The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4179

Application of SMOTE Method on Topic Based Question Classification Using Naïve Bayes Algorithm

Orvalamarva[1], Oktariani Nurul Pratiwi[2], Faqih Hamami [3]
orvalamarva@student.telkomuniversity.ac.id, onurulp@telkomuniversity.ac.id, faqihhamami@telkomuniversity.ac.id
[1, 2, 3] Telkom University, Bandung

Article Information

Received : 23 Jun 2024
Revised : 2 Jul 2024
Accepted : 25 Jul 2024

Keywords

Classification, Naïve Bayes, SMOTE, Cross Validation, Symbols

Abstract

In today's digital era, the utilization of technology in education is essential to support the learning process. This research discusses the classification of junior high school mathematics questions using the Naïve Bayes method. The use of an automated system in question classification helps reduce time and effort in grouping questions based on topics. The Naïve Bayes method was chosen because of its simplicity and ability to process data. The results showed that Naïve Bayes with SMOTE and Math symbols achieved 69% accuracy, while without SMOTE, the accuracy was lower. Cross-validation showed that the classification without symbols attained an accuracy of 89.35%, slightly superior to the classification using symbols, which was 88.79%. This result indicates that Naïve Bayes with SMOTE is more effective. Although the difference in accuracy with or without symbols is slight 0.56%, the performance is relatively equivalent, with an accuracy of 89%.

A. Introduction

Educators in today's digital era increasingly rely on technology as a supporting tool for the learning process. One effective approach is to use technology to compile and categorize questions that are relevant to the learning material [1]. In this way, the assessment of student's abilities can be more structured and meaningful in accordance with the learning objectives. This view is supported by [2], who emphasizes the importance of item analysis to ensure the quality of accurate assessment of students' understanding of the subject matter. Classifying questions based on topics is also a crucial aspect of education. Classification is the process of building a model from labeled data that is used to predict the class label of unlabeled data [3]. It aims to associate labels to questions based on specific criteria, such as difficulty level, expected answer type, or specific school topics [4]. Classification models, such as the one developed by [5], have proven their usefulness in assisting teachers and students in selecting questions that suit learning needs. Classification models aim to organize educational materials available online using machine learning technology to divide them by topic or type [6]. This grouping helps in simplifying access to educational resources, making it easier for learners to find the information they need. However, traditional classification is time-consuming and involves complex procedures [7]. This process requires in-depth analysis of various elements and terms in the dataset, which can be challenging. Therefore, to improve efficiency and overcome these obstacles, it is important to implement an automation system that can manage and classify data more efficiently.

According to the results of research by [8], this study aims to classify 500 hadiths of Sahih Bukhari into certain categories using the Naive Bayes Classifier. Methods include data collection from sunnah.com, the pre-processing stage, giving word weight with TF-IDF, and classification with Naive Bayes Classifier. Evaluation using the confusion matrix showed an increase in accuracy from 80.14% before pre-processing to 82.27% after pre-processing, demonstrating the effectiveness of pre-processing in improving classification accuracy. Naive Bayes Classifier can classify hadiths well, with the potential for further development by adding data attributes and exploring other methods. This research successfully produced an automated system for hadith search by category, making an important contribution to text processing and classification of religious documents.

Another research by [5] is to develop a classification model for Biology questions in grade 11 high school by grouping questions into nine different topic categories. The results demonstrated that the Naïve Bayes algorithm reached an accuracy of 72.72%. In addition, performance evaluation using the cross-validation method resulted in an accuracy of 73.09%, with precision, recall, and F1-Score values reaching 73%, 73%, and 70%, respectively. On the other hand, the C4.5 algorithm achieved an accuracy of 54.54%, with a cross-validation accuracy value of 54.09%. Precision, recall, and F1-Score for the C4.5 algorithm are 58%, 56%, and 55% respectively. This research provides convenience in the field of education by developing a classification method that can assist teachers in grouping and organizing questions based on several specific topics in Biology. The application of the Naïve Bayes algorithm proved to be more effective in accurately predicting question categories compared to the C4.5 algorithm in these results.

