2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)

Network Disruption Prediction Using Naïve Bayes Classifier

Shinta Oktaviana, Iklima Ermis
Dept. Computer and Informatics Engineering
Politeknik Negeri Jakarta
Indonesia
shinta.oktaviana@tik.pnj.ac.id,
iklimaermis.ismail@tik.pnj.ac.id

Mila Desi Anasanti
Section of Genomics of Common Disease,
Dept. Medicine
Imperial College London,
UK
m.anasanti15@imperial.ac.uk

Jehad Hammad
Dept. of Computer and Information Systems
Al-Quds Open University,
Bethlehem,
Palestine
jhammad35@hotmail.com

***Abstract*— The most crucial challenge of internet service providers is to assure the availability and reliability of their services to their customers. The companies should prevent the customer's complaint by recognizing a potential disruption for the customers, especially in the category 'under spec' condition (potentially impaired service). This study proposed and implemented a model using the Naïve Bayes classifier to classify and detect the potential disruption of network services to prevent customer's complaints about their service. The criteria for this model prediction are revenue number of each customer (REVENUE), recurrent disruption value of ODP (N_Q), attenuation value in ODP (OLT), and attenuation value in customer (ONU). The data classified into three classes or conditions, namely GREEN representing no network disruption, YELLOW is representing low-level disruption, and RED representing high-level disruption, which needs more attention to follow up. The result obtained 91.89% accuracy of the model performance using WEKA Tool.**

Keywords—network disruption detection, naïve Bayes classifier, attenuation, revenue, customer complaint, ODP

I. INTRODUCTION

The most crucial challenge of internet service providers is to assure the availability and reliability of their services to their customers [1]–[3]. As the number of internet users increases, the complaint about the services disruption and internet network speed become the highlight for the internet service provider companies. Therefore, companies should prevent their customer's complaints by identifying the potential disruption, especially for their customers in the category 'under spec' condition (potentially impaired service).

Customer's complaint is feedback to show their poor attitude towards their expectations of a product or service. It is the best indicator for companies to assess whether their services operate properly [4-6]. Companies should pay attention to the customer's complaint as it will be detrimental to them if the complaint has not appropriately handled in the shortest possible time. A customer complaint is a part of customer service related to Customer Relationship Management (CRM) that obligate to take care of customer's satisfaction, accommodate the customer's views and receive customer's complaint [7][8].

Identifying the cause of the complaint earlier is the best way to reduce the number of customer complaints. Prevention achieved by developing a model using the Naïve Bayes Classifier to predict, diagnose, and classify the causes of the network disruption. Naïve Bayes Classifier is widely used in many fields and environments, including in the network services. There has been much research into the detection of network disruption. However, there has been no research that uses fiber optic components as the affected factor to detect the potential of network disruption.

Naïve Bayes used to detect the failure of network equipment [9-11]. Research by T. Y. Fei et al. [9] focused on monitoring the trends of equipment's warning logs. The model was able to calculate the probability of failure earlier before the actual day of failure. The results achieved over 70% accuracy, and the model was better at capturing the pattern of the warnings as it was closer to the actual day the network equipment failed.

In a research conducted by R. W. J. Yang et al. [12], the Naïve Bayes algorithm had been modified based on Artificial Bee Colony Algorithm to detect network intrusion on Intrusion Detection System (IDS). Modified Naïve Bayes is used to calculate the accuracy of classification by adds the weights to the Naïve Bayes Classifier. The weights are adjusted using the Swarm intelligence algorithm. The higher weights are given to the individuals that have a more significant impact on the result, and the lower weight given to the lower individuals that have, a lower impact on the result. Then, this value will be used as the fitness value of the Artificial Bee Colony algorithm. The result showed that the accuracy of the model could reach above 91%.

In other paper [13], prevention and detection of the Distributed Denial of Service (DDoS) provided. Naïve Bayes used to classifying the incoming packet of the data as usual or attacked packets. The Analyzed features viz. MTI (Mean Time Intervals), POIP (Probability of Occurrence of IP), TTL (Time to Live), ACK value, SYN value, timestamp field, differentiated service field, and sequence number. The result achieved an accuracy above 90%, and the computation time was reduced by 46%.

