Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot

Muhammad Yusril Helmi Setyawan
Applied Bachelor Program of Informatics Engineering Politeknik Pos Indonesia
Bandung, Indonesia
yusrilhelmi@poltekpos.ac.id

Rolly Maulana Awangga
Applied Bachelor Program of Informatics Engineering Politeknik Pos Indonesia
Bandung, Indonesia
rolly@awang.ga

Safif Rafi Efendi
Applied Bachelor Program of Informatics Engineering Politeknik Pos Indonesia
Bandung, Indonesia
esafif637@gmail.com

Abstract— **Chatbot is software that communicates using natural language. chatbots such as machine conversation systems, Chatterbot, virtual agents, and dialogue systems. This software enables to simulate human conversations. In this research, the chatbot system that will be created must be able to understand the natural language of what is entered by the user, and the chatbot will answer according to what the user is expecting. The researcher proposes a classification method to identify intent rather than user input or called intent classification on the chatbot system; the researcher also wants to know the level of accuracy, precision, and recall on the evaluation results of both methods. The classification method applied in this research is the Naive Bayes method and compared with the Logistic Regression method to determine the class intention. The evaluation results show the level of accuracy precision and recall in the Logistic Regression model is higher than the Naive Bayes model.**

Keywords : Chatbot,Intent Classification, Naive Bayes, Logistic Regression.

I. Introduction

Internet chat is a popular application that allows text-based communication. Some people around the world use internet chat to exchange messages and discuss various topics online [1]. At present, it is essential for any company to have the infrastructure and services to listen to a social media platform, whether it's Twitter, Facebook, Messenger, e-mail for company applications, because most customers used social media to find information about companies because they need help with products or services. Companies must build a platform where customer service representatives reach social media users, to ensure that users get the information they need [2]. To reach out to the community in the social stream researchers propose to develop reporting applications on the chatbot platform for community reporting. This software is planned to responds common questions in a particular domain [3].

Chatbot system is a program which interacts with users. These interactions between users and the system are using natural languages, for example: machine conversation systems, virtual agents, dialogue systems, and Chatterbot. The purpose of this system is to simulate users conversations. Chatbot architecture is using language models and computational algorithms to mimic informal communication and interaction between users and computers using natural language processing [4]. On [5], the author suggests a conversation based on natural language understanding. Messages sent using chatbot are processed using Natural Language Processing techniques. [6]. NLP provides ways for users to communicate with computers using natural language. To understanding natural language there are three analyzes. These analyses are: decomposition, semantic interpretation, and knowledge-based structures [7]. Developments in knowledge management for the development of an intent management system on chatbot based on ontology [8][9]. and the evaluation process will produce performance values for each classification method applied [10].

A chatbot is a trending application created by Artificial Intelligence. It is used in humanoid robots, personal assistants, car assistance, etc., to facilitate human work [11]. The system can understand natural language that is input by the user, in this study the authors apply the classification method to be able to understand the text of the user, the researcher proposes to use classification to determine intent classification so that the system can provide answers according to intent classification.

II. Related Work

There are several text classifications; one of them is Naive Bayes. This classification is popular among the researchers for text categorizations. Naive Bayes classification is simple and efficient. Naive Bayes is a model-based classification method and offers to compete for classification performance for text categorization compared [12]. The Naive Bayes classification algorithm can be used widely in many cases because it has high efficiency and easy implementation. [13] This Bayesian classification is used as a probability learning method, and each feature of the algorithm that is classified is not dependent on the value of other features [14]. Naive Bayesian Chatbot is a simple classifier to find intent classification based on the application of Bayesian theorem with independent assumptions.[15] . Classification algorithms require data that has been trained and arranged into several classes. Naive Bayes requires a short time to build the model [16].

The Naive Bayes algorithm can be either adaptive and intelligent, and can be a function, but also fulfills personalized requirements., and therefore are wide or extensively used in commercial[17]. Naive Bayes is a simple technique for building classifiers: models that classify the problem instance, are described as feature value vectors, where class labels are taken from a limited number of circuits [18]. This machine learning can provide accurate results that are very efficient in use. Naive Bayes classification algorithm is a simple and straightforward

978-1-5386-8066-7/18/$31.00 ©2018 IEEE

structure algorithm, which has a class node [19]. Naive Bayesian is efficient in the term for time, CPU usage (Central Processing Unit) and memory. Naive Bayes can perform well even with small training sets and less extensive computing. [20]. Naive Bayes Classifier belongs to a family of simple probability analyzers based on the Bayes theorem and has an independent assumption that the value of specific features is always different from the value of other features [21]. This method is one of the supervised learning based on Bayesian rules on the statistical theory, which runs in the example of labeling training, and is given by a strong assumption that all attributes in the training data are interdependent, which is called the Bayes assumption [22].

