Seminar_Int_2020_ICIDM_IEEE_Mailia.pdf

by

Submission date: 12-Apr-2023 12:28PM (UTC+0700)

Submission ID: 2062296819

File name: Seminar_Int_2020_ICIDM_IEEE_Mailia.pdf (290.88K)

Word count: 4664

Character count: 24658

Hoax Information Detection System Using Apriori Algorithm and Random Forest Algorithm in Twitter

Mailia Putri Utami
Department of Information System
Diponegoro University
Semarang, Indonesia
mailiap2206@gmail.com

Oky Dwi Nurhayati
Department of Computer Engineering
Diponegoro University
Semarang, Indonesia
okydwinurhayati@lecturer.updip.ac.id

Budi Warsito
Department of Statistics
Diponegoro University
Semarang, Indonesia
budiwarsito@lecturer.updip.ac.id

***Abstract* — This research is based on the disturbance faced by Twitter users, related to the distribution of hoax information in the text form. One of the efforts to overcome such problem is by building a system to detect hoax news in Twitter application. A set of information in text from on social media, can capture the use of language in written or verbal (corpus) form. Based on the advantages of Apriori algorithm which it is able to mine the text data from many used datasets and able to find the relation pattern or itemset combination in a database as a recommendation of the raised pattern. Moreover, the detection system also needs a method to classifying information based on the classes of hoax and non-hoax. One of the algorithms used is Random Forest algorithm, which it is able to combine several models of decision trees to eliminate the problems of overfitting. The purpose of this research, apart from implementing and integrating the Apriori algorithm and the Random Forest algorithm, is to make it easier for researchers to analyze and evaluate the system's results to detect hoax information most optimal level of accuracy. The results show that the system can detect hoax and non-hoax news, whose data is integrated directly with the Twitter application. Accuracy level, precision and recall from the built system in detecting hoax news information reached 100% with minimum support value of 23.**

Keywords* — *apriori algorithm, text mining, random forest algorithm, machine learning, hoax detection

I. Introduction

The development of technology is currently rushing, especially in the use of social media. The development of social media can contribute indirectly, with both positive and negative impacts. However, social media is now often disrupted due to hoax circulating issues, such as the circulation issues of economic recession worldwide, issues related to politics, health, the economy, and other issues. As a result of the circulation of hoax information, the problem faced by social media users now is that users cannot distinguish between genuine news information or hoaxes, especially in the information that is widely spread in the Twitter application [1].

A set of various information text spread on social media, capture the use of Language in written or verbal (corpus) form, and has the tendency to grow rapidly and complex every time. Thus, it is in need of a method or algorithm that able to build a hoax message detection system in completing the complexity of the growing of hoax news in Twitter application [2].

The trend in classifying datasets was explained by scientists using machine learning methods, machine learning algorithms are one of the most effective methods in classifying or filtering the characteristics of hoax information data. One of the machine learning algorithms is the Random Forest algorithm. The advantage of the Random Forest algorithm is the averaging ensemble learning method that can be used for classification process problems [2]. In processing information text into data sets, techniques are needed to mine items that often arise from the many databases. Therefore in this study, it is necessary to add an algorithm that can overcome these problems. The Apriori algorithm is a data mining technique that can mine itemset data in many existing datasets and find the relationship pattern of a combination of items in a dataset in identifying data, the Apriori algorithm will be the initial process in the system development stage [3].

In this research, the authors used the Apriori algorithm and the Random Forest algorithm to overcome the increasing development of hoax news information. The use of two algorithms in this study aims to integrate the Apriori algorithm with the Random Forest algorithm. Apriori algorithm extracted the set of data through layer by layer of iteration in finding the recommendation of a set of items or patterns that often appears [4]. Next is the step of classification by applying the Random Forest algorithm, to create each class or group from the result of rule mining association, so that it able to create an accurate decision tree based on the data complexity. The most accurate detection system model is expected to make Twitter users easily distinguish between real news and hoaxes, allocating, and minimizing hoax news distribution.

II. Basic Theory

News information on Twitter (tweet) has a swift growth even in seconds and is collected continuously, as the amount of data from the transaction dataset is stored. The correlation of transactions for the number of datasets is similar [5]. Association of mining rules is one of the classic data mining algorithms useful for determining the relationship pattern of a combination of items in a dataset and can identify news data in the dataset. The process of mining items that often appear from the database is needed to make it easier to find patterns from the dataset in classifying news information data in one of the hoaxes [6].

