

Classification of Community Complaints Against Public Services on Twitter

1st Muqorobin, 2nd Siti Rokhmah, 3rd Isnawati Muslihah, 4th Nendy Akbar Rozaq Rais
1,2,3,4^{Inst} Institut Teknologi Bisnis AAS Indonesia Surakarta

1,2,3,4^{Jl} Slamet Riyadi No. 361 Windan, Makamhaji, Kartasura, Sukoharjo, Indonesia

¹robbyaullah@gmail.com, ²elfathiey@gmail.com, ³isnawatimuslihah12345@gmail.com, ⁴ab.terate@gmail.com

Abstract—The Surakarta Al-Islam Vocational School is a private educational institution that requires all students to pay school tuition fees. Education is an obligation for all Indonesian citizens. The cost of education is one of the most important input components in implementing education. Because cost is the main requirement in achieving educational goals. SPP School is a routine school fee that is carried out every month. Based on last year's School Admin report, many students were late in paying school tuition fees, around 60%. This is a very big problem because the income of school funds comes from school tuition. The purpose of this research is that the researcher will build a prediction system using the best classification method, which is to compare the accuracy level of the Naïve Bayes method with the K-K-Nearest Neighbor method. Because both methods can make class classifications right or late, in paying school fees. processing using dapodic data for 2017/2018 as many as 236 data. In improving accuracy, the researcher also applies feature selection with Information Gain, which is useful for selecting optimal parameters. System testing is carried out using the Confusion Matrix method. The final results of this study indicate that the Naïve Bayes Method + Information Gain Method produces the highest accuracy, namely 95% compared to the Naïve Bayes method alone, namely 85% and the K-NN method, namely 81%.

Keywords— Big Data, Twitter, Naïve Bayes, Classification, Complaint.

I. INTRODUCTION

The development of the internet today is very rapid and has a great influence on human life, one of which is the increasingly rapid dissemination of information. At present almost all information that we want is on the internet, both from websites, blogs, social media, and others. One of the most popular media in sharing information is Twitter social media. Twitter is a social network that allows each user to share information [1]. Besides being used as social media, Twitter is also used as a medium to read news. Users who want to get information from an account must first become a follower of the account. The large amount of information available on Twitter can also be used to conduct a study, especially in terms of sentiment analysis by categorizing information based on certain categories.

Text categorization methods that can be used at this time are quite large, including Bayes classification algorithm, K-Nearest Neighbor (KNN), Neural Network (NN), Support Vector Machine (SVM), The Decision Tree, K-Means, etc.

In this study the author will use the Naïve Bayes method. This method was chosen because Naïve Bayes was able to carry out the classification process, which would be able to predict complaints or not on public services based on the results of posts from Twitter of the Indonesian people. The Naïve Bayes method is also included in the Top 10 algorithms in data mining. Based on the paper of (ICDM) International Conference on Data Mining[2].

In a previous study entitled Sentiment Classification of Twitter Belonging Data to Saudi Arabian Telecommunication Companies, using the KNN method to cluster tweets into positive and negative classes based on twitter data, so that 80.1% accuracy is obtained[3].

In a previous study conducted by (Effendy, 2016), the SVM method was used to determine positive and negative opinions on public transportation via twitter, which then the results of the analysis will be used to determine the factors that are the main causes of the inability to use public

transportation and factors that make people choose to use this type of transportation[4]

The main aim of this study to predicting by classifying the public services that include on complaint category or not based on 3 parameters is Transportation, Hospital, and market. This analysis results can be used to predict is it a complaint or not in the future, so it will help the government to decide the general policy by using this information.

II. RESEARCH METHODS

The method used in classifying complaints and not complaints on public services uses the Naïve Bayes method. This method was chosen because it was able to carry out the classification process.

2.1 Method Naïve Bayes

Naïve Bayes algorithm is a classification method with probability and statistical methods. This method was put forward by British scientist Thomas Bayes, namely predicting opportunities in the future based on previous experience so that it is known as the Bayes theorem. The theorem is assumed to have mutually independent attributes[5].