Intrusion Detection using Naïve Bayes (IDNB) was proposed by [14]. Specifically, this model was designed to detect intrusion packet with large data streams. This model is capable of detecting known and unknown DDoS attacks by experiencing the pattern of legitimate network traffic. The result shows that the proposed model has achieved an accuracy of 92.34%.

Our study proposed and implemented a model using the Naïve Bayes classifier to classify and detect the potential disruption of network services to prevent customer's complaints. Features selection in this model is based on the values that affect network speed, and customer priority whose generate from some equipment. The accuracy of the model is studied using the Confusion Matrix. The paper organized as follows: Section I introduces the aim of this project and some related works. Section II describes the proposed model in this

project. Section III provides the experimental and discussion results before the conclusion in Section IV.

II. Methodology and Material

This prediction model was implemented using existing data from the X Company (Indonesia National Internet provider). This company has a big concern about its customer services, specifically to prevent disruption or delay of an internet network.

Certain conditions on some devices may cause disruption or delays in the Internet network. These devices include the Optical Distribution Point (ODP) and Optical Network Terminal (ONT). ODP is the distribution point of the distribution cable to multiple cables that have drop channels at the customer's home. ODP is a passive device whose installation is in the field. At the customer end, ONT is an active device, and its function is to convert optical signals into electronic signals.

Another factor that also affects disruption is the reduction in the bandwidth of the fiber-optic network, also known as attenuation. In fiber-optic communication, data transmitted through the light on fiber optic. Attenuation is a condition in which the signal is weakening as the distance that the signals should bridge increases, and the frequency of the signals is getting higher. The higher the attenuation, the lower the bandwidth.

Fig. 1 shows the process of how the proposed network disruption prediction works. The database used as inputs is the customer's database and the Optical Distribution Point (ODP) database, which is then processed with the Naive Bayes model prediction to obtain the prediction of network disruptions as output, which consequently sends an early warning to the technician via his mobile phone application. The technician then goes to the identified hotspot to resolve the network disruption problem. Once resolved, they can immediately update the progress status via the mobile application. Thus they can earn points for their mileage. A detailed description of each process provided in the next chapter.

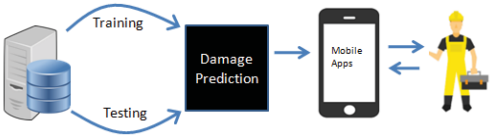

Fig. 1. Proposed network disruption prediction model

A. Naive Bayes Classification

Naive Bayes Algorithm is one of the supervised learning algorithms which can be used for preventing disruption of the telecommunication network [9]. The term considered naive because it assumes that all variables contribute to the classification and are correlated with each other [15]. Naive Bayes Algorithm used to classify the level of every network disruption, as shown (1) [16-17]:

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

where,
H is a class
X is a dataset
P(X) is a probability for dataset X
P(X|H) is a conditional probability dataset X belong to class H
P(H) is a probability for class H
P(X) is a probability for dataset X
To classify the data, Naive Bayes using (2)

$$P(H|X1 \dots Xn) = \frac{P(H)P(X1 \dots Xn|H)}{P(X1 \dots Xn)} \quad (2)$$

Where F1...Fn represents the characteristics of the instruction required to perform the classification. The formula explains that the probability of entering a sample of specific characteristics in class H (Posterior) is the probability of the emergence of class H (before the entry of the sample, often referred to as prior), multiplied by the probability of occurrence of characteristics sample in class H (also referred to as likelihood) with the probability of occurrence of sample characteristics globally (also referred to as evidence).

As a preliminary step, the preparation of the training data is the selected customer data, whose complaint due to the network disruption, which was previously processed by the technical team. Input is considered as the components that have the most affected network speed and customer priority. The input for this model prediction is the revenue number of each customer (REVENUE), recurrent disruption value of ODP (N_Q), attenuation value in ODP (OLT), and attenuation value in customer (ONU). The data classified into three classes, namely GREEN, YELLOW, and RED. GREEN represents a group of customers with no network disruption. YELLOW represents the group of customers with low-level disruption, while RED represents the group of customers having high-level disruption that will be given the highest priority to be followed up by technicians.

The process is run through 3 phases, the first is Data Cleaning, which is taken from multiple database inputs, followed by Data Training to obtain model predictions, and the third is Data Testing to evaluate model predictions.

