Sentiment Analysis for Zoning System Admission Policy Using Support Vector Machine and Naïve Bayes Methods

Reynaldy Aries Ariyanto¹, N Chamidah¹

¹Department of Mathematic, Universitas Airlangga, Indonesia

Corresponding email: nur-c@fst.unair.ac.id

Abstract. Indonesia have low quality of education. According Trend International Mathematics and Science Study, Indonesian student’s mathematical literacy is ranked 36 from 49 countries. For Science literacy, Indonesia is ranked 35 from 49 countries. To increase quality of education in Indonesia, Indonesia’s government make a new policy for new student admission called zoning system. Zoning system is a new student admission according distance from house to school. Zoning system is new policy in Indonesia and there are still many pros and cons of the zoning system. Sentiment analysis is used to know whether Indonesia’s people agree or disagree about zoning system policy. In this study, we use supervised statistical learning methods that are Support Vector Machine (SVM) and Naïve Bayes for sentiment analysis of zoning system admission policy. The results show that Indonesia’s people tend to disagree with the zoning system admission policy because negative opinion is greater than positive opinion. Furthermore, accuracy rates of SVM and Naïve Bayes are 92.93% and 79.86% respectively. So, SVM is better than Naïve Bayes for sentiment analysis of zoning system admission policy.

1. Introduction

Human resources with good quality is important element to improve the quality of the country. Education is needed to improve quality if human resources. Education is systematic process to improve holistic human dignity [1]. The purpose of education is to preparing children for life. Therefore, education is a tool to develop the potential of each individual [2]. Nowadays, Indonesia have bad quality of education. It can be seen from study result of Trends International Mathematics and Science Study which shows student in Indonesia do not have achievement. Based on that study, mathematic literacy Indonesian student is ranked 36 from 49 countries with score 405 and still below the international average score. Furthermore, Science literacy of Indonesian student in rank 35 from 49 countries with score 433 and still below the international average score. The results obtained are worse than Egyptian student which in rank 35 [3].

To improve quality of education in Indonesia, Indonesia’s government make a new policy called zoning system admission policy. Zoning system admission policy is student admission policy according distance between school and house [4]. The purpose of zoning system admission is to create equitable quality of education where student who have many achievements not gathered in favorite school [5]. Furthermore, zoning system admission can eliminate favorite school terms [6]. Zoning system admission policy is new policy in Indonesia and many people talk about this policy. There are Indonesian people who agree with zoning system and there are Indonesian people who disagree with zoning system admission policy. To know whether Indonesian people agree or disagree with zoning
system admission policy, we can use sentiment analysis. Sentiment analysis is field of study which analyze opinion, sentiment, assessment, evaluation, attitude, and emotion of someone related to a topic, public service, individual, organization, or some activity [7]. We need text mining for sentiment analysis that used text data from twitter. Text mining is the process of extracting information from unstructured data text [8]. Based on text mining, information about zoning system admission policy from twitter can be obtained.

For classifying data, we can use statistical learning that is a fundamental ingredient in the training of a modern data scientist. The statistical learning can be classified as supervised or unsupervised learnings. The supervised statistical learning builds a statistical model for predicting, or estimating, an output based on one or more inputs. The unsupervised statistical learning where there are inputs but no supervising output. Related to the supervised statistical learning for classifying medical data, [9] improved classification accuracy of cyst and tumor using local polynomial estimator, [10] identified Glaucoma on fundus retinal images using statistical modelling approach, [11] classified Choroidal Neovascularization on fundus retinal images by using Local Linear Estimator and [12] classified Lung Tumor on Human Chest X-Ray using statistical modelling approach. Based on [9-12] results, their accuracy rates by using statistical modelling approaches is higher than mathematical computation approaches. The other methods in supervised statistical learning are Support Vector Machine (SVM) and Naïve Bayes. Previous research [13] shown that by using SVM and Naïve Bayes method for classifying positive or negative sentiment analysis of political discourse in media social has accuracy rate 90.50% and 59.98%, respectively. It means that SVM method is better than Naïve Bayes method for classifying positive or negative sentiment of political discourse in media social.

Based on description above, in this research, we use SVM and Naïve Bayes methods to analyze opinion Indonesian people about zoning system admission from twitter. With sentiment analysis in zoning system admission policy can be known whether Indonesian people agree or disagree with zoning system admission policy. Furthermore, we compare between SVM and Naïve Bayes methods based on accuracy rate of classifying positive or negative sentiment for zoning system admission policy.