In the context of question classification, there are various methods, such as Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), and others [9]. However, this research utilizes the Naïve Bayes approach to categorize Math exam questions based on topics. Naïve Bayes is a frequently used data mining algorithm, which makes the simple assumption that all attributes are independent based on their class, thus simplifying probability computation [10]. This algorithm is known to process training and test data in various proportions efficiently [11]. In this study, the authors are interested in applying the Naïve Bayes algorithm to a dataset of Math exam questions from a junior high school question bank. This research not only classifies questions based on topics but also considers the influence of symbols in the classification process using the Naïve Bayes algorithm. Using this algorithm is expected to identify and classify questions based on topics without spending much time to increase efficiency in question classification.

B. Research Method

In this research, several phases must be passed to solve the problem. These phases include selection of data, preprocessing, transformation, modeling, and evaluation. The following is a compiled systematic diagram.

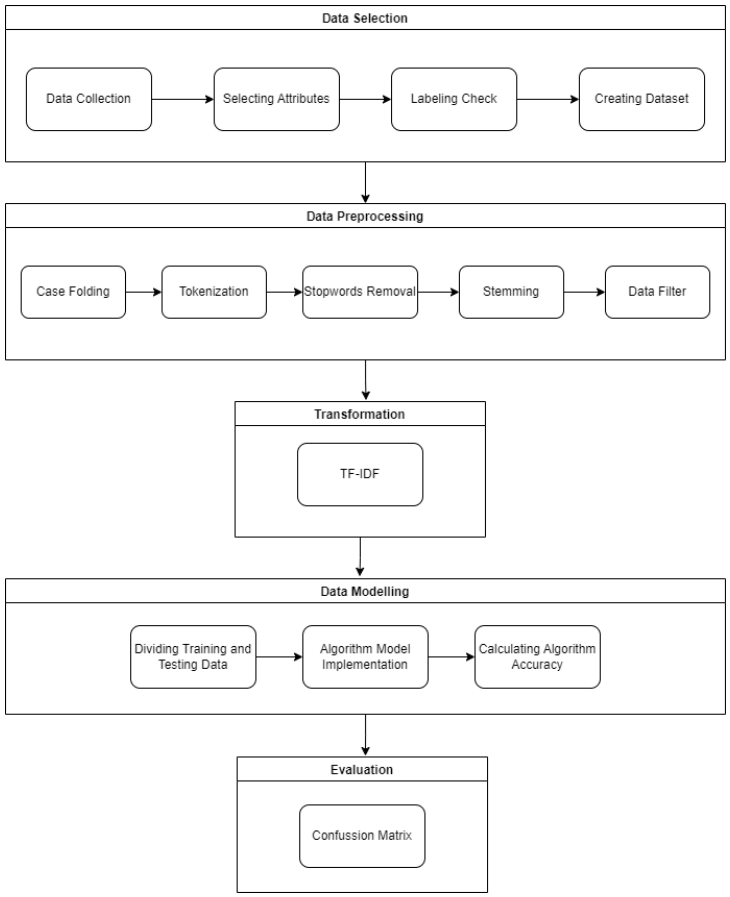


Figure1. Problem solving systematics

Data Selection

The data collection process is sourced from the question bank database of the Pinterran learning system, which contains a set of multiple choice questions about various junior high school level subjects, namely Indonesian Language, English, Mathematics, and Natural Sciences. This research only focuses on the classification of Mathematics questions. In this stage, data sorting includes selecting the required attributes, removing HTML elements in the question text, and changing the name of the "tag" attribute to "topic". The "topic" attribute was used as a label to categorize the questions, while the other attributes were used as variables. The cleaned question bank was exported back into an Excel file as input in the next stage, namely data preprocessing.

Data Preprocessing

In data preprocessing, several processes are undertaken. First, case folding converts uppercase letters to lowercase letters to equalize all letters, thus avoiding ambiguity in text analysis. For example, the words "Cube" and "cube" are considered the same after case folding. Second, tokenization separates the uniformed text into tokens, such as words or symbols, to facilitate further analysis. Third, stopword removal removes words that do not provide essential information, such as "given" and "following". In contrast, words such as "height" and "time" are retained because they are essential in math problems. Fourth, stemming converts the words in the text into their base word form by removing any affixes or suffixes from the words.