Naive Bayes one of the popular machines, an interesting framework in various tasks and reasonable performance is obtained in the task even though this learning is based on nonrealistic assumptions of independence. [23]. The Bayesian method is the most practical learning method for multiparameter by calculating the probability. This is very competitive compared to other methods for learning technique [24]. Bayesian is a statistical classification. Naive Bayes Classification is based on the Bayes Theorem which utilizes conditional probabilities to classify data into predetermined classes. This method is called "naive" because of independence assumption between various attribute values [25]. Naive Bayes method has been applied in various fields, especially in natural language processing and bioinformatics, including the classification of genre texts and author attributions, sentiment analysis, disease prediction from genomic data. Naive Bayes method is called "naive" because it assumes that all features are independent, depending on the class label [26].

Logistic regression can be considered as a general linear regression model. This type of logistic regression allows us to test the effect of numerical factors on binary responses [27]. Regression focuses on the relationship between the dependent variable called (Y) and one or more independent variables (x0, x1, ..., xn). In linear regression, 'Y' is an advanced value while logistic regression has a discrete value. The logistical function, also known as the sigmoid function, is used to calculate the logistic model in which each value from negative infinity to positive infinity is provided as limited input and output in the range of 0 and 1 [28]. Logistic regression can be used in machine learning applications. This algorithm can understand vector variables and evaluate the coefficients or weights for each input variable and then prediction the class expressed the value of the word vector [29]. Logistic regression is a technique used to study data sets where there are one or more independent variables that know the results [30].

III. RESEARCH METHOD

In this study, the researcher compare two models with different methods with the steps in Figure 1 to determine the intent classification of the chatbot system to be built

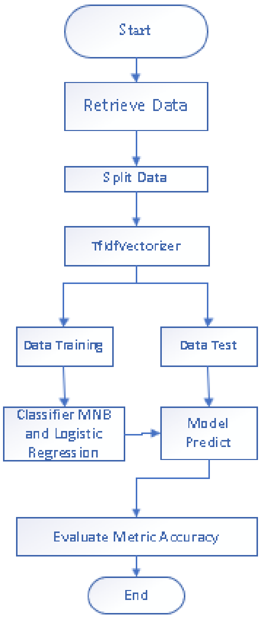


Figure 1. Proposed Evaluate Model

Retrieve the data in the study was reading the training data that had been labeled in the class intent that had been done by the researcher, and the researcher made the class as follows greet, report, info, point, trade-point, and thanks. In the training data, some data is taken to be used as a data test to test the level of accuracy and performance of each model at the evaluation stage. After retrieving the data the researcher processes the TF-IDF, TF is the feature items that appear in the document, IDF is the anti-document frequency, TF-IDF is calculated using the following formula [28].

$$w_{ij} = tf_{ij} \times idf = tf_{i} \times \log(\frac{N}{ni} + 0.001) \ldots.(1)$$

Formula 1. Equation TF-IDF

After being transformed using TF-IDF then doing the data sharing stage into training data and test data. in the training data section, the model was made by applying the Naive Bayes method and Logistic Regression to detection of classification. In making the model applied the Naive Bayes method with the following formula [18]:

$$P(ci|e) = \frac{P(e|ci)P(ci)}{\sum_{j=1}^{k} P(e|cj)P(cj)} \ldots,(i=1,2\ldots n)\ldots(2)$$

Formula 2. Equation Naïve Bayes Method

1. Find the previous probability class Ci.
2. According to the prior probability P (Ci), the posterior probability P (Ci | e) is obtained by the Bayesian formula.
3. According to the posterior probability, the test text is classified as a category with the highest probability value.

Likewise with the making of the model applied to the Logistic Regression method, logistic regression is widely used in machine learning model applications. This algorithm understands the vector variables and evaluates the coefficients or weights for each input variable and then prediction to expressed as a word vector [29]. Look for mathematical logistic regression functions estimating multiple linear functions defined as [26]:

$$\text{Logit}(s) = b_0 + b_1M_1 + b_2M_2 + b_3M_3 \ldots\ldots b_kM_k \ldots (3)$$

Formula 3. Equation Logistic Regression Method

S is the probability of the presence of an attractive feature. M1, M2 ... Mk is the Predictor value and b0, b1...bk is the intercept of the model. The assumptions for Logistic regression are as follows [29]:

1. A linear relationship between dependent and independent variables does not exist in Logistic Regression.
2. The dependent variable must be a dichotomy.
3. Independent variables must be linearly related; it is not usually distribution or the same variance in a group.
4. Grouping must be mutually exclusive.

