Classifying Medical Document in Bahasa Indonesia using
Semi-Supervised Learning

Dhomas Hatta Fudholi¹, Kiki Purnama Juwairi²

¹Department of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia,
²Master Program in Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia

E-mail: ¹hatta.fudholi@uii.ac.id, ²20917020@students.uii.ac.id

Abstract. The medical domain has always been an all-time important domain since healthiness is everyone's purpose. People find medical document resources in the sea of data and information, such as the web. To support information retrieval and knowledge dissemination through the web, we analyze the use of semi-supervised learning to classify medical-related documents. The semi-supervised learning technique is chosen to show the possibilities of creating good classifiers with limited human supervision. In this research, we use the Naïve Bayes and Pseudo Labeling technique. We analyze different labeled:unlabeled data ratios of the training dataset in the experiment, starting from 4:3, 3:4, 2:5, and 1:6, to see the semi-supervised learning performance with different levels of human supervision. We get a relatively similar result in terms of classification average accuracy (81%-83%). Interestingly, in one experiment, the highest accuracy of the 1:6 ratio (85%) outperforms the 2:5 ratio (82%) and has the same accuracy as the 4:3 (85%). However, the standard deviation of the accuracy in the 1:6 ratio is the highest, amongst others (4.183). Finally, semi-supervised learning can be used to create a great classifier model of the medical domain in Bahasa Indonesia with less human supervision.

Keywords: Semi-Supervised Learning, Medical Document, Document Classification.

1. Introduction

Dissemination of information related to health is widely spread. Classification of medical documents will make it easier to find specific topics in health as needed. The purpose of classifying text documents is to get useful information from a set of documents. So that information can be organized and stored in a structured format [1]. In machine learning, a classification is a form of supervised learning. Modeling the classifier in classification is done by determining the number of categories or classes from a set of documents affixed as labels on each document. With the large data growth, manual labeling of documents was very time consuming and inefficient [2]. Therefore, the semi-supervised document classification can be one alternative in learning classifiers using labeled and unlabeled data in the training process, where most of the data available in unlabeled data [3].

Semi-supervised learning is a method to enable the inclusion of large numbers of unlabeled data as part of training data. Semi-supervised learning techniques can also utilize unlabeled data and labeled data to improve performance [4]. However, the development of many natural language applications considers this to be a challenge where the unlabeled data is relatively abundant, while the labeled data is somewhat limited [5]. The problem found in classifying documents using semi-supervised learning techniques is the use of data that is not labeled. One technique in semi-supervised learning to cope with
the unlabeled data is to use the Pseudo Labeling method. The pseudo labeling method classifies unlabeled data by making pseudo labels using predictive results [6].

This study designed a classifier model to classify medical documents into health-related document categories or classes. The documents are written in Bahasa Indonesia and taken from trusted medical-related websites. Semi-supervised learning is used in this study to enable the possibilities to take a significant number of documents with minimum efforts of labeling tasks. The unlabeled data is pseudo-labeled, and the Multinomial Naive Bayes algorithm is used to model the classifier. The algorithm is chosen based on its ability to do multiclass classification through accurate yet straightforward and fast computation for a small dataset. We will also analyze different ratios of labeled:unlabeled data to understand the impact of different proportions in the accuracy of classification.

2. Literature Review

This research aims to create a classifier model to classify documents in the Indonesian language, especially in the medical field. Some of the techniques used for document classification in machine learning include Expectation-Maximization, Naive Bayes classifier, Support Vector Machine, Decision Trees, and Neural Network [7]. The Naive Bayes method is a simple probabilistic classification method. This method calculates a set of probabilities by adding up the frequency and value combinations from a given dataset. The Naive Bayes method assumes that all category attributes are independent of each other [8]. Research [9] used the Naive Bayes Multinomial, which is part of the Naive Bayes algorithm, for classifying news articles in Indonesian. Classification is done by combining TF-IDF features with the Naive Bayes Multinomial method. The highest precision and recall values in this research are 98.4%.