The theoretical basis used in carrying out this classification process is the Bayes theorem shown by the equation (1).

$$P(A|B) = (p(B|A) * p(A)) / p(B) \dots\dots\dots(1)$$

Explanation:

- P(A|B) : Posterior A value when B
- p(B|A) : B Likelihood value when A
- p(A) : Prior Value in class A
- p(B) : Evidence value of a class

At opportunity A as B, it is obtained from opportunity B when A multiplied by opportunity A and divided by

opportunity B. The use of Naïve Bayes on a data with more than one feature / attribute causes equation (1) to be more complex as shown in equation (2).

$$P(A|B_1, \dots, B_n) = \frac{P(A)P(B_1, \dots, B_n|A)}{p(B_1, \dots, B_n)} \dots\dots(2)$$

The value of P (B1 ... Bn) is constant for each experiment so that the maximum value of a class is determined by the maximum value between P (A) P (B.. Bn | A). The function of the equation formed into a maximum multiplication for the prior value and the likelihood function, the function is shown in the equation (3).

$$f_c(F) = \arg \max_{c \in F} P(A) (\prod_{i=1}^n P(B_i|A)) \dots\dots\dots(3)$$

2.2 Classification Of Twitter

The System Experiment conducted aims to create an analysis system in classifying tweets using the Naïve Bayes method. At this stage it is known that the suitability of the output produced by the system and the most suitable parameters to be used as input in the classification process. The analysis in this study is to classify tweets using Indonesian. The following are the stages in making the system as shown in Figure 1.

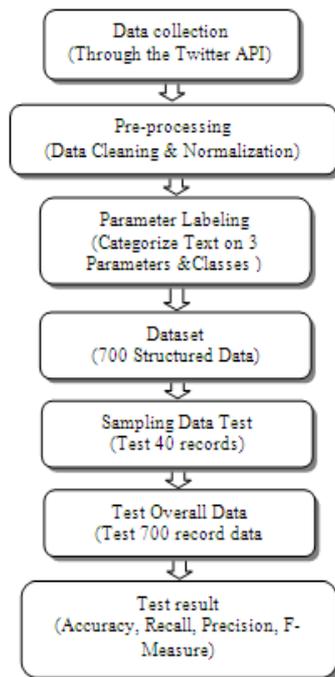


Figure 1. Stages of Data Processing

2.3 Classification Analysis

Analysis in this research the author will process the data sourced from the Twitter post. The purpose of this study is to predict complaints or not on public services based on activities in posts on Twitter. System built by applying the Naïve Bayes method[6]. At the complaint classification stage using training data from the results of previous experiments. Data processing in this study will use tweet data of 700 tweets or records. From 700 data records, the researcher will divide into two parts, 75% for training data, 525 records and 25% for data testing, which is 175 records.

Because in determining the good proportion for training data is 75% and testing data is 25%.

2.4 Performance Evaluation

System testing is useful to determine the level of accuracy of system performance. In Big Data for the classification of public service complaints there are several parameters that are mutually influential in predicting public service complaints. For this reason, in evaluating the performance of classifiers, the researchers used the Confusion Matrix test with Accuracy, Precision, Recall, F-Measure tests on the system that had been built. The system testing formula with the matrix confusion method is shown in the equation (4).

$$\begin{aligned}
 \text{Akurasi} &= \frac{\sum_{i=1}^I \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{I} \cdot 100\% \\
 \text{Presisi} &= \frac{\sum_{i=1}^I TP_i}{\sum_{i=1}^I (FP_i + TP_i)} \cdot 100\% \\
 \text{Recall} &= \frac{\sum_{i=1}^I TP_i}{\sum_{i=1}^I (TP_i + FN_i)} \cdot 100\% \\
 \text{F - measure} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \cdot 100\% \dots\dots\dots(4)
 \end{aligned}$$

The results of the Accuracy, Recall, Precision, F-Measure values produce a percentage value from 0% - 100% and good results are values above 70% and close to 100%[7]

III. RESULT AND DISCUSSION

Big Data in the classification of public service complaints is expected to make future predictions on public service complaints. The system was built by taking the twitter post data of the Indonesian community. From the results of twitter posts that have been post by the community, it generates large data collection so that the implementation of machine learning can help a model in classification of exclusion and not complaints on public services [8].

3.1 Data Analysis

The data used in this study comes from geolocation posts in the Indonesian region. Retrieving data using R studio tools by taking word categories based on parameters, namely transportation, hospital and market[9]. The results of data retrieval are still in the form of raw data that does not yet have a label, so before conducting the machine learning process the researcher conducts the cleaning data and labelling process, so that it becomes a dataset that can be used in predicting public service complaints. Retrieval of initial data as many as 1000 data, then after carrying out the cleaning process data obtained 700 data. The data is then labelled based on the types of parameters such as: transportation, hospitals and markets. For the results of class classification namely Complaints and Not Complaints.