A. Apriori Algorithm

The Apriori algorithm was first proposed by scientists named Rakesh Agrawal and R. Srikant in 1994 in the Boolean Association [6]. Discussion of the itemset frequency for boolean association rules. Association rule mining is a classic data mining algorithm for finding patterns of relationships between combinations of items in a data set. One of the

978-1-7281-4927-1/20/$31.00 ©2020 IEEE

exciting stages of analysis is its ability to produce efficient algorithms through frequent pattern mining analysis. There are two main parameters of the Apriori algorithm, namely, support value and confidence value. Support value is the percentage of item combinations in the database, while confidence value is the strength of the relationship between items in the associative rules. One example of associative rules can usually be expressed in the form {coffee, milk} -> {sugar} {support = 45%, confidence = 50%}, which means "someone who drinks coffee and milk has a 50% chance of adding sugar" [7].

The main processes involved in processing the Apriori algorithm include [4]:

1) Join: In this process, each item is combined with other items until the combination pattern is no longer formed.
2) Prune: In this process, the combined items' results are then trimmed to the minimum support value specified by the user.

Determining the minimum support value determined in the process is an essential point in processing data. The calculation of the itemset X value is calculated by (1). Support (X) is a measure of the frequency of the itemset in the database, N is a measure of the total frequency of the itemset in the database, and Count (X) is the number of occurrences of a series of X items in all transactions [8].

$$Support\ (X) = \frac{Count\ (X)}{N} \qquad (1)$$

While the value of support for two items (X and Y) is obtained from the following formula [8], on (2).

$$Support\ (X \cup Y) = \frac{Count\ (X\ and\ Y)}{N} \qquad (2)$$

B. Random Forest Algorithm

Random Forest was first discovered by American computer scientist Tin Kam Ho in 1995. Then the Random Forest algorithm was developed by members of the California Academy of Sciences Breiman [9]. Initially, Random Forest was one of the raw data mining techniques, namely a decision tree. In the decision tree process, the input in the form of data will be inputted into the treetop process in the form of a tree root, then it will be lowered down in the form of leaves in the process to determine what class includes the input data in the process. A random forest is a classifier consisting of structured tree classifications in which each tree issues a sound unit for popular class input. In other words, a random forest consists of a set of decision trees, where a collection of decision trees is used to classify data [10].

Decision trees use acquired information and Gini indexes for calculations to determine root nodes and rules. Likewise, Random Forest will use gain information and the Gini index for calculations in building trees, and it is just that Random Forest will build more than one tree [4]. Each tree is built using a dataset with attributes or variables taken randomly from the training data, where each tree will depend on the value of the independent vector sample with the same distribution on each tree built. Each tree will choose the most popular class [11].

The formula of the (3) in determining the Gini index, which is for a candidate (nominal) split attribute X_i, denotes possible levels as $L_1; \dots; L_J$. Gini indeks for this attribute is calculates as [12].

$$G(X_i) \coloneqq \textstyle\sum_{j=1}^{J} Pr(X_i = L_j)(1 - Pr(X_i = L_j)) \quad (3)$$

$$= 1 - \sum_{j=1}^{J} Pr(X_i = L_j)^2$$

The application of the Random Forest algorithm has been carried out in many studies. One example is predicting the accuracy of accidents at railroad-class road crossings by comparing the results of the decision tree model built [8]. Another example of implementing the Random Forest algorithm is detecting noise sources by classifying sensing (hearing) data based on maps as training data [13].

C. Undersampling

Resampling technique is one of many technique of preprocessing, where there is unbalance database distribution on labeling or from learning process. Method from resampling technique includes oversampling and undersampling [14]. Undersampling is a sampling technique to turn the majority class proportion from the database into a minority, for the purpose of balancing the minority class with the assumption that the result of the proportion will be 50:50. By using undersampling technique it can decrease the imbalance of proportion between majorities and minorities and also will make the data more precise [15].

III. Methodology

In this study, the authors present a method for building a hoax information detection system on the Twitter application that integrates and implementation two algorithms: the Apriori algorithm and the Random Forest algorithm (as the classification technique used). The method used has two main processes, namely the learning process and the prediction process.