After going through several processes, the data will be filtered. The filtering is necessary because, based on the clean data, there are a total of 1718 questions with 301 different topics, but most topics only have a few questions. Including these could lead to model breakage. Therefore, the top 10 topics with the highest number of questions were selected.

Table 1. Filtered Data

topic	number of questions
algebra	35
function	38
set	33
circle	107
probability	79
comparison	37
two_variable_equation	31
relation	33
statistics	30
angle	29
	452

Transformation

The next stage is to apply TF-IDF to assess the significance of each word within the document. This technique emphasizes calculating the frequency of each word in the document (TF: Term Frequency) and how the word is distributed throughout the text document (IDF: Inverse Document Frequency) [12]. Some terms, such as interrogative verbs and interrogative words, have greater importance in

classification tasks than others [13]. The following is the TF-IDF calculation formula [14]:

$$W_{xy} = tf_{dt} \times idf_t \quad (1)$$

Description:
W_{xy}: weight of document x against word y
Tf_{dt}: the frequency of a word in a document
Idf_t: inverse document frequency, calculated as log(N/df)
N: the total number of documents
df: the count of documents that contain the word being searched

Data Modelling

First, the division of training data and testing data that will be entered into the model. The data consists of a single label of 452 rows that have been filtered before. Next, the Naïve Bayes model is employed to execute direct classification. Before that, TF-IDF is applied first to convert text into numerical representations that can be understood by machine learning models such as Naïve Bayes. After TF-IDF values are generated for each word in each document, they are used as input to calculate Naïve Bayes probabilities. The category with the highest probability is used as the prediction of the Naïve Bayes model.

Evaluation

The metric used in this classification is a confusion matrix to see how well the model distinguishes between different labels. System performance evaluation is generally done by examining the data in the matrix [15].

Table 2. Confusion Matrix

		Prediction		
		C_1	C_2	C_3
Actual	C_1	A	B	C
	C_2	D	E	F
	C_3	G	H	I

According to [16], classification with multiple classes using TP, TN, FP, and FN differs slightly. In this classification's context, the discussion focuses on each label in turn. For example, when the focus is on label 1, TP is A The following are the TN, FP, and FN values when the focus is on label 1:

True Positive (TP): If label one is predicted correctly, then TP = A.

True Negative (TN): The sum of all cells except those in the column and row of label 1, TN = B + C + E + F + H + I.

False Positive (FP): The sum of all actual values other than label one predicted as label 1, FP = D + G.

False Negative (FN): The sum of all predicted values that are not labelled 1 when the actual value is 1, FN = B + C.

The results of this metric can be used for accuracy, precision, recall, and f1-measure calculations. In finding the final result of the calculation, values such as precision are obtained from the average precision of each label and the calculation of recall and F-measure [16]. Here are the formulas for classification evaluation [17].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F-measure} = \frac{2x(precision\ x\ recall)}{precision+recall} \quad (5)$$

C. Result and Discussion

Imbalance Handling

In the previous preprocessing stage, the author filtered the data that will be included in the model. To address the issue of data imbalance, the oversampling method SMOTE (Synthetic Minority Over-sampling Technique) was used.