3.2 Training and Testing

The dataset that has been collected is 700 data and has a label based on three parameter attributes and classes. From 700 data then 525 data were taken for Training and 175 Data Testing. Based on the data usage reference that In determining the proportion of training data and testing data is 75% training data and 25% testing data[10]. After the dataset is well formed, it can be used in the process of machine learning using the Naïve Bayes method. Based on the results of data collection, the researcher grouped the data into two parts, namely parameter data related to twitter text, which is shown in Table 1 and parameter data that has nothing to do with twitter text, shown in Table 2.

Table 1. Data Category Including Parameters

Category	Complaint Class	Not Complaint Class
Transportation	170	197
Hospital	124	119
Market	38	184
Total	382	500

Based on the data in table 1, the results of the classification of complaints have been obtained and not the complaints that the researcher made manually by giving a label Yes on each twitter text that has to do with the category of parameters and provides class values (complaints and not complaints). The above data can also be described in Figure 2 below.

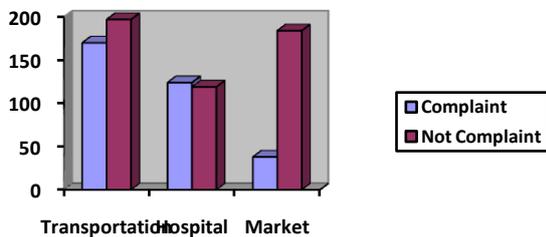


Figure 2. Data Category Including Parameters

Grouping of twitter text data that has nothing to do with parameters is shown in table 2.

Table 2. Data Category Excludes Parameters

Category	Complaint Class	Not Complaint Class
Transportation	128	205
Hospital	174	283
Market	160	218
Total	462	706

Based on the data in table 2, the results of the categorization of data from text that has no connection to the three parameters have been obtained. The above data can also be described in Figure 3 beside.

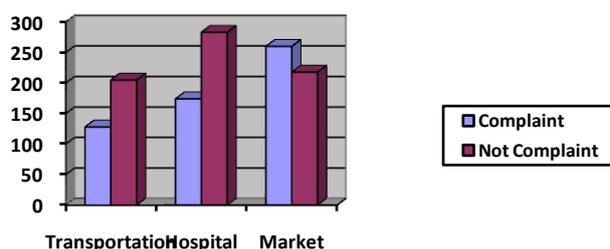


Figure 3. Data Category Excludes Parameters

Based on the results of comparisons in Figures 2 and 3, it can be concluded that the posts that discuss the extensibility of public services are quite small. Because the number of class grades that are not complaints is more than the amount of the complaint. This shows that the Indonesian people are quite satisfied with the public services provided by the government. To measure the system, an accuracy test will be conducted using the matrix confusion method.

3.3 Testing Result

In the system testing phase, the researcher conducted two test scenarios, namely testing data sampling as much as 40 data and testing the overall data of 700 data. In testing sampling data with 40 data, namely 30 training data and 10 data testing, we obtained the results of confusion matrix testing with True Positive (TP): 3, True Negative (TN): 5, Positive False (FP): 1 and False Negative (FN): 1. So that the test value is shown according to table 3 below.

Table 3. Test Results for 40 Data Sampling

Parameter	Value
Accuracy	80%
Precision	75%
Recall	75%
F-Measure	75%

Based on the results of testing 40 data sampling data obtained 80% accuracy this shows the system has been running well. For the test scenario both researchers used 700 data. With a distribution of 525 training data and 175 testing data. So that the test results obtained with confusion matrix with True Positive (TP): 38, True Negative (TN): 113, Positive False (FP): 15 and False Negative (FN): 9. So that the test results obtained are shown in table 4 below.

Table 4. Test Results for 700 Data Sampling

Parameter	Value
Accuracy	86%
Precision	72%
Recall	81%
F-Measure	76%

Based on the results of the two test scenarios that have been carried out, it can be concluded that by applying 700 data obtained an accuracy of 86%, better compared to 40 data which obtained an accuracy of 80%. However, the value of precision is 72% lower in the results of the second scenario if compared with the results of the first scenario, 75%. Based on the results of a comparison of testing of sampling data with the overall data, it can be described in Figure 4 below.