A. Data Collection Technique

Data is the main capital in a study process, where in this study the data used is differentiate based on two sources. First, in the learning process, it uses data from the website of Turnbackhoax.id with data collection technique process, researcher submit a request. Data of Turnbackhoax.id is a data that has been arranged by classification class of hoax and non-hoax, the data is categorized as the supervised learning data type (however this database classified as unbalance data). In the data of Turnback hoax, the base is learning process, using 5000 data in the period of 2016-2020.

Second, in the next process is prediction process, which in this process uses data source from Twitter app. The Data Crawling technique on Twitter is a data collection technique process from Twitter server by involving Application Programming Integration (API). The program will conduct data crawling in 5 minutes on Twitter app, just as news account such as Detik.com, Kompas.com, Berita Satu, Media Indonesia, Tempo.co, Breaking News, MNC News Channel, Kumparan, TMC Polda Metro, and Lewat Mana. Accounts from such news source can be added in accordance of how much news sources the user needed, however this study only focus on news source in Indonesia, so that it simplifies the researcher to review them manually and confirm their reliabilities.

B. *Learning Process*

The learning process is the first stage process of pattern creation or classification model from the system. Data training in this stage involves and uses the resampling method ecause the Turnbackhoax.id database is an unbalance database where hoax information classes are more dominant than non-hoax classes. The resampling method used is using the under-sampling, this method is a database sampling method in such way that the proportion of the majority class become smaller, with the purpose of balancing the minority class. The implementation of the Apriori algorithm here is to create a pattern from the key characteristic to simplify classification process (will be presented in the word occurrence which contains the training data dictionary is false and the dictionary is true). The classification process in this stage is using the Random Forest algorithm [16]. Fig. 1, is the system workflow of the learning process.

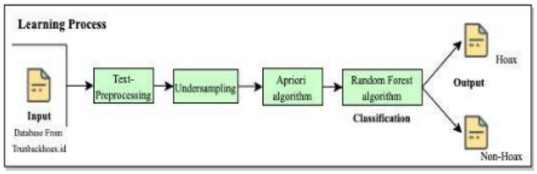

Fig. 1. Learning Process to Detect Hoax Information.

The database of Turnbackhoax.id still in the raw data form that not yet well structured, initial step of this study is to use several techniques of text-preprocessing, to simplify the program flow on the implementation of Apriori algorithm, Fig. 2, text-preprocessing. Case Folding is the process of converting document text letters into one standard usage form, the commonly used alphabet letters, and eliminating letters from the alphabet. [17]. Tokenizing is the process of separating or breaking a set of text into pieces of sentences and words and removing unnecessary characters in the process. Filtering is the process of removing inappropriate words and signs that have no meaningful meaning, such as hashtags (#), URLs, emoticons, and the stages of joining or taking words that are considered necessary from a sentence. Stemming is converting words into primary word forms so that they serve to reduce or reduce the number of different indexes of a text document [18].

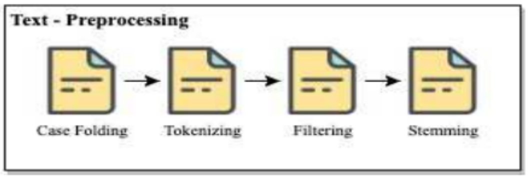

Fig. 2. Text-Preprocessing.

C. *Prediction Process*

The prediction process, starts after the learning process is carried out. The system model (the results of the learning process in the form of a table word occurrence) contains the occurrence of each false dictionary and the dictionary is true. The results of the training have accelerated the prediction process, where the prediction process uses a database of news information obtained from several Twitter application users.

The initial stage of the prediction process is the text preprocessing process. Furthermore, the data preparation results from the text-preprocessing process are followed by a data classification process, based on the hoax or non-hoax classes based on Twitter news, by applying the Random Forest algorithm. This stage of the prediction process is needed, which is useful when testing the system, whether the system is appropriate for tackling hoax information (Fig. 3, regarding the prediction process flow).

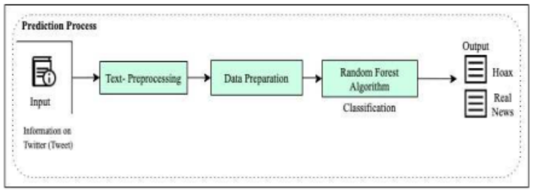

Fig 3. Prediction Process to Detect Hoax Information.

