Jurnal Teknik Komputer AMIK BSI
Volume 9, No.1, Januari 2023
P-ISSN 2442-2436, E-ISSN: 2550-0120
Akreditasi Ristekdikti, No: 36/E/KPT/2019 (Sinta 4)
DOI: 10.31294/jtk.v4i2

Prediction of PrivyID Application Comments Use as an Electronic Document (e-doc) using the Ensemble Vote method

Riza Fahlapi[1], Hermanto [2], Taufik Asra [3], Antonius Yadi Kuntoro[4],Ferry Syukmana[5] Ridatu Oca Nitra[6], Lasman Effendi [7]

[1,2,3,5,6,7] Universitas Bina Sarana Informatika
[4] Universitas Nusa mandiri

[1]e-mail: riza.rzf@bsi.ac.id
[2]e-mail: hermanto.hmt@bsi.ac.id
[3]e-mail: taufik.tas@bsi.ac.id
[4]e-mail: antonius.aio@nusamandiri.ac.id
[5]e-mail: ferry.fsk@bsi.ac.id
[6]e-mail: ridatu.rdo@bsi.ac.id
[7]e-mail: lasman.lef@bsi.ac.id

Diterima	Direvisi	Disetujui
22-10-2022	18-11-2022	07-01-2023

Abstract *- Indonesia is developing one of the more efficient and effective Financial Technology (Fintek) support services innovations by using electronic documents. The Electronic Document provider business that is used as a reference and utilized by fintech companies is PrivyID. In this study, how is the commentary aspect of using the PrivyID application for digital signature services to become a legal electronic document. Web-based application platforms and mobile applications in the community are indispensable for the use of Electronic Documents developed by PrivyID as a service provider in business and personal transactions that are needed by the community. More in-depth research regarding the Prediction of PrivyID Application Comments in Its Use as an Electronic Document (e-doc) taken from 818 data of PrivyID application users. The research was conducted by combining 3 (three) algorithms (k-Nearest Neighbor, Na¨ıve Bayes, and C4.5) in the Ensembles Vote method which resulted in the best Prediction Comment value with an accuracy of 86.80.*

Keywords: privyID, Ensemble Vote, Comments

INTRODUCTION

PrivyID as one of the Electronic Document Providers and Digital Identity Providers that has been certified at the Ministry of Communication and Information Technology of the Republic of Indonesia. In the development of Industry 4.0 where the industrial era connects all entities in it can communicate with each other in real time based on the use of internet technology and CPS (Cyber Physical System) in order to achieve the goal of creating new values or optimizing existing values from every process in the industry.

Support for business services using digital signatures will add value to services because they are more transparent, fast, easy, efficient and secure. The use of Digital Signatures has been accommodated in the Information and Electronic Transactions Act (UUITE) and its derivative regulations.In Indonesia, PrivyID has collaborated with the Ditjen Dukcapil and the Indonesian Ministry of Home Affairs for the process of verifying user data that proves the identity of the signer and maintains the integrity of the document contents.

The The legal basis and requirements for organizing electronic documents are listed in the Regulation of the Minister of Communication and Information Number 11 of 2018 concerning the implementation

of electronic certification Article 11-12 of the Law, Articles 13-14 of the Law, Articles 52-55,
Regulation of members of the board of governors number 19 /15/PADG/2017 concerning procedures for registration, submission of information, and monitoring of financial technology providers, Law 19 of 2016 (UUITE) and Articles 59-64 of Government Regulation Number 71 of 2019 concerning the Implementation of Electronic Systems and Transactions (PP PSTE) [4] Not only in the above provisions, technological progress also needs to meet the legal standards of the Electronic Certification Provider owned by PrivyID as amended by Law 19 of 2016 (UUITE).

In the use of electronic information and/or electronic documents and/or their printed results are valid legal evidence, electronicinformation and/or electronic documents and/or their printed results are extensions of valid evidence in accordancewith the procedural law in force in Indonesia, information electronic and/or electronic documents are declared valid if they use an electronic system in accordance with the provisions stipulated in the law.
Certified electronic signatures use a mathematical mechanism that is able to identify the identity of the signer and authenticate the contents of the signed electronic document. Digital certificates owned by PrivyID can be used as a guarantee for users of the privyID application because they contain information on the certificate owner, certificate issuer, certificate validity period, certificate use, and the cryptographic system used for application user personal data.