After getting the classifier model created using training data, the next step is to evaluate the data using test data taken by 20% of the training data [29]. at the time of retrieving the data, by predicting test data on training data in both methods, it produces an evaluation in the form of accuracy, precision, and recall value by using a confusion matrix. To calculate the confusion matrix using the following equation[29] .

Table 1. *Confusion Matrix.*

PREDICT VALUE	TRUE VALUE	
	TRUE	FALSE
TRUE	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
FALSE	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

To determine the precision, recall, and accuracy values as follows:

1. *Precision = TP / TP + FP*
2. *Recall = TP / TP + FN*
3. *Accuracy = TP + TN / TP + TN + FP + FN*

After doing all these steps, the research can be concluded from the two methods that affect the accuracy of predictions in each model.

IV. EXPERIMENT

Experiment in this study is the application of the Naive Bayes method that is compared using the Logistic Regression method applied to the chatbot, the training data in this study is data taken from reports that have been reported by the public, then the authors take some data samples and determine the class intention used as training data. To predict the text entered by the user the author uses the Naive Bayes method which is compared with the Logistic Regression method to determine the performance of the two methods in determining the Intent Classification there are several class intentions determined by the author namely greet, report, info, point, swap_point and thanks The author uses the library in the Python language to implement both methods to predict incoming text from chat in the classification that has been determined by the author as much as 55 data in training data, which are classified in the table as follows:

Table 2. Data Training.

No	Class	Value
1	Report	14
2	Greet	10
3	Info	9
4	Point	8
5	Thanks	7
6	Tukar_point	7
Sum Of Data		55

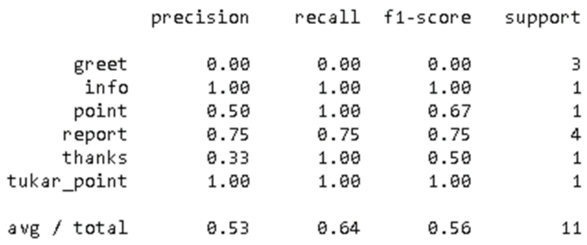
After the data is labeled in the training data, the next step is to do the TF-IDF process to calculate and weight the text that will be predicted before entering the model. At this step the text is evaluated using test data, referring to previous research to evaluate the model, extracted some data training of 10% - 50% for test data [29]. In this experiment was taken 20% of the training data for the test data, because of the limited amount of data. Calculating TF-IDF is explained in the proposed model and produces the following data in an experiment using the Python programming language[32]. to calculate data.

Table 3. Result Tf-Idf

Word	Calculate Tf-Idf
Kerusakan	0.527109669519
Nasution	0.527109669519
Jalan	0.438289146485
Di	0.386332471216
Ada	0.320874801684

After obtaining the training data and test data, the next step makes a model using the Naive Bayes model and Logistic Regression where to calculate the method using the equations (2) and (3), the model classifies training data, in this experiment the author uses libraries in python language to implement both methods that.

The next step evaluates the predictions of the test data on training data that has been obtained by the model of the two methods by using the Confusion Matrix. For the equation of the Coffusion Matrix, it has been explained in the proposed model. In this experiment, the author predicts the test data taken from training data on the model created and calculated it with the Python programming language[32]. From the evaluation results with the Coffusion Matrix, the accuracy results on Naive Bayes are 0.6363636363636364, and Logistic Regression is 0.7272727272727273 which results in the evaluation data as follows:

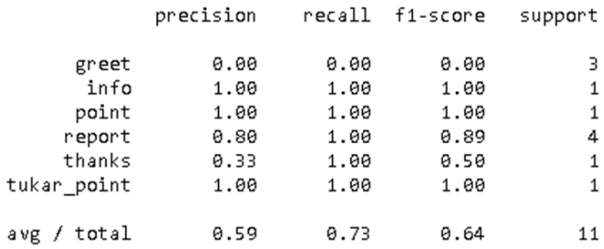


```
              precision    recall  f1-score   support

       greet       0.00      0.00      0.00         3
        info       1.00      1.00      1.00         1
       point       0.50      1.00      0.67         1
      report       0.75      0.75      0.75         4
      thanks       0.33      1.00      0.50         1
  tukar_point      1.00      1.00      1.00         1

avg / total        0.53      0.64      0.56        11
```