D. *Evaluation Model*

The detection system generally requires a stage where the system is evaluated from the entire work process, especially when it comes to classification. Confusion Matrix is the matrix used to summarize the correct/incorrect classification of data created by a method or classification model for a dataset. Confusion Matrix is the matrix for calculating performance matrix, which it can calculate the system accuracy value, precision value and recall value. The rows and columns of confusion matrix is related to the true classes (gold standard/actual class/real class) and predicted classes information, illustrated in Table I [19].

TABLE I. CLASSIFICATION MATRIX FOR THE TWO-CLASS MODEL

		Prediction Class	
		Positive Class	Negative Class
Actual Class	Positive Class	TP (True Positive)	FP (False Positive)
	Negative Class	FN (False Negative	TN (True Negative)

The accuracy value can be calculated by using the (4), which accuracy is the proportion of well classified data. Precision is the proportion of data that predicted as true positive class, calculated by (5). Next the Recall equation shows the proportion of data from the positive class classified as positive class, showed in (6) [19].

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \times 100\% \qquad (4)$$

$$Precission = \frac{Tp}{Fp+Tp} \times 100\% \qquad (5)$$

$$Recall = \frac{Tp}{Fn+Tp} \times 100\% \qquad (6)$$

IV. Results and Discussion

In this study, the programming used is Phyton version 3.8.5. The library or module used among others are, Libraries of NLTK, Sastrawi, Tweepy, Scikit-Learn, Openpyxl, Numpy and Tkinter.

A. Result Learning Process

First in learning process, the database used is database that came from the website of Turnbackhoax.id as data training with the number of data is 75% from the total of 5000, and data for testing process is 25%. Procedure of the two-stage process used is necessary in building a more optimum system. The implementation of the Apriori algorithm here is to create a pattern of key characteristics to simplify the classification process. TABEL II the word occurrence from the Turnbackhoax.id database and implementation of the Apriori algorithm, based on with and without undersampling method (example of a scenario of minimum support value 28).

TABLE II. One Example of Word Occurrence from The Apriori Algorithm

Without Undersampling Method		With Undersampling Method	
Dictionary false	Dictionary true	Dictionary false	Dictionary true
“Pesan beranta”	“Kait isu”	“Probowo dukung”	“Kait isu”
“Probowo dukung”	“Klarifikasi isu”	“Pesan beranta”	“Klarifikasi isu”
“Presiden Jokowi”	“Klasifikasi kait”	“Tinggal dunia”	“Klasifikasi kait”
“Sandi Prabowo”			
“Tinggal dunia”			
“Virus Corona”			

Based on the Table II, it is the initial result of the program after the text processing is done by Turnbackhoax.id database, it can be seen with the presence of additional from the under-sampling method can balance the data, compared to be unused. Aside from that advantage, there is one disadvantage found by researcher due to some erased random data (dominate), then for several case program, it will give false data, therefore, observation is needed from several scenarios for every time running the program so that the more precise result can be obtained.

Apart from the use of under-sampling method, there is another big influence that affected the result from the above table, which is the implementation of Apriori algorithm. The benefit of implementing the Apriori algorithm itself is in mining of hidden itemset combination from a database. The main process conducted in processing the Apriori algorithm is as follows [4]:

1. Join: in this process every item combined with other item until the combination pattern is no longer in form.
2. Prune: in this process the result of combined items then cut until reaching the pre-determined minimum support value by user. The determination of minimum support value in Apriori process is indirectly very influential since the bigger the value then the dataset chosen is the dataset that connected to each other.

B. Result Prediction Process

Prediction process is an indispensable process in this study, the result from the learning process can be the base in forming the classification model by using Random Forest algorithm. The result from Word Occurrence table can be a parameter to match the result of Table II with the prediction process data (as a step to integrate the Apriori process results with the Random Forest classification algorithm). The data used in prediction process is with the data from information news (tweet) in Twitter app. In Table III, it will contain the running time and mean absolute error from prediction process.

TABLE III. Running Time and Mean Absolute Error from Prediction Process

Mininum Support Value	With Undersampling		Without Undersampling	
	MAE	Mean Running Time (s)	MAE	Mean Running Time (s)
30	0.5 / 1	00:19	0 / 0.5	00:36
28	0.5 / 1	00:23	0 / 0.33	00:40
23	0 / 0.5	00:42	0.2 / 0.4	00:56
20	0 / 0.5	01:12	0.12 / 0.33	01:20
10	0.2 / 0.4 /0.6	09:14	0 / 0.005	02:55

The average time of running the program based on observation in 5 days period, with 30 scenarios of trial from every different form daily. The scenario of this study aims to ensure the most suitable minimum support value to be used in the to-be-built system, here the role of Mean Absolute Error (MAE) is presenting the incompatibility between the chosen dataset (training) with the data prediction (the tweet). So, the smaller the minimum support value then the bigger the unrelated data appeared, if there are a lot unrelated data appears then automatically the value of MAE will be bigger.