The use of electronic information and/or electronic documents and/or their printouts are valid legal evidence, electronic information and/or electronic documents and/or printouts are an extension of legal evidence under the applicable procedural law in Indonesia. electronic and/or electronic documents are declared valid if using an electronic system under the provisions stipulated in the Law. Certified electronic signatures use a mathematical mechanism that can identify the identity of the signer and authenticate the contents of the signed electronic document. Although the things above have become a guarantee for application users, support and the problem is that there are still many questions and comments from the public regarding the procedures for using this application platform.
From the information, legal basis and certification guarantee above, the researcher took 818 comments in the period from early January 2021 to March 2021 in the category of positive and negative comments that had been reviewed previously, the researcher predicted comments from the privyID application users using the k- Nearest Neighbor (k-Nearest Neighbor) algorithm. -NN), Naïve Bayes and C4.5 in the Ensembles Vote method which combines the 3 algorithms.
From the testing process, the existing dataset is entered into Rapidminer. Starting with the formation of a data model in the first part, textmining management, training data formation, and data testing to the accuracy value. The size of the confusion matrix, performance, accuracy and AUC results are displayed in the form of an ROC curve, to find out the best level of accuracy

METHODS

The PrivyID application which is available on the Web base and Mobile Application can be used and downloaded through the platforms found on the Appstore (Apple Inc Users) or Google Play (Android), where 818 data were taken from user comments. To test the dataset using the k-Nearest Neighbor (k-NN) Algorithm, Na¨ıve Bayes and C4.5 through the Ensembles Vote method which combines the 3 algorithms. The results of the algorithm test will present the accuracy and AUC values which show the predictions of PrivyID user comments.

a. Gata Framework
Basic processing of data processing through text mining process, text documents in conditions of irregular text data structures, require several initial stages, One of the implementations of text mining is the Text Preprocessing stage carried out through smoothing applications on the data structure also used online tools created by Dr. Windu Gata, M.Kom to convert unstructured text data into structured text data so that the data is ready to be used on predetermined modeling techniques Tools are accessed at http://www.gataframework.comstage.

b. RapidMiner
RapidMiner is software that uses the Java language which is open (open source) to perform analysis of data mining, text mining and predictive analysis. using a variety of descriptive and predictive techniques to provide insight to users so that they can make the best decisions. Has approximately 500 data mining operators, including operators for input, output, data preprocessing and visualization for data analysis and as a data mining engine that can be integrated into its own product so that it can work on all operating systems.
c. k-Nearest Neighbour (k-NN)
This algorithm classifies new objects based on attributes and training samples. To be matched and only based on memory given the query point, it will find a number of k objects or (training points) closest to the query point. The most voting among the classifications of k objects, the k-Nearest Neighbor (k-NN) algorithm uses the neighboring

classification as the predictive value of the new query instance. The K-Nearest Neighbor (k-NN) method algorithm works based on the shortest distance from the query instance to the training, sample to determine the k-NN.[9].

d. Algoritma C4.5
C4.5 is a flowchart-like structure where each internal node (a node that is not a leaf or a node is not the outermost). testing of the attribute variables of each branch is the result of the test, while the outermost node, namely the leaf, becomes the label.

e. Naïve Bayes
The method of Bayes' theorem, discovered by Thomas Bayes in the 18th century. is a statistical classification that can be used to predict the probability of membership of a class. The NB method takes two stages in the text classification process, namely the training stage and the classification stage. At the training stage, an analysis process is carried out on the sample document in the form of vocabulary selection, namely words that may appear in the sample document collection that as far as possible can be a document representation.Determination of the probability for each category based on a sample of documents. At the classification stage, the category value of a document is determined based on the terms that appear in the classified document.

f. Ensemble Vote
An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The hypotheses contained therein need not have the space of the models built. Thus, the ensemble can prove to have more flexibility in the functions that the algorithm can represent. This flexibility, in theory, allows them to fit more than one training data. ensemble modeling techniques tend to reduce problems related to fitting training data.

RESULT AND DISCUSSION

a. Business Understanding
An understanding of the object of research is done by digging up information through a review of the PrivyID application. Comment prediction is done to find a classification method that can help in determining positive and negative news article comments. There are several stages carried out in this research, including:

1. The process of collecting comment data related to PrivyID. as many as 818 data on public comments consisting of data on positive comments and data on negative comments;
2. The text preprocessing stage, in which the weighting and Vector Creation will be carried out using the Term Frequency Iners Document Frequency (TF-IDF);
3. The next stage uses the Ensembles Vote algorithm method in which there are k-NN, NB, C4.5 algorithms and for classification. The classification results will be re-optimized to get the best algorithm method.
4. Algorithm accuracy will be measured by Confusion Matrix and the results will be displayed in the form of ROC and Accuracy curves.accuracy.

