Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 6 No. 1, 2021, pp. 46 – 53

Classification of Student Aspiration Using Naïve Bayes Classifier

Ifan Rizqa*[1], Christy Atika Sari[2]
Dian Nuswantoro University, Semarang, Indonesia 50131
E-mail : ifan.rizqa@dsn.dinus.ac.id[1], christy.atika.sari@dsn.dinus.ac.id*[2]*
**Corresponding author*

Mohamed Doheir[3]
University Teknikal Malaysia Melaka, Melaka, Malaysia
E-mail : jeruslem.20088@gmail.com[3]

Abstract - Students aspiration are various demands from the student that packed in creative idea to propose changing process of a thing. Mostly, aspiration delivered in complaints and expectation. Aspiration is used for evaluating the laxity and early detection in university quality system for the better. This activity took place in Dian Nuswantoro University, and Student Representative Council (SRC) is the unit to manage the students aspiration. Aspiration is obtained through predetermined mechanism such as manual questionnaire distribution and or using google form. The provided questionnaire requires student to fill the content according to the provided aspiration categories. However, the problem is sometimes the student choose the wrong category according to the content. Therefore, it is needed to create an application that can classified the students aspiration automatically. Document text classification become the best way to determine the category based on the content of the students aspiration. Naïve bayes classifier method is used because it is capable to produce high accuracy. With 1000 data training document of each category, "facilities and infrastructure" (facilities), "lecturers" (attitudes, teaching methods, material delivered), "staffing and the academic system"(attitudes, ways of working, providing information), and "suggestions and feedback". This experiment achieved 90.20% accuracy. It can be said that this method is worth to implement in this research.

Keywords - Students Aspiration, Classification, Naïve Bayes

1. INTRODUCTION

Student aspirations are various guides from students that are packaged in creative ideas to propose a change process for something. Student aspirations are considered good when they are conveyed not only verbally but also in writing [1][2]. These aspirations are accompanied by scientific arguments and put forward scientific norms and rules [3]. Most of the aspirations submitted are in the form of complaints and hopes. Aspirations are useful as material for evaluation and early detection of weaknesses in the quality system for better universities [4]. This evaluation can be a guarantee of higher education quality so that students get satisfaction and there will be continuous improvement [5] [6]. Therefore, to improve campus standards, student aspirations are needed.

In general, in some universities, students can convey their aspirations through an online questionnaire. Through the questionnaire, aspirations will be recorded and structured properly [1]. Aspirations will be examined in order to filter which ones to convey. When an aspiration has been decided, it is then conveyed to the relevant parties for follow-up [1]. This

happened at Dian Nuswantoro University Semarang that there is a unit that specifically accommodates student aspirations which are then reviewed and conveyed to the relevant section by the Faculty Student Representative Council (SRC). Students' aspirations are obtained through a predetermined mechanism such as distributing questionnaires in the form of paper forms that have been prepared and also with a google form-based interface whose composition is adjusted exactly to the questionnaire format that has been distributed manually [5].

Student Representative Council often faces some difficulties. First, when students fill out the questionnaire, several choices of aspiration categories are provided. However, sometimes the categories selected and the aspirations do not match. So that Student Representative Council categorizes these aspirations manually. Second, the questionnaire that was still in paper form and the number of aspirations collected were not recorded and were not well structured, so that Student Representative Council finds it difficult to assess existing aspirations.

Of the several problems presented, classification of document text is the best way to determine categories based on the content of student aspirations. Using information retrieval can solve this problem [7]. There are several classification methods in it. However, the Naive Bayes Classifier will be used. The reason is that this method can classify document texts with various categories, for example the Naive Bayes Classifier method can classify Twitter posts about traffic jams in Bandung using 100 training data. This method is capable of producing 78% and 13106 test data capable of producing an accuracy of 91.6% [8]. A reference like this proves that the Naive Bayes Classifier method can provide high accuracy even with little training data. Research was also conducted by Dyarsa with the same method resulting in up to 91% accuracy by using 1000 training data in each category [9].

With some of the descriptions that have been submitted, the author wants to conduct research by creating a system for classifying student aspirations using the Naive Bayes Classifier method. In this system, students can fill in aspirations without choosing the category that is usually provided. Because later when the aspiration form is filled in, the program will detect or classify these aspirations. In addition, chart diagrams and recordings of student aspirations are provided which will make it easier and faster for Student Representative Council to study. Aspiration reports that have been reviewed will be submitted to the relevant parties.

2. RESEARCH METHOD

2.1. State of The Art

Related to this research, here are the relevant research that has been done previously as the author's reference. The research that has been done includes in Table 1.