Data Cleaning's stage is a process of combining, matching, and filtering features that are used as input for the next stage. The data Training stage used the Naive Bayes algorithm and was processed using the Weka Tool. In this stage, 50 datasets from 7 different cities, namely: South Jakarta, North Jakarta, Bekasi, Bandung, Cirebon, Surabaya, and Makassar, are used. Table I represents the detail numbers of data training from each city. Data testing will be explained in the next section.

TABLE I. Data For Training

Location	Training Data		
	Red	*Yellow*	*Green*
South Jakarta (SJ)	5	1	2
North Jakarta (NJ)	4	2	2
Bekasi (BKS)	2	2	3
Bandung (BDG)	1	3	1
Cirebon (CRB)	2	4	1
Surabaya (SBY)	3	1	6
Makasar (MKS)	2	1	2

Table II shows examples of several training data from customer data who have complaints about network disruption. The data represent the combination of four

criteria. The first row represents the condition for the user account 122218202835 on the area South Jakarta, and the device name of ODP is ODP-KMGFAC/11FAC/ D02/11.01. The revenue of the customer is 1474178, the number of recurring disruptions is 0.33333, the value of the attenuation in ODP is -12.84, and the value of the attenuation in ONT is -12.9.

TABLE II. TRAINING DATA

USER ACCOUNT	AREA	DEVICE NAME	REVE-NUE	N_Q	OLT	ONU	CONDITION
122218202835	SJ	ODP-KMGFAC/11FAC/D02/11.01	1474178	0.33333	-12.84	-12.9	GREEN
122218205346	SJ	ODP-KMGFAC/18FAC/D03/18.01	792140	0	-18.96	-19.95	GREEN
122502256099	NJ	ODP-TPRFAZ/18FAZ/D01/18.01	1049205	0	-24.71	-26.02	YELLOW
122844270150	BKS	ODP-PKYFG/28FG/D02/28.01	785558	0	-26.857	-25.688	YELLOW
122853303014	BKS	ODP-JBKFCH/34FCH/D03/34.01	2712968	1	-29.67	-31.65	RED
131161121867	BDG	ODP-GGKFBM/26FBM/D03/01.26	4026361	0	-26.842	-28.37	RED
152404236133	SBY	ODP-KBLFCA/12FCA/D02/12.01	771100	0	-21.8	-18.096	GREEN
172101810574	MKS	ODP-BALFC/24FC/D04/24.01	4258822	1	-26.4	-30.468	RED
172106206532	MKS	ODP-TMAFF/77FF/D07/77.01	1469503	0	-15.9	-14.596	GREEN
131236114794	CRB	ODP-PABFAR/03	3075193	0	-27.569	-27.96	RED

For example, we have one-row data testing, as is showed in Table III. This data is being tested using the proposed model, the means and the deviation values of the training data are needed, as shown in Table IV.

TABLE III. DATA TESTING

USER ACCOUNT	AREA	DEVICE NAME	REVENUE	N_Q	OLT	ONU	CONDITION
121114208944	EAST JAKARTA	ODP-GANFFF/16	2220629	0.66667	-23.67	-28.83	RED

TABLE IV. MEAN AND DEVIATION VALUES OF TRAINING DATA

	REVENUE			N_Q		
Condition	GREEN	YELLOW	RED	GREEN	YELLOW	RED
Mean	1773475	1348495	3473396	0.0583	0.0143	0.6714
Deviation	1700140	981574	1109727	0.1595	0.035	0.3452
	OLT			ONU		
Condition	GREEN	YELLOW	RED	GREEN	YELLOW	RED
Mean	-5.9664	-21.7223	-27.3522	-18.2634	-26.5874	-29.284
Deviation	3.8389	5.3431	2.1693	3.1912	2.333	1.3695

Equation (3) is implemented on each criterion within each class using the values in Table III.

$$P(X_i = x_i \,|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu ij)^2}{2\sigma^2 ij}} \quad (3)$$

where,
X_i is the criteria
Y is the condition (GREEN, RED, YELLOW)
σ is the deviation value
x_i is the value of each criterion from database
μ is mean value of each of the criteria