In classification, semi-supervised learning is an efficient method to add training data from unlabeled data [5] automatically. One of the advantages of semi-supervised learning techniques is unlabeled data to increase the accuracy value obtained from labeled data, often when labeling data manually [10]. Research [11] uses the Multinomial Naive Bayes method in semi-supervised learning to classify documents. One semi-supervised learning technique is to use Pseudo Labeling. Pseudo Labeling can create labels or pseudo labels for unlabeled data as learning processes [6]. Pseudo labels may be chosen based on the initial words which are representative of the category. It is considered that documents containing more core words from a category are more likely to be associated with specific categories [2]. However, other methods, such as clustering, are also used widely in semi-supervised learning. The clustering process is initially executed. Members in a cluster (unlabeled) follow labeled members within the cluster [12]. However, there are challenges in clustering techniques: (i) what if we have more than one class labeled member on one cluster, and (ii) what if we have no labeled member in one cluster [13].

Text or document classification may be highly valuable in the medical domain, such as medical coding [14]. Medical diagnosis needs to be assigned to a specific class. This information may be required to be available instantly. The probability is high that the information is presented in an unstructured form, resulting from the patient-physician communication through diagnosis and treatment [15]. Our research contributions are analyzing the use of semi-supervised learning techniques, combining Multinomial Naive Bayes and Pseudo Labeling to classify medical documents in Indonesian. Also, we divide the portions of labeled data and unlabeled data into several tests. Testing with different portions of labeled data and unlabeled data is intended to compare and analyze the performance of semi-supervised learning techniques based on classifier models that are built specifically in the Indonesian and medical domains.

3. Method

Classifying medical documents written in Bahasa Indonesia using semi-supervised learning and analyzing the model's performance trained with different labeled:unlabeled dataset ratios are the aims of this research. The document classification research method includes several phases: (i) data collection, (ii) preprocessing and feature extraction, (iii) modeling, and (iv) analysis, evaluation, and implementation.
3.1. Data Collection
The dataset used in this study consists of health-related articles taken from healthcare websites in Indonesia. We intentionally collect a small number of documents to see the performance of the model in this setup. We collect 800 documents in 10 categories. The ten categories are kesehatan bayi (baby health), diabetes, diet, jantung (cardiac), kecantikan (beauty), kehamilan (pregnancy), kesehatan gigi dan mulut (dental and oral), kolesterol (cholesterol), kulit (skin), and mata (eyes). The collected documents and categories are taken from health articles and their categories in alodokter.com, halodoc.com, sehatq.com, klikdokter.com, hellosehat.com, and doktersehat.com. The category of the article is based on the category given by the article sources. Every web page has different styles. Hence, the articles are saved without any automated scraping method to ensure the collected articles are within the correct category. 100 of 800 documents are taken out with an even amount of each document category, to be the test dataset. The rest of the 700 documents will then be divided into labeled and unlabeled data in different ratios during the experiments.

3.2. Preprocessing and Feature Extraction
Preprocessing is done to clean the documents. The documents often contain unnecessary characters that must be removed or commonly referred to as noise. All documents will go through the following preprocessing stages: (i) eliminating non-ASCII characters, (ii) eliminating punctuation, digits, URLs, white space; (iii) remove punctuation marks such as period punctuation, commas, exclamation points, and question marks; (iv) turn all letters into lowercase; (v) removing stopword which are non-descriptive words; and (vi) stemming.

Feature extraction uses information from a set of data to create features for the data before the training process. This study uses the TF-IDF weighting. TF-IDF gives high value if a word rarely appears in many documents, meaning that it has high discriminatory power in a document. On the other hand, it gives low value for the word that often appears in different documents, making the word unclear with the document category.

![Modeling steps](image)

3.3. Modelling
The semi-supervised learning algorithm used in this study is Multinomial Naïve Bayes combined with the Pseudo Labeling technique. Naïve Bayes is a fully supervised learning method that requires a learning phase to build a probabilistic model. The probabilistic model will later be used to calculate
prior and conditional probability in determining the documents' categories. The Pseudo Labeling technique uses a model built from the labeled data to predict unlabeled data. Finally, by utilizing existing labeled data and pseudo-labeled data, a new classifier model is created. This classifier model will be used as a classifier model for the medical documents. Figure 1 shows the modeling steps that are elaborated before.

