Single-task and Multi-task Learning on Research Protocol Document Classification," in *13th International Conference on Information & Communication Technology and System (ICTS)*, Surabaya, 2021.

[2] A. Binik and S. P. Hey, "A Framework for Assessing Scientific Merit in Ethical Review of Clinical Research," *ethics & Human Research*, vol. 41, no. 2, pp. 2-13, 2019.

[3] Pemerintah Republik Indonesia, "JDIH BPK RI: Database Peraturan," 2009. [Online]. Available: https://peraturan.bpk.go.id/Home/Details/38778/uu-no-36-tahun-2009. [Accessed 10 10 2021].

[4] Pemerintah Republik Indonesia, "Badan Pembina Hukum Nasional," [Online]. Available: http://www.bphn.go.id/data/documents/95pp039.pdf. [Accessed 10 10 2021].

[5] The Council for International Organizations of Medical Sciences (CIOMS), "International Ethical Guidelines for Health-related Research Involving Humans," Geneva, 2016.

[6] A. Benton, G. Coppersmith and M. Dredze, "Ethical Research Protocols for Social Media Health Research," in *Procedsings of the First Workshop on Ethics in Natural Language Processing*, Valencia, Spain, 2017.

[7] J. M. Barrow, G. D. Brannan and P. B. Khandar, "StatPearls [internet]," StatPerl Publishing, 28 8 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK459281/. [Accessed 10 10 2021].

[8] Ministry of Health of the Republic of Indonesia, *Pedoman dan Standar Etik Penelitian dan Pengembangan Kesehatan Nasional*, Jakarta, 2017.

[9] G. Mustafa, M. Usman, L. Yu, M. T. Afzal, M. Sulaiman and A. Shahid, "Multi-label classification of research articles using Word2Vec and identification of similarity threshold," *Sci Rep*, vol. 11, 2021.

[10] T. Pradhan and S. Pal, "A multi-level fusion based decision support system for academic collaborator recommendation," *Knowledge-Based Systems*, 2020.

[11] D. Chakrabarti, N. Patodia, U. Bhattacharya, I. Mitra, S. Roy, J. Mandi, N. Roy and P. Nandy, "Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support," in *Proceedings of TENCON 2018*, Jeju, 2018.

[12] W. K. Sari, D. P. Rini, R. F. Malik and I. S. B. Azhar, "Klasifikasi teks multilabel pada artikel berita menggunakan long short-term memory dengan Word2Vec," *RESTI*, vol. 5, no. 3, 2021.

[13] T. Baumel, J. N. Kassis, R. Cohen, M. Elhadad and N. Elhadad, "Multi-label classification of patient notes: case study on ICD code assignment," in *The Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] A. Omar, T. M. Mahmoud, T. A. El-Hafeez and A. Mahfouz, "Multi-label Arabic text classification in online social networks," *Information Systems*, vol. 100, 2021.

[15] S. Burkhardt and S. Kramer, "Online multi-label dependency topic models for text classification," *Mach learn*, vol. 107, pp. 859-886, 2018.

[16] A. Pal, M. Selvakumar and M. Sankarasubbu, "MAGNET: Multi-Label Text Classification using Attention-based Graph Neural Network," in *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, 2020.

[17] N. Isnaini, Adiwijaya, M. S. Mubarok and M. Y. Abu-Bakar, "A multi-label classification on topics of Indonesian news using K-Nearest Neighbor," in *The 2nd International Conference on Data and Information Science*, 2019.

[18] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[19] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, pp. 197-227, 2016.