Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. x (2022) 872 - 876 ISSN Media Electronic: 2580-0760

Application of The Naïve Bayes Classifier Algorithm to Classify Community Complaints

Keszya Wabang[1], Oky Dwi Nurhayati[2], Farikhin[3]
[1,2,3]Information System, Graduate School, Diponegoro University
[1]keszyaw1@gmail.com, [2]okydn@undip.ac.id, [3]farikhin.math.undip@gmail.com

Abstract

Unsatisfactory public services encourage the public to submit complaints/ reports to public service providers to improve their services. However, each complaint/ report submitted varies. Therefore, the first step of the community complaint resolution process is to classify every incoming community complaint. The Ombudsman of The Republic of Indonesia annually receives a minimum of 10,000 complaints with an average of 300-500 reports per province per year, classifies complaints/ community reports to divide them into three classes, namely simple reports, medium reports, and heavy reports. The classification process is carried out using a weight assessment of each complaint/ report using 5 (five) attributes. It becomes a big job if done manually. This impacts the inefficiency of the performance time of complaint management officers. As an alternative solution, in this study, a machine learning method with the Naïve Bayes Classifier algorithm was applied to facilitate the process of automatically classifying complaints/ community reports to be more effective and efficient. The results showed that the classification of complaints/ community reports by applying the Naïve Bayes Classifier algorithm gives a high accuracy value of 92%. In addition, the average precision, recall, and f1-score values, respectively, are 91%, 93%, and 92%.

Keywords: *classification, complaints/ community reports, Naïve Bayes Classifier*

1. Introduction

One of the essential things in the implementation of the state is the existence of public services to meet the needs of every community for goods, services, and administrative services organized by public service providers [1]. However, the organizers are not always optimal in carrying out public services. They can make mistakes that result in dissatisfaction and harm the community materially and immaterially [2]. This encourages the public to provide complaints or complaints so that the organizers improve their services [3]. In general, public service providers usually provide a forum for the public to submit their complaints directly by visiting the complaint service counter or indirectly through the contact number of the complaint officer [4][5].

The public service problems experienced by each community are different, so each complaint/report submitted varies. Therefore, the first step of the community complaint resolution process is classifying every community complaint. This is intended to make complaint handling more effective and efficient [6]. The classification of complaints is carried out by several government agencies when receiving complaints/ community reports. One of which is the Ombudsman of the Republic of Indonesia as a State Institution for Public Service Supervisors in Indonesia, which annually receives a minimum of 10,000 complaints with an average of 300-500 reports per province per year. The Indonesian Ombudsman classifies complaints/community reports into three classes: simple reports, medium reports, and heavy reports. The classification process uses a manual weight assessment of each complaint/report using 5 (five) attributes.

Classifying each complaint/report that comes in is undoubtedly a big job, especially if, at one time, the number of complaints that come in is vast. Unfortunately, this can result in less efficient performance for complaint management officers [7]. Therefore, it is necessary to automatically classify complaints/community reports so that the process becomes more accessible, faster, and more precise, making complaint management officers' performance time more efficient. This can be done by applying machine learning methods using specific algorithms [8][9].

In machine learning, classification identifies a database and then organizes and collects it into one of the

Accepted: 12-09-2022 | Received in revised: 16-10-2022 | Published: 31-04-2022

predetermined classes. In classification, the classifier algorithm trains a data set with predefined classes to form a classification pattern that will be applied to new data as inputs. This is also called supervised learning. Classification is applied by involving attributes as parameters or criteria to define data classes. This process is carried out systematically to find information based on the data records obtained [10].

Several previous studies have classified complaints but carried out a sentiment analysis on complaints' text using a semi-supervised min cuts algorithm. It classifies the text of complaints/community reports into two classes, namely reports and non-reports, with an accuracy of 83.8% [11]. Meanwhile, in this study, the Naïve Bayes Classifier algorithm will be applied to classify complaints/reports into three classes: simple reports, medium reports, and heavy reports involving 5 (five) attributes. Naïve Bayes Classifier is a simple algorithm but can classify with high accuracy [12]. Moreover, Naïve Bayes Classifier is known to have a degree of accuracy and works better than other classifier algorithms [13].

2. Research Methods

In this study, the method used was machine learning with the Naïve Bayes Classifier algorithm. The research flow is presented in figure 1.

