Text Mining in Healthcare for Disease Classification using Machine Learning Algorithm

Ghulam Asrofi Buntoro
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
Informatics Engineering
Universitas Muhammadiyah Ponorogo
Ponorogo, Indonesia
ghulam.207022@mhs.its.ac.id
ghulam@umpo.ac.id

Adhi Dharma Wibawa
Department of Electrical Engineering
Department of Computer Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
adhiosa@te.its.ac.id

Mauridhi Hery Purnomo
Department of Electrical Engineering
Department of Computer Engineering
University Center of Excellence on Artificial Intelligent for Healthcare and Society (UCE AIHeS)
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

2021 International Electronics Symposium (IES) | 978-1-6654-4346-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IES53407.2021.9593998

Abstract**—**The development of information technology and smartphones has caused production of many data around us. In every second million of new data is created in the form of text, audio, image and even videos. This environment then has triggered big data analytics demand. One of big data that is produced daily is data on the history of healthcare services in hospitals. Important new information can be retrieved through this huge dataset, especially concerning the patient symptoms, drug usage and new diseases report. In this study, text processing technique is applied on text data of patient medical record data from public hospital during 2017 till 2019 regarding the patient symptoms and the disease classification. Naïve Bayes Classifier and Random Forest algorithms are used to classify diseases in medical record data with 19 diseases in preprocessing data. A list of modified Indonesian stop words was used to filter the symptom sentences. The result indicates that the Random Forest classification algorithm can achieve the highest accuracy of around 99.9%, better and more accurate than the Naïve Bayes classification algorithm. This experiment shows that our proposed method provides a robust system and good accuracy for classifying medical record data with many diseases.

Keywords—text mining, healthcare, disease, naïve Bayes classification, random forest

I. INTRODUCTION

History of public health care data is widely available in hospitals, community health centers, health clinics, or practicing doctors [1] [2]. Unfortunately, from those data, there is less that have been used to help doctors or other to get new insights [3] [4] and gaining knowledge for understanding their Health condition [5] in the society in general. Health care history data has the potential to be used as a knowledge base for Expert System [6] and Decision-Making [7] to determine a person's health condition with predictions [8], [9] based on health care history data. Many new strategies can also be retrieved from processing those data such as which treatment resulting the best recovery in time, or what disease that have been handle by the hospital and how do they spread, and many more new insights. Artificial Intelligence research, especially Text Mining, has penetrated the health sector recently; this Text Mining processes, analyzes, recognizes, and provides predictions with historical health care data using Text Mining methods, namely Fuzzy c-means algorithm [10], fuzzy c- research. The means algorithm aims to classify Health care history data so that decision-making is easier and more accurate. Previous research using the multilayer perceptron NN and SVM [11] resulted in the accuracy of the support vector machines method being better than the multilayer perceptron neural network method. The utilization of medical history data can be very useful to help clinicians work because there are many important information and new insights that can be explored from those data. In general, hospitals and community can give better service when they can make use of medical record history data. This study will present text processing on medical record data based on patient symptoms and diseases to be classified using two algorithms namely Naïve Bayes and Random Forest. This study will contribute to build a model for classifying several diseases based on the patient's symptoms.

II. RELATED WORKS

Text Mining aims to dig and find information or knowledge from a text. The development of information technology makes today much information scattered and stored in personal storage or in the virtual world of the internet. Most of the disseminated text information is not structured, so that it becomes a challenge in text mining research [12], so it needs new techniques in text mining to produce better accuracy. Electronic medical record data is one example of unstructured data in the form of text in healthcare service.

Public healthcare is defined as a public health service either in a hospital or community health centers. The more people who use health services, the more Health care history data is stored in the hospital or Puskesmas. Text mining can be a useful tool for clinical management and predicting health diagnoses [13] [14]. Due to the large number of medical record files for each patient currently available in clinics, machine learning or other data processing techniques can be used to extract information and create knowledge [15]. This research will begin with data collection then processed and analyzed with a machine learning classification model. This study aims to find the best classification model to make disease prediction using Health Care History data more accurate. This paper discusses more experimental data testing and evaluation of classification results.