The result obtained is shown below:

$$P(revenue|Green) = \frac{1}{\sqrt{2\pi\, x\, 1700140}} e^{\frac{(2220629-1773475)^2}{2\,x\,(1\,700140)}}$$

$$= 0.000000226732831526203$$

Other values for each criteria on the GREEN condition are:
N_Q = 0.00173429132
OLT = 0.013879604874084
ONU = 0.000520354901
GREEN = 0.016134478
The value for GREEN condition is obtained from the total of all criteria.
All values for YELLOW condition are:
REVENUE = 0.000000273950368343826
N_Q = 4.13208E-75
OLT = 0.013879604874084
ONU = 0.0698831527219061
YELLOW = 0.17764412
All values for RED condition are:
REVENUE = 0.0000001901384327 6255
N_Q = 1.1558690142619
OLT = 0.043556491859750
ONU = 0.275799968695395
RED = 1.475225665

From the values above, it found that the highest values are in the RED condition. The RED condition matched to the condition labeled before for the row of data testing in Table III.

III. EXPERIMENT AND RESULT

It is essential to evaluate the performance of a model after creating a proposed model. A model should have a few mistakes as possible. The training data is classified with the Weka Tool by10-fold cross-validation to evaluate the trained classifier model[18]–[20]. The result of the mean and deviation of the data for each criterion was calculated using Weka Tool, as shown in Table VI. The summarizing classification result shows that the mode achieved its accuracy of 91.89% for 74 instances. The 68 instances are correctly classified, and six instances are incorrectly classified.

TABLE V. MEAN AND DEVIATION VALUES USING WEKA TOOL

	RED	YELLOW	GREEN
REVENUE	(3473396)	(1348495)	(1773475)

N_Q	(0.6714)	(0.0143)	(0.0583)
OLT	(-27.35)	(-21.72)	(-15.96)
ONU	(-29.28)	(-26.58)	(-18.26)

Confusion matrix as a routine evaluation was used to measure the proposed model, as it was adopted in this research to study the model performance. Table VI defines the confusion matrix [21].

TABLE VI. CONFUSION MATRIX

Confusion Matrix		**Actual**	
		Positive	***Negative***
Predicted	***Positive***	True Positive	False Positive
	Negative	False Negative	True Negative

True - Positive is where the model could recognize all valid data correctly as a true class. False Positive is where false data recognize as a true class by the model. A false negative is where the model could not recognize all valid data in true class. True Negative is where the model could recognize as a false class from false cases.

Precision is the proportion of positive cases that recognized as positive overall cases classified as positive, and it expressed in (4)[21].

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

A recall is the proportion of relevant classes that are successfully recognized. It is shown in (5)[21].

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

Accuracy is the proportion of correctly classified cases overall cases, and it expressed in (6) [21].

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

Error is the proportion of incorrectly classified cases overall cases; it can calculate in (7) [21].

$$Error = \frac{(FP+FN)}{(TP+TN+FP+FN)} \quad (7)$$

Data testing for model prediction is done with 100 datasets from 7 locations. They are from customer and ODP data from X Company database. Table VII shows the result of the experiment.

TABLE VII. EXPERIMENT RESULTS

	Testing Data			
	Red	***Yellow***	***Green***	***Average***
Precision	0.88	0.8	0.87	0.85
Recall	0.86	0.72	0.92	0.83
Accuracy	0.84	0.75	0.86	0.82
Error	0.15	0.25	0.14	0.18

According to Table VII, the highest score for precision is in the RED class, which is 0.88. It can happen as the number of data for the RED class was more significant than the other classes in the training process. On the other hand, the highest score for accuracy is in the GREEN class, which is 0.86 as the number of valid data testing in the GREEN class is the biggest one. The maximum error is in the YELLOW class, which has a score of 0.25 as the number of data from YELLOW class in the data training is lower than the other classes.

IV. CONCLUSION

The proposed model for prediction and detection of the disruption network has been implemented using Naïve Bayes Classifier. According to the result, the proposed model achieved good accuracy for the entire training data, indicated by 91.89% accuracy using WEKA Tool. Whereas the average accuracy value for all classes of the test data is 82%, which is lower than the training data value. The accuracy value still depends on the number of data. The upcoming study of this research is to use the model for handling customer complaints in X Company to verify whether the implementation of the model can minimize customer complaints about network disruption.