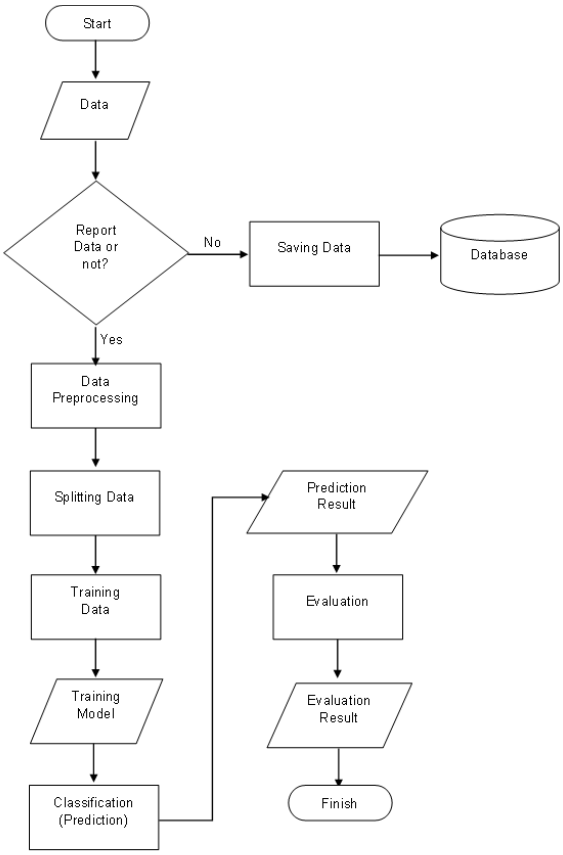


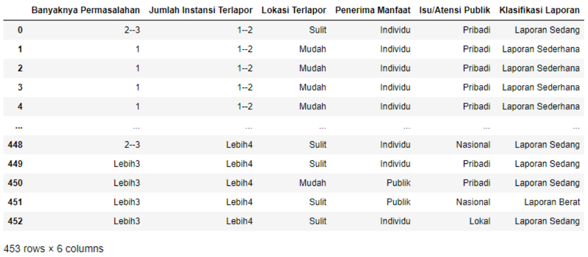
Figure 1. Research Flow

2.1 Data

The data for this study was obtained from the the Ombudsman of the Republic of Indonesia, which was limited to data on community complaints/reports in East Nusa Tenggara Province (NTT). The data was 453 records from January 1, 2019, to December 31, 2020. The complaint data/report consists of the Reporter's identity and the report data.

2.2 Selection Data

The data obtained is then processed using the Python programming language version 3.10.0. The data that is further processed is only part of the report data. The Reporter's identity containing the name, NIK, gender, occupation, and the mobile number does not need to be processed. The data in this study includes 5 (five) attributes as free variables (x) consisting of the number of problems, the number of reported agencies, the location of the reported, the beneficiaries, and public issues/attention. The label/class used is the report's classification as a bound variable (y). A snippet of the data to be further processed is shown in the following figure 2.



	Banyaknya Permasalahan	Jumlah Instansi Terlapor	Lokasi Terlapor	Penerima Manfaat	Isu/Atensi Publik	Klasifikasi Laporan
0	2--3	1--2	Sulit	Individu	Pribadi	Laporan Sedang
1	1	1--2	Mudah	Individu	Pribadi	Laporan Sederhana
2	1	1--2	Mudah	Individu	Pribadi	Laporan Sederhana
3	1	1--2	Mudah	Individu	Pribadi	Laporan Sederhana
4	1	1--2	Mudah	Individu	Pribadi	Laporan Sederhana
...	...	...	...	...	...	...
448	2--3	Lebih4	Sulit	Individu	Nasional	Laporan Sedang
449	Lebih3	Lebih4	Sulit	Individu	Pribadi	Laporan Sedang
450	Lebih3	Lebih4	Mudah	Publik	Pribadi	Laporan Sedang
451	Lebih3	Lebih4	Sulit	Publik	Nasional	Laporan Berat
452	Lebih3	Lebih4	Sulit	Individu	Lokal	Laporan Sedang

453 rows × 6 columns

Figure 2. Data Snapshot after Data Selection on Jupyter Notebook

2.3 Data Pre-processing

Data pre-processing is carried out to ensure that the data to be processed further is quality data. This is because quality data will produce a learning process (training), classification, information, and quality decisions [14]. In this study, the data pre-processing consisted of checking the missing data (missing value) and transforming the data, as shown in figure 3. The results of checking the missing data (missing value) are shown in figure 4.