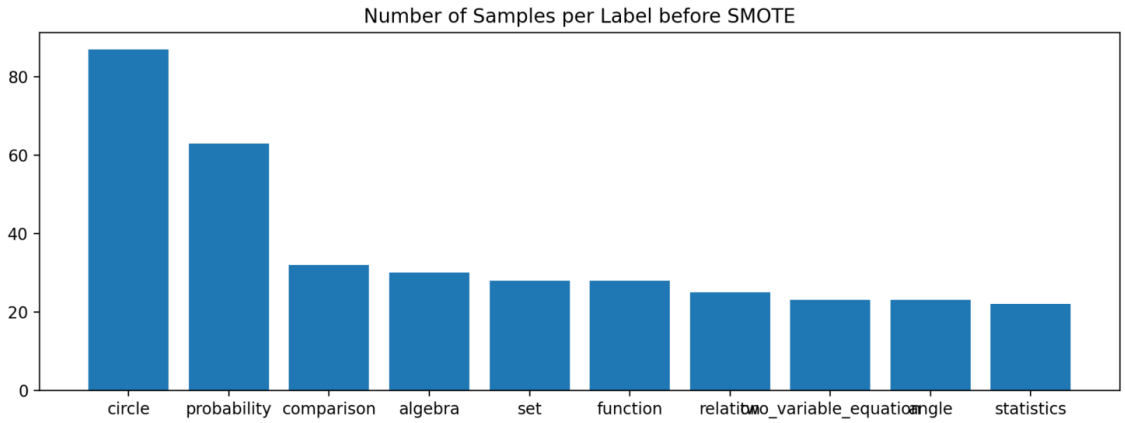


Figure2. Data Distribution Before SMOTE

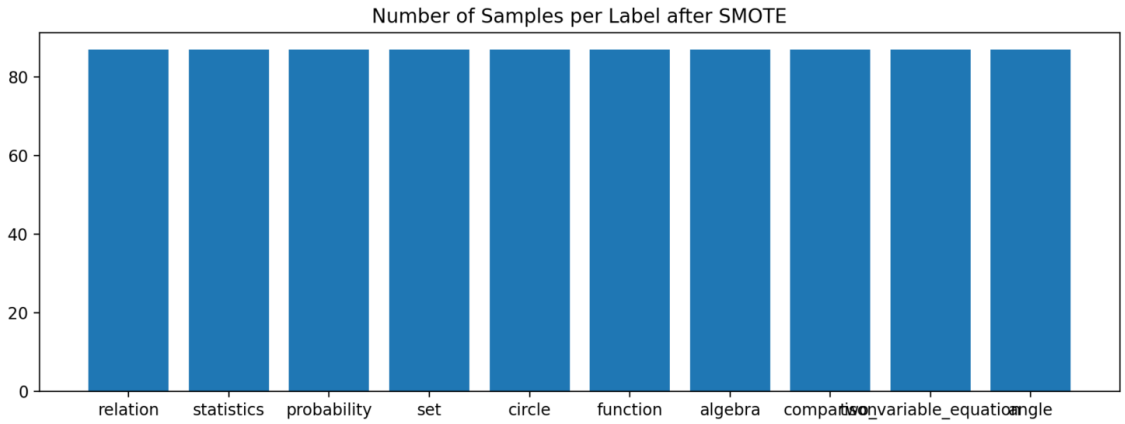


Figure3. Data Distribution After SMOTE

Before classification, the data was split into training and testing sets with a ratio of 80:20, resulting in 361 data for training and 91 data for testing. After SMOTE was applied, the number of samples for each label became more evenly distributed, allowing the model to learn better from the training data. After SMOTE, the total data was 961, with 870 data used for training and 91 for testing.

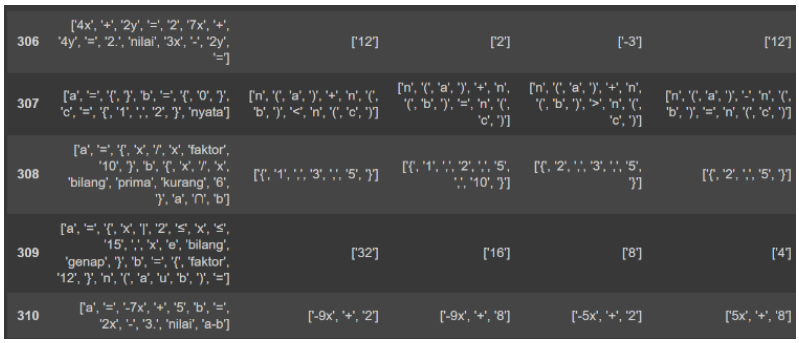
Preprocessing math symbols

In this classification process, a special approach is needed for data that contains many numbers, such as in Mathematics problems. The goal is to determine whether the model can learn from the symbols in the data. The stemming process in this research uses the Literature library. In the first scenario, stemming is performed directly on all tokens using a stemmer. The output is that all words are converted into base words, and tokens in the form of symbols or operators become empty tokens because Sastrawi cannot process them correctly.