III. Material and Methods

A. Collection of Data

This study uses medical record data from the general hospital dr. Soetomo Surabaya. The data period used in this study was from February 1, 2017, to February 1, 2019, with a total of 2271 medical records with 19 disease categories.

B. Methodology

The methodology proposed in this study consists of several stages, namely the Collection Data, Preprocessing Data, Classification Processes, and Evaluation Results. Figure 1. is a picture of the methodology proposed in this study.

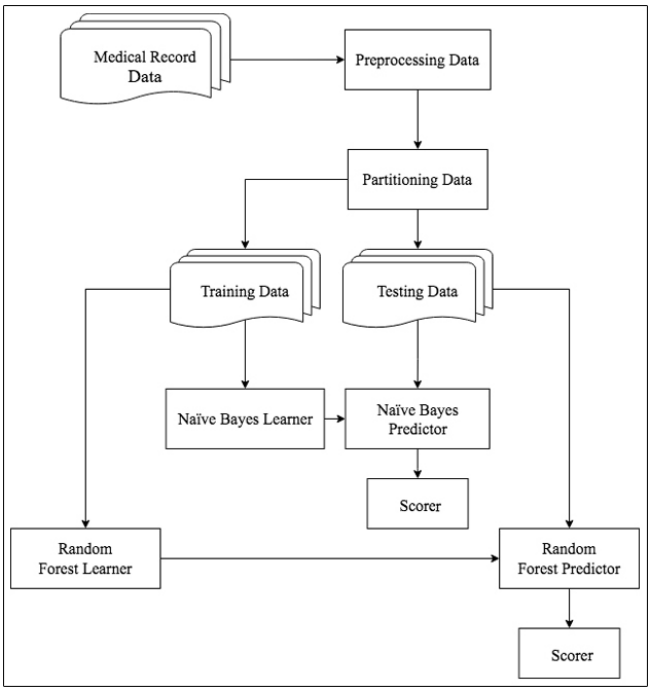

Fig. 1. Propose methods

The step-by-step disease classification design in KNIME is shown in Figure 2.

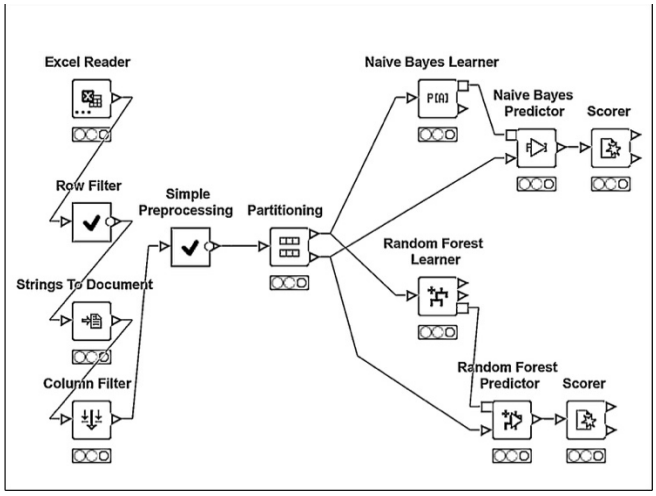

Fig. 2. KNIME Workflow for Disease Classification

1) Collection Data

This study uses Electronic Medical Record (EMR) data from the general hospital dr. Soetomo Surabaya. The data is retrieved from the Database in the form of unstructured text data. The total of 2271 records of text as patient's subjective symptoms with 19 disease categories were analyzed. The data period used in this study was from February 1, 2017, to February 1, 2019.