b. Data Understanding
At the data understanding stage, the raw data collection process is carried out according to the required attributes. By using the source of the data obtained, a dataset is created with attributes, namely data comments from each data. The comment data is combined and stored in the form of an .xls extension.

c. Data Preparation
The data preparation stage is the stage with the data preparation process that aims to obtain clean and ready data for use in research. In text mining, the initial stage to be carried out is the text preprocessing stage, at this stage the researcher uses the Gata Framework Tools. The following are the steps carried out in text preprocessing.

d. Remove Duplicate
Delete Duplicates is used to delete the same data in the row data text. thus avoiding data double that is in the line text.

e. Nominal to Text
Nominal to text is an operator in rapidminer that functions to convert all numbers in the text into a text. So that the existing numbers will be considered as text data types, not numeric or nominal.text.

f. Transform Case
The operator used is to convert capital letters to lowercase letters. for uniformity of letters and there is no error in the tokenize process.

g. Filter Token (By Length)
In data preparation to eliminate a number of words (after the tokenize process) with a certain character length. In this study, the minimum length of the characters used is 4 characters and the maximum length is 25 characters. Words longer than 25 characters will be omitted.

h. Filter Stopword (Dictionary)
Next is the use of the Stopword Removal operator (by Directory) which functions to remove words that are not related to the content of the text. In the previous stage using the Gata framework text mining service has been done, for some words that cannot

be removed by the service with the Stopword Removal (by Directory) operator, researchers can register words that should be removed from the text.process.

i. Modelling Stages
Is the stage of selecting mining techniques by determining the algorithm to be used. The tool used is RapidMiner
version 9.1. The results of the model testing carried out are to correctly classify complain email and not complain
email using the Naive Bayes algorithm and Support Vector Machine to get the best accuracy value.

j. Model Testing with Algorithms
Is The settings and use of operators and parameters in the Rapid Miner framework greatly affect the accuracy and the model formed, more clearly the testing of the algorithm value.

k. Evaluation Model
The evaluation stage aims to determine the usability value of the model that has been successfully created in the previous step. The Ensembles Vote algorithm is used to train the input dataset and then the results are used for testing using the same dataset. The following is the design process for testing the Ensembles Vote method model used.

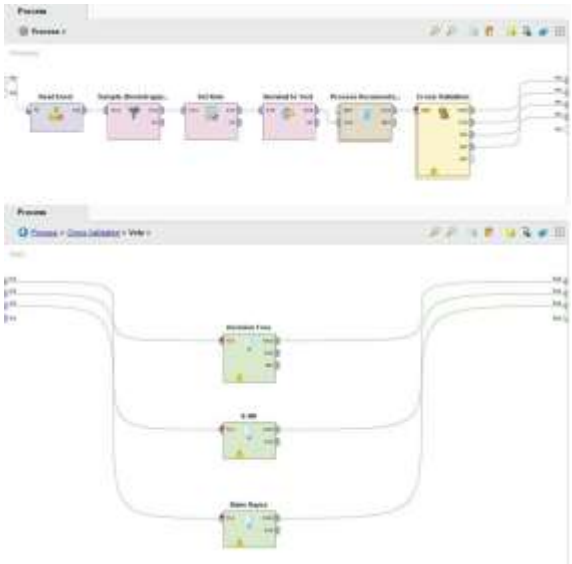

Figure 1. Testing Stages

Figure 2. Design Process Model for Vote Ensembles

namely testing the data used is clean data that has gone through preprocessing. the data is taken from the Read Excel operator, this is done because the dataset is stored in Excel form. Process documents from files to convert files into documents. From the results of the modeling that has been done previously. The following will explain the ROC Curve and Confusion Matrix of each algorithm.value.

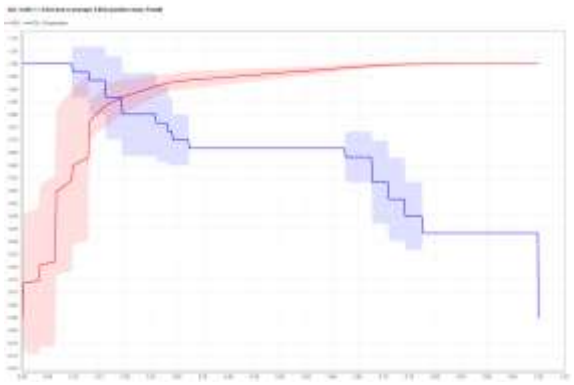

Figure 3. Testing Stages

l. Confusion Matrix Ensembles Vote
In this research, the calculation results of the Ensemble Vote method get the value that produces the best Comment Prediction value with an accuracy of 86.80% and an AUC value of 0.883 which is used to make it easier to find out positive comments and negative comments. Based on the comment data processed using the Rapid Miner Tools, the comment data will be separated into words that weigh each word. These words will be used to see words related to words that appear frequently and have the highest weight and can be used to find out positive comments and negative comments.