306	['4x', '', '2y', '', '2', '7x', '', '4y', '', '2', 'nilai', '3x', '-', '2y', '']	['12']	['2']	['-3']	['12']
307	['a', '', '', '', 'b', '', '', '0', '', 'c', '', '', '1', '', '2', '', 'nyata']	['n', '', 'a', '', '', 'n', '', 'b', '', '', 'n', '', 'c', '']	['n', '', 'a', '', '', 'n', '', 'b', '', '', 'n', '', 'c', '']	['n', '', 'a', '', '', 'n', '', 'b', '', '', 'n', '', 'c', '']	['n', '', 'a', '', '-', 'n', '', 'b', '', '', 'n', '', 'c', '']
308	['a', '', '', 'x', '', 'x', 'faktor', '10', '', 'b', '', 'x', '', 'x', 'bilang', 'prima', 'kurang', '6', '', 'a', '', 'b']	['', '1', '', '3', '', '5', '']	['', '1', '', '2', '', '5', '', '10', '']	['', '2', '', '3', '', '5', '']	['', '2', '', '5', '']
309	['a', '', '', 'x', '', '2', '', 'x', '', '15', '', 'x', 'e', 'bilang', 'genap', '', 'b', '', '', 'faktor', '12', '', 'n', '', 'a', 'u', 'b', '', '']	['32']	['16']	['8']	['4']
310	['a', '', '-7x', '', '5', 'b', '', '2x', '-', '3', 'nilai', 'a-b']	['-9x', '', '2']	['-9x', '', '8']	['-5x', '', '2']	['5x', '', '8']

Figure4. Stemming Without Symbols

In the second scenario, a special function is created to process the tokens by involving Regular Expressions (regex). This function checks each token; if the token is a symbol, number, or operator, it is not changed, but if the token is plain text, it will be changed using a stemmer. The output of this scenario is that tokens in the form of text are converted into basic words while tokens in the form of numbers, symbols, or operators such as ∏ α β ° will still be included.



306	['4x', '+', '2y', '=', '2', '7x', '+', '4y', '=', '2.', 'nilai', '3x', '-', '2y', '=']	['12']	['2']	['-3']	['12']
307	['a', '=', '{', '}', 'b', '=', '{', '0', '}', 'c', '=', '{', '1', ',', '2', '}', 'nyata']	['n', '(', 'a', ')', '+', 'n', '(', 'b', ')', '<', 'n', '(', 'c', ')']	['n', '(', 'a', ')', '+', 'n', '(', 'b', ')', '=', 'n', '(', 'c', ')']	['n', '(', 'a', ')', '+', 'n', '(', 'b', ')', '>', 'n', '(', 'c', ')']	['n', '(', 'a', ')', '-', 'n', '(', 'b', ')', '=', 'n', '(', 'c', ')']
308	['a', '=', '{', 'x', '/', 'x', 'faktor', '10', '}', 'b', '{', 'x', '/', 'x', 'bilang', 'prima', 'kurang', '6', '}', 'a', 'n', 'b']	['{', '1', ',', '3', ',', '5', '}']	['{', '1', ',', '2', ',', '5', ',', '10', '}']	['{', '2', ',', '3', ',', '5', '}']	['{', '2', ',', '5', '}']
309	['a', '=', '{', 'x', '|', '2', '≤', 'x', '≤', '15', ',', 'x', 'e', 'bilang', 'genap', '}', 'b', '=', '{', 'faktor', '12', '}', 'n', '(', 'a', 'u', 'b', ')', '=']	['32']	['16']	['8']	['4']
310	['a', '=', '-7x', '+', '5', 'b', '=', '2x', '-', '3.', 'nilai', 'a-b']	['-9x', '+', '2']	['-9x', '+', '8']	['-5x', '+', '2']	['5x', '+', '8']

Figure5. Stemming With Symbols

Evaluation

The confusion matrix measurement is done for each label in the classification. In this case, ten topics are targeted for classification.