2. Literature Review

2.1. K fold Cross Validation
K fold cross validation is method where data is randomly divided into k parts then do training on classifier using some part and do testing using another parts. The usual number of k is 10 and it called 10-fold cross validation. In 10-fold cross validation, data divided into 10 parts \((S_1, S_2, ..., S_{10})\). In first iteration, training is done using \(S_1\) until \(S_9\) classifier result from training then tested using \(S_{10}\). In second iteration, training is done using \(S_1\) until \(S_9\) and \(S_{10}\), classifier result from training then tested using \(S_{10}\). Illustration of this method can be explained in figure 1:

![Figure 1. Illustration of 10-fold cross validation](image-url)
Final accuracy of classifier is average from $k$ process [14].

2.2. Support Vector Machine

Support Vector Machine (SVM) is non probabilistic binary linear classifier. SVM is used to analyze vector data based on word to determine hyperplane [15]. The best hyperplane can be determined with measure margin of hyperplane and find maximum point. Margin is distance between hyperplane with closest pattern from each class. The closest pattern is called support vector [16]. Illustration of hyperplane margins in 2-dimension room can be explained in figure 2:

![Figure 2. Illustration of Hyperplane Margin](image)

In two linearly separated class category, hyperplane with bigger room is better than with hyperplane which too closest with one of category or with both category class. Criteria for determining the decision area in SVM are determined through maximum distance from a point to the hyperplane [17]. Separation vector which have two different classes in some $m$ documents can be formulated as follows:

\[
(y_1, x_1), \ldots, (y_m, x_m), x \in R^n, y \in [-1 + 1]
\]  

With definition of hyperplane is:

\[
(w, x) + b = 0
\]  

$w$ is weighting vector with $w = \{w_1, w_2, w_3, \ldots, w_n\}$. $n$ is number of attributes, $x$ is input dataset which used as attribute, and $b$ is scalar for additional weighting. If two delimiters field represented on inequality, so that obtained equation as follows:

\[
y_i[(w, x_i) + b] - 1 \geq 0
\]  

Distance between each hyperplane according formula of distance line to center point is:

\[
\frac{2}{||w||}
\]  

This distance is maximized with still fulfilling equation (2). Maximized $\frac{1}{||w||}$ is same with minimize $||w||^2$. Furthermore, searching best hyperplane with biggest margin can be formulated into constrain optimization as follow:

\[
\min \frac{1}{2} ||w||^2
\]  

which $y_i[(w, x_i) + b] - 1 \geq 0$ and limitation function as follow:

\[
\sum_{i=1}^{m} a_i[y_i(w, x + b) - 1]
\]
The problem can be solved easily use Lagrange multiplier method. Equation (6) can be transformed into form of Lagrange multiplier function as follow:

\[
L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} a_i [w \cdot x_i + b] y_i - 1
\]

(7)

With \(a_i \geq 0\) is lagrange coefficient. Based on equation (2) obtained equation as follow:

\[
L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} a_i [w \cdot x_i + b] y_i - 1 + \sum_{i=1}^{m} a_i
\]

(8)

To calculate value of Lagrange coefficient, need to calculate L optimum with minimize \(L\) w and b. That value obtains with calculate partial derivative L to \(w\) and \(b\). so that obtain equation as follow:

\[
\min \left( \frac{\partial}{\partial w} L(w, b, a) = 0 \right) \Rightarrow \sum_{i=1}^{m} a_i x_i y_i = 0
\]

(9)

Substitute equation (9) into equation (8) and obtained equation as follow:

\[
L(\alpha) = \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} a_i a_j y_i y_j (x_i \cdot x_j)
\]

(10)

Best hyperplane can be formulated as follow:

\[
\max (L(\alpha)) = \sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} a_i a_j y_i y_j (x_i \cdot x_j)
\]

(11)

With constraint function \(\sum_{i=1}^{m} a_i y_i = 0\) and \(a_i \geq 0\) for \(i = 1, 2, \ldots, m\) equation (11) became:

\[
\hat{w} = \sum_{i=1}^{m} \hat{a}_i x_i y_i
\]

(12)

Furthermore, [18] obtain value of \(b\) as follow:

\[
\hat{b} = -\frac{1}{2} \hat{w} \cdot [x_r + x_s]
\]

(13)

where \(x_r\) and \(x_s\) are support vectors that fulfil each class, for \(\hat{a}_r, \hat{a}_s > 0, y_r = 1, y_s = -1\). Decision function obtained based on equation (12) and (13) is:

\[
f(x) = \text{sign}(\hat{w} \cdot x + \hat{b})
\]

(14)

2.3. Naive Bayes

Naïve Bayes is classification method based on probability. This method calculates a set of probabilities with add up the frequency and value combination from dataset. This method using Bayes theorem and assume all independent attribute [14].