Based on figure 4 above, it is known that no value is missing or all data records have been filled, so there is no need to impute values on existing data.

Furthermore, based on figure 4, it is also known that each attribute and label/class has an object or categorical data type. However, in the application of machine learning, computers cannot process categorical type data, so existing data needs to be converted into numerical data. This is also called changing the shape of the data or transforming the data [15]. The results of the data transformation are shown in the following figures 5 and 6.

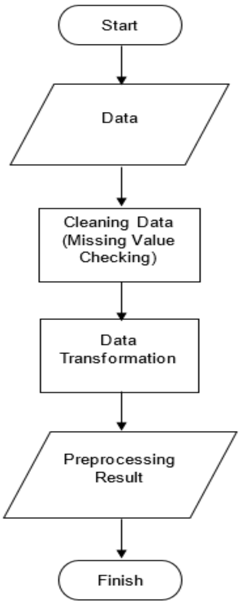
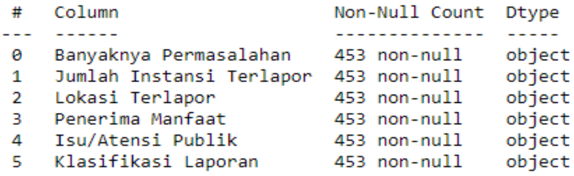
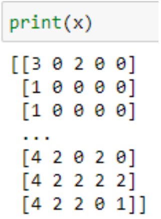


Figure 3. Data Pre-processing Flow



```
#   Column                     Non-Null Count  Dtype
---  ------                    --------------  -----
0   Banyaknya Permasalahan     453 non-null    object
1   Jumlah Instansi Terlapor   453 non-null    object
2   Lokasi Terlapor            453 non-null    object
3   Penerima Manfaat           453 non-null    object
4   Isu/Atensi Publik          453 non-null    object
5   Klasifikasi Laporan        453 non-null    object
```

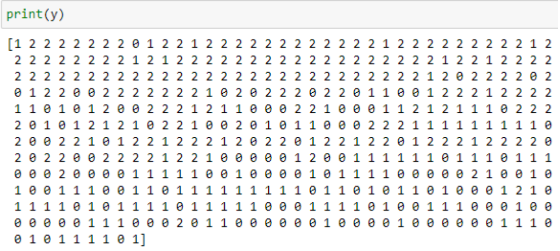
Figure 4. Missing Value Checking Results



```
print(x)
```

```
[[3 0 2 0 0]
 [1 0 0 0 0]
 [1 0 0 0 0]
 ...
 [4 2 0 2 0]
 [4 2 2 2 2]
 [4 2 2 0 1]]
```

Figure 5. Data Transformation Results on 5 (five) Attributes



```
print(y)
```

```
[1 2 2 2 2 2 2 2 0 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2
 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 0 2 2 2 2 0 2
 0 1 2 2 0 0 2 2 2 2 2 2 2 1 0 2 0 2 2 2 0 2 2 0 1 1 0 0 1 2 2 2 1 2 2 2 2
 1 1 0 1 0 1 2 0 0 2 2 2 1 2 1 1 0 0 0 2 2 1 0 0 0 1 1 2 1 2 1 1 1 0 2 2 2
 2 0 1 0 1 2 1 2 1 0 2 2 1 0 0 2 0 1 0 1 1 0 0 0 2 2 2 1 1 1 1 1 1 1 1 1 0
 2 0 0 2 2 1 0 1 2 2 1 2 2 2 1 2 0 2 2 0 1 2 2 1 2 2 0 1 2 2 2 1 2 2 2 2 0
 2 0 2 2 0 0 2 2 2 2 1 2 2 1 0 0 0 0 0 1 2 0 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1
 0 0 0 2 0 0 0 0 1 1 1 1 1 0 0 1 0 0 0 0 1 0 1 1 1 1 0 0 0 0 0 2 1 0 0 1 0
 1 0 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 1 0 1 0 0 0 1 2 1 0
 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 0 1 0 0 1 1 1 0 0 0 1 0 0
 0 0 0 0 0 1 1 1 0 0 0 2 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 1 1 0
 0 1 0 1 1 1 1 0 1]
```

Figure 6. Data Transformation Results on Labels/Classes

After the data is successfully transformed into numerical data, it can be seen that the existing data do not have a long distance of values between one data and another, so there is no need for a scaling process. Therefore, existing data can be processed at the next stage, namely, conducting data training.