Table 3. Confusion Matrix Results

Topic	TN	FP	FN	TP
algebra	84	2	1	4
function	78	3	2	8
set	84	2	2	3
circle	69	2	1	19
probability	75	0	5	11
comparison	80	6	3	2
two_variable_equation	82	1	4	4
relation	79	4	6	2
statistics	79	4	4	4
angle	81	4	0	6

Calculations such as accuracy, precision, recall, and f-measure can also be calculated for each label individually. Furthermore, the results of these calculations can be averaged to get an idea of the model's overall performance in classifying the data. Here are the overall results:

Table 4. Evaluation Results

	Accuracy	Precision	Recall	F-measure
Result	69%	72%	69%	69%

The results show an accuracy of 69%, precision of 72%, recall of 69%, and f-measure of 69%. Although these metric values indicate a fairly good overall performance, the results are still not optimal. The suboptimal performance is attributed to the unique characteristics of the math problem data, which may be challenging for the model to handle despite the special approach taken during the preprocessing stage.

The next evaluation process used cross-validation with k=10. There are four scenarios in applying this evaluation, namely:

1. Before SMOTE, use data that retains the Math symbols

2. Before SMOTE, use data without Math symbols

3. After SMOTE, use data that retains the Math symbols

4. After SMOTE, use data without Math symbols

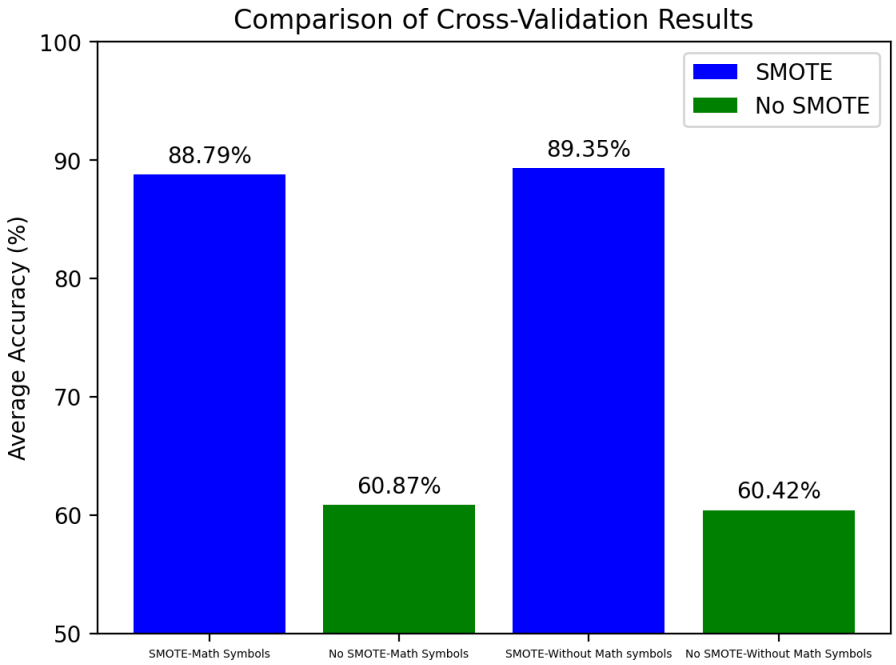


Figure6. Comparison of Cross-Validation Results

Based on the cross-validation results in Figure6, the use of SMOTE and math symbols affects the performance of the model. Classification with SMOTE and without math symbols achieved 89.35% accuracy, while with math symbols, it reached 88.79%. Although there is a slight decrease when using math symbols, SMOTE still effectively balances unbalanced data. On the other hand, removing math symbols helps the model focus on more relevant keyword information, reduces potential confusion caused by math symbols, and improves the model's understanding of the core problem.