More concrete if assume have document \(D = \{d_1 | i = 1, 2, \ldots, n_d\} = \{d_1, d_2, \ldots, d_{n_d}\}\) and categories \(V = \{v_l | i = 1, 2, \ldots, n_v\} = \{v_1, v_2, \ldots, v_{n_v}\}\). Naive Bayes done by find probabilities \(P(V = v_j | D = d_i)\), is probabilities of categories \(v_j\) if known document \(d_i\). The document \(d_i\) seen as tuple from word in document, that is \(A = \{a_k | k = 1, 2, \ldots, n\} = \{a_1, a_2, \ldots, a_n\}\) which frequency of emergence assume as random variable with multinomial distribution [19].

Determining the best category is based on high posterior probability or called maximum a posteriori (MAP) value which is calculated from the following optimization:

\[
V_{\text{MAP}} = \arg \max_{v_j \in V} P(v_j | a_1, a_2, \ldots, a_n)
\]

(15)

Bayes theorem states about the conditional probability that:
By applying Bayes theorem concept, we can write equation (16) as follows:

\[ P(B|A) = \frac{P(A|B)P(B)}{P(A)} \]  

\[ (16) \]

\[ V_{MAP} = \max_{v_j \in V} \frac{P(a_1, a_2, ..., a_n|v_j)P(v_j)}{P(a_1, a_2, ..., a_n)} \]  

\[ (17) \]

Because value of \( P(a_1, a_2, ..., a_n) \) to all \( v_j \) is same, so its value can be ignored and equation (17) can be written as follows:

\[ V_{MAP} = \max_{v_j \in V} P(a_1, a_2, ..., a_n|v_j)P(v_j) \]  

\[ (18) \]

With assume that each word in \( \{a_1, a_2, ..., a_n\} \) is independent, so that \( P(v_j|a_1, a_2, ..., a_n)P(v_j) \) in equation (19) can be written as follows:

\[ P(a_1, a_2, ..., a_n|v_j) = \prod_i P(a_i|v_j) \]  

\[ (19) \]

So that equation (18) can be written:

\[ V_{MAP} = \max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \]  

\[ (20) \]

Value of \( P(v_j) \) is approach with :

\[ P(v_j) = \frac{N_{v_j}}{N} \]  

\[ (21) \]

With \( N_{v_j} \) is the number of documents which have \( j \) category in training. Furthermore, \( N \) is the number of document which used for training.

According [20], term does not always appear in one of category when classification so the value \( P(a_k|v_j) \) result is 0. To solve that problem, used add-one smoothing or Laplace smoothing, that is adding term frequency as much as 1. So that calculation \( P(a_k|v_j) \) became:

\[ P(a_k|v_j) = \frac{N_{jk} + 1}{N_j + B} \]  

\[ (22) \]

With \( N_{jk} \) is frequency of occurrence of \( a_k \) in document which category \( v_j \). Furthermore, value of \( N_j \) is the total number of all word in document with category \( v_j \), and \( B \) is many variation word in training.

**2.4. Performance Evaluation**

Performance evaluation is method to know accuracy rate from classification. Below is calculation method used to performance evaluation:

**2.4.1. Apparent Error Rate**

Apparent Error Rate (APPER) is a method to see the probability of error in classifying object. Calculation of APPER based on result of classification which has been summarized into clarification table. Below is clarification table for 2 categories:

| Observation | Prediction | Category 1 | Category 2 | Total |
|-------------|------------|------------|------------|-------|
| Category 1  |            | \( n_{11} \) | \( n_{12} \) | \( n_{11} + n_{12} \) |
| Category 2  | \( n_{21} \) | \( n_{22} \) | \( n_{21} + n_{22} \) |
| Total       | \( n_{11} + n_{21} \) | \( n_{12} + n_{22} \) | \( n_{11} + n_{12} + n_{21} + n_{22} \) |
According Table 1, error in classification can be calculated with using APPER [21] which is defined as follow:

\[
APPER = \frac{n_{22} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\% \quad (23)
\]

Formula to see the chance of accuracy in classification can be written as follow:

\[
Accuracy\ of\ classification = 100\% - APPER \quad (24)
\]

The accuracy of classification value in Eq. (24) show result of evaluation to measure accuracy rate of classification.

3. Research Method

3.1. Data Resources

Data used for this study is tweet about zoning system admission policy with keyword “zonasi”. Retrieval of tweets using website crawling with Application Program Interface (API) which has been already provided by Twitter. Tweet about zoning system admission policy around 1000 tweet. Tweet that obtained are cleared from ethnic and racial issues. The amount of tweet after cleared from ethnic and racial issues is 440 tweets. The period of tweets taken from July 2019 until September 2019.