Figure 3. Evaluate Naive Bayes Model



```
              precision    recall  f1-score   support

       greet       0.00      0.00      0.00         3
        info       1.00      1.00      1.00         1
       point       1.00      1.00      1.00         1
      report       0.80      1.00      0.89         4
      thanks       0.33      1.00      0.50         1
  tukar_point      1.00      1.00      1.00         1

avg / total        0.59      0.73      0.64        11
```

Figure 4. Evaluate *Logistic Regression* Model

V. RESULT AND ANALYSIS

From the evaluation results of the experimental classification model to determine the intent classification on chatbot, the accuracy of the Naive Bayes model is 0.6363636363636364, and the Logistic Regression model is 0.7272727272727273. This shows that there is an accuracy distance between the two models of 0.0909090909090909 or a classification model with logistic regression which is more accurate at 12.5%. In the evaluation results of the experiment also obtained the value of precision and recall. Logistic Regression model produces data on the average total precision of 0.59 and recall of 0.73 while the Naive Bayes method produces data on the average precision of 0.53 and recall of 0.64.

VI. DISCUSSION

In this study, researchers still use training data with a reasonably limited amount, the amount of data is not the same in each class of intent, so it is possible that errors will occur when predicting class intent with less training data. The accuracy of the two methods will experience differences in accuracy distance when the training data for each intent class has the same amount of data in each class. The researcher found a decrease in the accuracy of the logistic regression model when the intent class had the same amount of data in each class.

VII. CONCLUSION

From the experiments results to determine the class intention obtained the following conclusions::

1. The Naive Bayes classification method or the logistic system can be used for the chatbot system.
2. The model using the Logistic Regression method shows a higher level of accuracy and higher value of precision compared to the Naive Bayes method. This experiment proves that the performance of the Logistics Regression model has a better performance compared to the Naive Bayes model.

VIII. REFERENCES

[1]. S. Ghose and J. J. Barua, “Toward the implementation of a topic-specific dialogue based natural language chatbot as an undergraduate advisor,” in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, may 2013. [Online]. Available: https://doi.org/10.1109/iciev.2013.6572650.

[2]. N. Thomas, “An e-business chatbot using aiml and lsa,” in *Advances in Computing, Communications, and Informatics (ICACCI), 2016 International Conference on*. IEEE, 2016, pp. 2740–2742.

[3]. M. N. Kumar, P. C. L. Chandar, A. V. Prasad, and K. Sumangali, “Android based educational chatbot for visually impaired people,” in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, Dec 2016. [Online]. Available: https://doi.org/10.1109/iccic.2016.7919664.

[4]. G. M. D’silva, S. Thakare, S. More, and J. Kuriakose, “Real world smart chatbot for customer care using a software as a service (saas) architecture,” in *I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC), 2017 International Conference on*. IEEE, 2017, pp. 658–664.

[5]. C. J. Baby, F. A. Khan, and J. Swathi, “Home automation using IoT and a chatbot using natural language processing,” in *Power and Advanced Computing Technologies (i-PACT), 2017 Innovations in*. IEEE, 2017, pp. 1–6.

[6]. B. Setiaji and F. W. Wibowo, “Chatbot using knowledge in database: Human-to-machine conversation modeling,” in *Intelligent Systems, Modelling, and Simulation (ISMS), 2016 7th International Conference on*. IEEE, 2016, pp. 72–77.

[7]. O. Efraim, V. Maraev, and J. Rodrigues, “Boosting a rule-based chatbot using statistics and user satisfaction ratings,” in *Conference on Artificial Intelligence and Natural Language*. Springer, 2017, pp. 27–41.

[8]. Setyawan, Muhammad Yusril Helmi, Rolly Maulana Awangga, and Rezka Afriyanti. "Analysis and Design of Feature Application Setting Dashboard on Svara Applications Using Ucd Method (User-Centred Design) at PT. Zamrud Khatulistiwa Technology." TELKOMNIKA (Telecommunication Computing Electronics and Control) 17.1 2018.

[9]. Awangga, Rolly Maulana, Muhammad Yusril, and Helmi Setyawan. "Ontology Design of Influential People Identification Using Centrality." Journal of Physics: Conference Series. Vol. 1007. No. 1. IOP Publishing, 2018.

[10]. Awangga, R. M. "Sampeu: Servicing Web Map Tile Service over Web Map Service to Increase Computation Performance." IOP Conference Series: Earth and Environmental Science. Vol. 145. No. 1. IOP Publishing, 2018.

[11]. B. Tang, S. Kay, and H. He, “Toward optimal feature selection in naive Bayes for text category ration,” *IEEE transactions on knowledge and data*

engineering, vol. 28, no. 9, pp. 2508–2521, 2016.