Application of the classification model

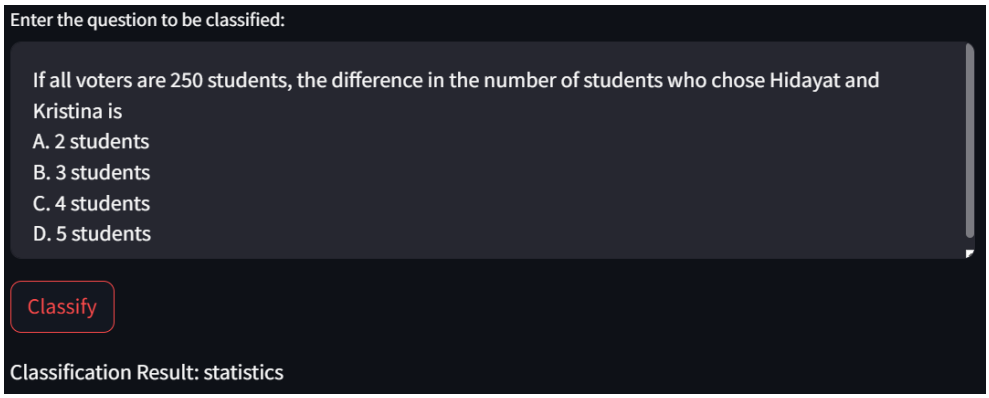


Figure7. Example 1 Model Application

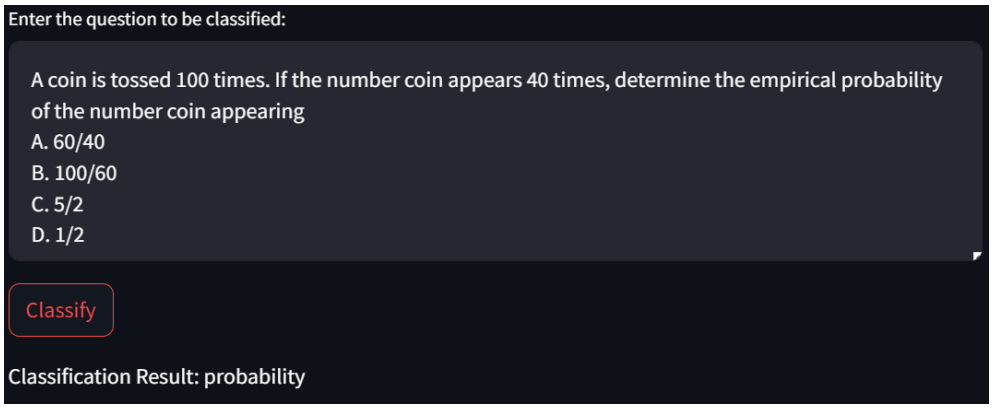


Figure8. Example 2 Model Application

The Naive Bayes model successfully predicts questions with or without the keywords corresponding to their labels. For example, a question labelled statistics can be correctly predicted without the keyword "statistics", and a question labelled probability can be correctly predicted even with the keyword "probability". However, misclassification still occurs due to the inconsistent occurrence of keywords.

D. Conclusion

Naïve Bayes method achieves 69% accuracy in classifying junior high school math problems by topic. The application of SMOTE improved the accuracy from 62% to 69% by addressing data imbalance. Cross-validation revealed a higher performance, with an accuracy of around 89%, regardless of including mathematical symbols, indicating that these symbols have a minimal impact on model accuracy. Removing symbols slightly improved the model's focus on key textual information.

E. Acknowledgment

I would like to thank all relevant parties who have provided support in this research. This journal will be a useful reference for future researchers, especially in classifying mathematics problems. Although this research has many limitations and shortcomings, it can be further developed in subsequent studies.

F. References

[1] Kuswandi, Setiawan, And A. Setiabudi, "Optimalisasi Examview Dalam Pengelolaan Bank Soal Sebagai Upaya Pengembangan Keterampilan Guru Di SMAN 107 Jakarta," In *The 1st LP3I National Conference Of Vocational Business And Technology (LICOVBITECH)* , Nov. 2022.