The condition assumption of assessment of Mean Absolute Error is, first class with false condition (non-hoax) $0 \geq \text{MAE} \geq 0.05$ and second class with true condition (hoax) $0.05 < \text{MAE} \leq 1$. Mean Absolute Error (MAE) is the average error in a data training. X_{ip} is the predicted result, X_{im} is the measured value from an $-i$ index, n is the total of data used, with formulation formula, (7) [20].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_{im}-X_{ip}|}{X_{im}} \tag{7}$$

C. System Evaluation

Table IV below is one of the data (tweet) that successfully predicted (by using the classification algorithm of Random Forest) and one of the trial scenarios as one of the observational attempts to see the accuracy result of the built model.

TABLE IV. ONE EXAMPLE OF THE RESULT FROM CONFUSION MATRIX TESTING

News	KPK menangkap Menteri Kelautan dan Perikanan (KKP) Edhy Prabowo. Penangkapan Edhy Prabowo terkait dengan ekspor benur / benih lobster. (25 / November / 2020).				
Confusion Matrix	$\begin{bmatrix} 2 & 0 \\ 3 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}$
Acurracy	40%	50%	100%	50%	0%
Precission	40%	0%	100%	0%	0%
Recall	100%	0%	100%	0%	0%
MAE	0.6	0.5	0	0.5	1
Minimum Support Value	10	20	23	28	30
Result Checked Manually	The classification process is not precise	The classification process is not precise	The classification process is accurate	The classification process is not precise	The classification process is not precise

Based on Table IV, it shows the result of confusion matrix form and a presence of manually checking process to prove the validity of system prediction process, so it can simplify the researcher in determining the suitable and appropriate minimum support value. The result of the table and from several observations with the same steps concludes that the most suitable and appropriate minimum support value is 23, which the value of True Positive (TP) estimates that the content of such news is predicted to be true and that the news itself is true. The minimum support value of 23 states the result from Mean Absolute Error program which is 0 and accompanied with manual checking which the news example has authentic information (non-hoax). The base manual checking parameters that is by looking at a valid resource of news, where many news sources discuss the related news, and also there is evidence to support the news as well as photos, or video.

V. CONCLUSION

In this study based on the analysis result towards the hoax information detection system on Twitter app by implementing Apriori and Random Forest algorithms, then it can be concluded that the process of integrating the result of Apriori algorithm implementation can simplify the prediction process (prediction process is using the algorithm from Random Forest), the result of the Apriori algorithm then formed a frequent itemset from dataset of Turnbackhoax.id, this frequent itemset became the important reference in prediction process from the system to create a more precise final result, this result is quite good when compared to using only one algorithm in building information detectors in the form of text data that grows very fast (real time data). This system formed two main classes as a description of the result of prediction process; first class with false condition (non-hoax) 0 ≥ Mean Absolute Error ≥ 0.05 and second class with true condition (hoax) 0.05 < Mean Absolute Error ≤ 1, where Mean Absolute Error (MAE) represents the incompatibility between each chosen dataset and tweet data. The unsuitable classification process in this study may be affected by several factors, such as unsuitable minimum support value, inappropriate data training used, missing pattern from Apriori process and possibility of source differences of training data and prediction data. However, the attempt to find minimum support value and appropriate system model must be accompanied by manual check to obtain the most precise result. Accuracy level, precision and recall from the built system in detecting hoax news information reached 100% with minimum support value of 23 (the most appropriate).

According to researcher, this study still experiencing several obstacles and deficiency. Recommendation for next research is first to do the needed trial with the semi-unsupervised learning. This technique will help researcher to process more samples. Second, conduct the calculation of confidence value from Apriori algorithm process which aims to obtain the exact minimum support value. Third, it is in need of a method to avoid further overfitting, probably by using a method from cross validation.

REFERENCES

[1] Atodiresei, C. S., Tănăselea, A., & Iftene, A. (2018). Identifying Fake News and Fake Users on Twitter. *Procedia Computer Science*, *126*, 451–461. https://doi.org/10.1016/j.procS.2018.07.279.