CONCLUSION

Conclusions that can be summarized regarding the prediction of comments on PrivyID, the following conclusions can be drawn:

a. To predict PrivyID Application comments, use the k- Nearest Neighbor (k-NN), Na¨ıve Bayes, and C4.5 algorithm using the Ensemble Vote method with accuracy of 86.80% and AUC value of 0.883
b. Positive comments are the result of confidence in comment predictions based on the values in the calculations using Rapidminer because they have a high percentage. Meanwhile, for validation that is more important and determines predictions in the above conditions, it has a good classification interpretation.

REFERENCES

Yuniati, T., Sidiq, M. F. (2020). Literature Review: Legalisasi Dokumen Elektronik Menggunakan

Tanda Tangan Digital sebagai Alternatif Pengesahan Dokumen di Masa Pandemi. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(6).

Informatika, M. K. dan. 2018. penyelenggaraan sertifikasi elektronik. 283.

Finansial, T., Dewan, A., Bank, G. Peraturan anggota dewan gubernur nomor 19/15/PADG/2017. , (2021).

Indonesia, R. (2019). PP Nomor 71 Tahun 2019. 1–90.

Fay, D. L. (1967). PrivyID Company Overview en. Angewandte Chemie International Edition, 6(11), 951–952.

Fahlapi, R., Hermanto, H., Kuntoro, A. Y., Effendi, L., Nitra, R. O., Nurlela, S. (2020). Prediction of Employee Attendance Factors Using C4.5 Algorithm, Random Tree, Random Forest. Semesta Teknika, 23(1), 39–53.

Gata, W.2017. Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS. 6, 1–13.

Deviyanto, A., Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. JISKA (Jurnal Informatika Sunan Kalijaga), 3(1), 1.

Sulaeman, F. S., Rilmansyah, M. A. 2021. Aplikasi Penerapan Algoritma C45 untuk Memprediksi Kelulusan Mahasiswa Berbasis Web. Jurnal Media Teknik Sistem Industri, 5, 41–54.

Hermanto, Kuntoro, A. Y., Asra, T., Pratama, E. B., Effendi, L., Ocanitra, R. (2020). Gojek and Grab User Sentiment Analysis on Google Play Using Naive Bayes Algorithm and Support Vector Machine Based Smote Technique. Journal of Physics: Conference Series,1641(1).https://doi.org/10.1088/1742-6596/1641/1/012102

Alhamad, A., Azis, A. I. S., Santoso, B., Taliki, S. (2019). Prediksi Penyakit Jantung Menggunakan Metode-Metode Machine Learning Berbasis Ensemble – Weighted Vote. 5(3), 352–360

Hermanto, B., SN, A. (2017). Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 11(1), 43. https://doi.org/10.22146/ijccs.15946

Ibrahim, D. (2017). Analisis Hubungan antar Faktor dan Komparasi Algoritma Klasifikasi pada Penentuan Penundaan Penerbangan. Senit, (September), 15–17.

Rosyidi, R. (2019). Perbandingan Algoritma K-Nn Dan Cart Pada Data. 4(2), 169–177.

Salini, A., Jeyapriya, U., College, S. M., College, S. M. (2018). A Majority Vote Based Ensemble

Fauziah, S., Sulistyowati, D. N., Asra, T. (2019). Optimasi Algoritma Vector Space Model Dengan Algoritma K-Nearest Neighbour Pada Pencarian Judul Artikel Jurnal. Jurnal Pilar Nusa Mandiri, 15(1), 21–26. https://doi.org/10.33480/pilar.v15i1.27

Vulandari Tri Retno. (2017). Data Mining Teori dan Aplikasi Rapidminer.

R. A. Anggraini, G. Widagdo, A. S. Budi, and M. Qomaruddin, “Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Na¨ıve Bayes,” J. Sist. dan Teknol. Inf., vol. 7, no. 1, p. 47, 2019, doi: 10.26418/justin.v7i1.30211.

Pareza Alam Jusia, “Analisis komparasi pemodelan algoritma decision tree menggunakan metode particle swarm optimization dan metode adaboost untuk prediksi awal penyakit jantung,” Semin. Nas. Sist. Inf. 2018, pp. 1048–1056, 2018.