[2] L. Octiviyani, "Analisis Butir Soal Pilihan Ganda Ujian Akhir Semester Ganjil Mata Pelajaran Sejarah Peminatan Kelas XI IPS Di SMA Negeri 10 Tanjung Jabung Barat Tahun Ajaran 2020/2021," Universitas Jambi, Jambi, 2021.

[3] J. Han, J. Pei, And H. Tong, *Data Mining: Concepts And Techniques*. 2023.

[4] V. A. Silva, I. I. Bittencourt, And J. C. Maldonado, "Automatic Question Classifiers: A Systematic Review," *IEEE Transactions On Learning Technologies*, Vol. 12, No. 4, Pp. 485–502, Oct. 2019, Doi: 10.1109/TLT.2018.2878447.

[5] L. A. Muhaimin, O. N. Pratiwi, And R. Y. Fa'rifah, "Klasifikasi Soal Berdasarkan Kategori Topik Menggunakan Metode Algoritma Naïve Bayes Dan Algoritma C4.5," *E-Proceeding Of Engineering* , Vol. 10, No. 2, Apr. 2023.

[6] T. B. Lalitha And P. S. Sreeja, "Personalised Self-Directed Learning Recommendation System," In *Procedia Computer Science*, Elsevier B.V., 2020, Pp. 583–592. Doi: 10.1016/J.Procs.2020.04.063.

[7] A. N. Liyantoko, I. Candradewi, And A. Harjoko, "Klasifikasi Sel Darah Putih Dan Sel Limfoblas Menggunakan Metode Multilayer Perceptron Backpropagation," *IJEIS (Indonesian Journal Of Electronics And Instrumentation Systems)*, Vol. 9, No. 2, 2019, Doi: 10.22146/Ijeis.49943.

[8] Emrald And K. M. Lhaksmana, "Klasifikasi Kategori Hadits Menggunakan Naive Bayes Classifier," *E-Proceeding Of Engineering*, Vol. 6, No. 2, 2019.

[9] N. Çolakoğlu And B. Akkaya, "Comparison Of Multi-Class Classification Algorithms On Early Diagnosis Of Heart Diseases," *Y-BIS 2019 Conference Book: Recent Advances In Data Science And Business Analytics*, No. March, 2019.

[10] L. K. Foo, S. L. Chua, And N. Ibrahim, "Attribute Weighted Naïve Bayes Classifier," *Computers, Materials And Continua*, Vol. 71, No. 1, 2022, Doi: 10.32604/Cmc.2022.022011.

[11] A. S. Andini, D. T. Murdiansyah, And K. M. Lhaksmana, "Topic Classification Of Islamic Question And Answer Using Naïve Bayes And TF-IDF Method," *Computer Engineering And Applications Journal*, Vol. 10, No. 3, 2021, Doi: 10.18495/Comengapp.V10i3.385.

[12] B. Bramantyo, M. P. K. Putra, And N. Hendrastuty, "Implementasi Recurrent Neural Network Pada Multiclass Text Classification Judul Berita," *Jurnal Media Borneo*, Vol. 1, No. 1, 2023, Doi: 10.58602/Mediaborneo.V1i1.6.

[13] H. Balla, M. L. Salvador, And S. J. Delany, "Arabic Question Classification Using Deep Learning," In *ACM International Conference Proceeding Series*, 2022. Doi: 10.1145/3562007.3562024.

[14] R. Melita, V. Amrizal, H. B. Suseno, And T. Dirjam, "Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis

Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *Jurnal Teknik Informatika*, vol. 11, no. 2, Nov. 2018, doi: 10.15408/jti.v11i2.8623.

[15] A. Afif, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus di Rumah Sakit Aisyiah," *Jurnal Ilmu Komputer dan Matematika*, vol. 1, no. 2, 2020.

[16] R. B. Widodo, *Machine Learning Metode K-Nearest Neighbors-Klasifikasi Angka Bahasa Isyarat*. Malang: Media Nusa Creative, 2022.

[17] S. Verma and C. Prakash, "IOT Security Against Network Anomalies through Ensemble of Classifiers Approach," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7 s, 2023, doi: 10.17762/ijritcc.v11i7s.6976.