2) Preprocessing Data

This node functions for preprocessing data, stage in this study is Punctuation Erasure to remove punctuation marks in the raw data, the Number Filter to erase numbers in the raw data, the N Chars Filter node to delete words with the number n-chars, for example, n = 2 then the word with two letters will be deleted. Node Case Converter to convert all sentences to uppercase or lowercase letters, Stop Word Filter functions to remove words that are less important in symptomatic sentences and words that do not affect the classification results later. The Stop Word List used in Indonesian [16], This study modifies the Indonesian Stop Word List by including words in the symptom sentence that are not a symptom of the disease. Node Bag Of Words Creator functions to search for keywords in a symptom sentence.

3) Partitioning Data

Node Partitioning is applied for dividing the dataset into training data and testing data. In this study, 50% of the data is used for building the model in learning stage, while the other 50% is applied in the classification mode. [17].

4) Naïve Bayes (NB)

Naive Bayes Classification model, the node utilized in this examination are the Naïve Bayes Learner node and the Naïve Bayes Predictor node. Node Naïve Bayes Learner is responsible for building the Bayesian model from the preparation information. The Naive Bayes Algorithm is a probabilistic-based grouping calculation that utilizes the Bayes hypothesis. This calculation depends with the understanding of freedom among indicators and requires little preparing information for the preparation model [18]. We calculated the posterior probabilities using (1).

$$P(c \text{ — } x) = P(x \text{ — } c) * P(c) / P(x) \quad (1)$$

Where (x) is the predictor, which is the client survey in our trial, in a given class (c), which is the sentiment of the yield variable (Positive, Negative or Neutral), P (c - x) is the posterior probability of the predictor of a particular class, P (c) is the prior class probability, P (xc) is the probability which is the predictor probability of a certain class, and P (x) is the prior probability of the predictor [19]. The result of probability is the probability of the class and the probability of the class trait itself. The normal distribution per attribute is that the chance calculated from the numerical worth, whereas the amount of occurrences of the category value divided by the entire number of occurrences of the class value is the probability of the nominal value.

5) Random Forest (RF)

The model accustomed build the Random Forest Classification model during this study is that the Random Forest Learner node and also the Random Forest Predictor node. The Random Forest Learner node is accountable for learning from many decision trees [20]. every style of the choice tree is studied on a distinct set of rows and a different set of columns, which might later be a bit-vector or byte-vector descriptor [21]. The row set for every decision tree is bootstrapped and the same size because the table' original input. every node of the decision tree can confirm a replacement attribute by taking a random sample of the scale of the root (m), wherever m is that the total range of attributes. The output model depicts a Random Forest enforced at the suitable predictor nodes, employing a easy majority vote. The Random Forest Predictor node predicts a pattern supported the aggregation of the predictions of individual trees within the Random Forest model for outlined criteria information Gain.

6) Scorer

The scorer node is used for the evaluation stage in this study; the scorer will display the performance of True Positive (TP) Rate, False Positive (FP) Rate, True Negative (TN) Rate, False Negative (FN) Rate, Precision, Recall, F-Measure, Accuracy, and Cohen's kappa from the experiments that have been done. The process of evaluating the results was carried out using the Confusion Matrix [22].

IV. Result and Discussion

A classification study was carried out using the KNIME machine learning technique in-text mining research on health care data. The performance measurement of the classification method used is the Confusion Matrix to discuss Precision, Recall, F-Measure, Accuracy, Cohen's kappa.

Table 1 shows the comparative performance of Naïve Bayes (NB) and Random Forest (RF) by showing the correct classification values with incorrect classifications of the NB and RF classification models. The classification results for 2271 medical record data with 19 disease categories divided into 50% training data and 50% test data showed that RF obtained better and higher accuracy results compared to NB. The Random Forest Classification Model correctly classifies the various responses given by the individual model and can overcome any constraints that may occur due to the resulting problem due to missing values. The accuracy of the Random Forest classification model is higher for all values of Precision, Recall, F-measure, and Cohen Kappa (k) than the Naïve Bayes classification model.

Table 1 shows the performance comparison of the Naïve Bayes and Random Forest classification models.