2.4 Splitting Data and Training Data

Before conducting data training, it is necessary to share data trains and test data with a ratio of 80% of train data, namely 362 records and 20% of test data, which is 91 records. Furthermore, conduct a training process on the data train with the Naïve Bayes Classifier algorithm to obtain a classification pattern that will be applied to the test data.

2.5 Naïve Bayes Classifier

The Naïve Bayes Classifier is a statistical classification algorithm based on Bayes' theorem initiated by Thomas Bayes, a conformist English clergyman. He conducted early studies on probability and decision theory during the 18th century—Bayes' theorem formula as equation (1).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (1)$$

P(H|X) is the posterior probability (posterior), or the final probability of the condition *H* on *X*. *P(H)* is also called the prior probability (prior) or the initial probability of *H*. Similarly, *P(X|H)* is the posterior probability (posterior), or the final probability of the condition *X* on *H*. Whereas *P(X)* is the prior probability (prior), or the initial probability of *X*. Bayes' theorem is beneficial because it provides a calculation of the final probability of *P(H|X)* of *P(H)*, *P(X|H)*, and *P(X)* [16]. Furthermore, studies comparing classification algorithms have obtained findings, namely a simple Bayesian classification of Bayesian theorem known as the Naïve Bayes Classifier that provides high accuracy and performance for its application to large databases. Naïve Bayes Classifier is one of the supervised learning algorithms to perform classification with a probability approach that calculates the likelihood of each attribute so that effective results can be obtained in a short (efficient) way [17]. Naïve Bayes Classifier assumes that an attribute's value in a class does not depend on the value of another attribute. This assumption is called conditional class independence. This is done to simplify complex calculations and is considered "naïve".

In the Naïve Bayes Classifier, suppose there is an *m* class, $C_1, C_2, \ldots, C_m$. Given a tuple *X*, the classifier will predict that *X* will belong to the class with the highest posterior probability value, conditioned on *X*. This means the Naïve Bayes Classifier predicts that tuple *X* belongs to a class C_i if and only if $P(C_i\,|\boldsymbol{X}) > P(C_j\,|\boldsymbol{X})$ for $1 \le j \le m, j \ne i$. Thus, $P(C_i\,|\boldsymbol{X})$ maximized. Class C_i for $P(C_i\,|\boldsymbol{X})$ maximized the so-called maximum posterior hypothesis. Based on Bayes' theorem as equation (1), equation (2) is obtained.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (2)$$

DOI: https://doi.org/10.29207/resti.v6i5.4498

If there is a set of data with many attributes, it also takes a lot of computation to calculate $P(X|C_i)$. For the calculation to be reduced to evaluating $P(X|C_i)$, the naïve assumption of "class independence - conditional" is carried out. It is created by assuming that the values of the attributes are conditionally mutually independent. Thus obtained, equation (3).

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (3)$$

2.6 Classification

After conducting data training with the Naïve Bayes Classifier algorithm and obtaining a classification model, the next stage is to classify the Naïve Bayes Classifier algorithm on the existing test data.

2.7 Evaluation

After the classification process is completed, the next stage is to evaluate the classification results. At this stage, accuracy, precision, recall, and f1-score values are calculated by utilizing the confusion matrix. The confusion matrix is a matrix that shows the results of actual classification and predictions with the size of LxL, where L is the number of labels/classification classes. In this study, the confusion matrix used was 3x3 because it had 3 (three) labels/classes [18]. The terms of the 3x3 confusion matrix are shown by the following figure 7.

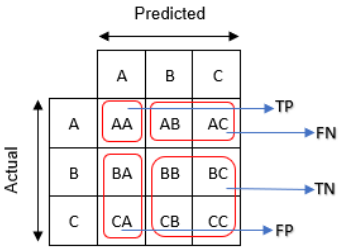


Figure 7. Confusion Matrix 3x3

The scores of accuracy, precision, recall, and f1-score can be calculated by the following equations (4) – (7).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP\ Total}{Dataset\ Total} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{Prediction\ Total} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{Actual\ Total} \quad (6)$$

$$f1-score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (7)$$

3. Results and Discussions

The data training process carried out on the data train as many as 362 records using the Naïve Bayes Classifier algorithm, and the classification applied to the test data as many as 91 records produced accuracy as presented in figure 8 below.