Table 1. State of The Art

Year	Researcher	Tittle	Decription	Result
2013	Sandi F.R and Edi W [8]	Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification	Review traffic jams by classifying tweets.	lowest accuracy is 78% and highest accuracy is 91.06%
2014	Amir Hamzah [10]	Sentimen Analisis untuk Memanfaatkan Sarana Kuesioner dalam Evaluasi Pembelajaran dengan Menggunakan Naïve Bayes	Sentiment analysis is used to evaluate the success of the learning process from student opinions obtained from	Opinion classification can be done automatically,

		Classifier	distributed questionnaires.	with an accuracy of 85.96%.
2014	Shruti B.M and Dr. Vishal Goyal [11]	Text News System Using Naïve Bayes Classification	News grouping in Punjabi which was originally still manual.	0.72 recall, 0.78 precision, 0.74 fi-measure.
2015	Zulfany Erlisa Rasjid and Reina Setyawan [12]	Document Classification of Assamese Text Using Naïve Bayes Approach	Developing the NBC method by classifying an online news piece.	94,41% precision and 94,69% recall.
2015	Dyarsa S.h, Noor A.S and Erlin Dolphina [9]	Analisis Sentiment pada Sosial Media Twitter Menggunakan Naïve Bayes Classifier Terhadap Kata Kunci “Kurikulum 2013”	Conduct sentimentation on tweets related to the "2013 curriculum"	Produced accuracy 91%
2020	Raghavendra Vijay Bhasker Vangara, et al. [13]	Opinion Mining Classification using Naive Bayes Algorithm	focus on the aspect of reducing the time and effort for the user by recommending the best product to him.	Produced 85.7% better than previous paper in 64.5%

Based on the six studies that have been described above, the Naïve Bayes Classifier method can categorize documents with high accuracy with few documents. However, the more documents or training data, the higher the resulting accuracy. Therefore, this research will use the Naïve Bayes Classifier method as a classification method for student aspirations which is grouped into 5 categories, namely: "facilities and infrastructure" (facilities), "lecturers" (attitudes, teaching methods, material delivered), "staffing and the academic system"(attitudes, ways of working, providing information), and "suggestions and feedback".

2.2. Naïve Bayes

Classification is a learning function that maps or classifies a data element (item) into one of several predefined classes [14]. In this research, the Naïve Bayes method involves training data and test data using the following (1). Naïve Bayesian Classifier can also be defined as a classification method based on probability theory and Bayesian theorem with the assumption that each variable or parameter determines the decision to be independent [15], so that the existence of each variable has nothing to do with the existence of other attributes.

$$P(w_i|C) = \frac{count(W_i,C)+1}{count(C)+|V|} \qquad (1)$$

Where *C* is class, W_i is the i^{rd} word *W*, Count $(W_i\ C)$ is number of words W_i in *C*, Count *(C)* is number of words in class *C* and $|V|$ is number of vocabulary. Based on [16], [17], this algorithm has a process of calculating the number of word equations between training data and test data. From the results of these calculations yield probability values for each category. The maximum value obtained is the result of the category obtained or detected .

2.3. Preprocessing

Prepocessing is done on training data and test data in order to convert the text to the term index. So a set of term indexes that can represent documents and are more structured. In this process, preposeccing uses literary libraries. The library is obtained by installing the vendor's composer, which includes literature. This makes it easier to prepare documents easily and quickly. Before prepocessing: "for the facility, please enlarge the GALLERY ROOM SO THAT DOES NOT LIMIT STUDENT CREATIVITY". After prepocessing with literary libraries: "for the facilities, please use the large gallery space again so that there is no limit to student creativity".

2.4. Training Data

This study uses the Naive Bayes Classifier method as a classification process for student aspirations. The reason for using this method is because previous research has proven that it has quite high accuracy even though only a few documents are used. In applying this method, the classification process is divided into 2 parts, namely processing training data and test data.

Training data obtained from Student Representative Council (SRC) faculty at Dian Nuswantoro University. The problems that occur in each faculty have different categories. Therefore it is necessary to equalize the categories to make it easier to test the algorithm on student aspirations. From the training data obtained, 1000 documents of student aspirations are valid. The document is said to be valid because it already has the specified categories, so that the documents from the training data are not grouped manually. Table 2 is a collection of training data obtained.