TABLE I. Comparative Performance of Naïve Bayes and Random Forest

Classification Models	Correctly Classified	Incorrectly Classified
Naïve Bayes	97.2%	2.76%
Random Forest	99.9%	0.03%

Based on the evaluation process of confusion matrix results that see the TP Rate value - positive samples that are predicted correctly; TN Rate - negative sample correctly predicted; FP Rate and FN Rate - respectively positive and negative samples that are incorrectly predicted, from these formulas we can find the values of Accuracy, Precision, Recall, F-measure and Cohen Kappa for Naive Bayes and Random Forest as shown in Table 2.

TABLE II. Precision, Recall, F-measure and Cohen Kappa for Naïve Bayes and Random Forest

Classification Models	Precision	Recall	F-measure	Cohen Kappa(k)
Naïve Bayes	0.931	0.523	0.687	0.968
Random Forest	0.979	0.955	0.977	1

Accuracy value is one of the main parameters in text mining research for disease classification using a classification model. The formula for the accuracy value in this study is the number of symptom data that has been successfully classified according to the disease diagnosis for the total number of classified symptom data. Therefore, the greater the amount of symptom data that is correctly classified according to the diagnosis of the disease, the higher the accuracy value will be.

In this research, the Random Forest classification model obtains the highest accuracy value, which has an accuracy value of 99.9%. At the same time, the value of the Naïve Bayes classification model is 97.2%. The Random Forest classification model produces the highest accuracy because this model is in the form of decision trees that allow much knowledge to make predictions accurate.

To find out the Precision value of a classification result data, the formula is (2)

$$TP / TP + FP \quad (2)$$

This means that the number of symptoms of a correctly classified disease according to the disease class is divided by the total data classified as symptom data. The Random Forest classification model obtains the highest precision value with a value of 97.9%; the precision value obtained is very high because many disease data with symptoms are classified according to the symptoms and disease. For the Naïve Bayes classification model, the Precision value is 93.1%.

Furthermore, the highest recall value is 95.5% by the Random Forest classification model. Meanwhile, the Naïve Bayes classification model scores 52,3%. The formula for the recall value is (3)

$$TP / TP + FN \quad (3)$$

This means the number of disease classes that are correctly classified as a said class of disease divided by the number of disease classes in question.

The F-measure is the harmonic mean of recall and precision and can be calculated by the formula (4)

(2 * Precision * Recall) / Precision + Recall (4)

The highest F-measure value is obtained by the Random Forest classification model with a value of 97.7%, while the Naïve Bayes classification model gets a value of 68.7%.

Comparison of Precision, Recall, F-measure and Cohen Kappa for Naive Bayes and Random Forest is shown in Figure 3.

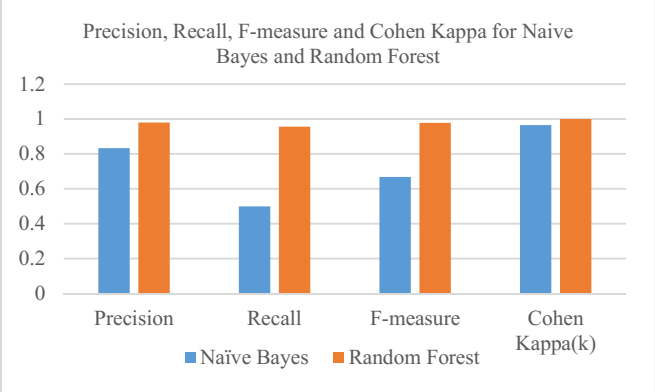

Fig. 3. Precision, Recall, F-measure, and Cohen Kappa for Naive Bayes and Random Forest.

As shown in Figure 3, the higher the accuracy value, the higher the Precision, Recall, and F-measure values. Conversely, if the accuracy value is low, then the three values will also be lower. The mean of the Precision, Recall, and F-measure values are almost the same as the difference, indicating that the model is balanced. With almost equal values, it shows that the classification model can classify positive samples correctly and its ability to classify negative samples correctly [23].