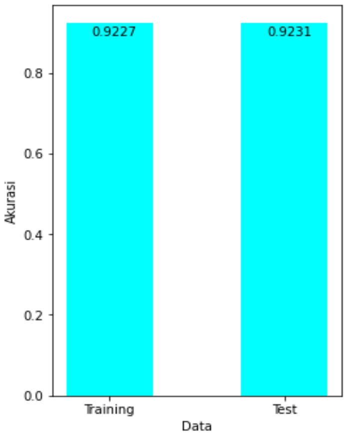


Figure 8. Training Data Accuracy and Classification Graph

Figure 8 shows that the training data and classification of the test data with the Naïve Bayes Classifier algorithm both show good results by achieving an accuracy of 92% and are only slightly different. The following is shown the confusion matrix for evaluating classification results that have been carried out with the Naïve Bayes Classifier algorithm, as shown in figure 9.

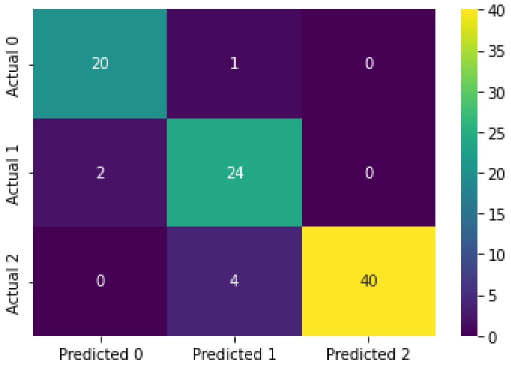


Figure 9. Confusion Matrix Classification Results with Naïve Bayes Classifier Algorithm

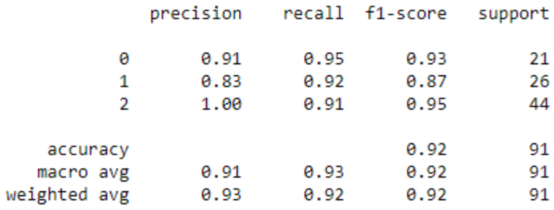
Based on the 3x3 confusion matrix in figure 9, predictions to determine the classification of complaints/community reports were made on 91 records. Prediction results for the classification of simple reports labeled 0 by 20, the classification of medium reports labeled by 1 by 24, and the classification of heavy reports labeled by 2 by 40.

Furthermore, to determine the performance of the Naïve Bayes Classifier algorithm to classify complaints/ community reports, an evaluation is carried out by calculating accuracy, precision, recall, and f1-score scores. The results obtained are shown in the following figure 10.

In figure 10, the resulting accuracy is 92%, with the average scores for precision, recall, and f1-scores, respectively, being 91%, 93%, and 92%. This result shows that the Naïve Bayes Classifier algorithm

DOI: https://doi.org/10.29207/resti.v6i5.4498

performs well in classifying community complaints/reports based on the classification pattern of data. It has been obtained into the class of superficial, medium, and heavy reports involving 5 (five) attributes of provisions.



```
              precision    recall  f1-score   support

           0       0.91      0.95      0.93        21
           1       0.83      0.92      0.87        26
           2       1.00      0.91      0.95        44

    accuracy                           0.92        91
   macro avg       0.91      0.93      0.92        91
weighted avg       0.93      0.92      0.92        91
```

Figure 10. Scores Accuracy, Precision, Recall, and f1-Score for Each Label/Class Predicted with The Naïve Bayes Classifier Algorithm

4. Conclusion

Classification of complaints/community reports can be done with the Naïve Bayes Classifier algorithm. The Naïve Bayes Classifier algorithm has shown good performance by producing a high accuracy value of 92%. In addition, the average scores of precision, recall, and f1-scores obtained are 91%, 93%, and 92%, respectively.

This research can be developed by applying other classification algorithms to compare the best performance in classifying complaints/community reports.

Acknowledgment

The authors expressed gratitude to the Ombudsman of the Republic of Indonesia for contributing by providing research data. In addition, a thank you was also conveyed to the Indonesia Endowment Fund for Education (LPDP) for providing financial support to implementing this research.