3.2. Research Variables

Research variable divided into predictor variable and response variable. Predictor variable is text data about zoning system admission policy. Furthermore, response variable is sentiment classification (positive or negative). Text data about zoning system admission divided into training data and testing data using method in section 2.1.

3.3. Data Analysis Procedure

The procedure to sentiment analysis about zoning system admission policy using SVM and Naïve Bayes is given as follow:

1. Get tweets about zoning system admission policy from twitter with period of tweet July 2019 until September 2019.
2. Cleaning the text data from hashtag (#), url, emoticon, number, symbol, slang words, and stop words.
3. Labelling text data with negative or positive manually.
4. Divided text data that has been cleaned into training data and testing data using k fold cross validation in section 2.1 which number of k is 10.
5. Classifying training data using SVM and Naïve Bayes method in section 2.2 and section 2.3 based on negative or positive sentiment.
6. Testing the result of classification from training data in testing data
7. Calculating accuracy of classification by using equation (24)

4. Results and Discussion

The text data is tweet about zoning system admission policy from twitter with period July 2019 until September 2019. The amount of text data obtained is 440 tweets. After get text data, text data cleaned from hashtag (#), url, emoticon, number, symbol, slang word, and stop word to get key word about negative and positive tweets. The text data that has been cleaned, Text data is given a negative or positive label according keyword that has been obtained by manually. The text data is divided into 10 part and text data divided into data training and data testing based on Fig.1. The text data is classified using SVM and Naïve Bayes for 10 iterations. Each iteration is calculated accuracy of classification using equation 24 and calculated average of accuracy for determine accuracy of classification method.
The results of sentiment analysis for zoning system admission policy by using SVM method can be seen in Table 2.

Table 2. Result of Sentiment Analysis Using SVM

| Iteration | Accuracy | Category Majority | Total Accuracy |
|-----------|----------|-------------------|---------------|
|           | Training | Testing          |               |
| Iteration 1 | 95.71% | 77.27% | Negative | 93.86% |
| Iteration 2 | 95.71% | 63.64% | Negative | 92.5% |
| Iteration 3 | 95.45% | 70.45% | Negative | 92.95% |
| Iteration 4 | 95.45% | 59.09% | Negative | 91.82% |
| Iteration 5 | 95.71% | 63.64% | Negative | 92.5% |
| Iteration 6 | 96.21% | 63.64% | Negative | 92.95% |
| Iteration 7 | 96.21% | 75% | Negative | 93.64% |
| Iteration 8 | 95.45% | 70.45% | Negative | 92.95% |
| Iteration 9 | 95.71% | 72.73% | Negative | 93.41% |
| Iteration 10 | 95.71% | 65.91% | Negative | 92.79% |

Average Accuracy 92.93%

Table 2 shown that category majority for zoning system admission using SVM is negative. It means the majority of Indonesian people who use Twitter have bad opinion about zoning system admission policy. The average accuracy rate of sentiment analysis using SVM for zoning system admission policy is 92.93%. It shown that sentiment analysis using SVM for zoning system admission policy can classify text data precisely as much 409 text data.

Below the result table of sentiment analysis for zoning system admission policy using Naïve Bayes method:

Table 3. Result of Sentiment Analysis Using Naïve Bayes

| Iteration | Accuracy | Category Majority | Total Accuracy |
|-----------|----------|-------------------|---------------|
|           | Training | Testing          |               |
| Iteration 1 | 81.57% | 77.27% | Negative | 81.14% |
| Iteration 2 | 82.07% | 61.36% | Negative | 80.00% |
| Iteration 3 | 81.57% | 63.64% | Negative | 79.77% |
| Iteration 4 | 81.57% | 58.14% | Negative | 79.09% |
| Iteration 5 | 80.56% | 63.64% | Negative | 78.86% |
| Iteration 6 | 81.82% | 61.36% | Negative | 79.77% |
| Iteration 7 | 80.30% | 75% | Negative | 79.77% |
| Iteration 8 | 81.57% | 75% | Negative | 80.91% |
| Iteration 9 | 81.82% | 71.11% | Negative | 80.91% |
| Iteration 10 | 80.56% | 59.09% | Negative | 78.41% |

Average Accuracy 79.86%

Table 3 shown that category majority for zoning system admission using Naïve Bayes is negative. It means the majority of Indonesian people who use Twitter have bad opinion about zoning system admission policy. The average accuracy rate of sentiment analysis using Naïve Bayes for zoning system admission policy is 79.86%. It shown that sentiment analysis using Naïve Bayes for zoning system admission can classify text data precisely as much as 351 text data.