The next assessment parameter in this study is Cohen's Kappa, the Cohen Kappa way of assessing by considering the class distribution; this parameter is related to accuracy. The highest Cohen Kappa score was obtained by the Random Forest classification model with a value of 99.9%, while the Naïve Bayes classification model received a value of 96.9%. The function of Cohen Kappa is to handle multiclass class problems and unbalanced classes.

Cohen Kappa (k) typically ranges from lower than or adequate to 1. a value of 0 or less, indicating that the classifier is useless. there's no customary way of deciphering the values. Some way to characterize values. consistent with their scheme, a value less than 0 implies no agreement, 0-0.20 slight, 0.21-0.40 reasonable, 0.41-0.60 moderate, 0.61-0.80 substantial and 0.91-1 nearly good [24].

The confusion matrix is a short representation of the model's performance and gives us the corresponding values for true positives, true negatives, false positives, and false negatives. Our research on disease prediction based on medical record data shows that the Random Forest classification model has the highest accuracy of 99.9%, while the Naïve Bayes classification model has an accuracy of 97.2%. It can be concluded that in this study, the Random Forest classification model obtained the highest accuracy when applied in disease prediction based on medical record data.

V. Conclusion

From the experiments that have been carried out in this study, it can be concluded that text mining can be used to classify text symptoms with the disease. The purpose of this study was to build a classification model for medical record data by testing the Naïve Bayes and Random Forest classification algorithms to classify diseases in medical record data with 19 diseases. In this study, we also tested the effectiveness of the Indonesian stop words that we have modified by including words from symptom sentences that are not symptoms of the disease. The results obtained in this study proves that the performance of the Random Forest algorithm is better for classifying EMR data than Naïve Bayes classification algorithm. We opinioned that this is due to the form of unstructured text that would cause more ambiguities condition, and thus resulting the probability pattern calculated by Naïve Bayes is then causing more false positive aspects in confusion matrix or false negatif. It is also proven that the Indonesian language stop word list that we have modified affects the accuracy of the built classification model. Therefore, the Random Forest classification algorithm is highly recommended as a classification algorithm for medical record data that contains symptomatic text with many diseases. For further studies, it still needs much improvisation in the application and trials to understand better the classification model's performance built in this research. One of the applications is to classify more and real-time data, besides how to generated precised stop-word list based on the characteristic of text raw data.

Acknowledgment

There is no conflict of interest in writing this article.

References

[1] L. Li and G. Liu, “In-hospital Mortality Prediction for ICU Patients on Large Healthcare MIMIC Datasets Using Class Imbalance Learning,” *2020 5th IEEE Int. Conf. Big Data Anal. ICBDA 2020*, pp. 90–93, 2020, doi: 10.1109/ICBDA49040.2020.9101272.

[2] B. Ru, Q. Wu, X. Wang, L. Yao, and Y. Jia, “Integration of Accountable Care Organization and Additional Hospital Data into CMS Referral Analytics System,” *Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017*, pp. 357–361, 2017, doi: 10.1109/ICHI.2017.44.

[3] H. Cui, Q. Li, H. Li, and Z. Yan, “Healthcare fraud detection based on trustworthiness of doctors,” *Proc. - 15th IEEE Int. Conf. Trust. Secure. Prev. Comput. Commun. 10th IEEE Int. Conf. Big Data Sci. Eng. 14th IEEE Int. Symp. Parallel Distrib. Price*, pp. 74–81, 2016, doi: 10.1109/TrustCom.2016.0048.

[4] S. Chen, X. Guo, T. Wu, and X. Ju, “Exploring the Online Doctor-Patient Interaction on Patient Satisfaction Based on Text Mining and Empirical Analysis,” *Inf. Process. Manag.*, vol. 57, no. 5, p. 102253, 2020, doi: 10.1016/j.ipm.2020.102253.

[5] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, “Text mining in healthcare. Applications and opportunities.,” *J. Healthc. Inf. Manag.*, vol. 22, no. 3, pp. 52–56, 2008.