3.4. Analysis, Evaluation, and Implementation

The final phase of the study is to analyze, evaluate, and implement the model. The test dataset is used to measure the accuracy of the models. We design four different scenarios of semi-supervised training. The scenarios are divided based on the different ratio of training dataset between labeled and unlabeled data. In this study, the labeled:unlabeled ratios used are 4:3, 3:4, 2:5, and 1:6. The chosen ratios are based on the amount of the data and shifted slowly so that unlabeled data is getting higher than the labeled ones. The high number of unlabeled data becomes challenges in semi-supervised learning [5], and we tried to show the analytical finding using the shifted ratios. Table 1 shows the detailed four scenarios of training based on different training dataset combinations. The created final models from different scenarios are then implemented in a web-based application as a proof of concept in classifying medical documents written in Bahasa Indonesia.

Accuracy is the matrix used to evaluate the document classification model. The equation to measure the accuracy of the model is listed as Equation (1) as follows:

\[
\text{Accuracy} = \frac{\text{correctPrediction}}{\text{totalNumberOfPrediction}} \times 100\%
\]  

Equation (1)

correctPrediction refers to the correct classification, which consists of true positive, and true negative classifications. totalNumberOfPrediction refers to the number of all classification occurred during inferencing.

| Scenario | # labeled data | # unlabeled data | Ratio |
|----------|----------------|------------------|-------|
| SC1      | 400            | 300              | 4:3   |
| SC2      | 300            | 400              | 3:4   |
| SC3      | 200            | 500              | 2:5   |
| SC4      | 100            | 600              | 1:6   |

4. Result and Discussion

One of the focuses of this study is to analyze the performance of semi-supervised learning in various scenarios. The analysis was carried out to determine whether the classifier model built by applying the Pseudo Labeling method using the Naïve Bayes Multinomial algorithm was able to produce a fairly good accuracy in classifying medical documents. The study has a total of 800 documents in 10 categories. For each scenario presented in Table 1, we take 100 documents to be the test dataset with an equal number of documents for each document category. Also, for each scenario, we run five different training processes. In each training process, we divide randomly labeled data and unlabeled in a certain ratio. Hence, we have a different accuracy score for each process. The highest and the lowest accuracy of each process are documented.

Table 2 shows the highest, the lowest, and the average accuracy of the final model created from labeled and pseudo-labeled data from five different training processes within each scenario. As seen in the table, the overall lowest and average accuracy values are pretty much anticipated, since the lower the number of labeled data (the higher the number of pseudo-labeled data) used in training decreases the accuracy value. In the average accuracy column, the decrement of value is quite small. The largest gap (between SC2 and SC4) is only 2.4%. We need to point out that the highest accuracy value of all
scenarios is pretty similar, ranging from 82% to 87%. Interestingly, the model with the smallest number (SC4) of labeled data holds the same highest accuracy value (85%) with the model with the highest number of labeled data (SC1). However, SC4 has the highest standard deviation value, among others, which is 4.183. From the given analysis, we could say that semi-supervised learning using pseudo-labeled data has a big potential to be used as one alternative in designing document classification models, especially in Indonesian medical documents. We can reduce the number of labeled data to save resources during the data preparation without losing much accuracy and even have a chance to get higher accuracy.

Table 2. Accuracy of Classification Model

| Scenario | Ratio | Highest accuracy | Lowest accuracy | Average accuracy | Standard Deviation |
|----------|-------|------------------|-----------------|-----------------|-------------------|
| SC1      | 4:3   | 85.00%           | 81.00%          | 83.20%          | 1.643             |
| SC2      | 3:4   | 87.00%           | 81.00%          | 83.40%          | 2.509             |
| SC3      | 2:5   | 82.00%           | 79.00%          | 81.00%          | 1.581             |
| SC4      | 1:6   | 85.00%           | 74.00%          | 81.00%          | 4.183             |

From the ten categories of the medical document used in this study, we found that the wrong predicted label is quite visible in some categories. Figure 2 shows the confusion matrix of one training process in SC4. We noticed that in many modeling/training processes, misclassification often happens between kecantikan (beauty) and kulit (skin), and jantung (cardiac) and kolesterol (cholesterol). In common sense, we can notice that both of them are related. Skincare is one factor in beauty and cholesterol affects the health of the heart (cardiac). Hence, the content of the document might be similar.