Table 2. Training data

No	Document	Categories
1	students need more parking space than cars and motorbikes. And students need a resting place after finishing college. Lecturers do not just give replacement hours	Facilities and infrastructure
2	the H white borders building is small, and noisy due to the unfinished building construction. the spread of the wi-fi area is less wide at a slow speed.	Facilities and infrastructure
3	lecturer: the lecturers often come on time and on time, so they can't make unlimited wifi	Lecture
4	many of Udinus' lecturers did not have fun and were difficult to understand how to teach	Lecture
5	the parking attendant is less detailed (The motorbike still has land but cannot enter it should be neatly arranged)	Staffing and Academic System
6	sometimes scholarship information is notified to students at the close of the grace period	Staffing and Academic System
…	…	…
…	...	…
249	improve the existing and often complain about students being accommodated and implemented	Suggestions and Feedback
250	you should have your own building for the health faculty, like other university	Suggestions and Feedback

2.4. Testing Data

The test data is used to determine which algorithms can be used to detect categories automatically. The test data used does not have any previous categories. By including the Student ID Number, the training data is valid for use. Therefore, training data is obtained when the researcher has completed the automatic classification system for student aspiration categories by printing the Naïve Bayes Classifier algorithm in it. The web-based system is hosted so that all Dian Nuswantoro University students can access and try the program at the same time.

Table 3. Testing data

No	Student ID Number	Document
1	D22.2015.01612	requires learning that is more exciting, not monotonous, for example learning outside the campus
2	D22.2015.01622	LCD Monitors have several classes that are difficult to connect with a laptop
…	…	
…	…	
118	B11.2014.03365	Hot, less trees.
119	B11.2014.03365	Why do the elevators often break down?

2.5. Proposed Scheme

The training data processing process is provided with a flowchart and an explanation as follows in Figure 1. Based on Figure 1, training data has been processed using several steps here :

1. Training data has been categorized and stored in the database.
2. All training data go through a preprocessing process to make the data more structured. The results of preprocessing produce the appearance of the word all training data (term).
3. Of all the words that appear, the total is counted (|V|).
4. Then count all the words in each category Count (C).
5. The training process is complete.