REFERENCES

[1] G. Vennila, N. S. Shalini, and M. S. K. Manikan, "Naive Bayes intrusion classification system for VoIP network using honeypot," *Int. J. Eng. Trans. A-Basics*, 2015.

[2] Z. B. Abu, F. E. B., Shahbudin, M. B. Mansor, N. Z. B. A. Rahim, and N. A. B. Norwahi, "Improving user complaint management system and satisfaction level via reader-friendly linguistic features," in *2015 International Symposium on Mathematical Sciences and Computing Research, iSMSC 2015 - Proceedings*, 2016.

[3] Y. Tang, H. P. Hu, X. C. Lu, and J. Wang, "HonIDS: Enhancing honeypot system with intrusion detection models," in *Proceedings - Fourth IEEE International Workshop on Information Assurance, IWIA 2006*, 2006.

[4] A. J. C. Trappey, C. H. Lee, W. P. Chen, and C. V. Trappey, "A framework of customer complaint handling system," *2010 7th Int. Conf. Serv. Syst. Serv. Manag. Proc. ICSSSM' 10*, pp. 879–884, 2010.

[5] M. Davidow, "Organizational Responses to Customer Complaints: What Works and What Does Not," *J. Serv. Res.*, 2003.

[6] J. Cambra-Fierro, I. Melero, and F. J. Sese, "Managing complaints to improve customer profitability," *J. Retail.*, 2015.

[7] P. Kormpho, P. Liawsomboon, N. Phongoen, and S. Pongpaichet, "Smart complaint management system," *Proceeding 2018 7th ICT Int. Student Proj. Conf. ICT-ISPC 2018*, pp. 1–6, 2018.

[8] R. Chalmeta, "Methodology for customer relationship management," *J. Syst. Softw.*, 2006.

[9] T. Y. Fei, L. J. Yan, L. H. Shuan, G. Xiaoning, and S. W. King, "Detection on network equipment failure using Naïve Bayes classification," *Proc. 2017 IEEE 2nd Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2017*, vol. 2018-Janua, pp. 286–290, 2018.

[10] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *Int. J. Sci. Res.*, vol. 5, no. 1, pp. 1842–1845, Jan. 2016.

[11] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *Performance Evaluation Review*, 2005.

[12] R. W. Juan Yang, Zhiwei Ye, Lingyu Yan, Wei Gu, "Modified Naive Bayes Algorithm for Network Intrusion Detection based on Artificial Bee Colony Algorithm," 2018, pp. 35–40.

[13] N. A. Singh, K. J. Singh, and T. De, "Distributed denial of service attack detection using Naive Bayes Classifier through Info Gain Feature Selection," in *Proceedings of the International Conference on Informatics and Analytics - ICIA-16*, 2016, vol. 25-26-Augu, pp. 1–9.

[14] V. Hema and C. E. Shyni, "DoS Attack Detection Based on Naive Bayes Classifier," *Middle-East J. Sci. Res. Signal Process. Secur.*, vol. 23, pp. 398–405, 2015.

[15] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Mach. Learn.*, 1997.

[16] Venkatesh and K. V. Ranjitha, "Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier," *2018 IEEE World Symp. Commun. Eng. WSCE 2018*, pp. 33–36, 2019.

[17] D. Buzic and J. Dobsa, "Lyrics classification using Naive Bayes," *2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc.*, pp. 1011–1015, 2018.

[18] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "Automatic Speaker Recognition

System based on Machine Learning Algorithms," *Proc. - 2019 South. African Univ. Power Eng. Conf. Mechatronics/Pattern Recognit. Assoc. South Africa, SAUPEC/RobMech/PRASA 2019*p. 141–146, 2019.

[19] S. Choudhury and A. Bhowal, "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection," *2015 Int. Conf. Smart Technol. Manag. Comput. Commun. Control. Energy Mater. ICSTM 2015 - Proc.*, no. May, pp. 89–95, 2015.

[20] S. Sharma, R. Purohit, and P. S. Rathore, "Algorithm," no. Icces, pp. 386–390, 2017.

[21] T. Garg and S. S. Khurana, "Comparison of classification techniques for intrusion detection dataset using WEKA," *Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2014*, 2014.