Figure 3 shows the prediction page in the classification app that is developed using the created model. The app shows the predicted results' output with its probability value made by the selected classifier model. Information on each category's probability value is sorted from the highest to the lowest value of the probability obtained.

Figure 2. Confusion matrix in one training process in SC4.
5. Conclusion
The classification of medical documents using the semi-supervised learning technique utilizes two types of data: labeled and unlabeled data. In this study, the Naïve Bayes Multinomial method, combined with the Pseudo Labeling technique, is used. The domain taken as the case study in this study is medical document classification since the health domain is an all-time important domain. The classification of the medical documents supports information discovery, which is useful to various users. Creating the semi-supervised model involves training the labeled data to get a model to label the unlabeled data, which then becomes pseudo-labeled data. Both labeled data and pseudo-labeled data are then combined to get the final classifier model. We design different scenarios to show the method's performance in different ratios of labeled and unlabeled data. The accuracy value results show that even though the lower labeled data used in training might reduce the accuracy, the gap, or the difference is very small (less than 3% in our study). Moreover, the scenario with a lower number of labeled data still has a chance to bring higher accuracy rather than the model with a higher number of labeled data.

References
[1] Li X and Yang B 2018 A pseudo label based dataless naive bayes algorithm for text classification with seed words Proc. Int. Conf. on Computational Linguistics pp 1908–17
[2] Elhadad M K, Badran K M and Salama G I 2017 A novel approach for ontology-based dimensionality reduction for web text document classification Int. Conf. on Computer and Information Science pp 373–8
[3] Da Silva D F F, Coletta L F S, Hruschka E R and Jr Hruschka E R 2016 Using unsupervised information to improve semi-supervised tweet sentiment classification J. Inf. Sci. 355–356 pp 348–65
[4] Cascante-Bonilla P, Tan F, Qi Y and Ordonez V 2020 Curriculum Labeling: Self-Paced Pseudo Labeling For Semi-Supervised Learning
[5] Qiu Z, Cho E, Ma X and Campbell W 2019 Graph-based semi-supervised learning for natural language understanding Proc. of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing pp 151–8
[6] Arazo E, Ortego D, Albert P, O’Connor N E and McGuinness K 2019 Pseudo-labeling and confirmation bias in deep semi-supervised learning Preprint arXiv:1908.02983
[7] Basarkar A 2017 Document Classification Using Machine Learning Master Project (California: San José State University) DOI: https://doi.org/10.31979/etd.6jmu-9xdt
[8] Nafalski A and Wibawa A P 2016 Machine translation with javanese speech levels’ classification J. of Informatics, Control. Meas. Econ. Environ. Prot vol 6 no 1 pp 21–5
[9] Wongso R et al 2017 News article text classification in Indonesian language 2nd. Int. Conf. on
Leonandya R A, Distiawan B and Praptono N H 2016 A semi-supervised algorithm for Indonesian named entity recognition Proc. - 2015 3rd Int. Symp. Comput. Bus. Intell. ISCBII 2015 pp 45–50

Zhao L, Huang M, Yao Z, Su R, Jiang Y and Zhu X 2016 Semi-supervised multinomial naive bayes for text classification by leveraging word-level statistical constraint Proc. of the Thirtieth AAAI Conference on Artificial Intelligence pp 2877–83

Gowda H S, Suhil M, Guru D and Raju L N 2016 Semi-supervised text categorization using recursive k-means clustering Int. Conf. on Recent Trends in Image Processing and Pattern Recognition (Springer: Berlin/Heidelberg, Germany) pp 217–27

Kowsari K, Yammahi M, Bari N, Vichr R, Alsaby F and Berkovich S Y 2015 Construction of fuzzy find dictionary using golay coding transformation for searching applications Preprint arXiv:1503.06483

Kowsari K, Meimandi K J, Heidarysaf M, Mendu S, Barnes L and Brown D 2019 Text classification algorithms: A survey (MDPI) Information 10 150

Lauria E J and March A D 2011 Combining bayesian text classification and shrinkage to automate healthcare coding: a data quality analysis J. Data Inf. Qual. (JDIQ) 2 13