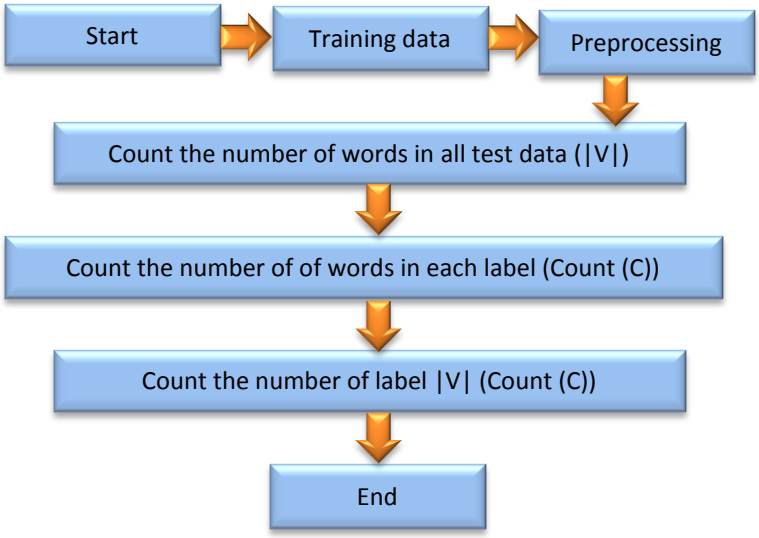

Figure 1. Steps of training data

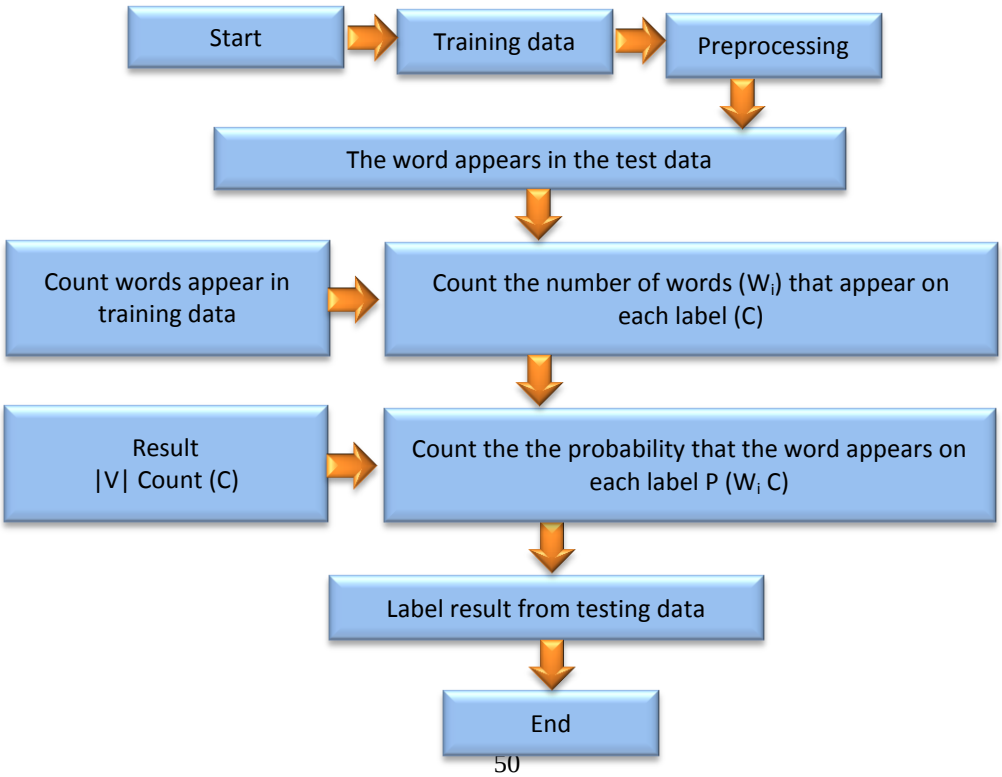

Figure 2. Steps of testing data

Based on Figure 2, testing data has been processed using several steps here :

1. Input the test data to determine the category.
2. The test data is entered into the preprocessing stage to eliminate meaningless characters.
3. The preprocessing results will show the terms from the test data. Then count the number of words that appear.
4. Count the number of test data terms that appear in the training data terms for each category.
5. Calculate the probability $P(W_i|C)$ by involving the sum of all terms in the training data $(|V|)$ and the number of words from the training data for each category *(Count (C))*
6. Find the result of the highest probability $P(W_i|C)$.
7. Then the output of the test data is known for the category.
8. The test data process is complete.

3. RESULTS AND DISCUSSION

The following is a table of the results of the probability value and category detection from the test documents that have been carried out as follow in Table 4.

Table 4. Experiment Results

No	Document	Probabllity				Result (max)
		FI	L	SAS	SF	
1	Easy input, but sometimes runs out of class	1,77E-09	6,27E-10	2,50E-07	1,04E-12	SAS
2	Servers are often down	5,09E+02	2,54E+02	5,69E+03	4,30E+02	SAS
3	Sometimes the material provided is not conveyed well, hopefully in the future it can explain it better so that students understand and understand	2,19E-42	1,54E-33	1,03E-39	3,96E-40	L
…	…					
…	…					
197	prayer hours and lecture hours are very tight, especially evening prayers	1,07E-19	2,25E-19	1,84E-19	1,36E-19	SAS
199	If possible, repair the elevator in building c. The problem is often jammed, and also given a fragrance. Because the place is damp, it smells easy. Thank you	2,13E-48	7,80E-54	2,32E-53	3,79E-50	FI

Information :
FI is Facilities and infrastructure
L is Lecture
SAS is Staffing and Academic System
SF is Suggestions and Feedback

Based on the data obtained, this accuracy calculation uses training data of 1000 documents. Here are the results of the calculations as follow in Table 5.

Table 5. Confussion Matrix

	Actual of FI	Actual of L	Actual of SAS	Actual of SF
Prediction of FI	188	57	2	3
Prediction of L	1	249	0	0
Prediction of SAS	3	17	226	4
Prediction of SF	5	5	1	239

Table 5 above shows that not all predicted documents are in accordance with the actual results. With a configuration matrix table of 1000 training data documents, each produces 188 "Facilities and Infrastructure", 249 "Lecturers", 226 "personnel and academic systems" and 239 "Suggestions and Feedback" which match between predictions and actuals. Therefore the accuracy value obtained is as follows :

$$\text{Accuracy} = \frac{188 + 249 + 226 + 239}{1000} \times 100\% = 90{,}2\ \%$$

The resulting accuracy is 90.2%. According to Triowali, if the accuracy value is 0.90-1.00, it can be said to be excellent classification . So that the conclusion is this training data has excellent classification accuracy or very good.

4. CONCLUSION

Based on the results of the tests that have been done in this final project, it can be concluded that:

1. The Naïve Bayes Classifier method can be implemented to categorize the aspirations of Dian Nuswantoro University students.
2. Using training data as many as 1000 documents, each in the category of 250 documents resulting in an accuracy of 90.2% which can be called an excellent classification diagnostic value or very good with the Confusion Matrix testing technique.
3. This system can classify student aspirations automatically and systematically. Because information is provided on the amount of aspiration data submitted by each category. This existence makes it easier for Student Representative Council to manage student aspirations.
4. Data reports provide chart information on the amount of aspiration data for each category in each faculty. With a view to analyzing and comparing which categories have the highest number of aspirations in each faculty.