5. Conclusion
Based on both methods, majority classification for zoning system admission policy is negative. It means majority of Indonesian people who use Twitter have bad opinion and tend to disagree with zoning system admission policy for improve quality of education in Indonesia. The average of
accuracy rates for SVM and Naïve Bayes are 92.93% and 79.86%, respectively. It means that the analysis sentiment for zoning system admission policy by using SVM is better than Naïve Bayes.

References
[1] Ministry of Education 2005 Rencana Strategis Departemen Pendidikan Nasional 2005-2009 (Indonesia: Information and Public Relation Center)
[2] Tjalla A 2010 Potret Mutu Pendidikan Indonesia Ditinjau dari Hasil-Hasil Studi International. Prosiding Seminar Temu Ilmiah Nasional Guru II 300 pp 370-379.
[3] Martin and Sala X 2008 The Global Competitiveness Report 2008-2009 (United States of America: World Economic Forum)
[4] Purwanti D, Irawati I, Adiwisastra J, and Bekti H 2009 Implementasi Kebijakan Penerimaan Peserta Didik Baru Berdasarkan Sistem Zonasi di Kota Bandung Jurnal Governansi 5 pp 12- 23
[5] Zainal M A, and Asrori 2018 Peranan Sekolah Kawasan Berabasis Sistem Zonasi dalam Pembentukan Karater di SMP Negeri 15 Kedung Cowek Surabaya Tadarus: Jurnal Pendidikan Islam 7
[6] Wahyuni, and Dinar 2018 Pro Kontra Sistem Zonasi Penerimaan Peserta Didik Baru Tahun Ajaran 2018/2019 Jurnal Info Singkat 10
[7] Liu L C, Hsaion W H, Lee C H, Lu G C, and Jou E 2012 Movie Rating and Summarization in Mobile Environment IEEE Transaction on System, Man, and Cybernetic Part C (Applications and Reviews) 42 pp 397-407
[8] Tan A H 199 Text Mining: The State of the Art and the Challenges Proceedings of the PAKDD 199 Workshop on Knowledge Discovery from Advanced Databases 8 pp 65-70
[9] Chamidah N, Gusti K H, Tjahjono E and Lestari B 2019 Improving of classification accuracy of cyst and tumor using local polynomial estimator TELKOMNIKA, Vol.17, No.3, pp.1492-1500
[10] Anwar A.E and Chamidah N 2019 Glaucoma Identification on Fundus Retinal Images Using Statistical Modelling Approach IOP Conf. Series: Materials Science and Engineering 546 052010
[11] Puspitawati A and Chamidah N 2019 Choroidal Neovascularisation Classification on Fundus Retinal Images Using Local Linear Estimator IOP Conf. Series: Materials Science and Engineering 546 052056
[12] Rizka N and Chamidah N 2019 Lung Tumor Classification on Human Chest X-Ray Using Statistical Modelling Approach IOP Conf. Series: Materials Science and Engineering 546 052065
[13] Andi N H 2015 Analisis Sentimen Terhadap Wacana Politik pada Media Masa Online Menggunakan Algoritma Support Vector Machine dan Naïve Bayes Jurnal Elektronik Sistem Informasi dan Komputer 1
[14] Pramana S, Yuniarto B, Mariyah S, Santoso I, and Nooraeni R 2018 Data Mining dengan R Konsep Serta Implementasi (Indonesia: In Media)
[15] Tripathy A, Agrawal A, and Santau K R 2015 Classification of Sentimental Reviews Using Machine Learning Techniques 3rd International Conference on Recent Trend in Computing 2015 (ICRTC-2015) Procedia Computer Science 57 pp 821-829
[16] Athoiilah M, Irawan M I, and Imah E M 2015 Support Vector Machine untuk Image Retrieval Prosiding Seminar Nasional Matematika dan Pendidikan Matematika 2015 ISBN No 978-979 028-728-0
[17] Manning C D, Raghavan P, and Schutze H 2009 An Introduction to Information Retrieval (England: Cambridge University Press)
[18] Gunn S R 1998 Support Vector Machine for Classification and Regression (Southampton: University of Southampton)
[19] McCalum A, and Nigam K 1998 A Comparison of Event Model for Naïve Bayes Text Classification DOI: 10.1.1.65.932
[20] Pin L V, and Luo Z 2012 Naïve Bayes Text Classifier Based on Different Probability Model
   *International Journal of Digital Content Technology and its Application* 6

[21] Meyers L S, Glenn G, and Guarino A J 2016 *Applied Multivariate Research, 3rd Edition* (United
   State of America: SAGE Publication)