References

[1] N. Safarov, “Personal experiences of digital public services access and use: Older migrants’ digital choices,” *Technol. Soc.*, vol. 66, no. May, p. 101627, 2021, doi: 10.1016/j.techsoc.2021.101627.

[2] S. Shokouhyar, S. Shokoohyar, and S. Safari, “Research on the influence of after-sales service quality factors on customer satisfaction,” *J. Retail. Consum. Serv.*, vol. 56, no. May, p. 102139, 2020, doi: 10.1016/j.jretconser.2020.102139.

[3] J. L. Stevens, B. I. Spaid, M. Breazeale, and C. L. Esmark Jones, “Timeliness, transparency, and trust: A framework for managing online customer complaints,” *Bus. Horiz.*, vol. 61, no. 3, pp. 375–384, 2018, doi: 10.1016/j.bushor.2018.01.007.

[4] H. Moon, W. Wei, and L. Miao, “Complaints and resolutions in a peer-to-peer business model,” *Int. J. Hosp. Manag.*, vol. 81, no. May, pp. 239–248, 2019, doi: 10.1016/j.ijhm.2019.04.026.

[5] N. Yan, X. Xu, T. Tong, and L. Huang, “Examining consumer complaints from an on-demand service platform,” *Int. J. Prod. Econ.*, vol. 237, no. March, 2021, doi: 10.1016/j.ijpe.2021.108153.

[6] D. C. Ferreira, R. C. Marques, A. M. Nunes, and J. R. Figueira, “Customers satisfaction in pediatric inpatient services: A multiple criteria satisfaction analysis,” *Socioecon. Plann. Sci.*, 2021, doi: 10.1016/j.seps.2021.101036.

[7] S. Khedkar and S. Shinde, “Deep Learning and Ensemble Approach for Praise or Complaint Classification,” *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 449–458, 2020, doi: 10.1016/j.procs.2020.03.254.

[8] A. Ghazzawi and B. Alharbi, “Analysis of Customer Complaints Data using Data Mining Techniques,” *Procedia Comput. Sci.*, vol. 163, pp. 62–69, 2019, doi: 10.1016/j.procs.2019.12.087.

[9] Y. HaCohen-Kerner, R. Dilmon, M. Hone, and M. A. Ben-Basan, “Automatic classification of complaint letters according to service provider categories,” *Inf. Process. Manag.*, vol. 56, no. 6, pp. 1–20, 2019, doi: 10.1016/j.ipm.2019.102102.

[10] E. K. Jacob, “Classification and categorization: A difference that makes a difference,” *Libr. Trends*, vol. 52, no. 3, 2004.

[11] A. Singh, S. Saha, M. Hasanuzzaman, and A. Jangra, “Identifying complaints based on semi-supervised mincuts,” *Expert Syst. Appl.*, vol. 186, no. November 2020, p. 115668, 2021, doi: 10.1016/j.eswa.2021.115668.

[12] H. Zhang *et al.*, “Developing novel computational prediction models for assessing chemical-induced neurotoxicity using naïve Bayes classifier technique,” *Food Chem. Toxicol.*, vol. 143, no. July, 2020, doi: 10.1016/j.fct.2020.111513.

[13] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, “Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier,” *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[14] C. Patgiri and A. Ganguly, “Adaptive thresholding technique based classification of red blood cell and sickle cell using Naïve Bayes Classifier and K-nearest neighbor classifier,” *Biomed. Signal Process. Control*, vol. 68, no. April, 2021, doi: 10.1016/j.bspc.2021.102745.

[15] C. Zhou *et al.*, “Recognizing black point in wheat kernels and determining its extent using multidimensional feature extraction and a naive Bayes classifier,” *Comput. Electron. Agric.*, vol. 180, no. November 2020, 2021, doi: 10.1016/j.compag.2020.105919.

[16] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining concepts and techniques*, Morgan Kaufmann Publishers, 2014.

[17] S. Kapoor, R. Verma, and S. N. Panda, “Detecting kidney disease using Naïve bayes and decision tree in machine learning,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 498–501, 2019, doi: 10.35940/ijitee.A4377.119119.

[18] Y. Ghatas, M. Fayek, and M. Hadhoud, “A hybrid deep learning approach for musical difficulty estimation of piano symbolic music,” *Alexandria Eng. J.*, vol. 61, no. 12, pp. 10183–10196, 2022, doi: 10.1016/j.aej.2022.03.060.