After conducting this research, it was found that there were some things that were imperfect due to limited time and abilities. So suggestions that are appropriate for the perfection of this research will be explained as follows:

5. Classifying student aspirations using other methods that have higher accuracy.
6. The more training data used, the higher the probability of accuracy being obtained.
7. Because previously the categories have been determined by the manager, and according to the analysis there are similarities between the documents from one category to another.
8. Therefore it is also necessary to do training data clustering with certain methods in order to know the exact categories.
9. In the system, there are sub-categories that are still filled in by students with manuals and the sub-categories provided are determined by the researcher based on field analysis. Therefore, it can use clustering and classification methods for future research.

REFERENCES

[1] V. Dhanalakshmi and D. Bino, “Opinion mining from student feedback data using supervised learning algorithms,” in *2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016*, 2016, pp. 1–5.

[2] L. Khanna, S. N. Singh, and M. Alam, “Educational data mining and its role in determining factors affecting students academic performance: A systematic review,” 2017, doi: 10.1109/IICIP.2016.7975354.

[3] F. F. Balahadia, M. C. G. Fernando, and I. C. Juanatas, “Teacher’s performance evaluation

tool using opinion mining with sentiment analysis," in *Proceedings - 2016 IEEE Region 10 Symposium, TENSYMP 2016*, 2016, pp. 95–98, doi: 10.1109/TENCONSpring.2016.7519384.

[4] S. Dey Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A Novel Feature Selection Technique for Text Classification Using Naïve Bayes," *Int. Sch. Res. Not.*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/717092.

[5] F. C. Permana, Y. Rosmansyah, and A. S. Abdullah, "Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media," *J. Phys. Conf. Ser.*, vol. 893, no. 1, 2017, doi: 10.1088/1742-6596/893/1/012051.

[6] F. F. Balahadia and B. E. V. Comendador, "Adoption of Opinion Mining in the Faculty Performance Evaluation System by the Students Using Naïve Bayes Algorithm," *Int. J. Comput. Theory Eng.*, vol. 8, no. 3, pp. 255–259, 2016, doi: 10.7763/ijcte.2016.v8.1054.

[7] M. L. Barrón Estrada, R. Zatarain Cabada, R. Oramas Bustillos, and M. Graff, "Opinion mining and emotion recognition applied to learning environments," *Expert Syst. Appl.*, vol. 150, 2020, doi: 10.1016/j.eswa.2020.113265.

[8] S. F. Rodiyansyah and E. Winarko, "Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification," *IJCCS*, vol. 6, no. 1, pp. 91–100, 2012, doi: 10.1163/ej.9789004182127.i-302.6.

[9] D. S. Pamungkas, N. A. Setiyanto, and E. Dolphina, "Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap," *Techno.COM*, vol. 14, no. 4, pp. 299–314, 2015.

[10] A. Hamzah, "SENTIMENT ANALYSIS UNTUK MEMANFAATKAN SARAN KUESIONER DALAM EVALUASI PEMBELAJARAN DENGAN MENGGUNAKAN NAIVE BAYES CLASSIFIER (NBC)," in *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, 2014, no. November, pp. 211–216.

[11] B. Shruti, And, and G. Vishal, "Text News Classification System using Naïve Bayes Classifier," *Int. J. Eng. Sci.*, vol. 3, no. December, pp. 209–213, 2014, [Online]. Available: http://www.ijoes.vidyapublications.com.

[12] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.

[13] R. Vijay, B. Vangara, K. Thirupathur, and S. P. Vangara, "Opinion Mining Classification using Naive Bayes Algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 495–498, 2020, doi: 10.35940/ijitee.e2402.039520.

[14] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," 2017, doi: 10.1109/CITSM.2017.8089231.

[15] P. Liu, H. han Zhao, J. yu Teng, Y. yan Yang, Y. feng Liu, and Z. wei Zhu, "Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark," *J. Cent. South Univ.*, vol. 26, no. 1, pp. 1–12, 2019, doi: 10.1007/s11771-019-3978-x.

[16] K. M. A. Hasan, M. S. Sabuj, and Z. Afrin, "Opinion mining using Naïve Bayes," in *2015 IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE 2015*, 2016, pp. 511–514, doi: 10.1109/WIECON-ECE.2015.7443981.

[17] V. M., J. Vala, and P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," *Int. J. Comput. Appl.*, vol. 133, no. 9, pp. 7–11, 2016, doi: 10.5120/ijca2016907977.