[12]. P. Liu, H. Yu, T. Xu, and C. Lan, "Research on archives text classification based on naive Bayes," in *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*. IEEE, Dec 2017. [Online]. Available: https://doi.org/10.1109/itnec.2017.8284934

[13]. S. R. Gomes, S. G. Saroar, M. M. A. Telot, B. N. Khan, A. Chakrabarty, and M. M. Mostakim, "A comparative approach to email classification using naive Bayes classifier and hidden Markov model."

[14]. Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in *Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on*. IEEE, 2017, pp. 635–638.

[15]. M. D. N. Arusada, N. A. S. Putri, and A. Alamsyah, "Training data optimization strategy for multiclass text classification," in *Information and Communication Technology (ICoIC7), 2017 5th International Conference on*. IEEE, 2017, pp. 1–5.

[16]. L. Li and C. Li, "Research and improvement of a spam filter based on naive Bayes," in *Intelli gent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on*, vol. 2. IEEE, 2015, pp. 361–364.

[17]. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *Electrical and Computer Engineering (UKRCON), 2017 IEEE First Ukraine Conference on*. IEEE, 2017, pp. 900–903.

[18]. M. Shafiq, X. Yu, and A. A. Laghari, "Wechat text messages service flow traffic classifica tion using machine learning technique," in *IT Convergence and Security (ICITCS), 2016 6th International Conference on*. IEEE, 2016, pp. 1–5.

[19]. A. Rahman and U. Qamar, "A bayesian classifiers based combination model for automatic text classification," in *Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on*. IEEE, 2016, pp. 63–67.

[20]. N. Sharma and M. Singh, "Modifying naive bayes classifier for multinomial text classification," in *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*. IEEE, 2016, pp. 1–7.

[21]. X.-R. Yu, Z.-L. Xiang, and D.-K. Kang, "Classification of chinese-to-english translated social network timelines using naive bayes," in *Advanced Communication Technology (ICACT), 2015 17th International Conference on*. IEEE, 2015, pp. 296–299.

[22]. S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE transactions on knowledge and data engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[23]. E. Pramunanto, S. Sumpeno, and R. S. Legowo, "Classification of hand gesture in indonesian sign language system using naive bayes," in *2017 International Seminar on Sensors, Instrumentation, Measurement and Metrology (ISSIMM)*. IEEE, aug 2017. [Online].

[24]. F. Burdi, A. H. Setianingrum, and N. Hakiem, "Application of the naive bayes method to a decision support system to provide discounts (case study: Pt. bina usaha teknik)," in *Information and Communication Technology for The Muslim World (ICT4M), 2016 6th International Conference on*. IEEE, 2016, pp. 281–285.

[25]. L. Sayfullina, E. Eirola, D. Komashinsky, P. Palumbo, Y. Miche, A. Lendasse, and J. Karhunen, "Efficient detection of zero-day android malware using normalized bernoulli naive bayes," in *Trustcom/BigDataSE/ISPA, 2015 IEEE*, vol. 1. IEEE, 2015, pp. 198–205.

[26]. B. Pavlyshenko, "Machine learning, linear and bayesian models for logistic regression in failure detection problems," *arXiv preprint arXiv:1612.05740*, 2016.

[27]. J. Isaac and S. Harikumar, "Logistic regression within dbms," in *Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on*. IEEE, 2016, pp. 661–666.

[28]. A. Prabhat and V. Khullar, "Sentiment classification on big data using na¨ıve bayes and logistic regression," in *Computer Communication and Informatics (ICCCI), 2017 International Conference on*. IEEE, 2017, pp. 1–5.

[29]. C. Prathibhamol, K. Jyothy, and B. Noora, "Multi label classification based on logistic regression (mlc-lr)," in *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*. IEEE, 2016, pp. 2708–2712.

[30]. A. Guo and T. Yang, "Research and improvement of feature words weight based on tfidf algorithm," in *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*. IEEE, 2016, pp. 415–419.

[31]. W. Ramadhan, S. A. Novianty, and S. C. Setianingsih, "Sentiment analysis using multinomial logistic regression," in *Control, Electronics, Renewable Energy and Communications (ICCREC), 2017 International Conference on*. IEEE, 2017, pp. 46–49. Available: https://doi.org/10.1109/issimm.2017.8124288

[32]. Awangga, R. M. "Peuyeum: A Geospatial URL Encrypted Web Framework Using Advance Encryption Standard-Cipher Block Chaining Mode." IOP Conference Series: Earth and Environmental Science. Vol. 145. No. 1. IOP Publishing, 2018.