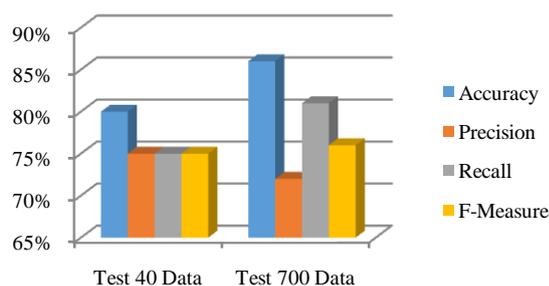


Figure 4. Comparison of Test Results

Based on the results of the two system tests the results are above 50% on each test parameter both from the value of Accuracy, Precision, Recall and F-Measure. While the best results are carried out using 700 data obtained the value of Accuracy: 86%, Precision: 72%, Recall: 81% and F-Measure: 76%. From all parameter test values the results are above 70%. So with this shows the system performance has run very well in conducting data classification.

IV. CONCLUSION

In this study has produced a data mining model that is able to process the classification of complaints or not complaints based on three parameters, namely transportation, hospitals and markets. System testing is done with 2 test scenarios, namely the first test scenario with 40 data, then produces an accuracy value: 80%, precision: 75%, recall: 75% and f-measure: 75%. While the second test scenario with 700 data, it produces an accuracy value: 86%, precision: 72%, recall: 81% and f-measure: 76%. Comparison of the two test scenarios that have been done, the best test results are obtained using 700 data. Because 86% accuracy is obtained.

REFERENCES

- [1] B. Kurniawan, M. A. Fauzi, and A. W. Widodo, "Twitter News Classification Used Metode Improved Naïve Bayes," vol. 1, no. 10, pp. 1193–1200, 2017.
- [2] K. Yu, "Toward an Incremental Democracy and Governance : Chinese Theories and Assessment Criteria," *New Polit. Sci.*, vol. 24, no. 2, pp. 181–199, 2002.
- [3] V. Effendy, A. Novantirani, and M. K. Sabariah, "Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method," *Intl. J. ICT*, vol. 2, no. 1, pp. 57–66, 2016.
- [4] A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 1, pp. 769–773, 2018.
- [5] A. M. Qamar, S. A. Alsuhibany, and S. S. Ahmed, "Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 1, pp. 395–401, 2017.
- [6] X. Li, Z. Ma, and H. Chen, "QODM: A query-oriented data modeling approach for NoSQL databases," in *Proceedings - 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications*, WARTIA 2014, 2014, pp. 338–345.
- [7] S. K. Kim, M. J. Park, and J. J. Rho, "Effect of the Government ' s Use of Social Media on the Reliability of the Government : Focus on Twitter," no. April 2015, pp. 37–41, 2015.
- [8] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," pp. 151–160, 2011.
- [9] K. Lee, D. Palsetia, R. Narayanan, M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter Trending Topic Classification," pp. 251–258, 2011.
- [10] R. Amalia, M. A. Bijaksana, and D. Darmantoro, "A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter," in *International Conference on Computing and Applied Informatics*, 2017.
- [11] Abdullah, Robi W., et al. "Keamanan Basis Data pada Perancangan Sistem Kepakaran Prestasi Sman Dikota Surakarta." *Creative Communication and Innovative Technology Journal*, vol. 12, no. 1, 2019, pp. 13-21.
- [12] Muqorobin, M., Apriliyani, A., & Kusriani, K. (2019). Sistem Pendukung Keputusan Penerimaan Beasiswa dengan Metode SAW. *Respati*, 14(1).
- [13] Muqorobin, M., Hisyam, Z., Mashuri, M., Hanafi, H., & Setiyantara, Y. (2019). Implementasi Network Intrusion Detection System (NIDS) Dalam Sistem Keamanan Open Cloud Computing. *Majalah Ilmiah Bahari Jogja*, 17(2), 1-9.
- [14] K. Kusriani, E. T. Luthfi, M. Muqorobin and R. W. Abdullah, "Comparison of Naive Bayes and K-NN Method on Tuition Fee Payment Overdue Prediction," 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2019, pp. 125-130, doi: 10.1109/ICITISEE48480.2019.9003782.