[6] J. Feng, R. Zhang, D. Chen, and W. Zhang, “Extracting Meaningful Correlations among Heterogeneous Datasets for Medical Question Answering with Domain Knowledge,” *2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. ICON 2018*, no. 71532002, pp. 297–301, 2019, doi: 10.1109/IEMCON.2018.8615045.

[7] X. Geng *et al.*, “Online and intelligent route decision-making from the public health dataset,” *Proc. 2012 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2012*, pp. 453–456, 2012, doi: 10.1109/CyberC.2012.82.

[8] M. Hao, H. Li, G. Xu, Z. Liu, and Z. Chen, “Privacy-aware and Resource-saving Collaborative Learning for Healthcare in Cloud Computing,” *IEEE Int. Conf. Commun.*, vol. 2020-June, 2020, doi: 10.1109/ICC40277.2020.9148979.

[9] Q. Liu, K. Liao, and Z. Wei, “Automatic Acceptance Prediction for Answers in Online Healthcare Community,” *Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018*, pp. 1262–1265, 2019, doi: 10.1109/BIBM.2018.8621209.

[10] B. R. Reddy, Y. Vijay Kumar, and M. Prabhakar, “Clustering large amounts of healthcare datasets using fuzzy c-means algorithm,” *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 93–97, 2019, doi: 10.1109/ICACCS.2019.8728503.

[11] P. Naraei, A. Abhari, and A. Sadeghian, “Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data,” *FTC 2016 - Proc. Futur. Technol. Conf.*, no. December, pp. 848–852, 2017, doi: 10.1109/FTC.2016.7821702.

[12] M. Sukanya and S. Biruntha, “Techniques on text mining,” *Proc. 2012 IEEE Int. Conf. Adv. Commun. Control Comput. Technol. ICACCCT 2012*, no. 978, pp. 269–271, 2012, doi: 10.1109/ICACCCT.2012.6320784.

[13] A. Begum and A. Parkavi, “Prediction of thyroid Disease Using Data Mining Techniques,” *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, no. August 2016, pp. 342–345, 2019, doi: 10.1109/ICACCS.2019.8728320.

[14] P. Jatunarapit, K. Piromsopa, and C. Charoeanlap, “Development of Thai text-mining model for classifying ICD-10 TM,” *Proc. 8th Int. Conf. Electron. Comput. Artif. Intell. ECAI 2016*, pp. 2–7, 2017, doi: 10.1109/ECAI.2016.7861163.

[15] I. A. Tache, M. Dragoicea, E. S. Apostol, and C. O. Taurica, "Text mining of medical records," *2019 7th E-Health Bioeng. Conf. EHB 2019*, pp. 2019–2022, 2019, doi: 10.1109/EHB47216.2019.8969943.

[16] F. Z. Tala, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia,” *M.Sc. Thesis, Append. D*, vol. Pp, pp. 39–46, 2003.

[17] M. Wiley and J. F. Wiley, *Advanced R statistical programming and data models: Analysis, machine learning, and visualization*. Apress Media LLC, 2019.

[18] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed. Chapman & Hall/CRC, 2010.

[19] E. Keogh, “Naïve Bayes Classifier,” 2006.

[20] A. Khanna and N. Dey, “A comparative study on decision tree and random forest using konstanz information miner (KNIME),” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, pp. 2365–2376, 2020.

[21] C. Chauhan and S. Sehgal, “Sentiment classification for mobile reviews using KNIME,” *2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018*, pp. 548–553, 2019, doi: 10.1109/GUCON.2018.8674946.

[22] "Confusion Matrix for Your Multi-Class Machine Learning Model." https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826 (accessed June 27, 2020).

[23] "machine learning - What does it imply if accuracy and recall are the same? - Cross Validated." https://stats.stackexchange.com/questions/99694/what-does-it-imply-if-accuracy-and-recall-are-the-same (accessed April 15, 2021).

[24] "Performance Measures: Cohen's Kappa statistic - The Data Scientist." https://thedatascientist.com/performance-measures-cohens-kappa-statistic/ (accessed April 15, 2021).