[2] Amir, S. N. N., Mohd, A. N. F., Chuprat, S., Sarkan, H. M., Yahya, Y., & Sam, S. M. (2019). SMS spam message detection using term frequency-inverse document frequency and random forest algorithm. *Procedia Computer Science*, *161*, 509–515. https://doi.org/10.1016/j.procs.2019.11.150.

[3] John, M., & Shaiba, H. (2019). Apriori-Based Algorithm for Dubai Road Accident Analysis. *16th International Learning & Technology Conference*, 218–227.

[4] Han, J., Kamber, M., & Pei, J. (2012). 8 - Classification: Basic Concepts. In *Data Mining Concepts and Techniques*. https://doi.org/10.1016/B978-0-12-381479-1.00008-3.

[5] Ahmed, I., Guan, D., & Chung, T. C. (2014). SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset. *International Journal of Machine Learning and Computing*, *4*(2), 183–187. https://doi.org/10.7763/ijmlc.2014.v4.409.

[6] Robu, V., & Dos Santos, V. D. (2019). Mining frequent patterns in data using apriori and eclat: A comparison of the algorithm performance and association rule generation. *2019 6th International Conference on Systems and Informatics, ICSAI 2019*, *Icsai*, 1478–1481. https://doi.org/10.1109/ICSAI48974.2019.9010367.

[7] Bhandari, A., Gupta, A., & Das, D. (2015). Improvised apriori algorithm using frequent pattern tree for real time applications in

data mining. *Procedia Computer Science*, *46*(Icict 2014), 644–651. https://doi.org/10.1016/j.procs.2015.02.115.

[8] Zeng, N., & Xiao, H. (2020). Inferring implications in semantic maps via the Apriori algorithm. *Lingua*, *239*, 102808. https://doi.org/10.1016/j.lingua.2020.102808.

[9] Breiman, L. (2001). ST4_Method_Random_Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004.

[10] Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, *1*(1), e9. https://doi.org/10.1002/spy2.9.

[11] Mother, A., & Alwahedi, A. (2018). Detecting Fake News in Social Media Networks. *Procedia Computer Science*, *141*, 215–222.

[12] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*(Itqm 2019), 503–513. https://doi.org/10.1016/j.procs.2019.12.017.

[13] Maas, A. E., Rottensteiner, F., & Heipke, C. (2019). A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. *Computer Vision and Image Understanding*, *188*(July), 102782. https://doi.org/10.1016/j.cviu.2019.07.002.

[14] Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, *193*, 115–122. https://doi.org/10.1016/j.neucom.2016.02.006.

[15] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3 PART 1), 5718–5727. https://doi.org/10.1016/j.eswa.2008.06.108.

[16] Tian, M., Zhang, L., Guo, P., Zhang, H., Chen, Q., Li, Y., & Xue, A. (2020). Data Dependence Analysis for Defects Data of Relay Protection Devices Based on Apriori Algorithm. *IEEE Access*, *8*, 120647–120653. https://doi.org/10.1109/ACCESS.2020.3006345.

[17] Miner, G., Nisbet, R., Fast, A., Delen, D., & Hill, T. (2012). Conceptual Foundations of Text Mining and Preprocessing Steps. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, 43–51. https://doi.org/10.1016/B978-0-12-386979-1.00003-7.

[18] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006.

[19] Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*, 772–794. https://doi.org/10.1016/j.ins.2019.06.064.

[20] Akdemir, B., & Çetinkaya, N. (2012). Long-term load forecasting based on adaptive neural fuzzy inference system using real energy data. *Energy Procedia*, *14*, 794–799. https://doi.org/10.1016/j.egypro.2011.12.1013.

Seminar_Int_2020_ICIDM_IEEE_Mailia.pdf

ORIGINALITY REPORT

6%	4%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Hepatitis Patient Classification using Random Forest Algorithms with Cost-Sensitive Method", International Journal of Engineering and Advanced Technology, 2020 Publication	2%
2	www.semanticscholar.org Internet Source	2%
3	Arif Dwi Laksito, Kusrini. "Apriori Algorithm Optimization using Temporary Table", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Publication	1%
4	www.spiedigitallibrary.org Internet Source	1%
5	Oxapisi Adikhresna, Retno Kusumaningrum, Budi Warsito. "Comparative Experimental Study of Multi Label Classification using Single Label Ground Truth with Application to Field Majoring Problem", 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019	1%

Publication

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		