Augma Bot: An AI Chatbot Consultant for Higher Education Degree Programs

Edbert Felix Fangasasdha

*Computer Science Department*

*School of Computer Science*

*Bina Nusantara University*

*Jakarta, Indonesia 11480*

edbert.fangasadha@binus.ac.id

Steffi Soeroredjo

*Computer Science Department*

*School of Computer Science*

*Bina Nusantara University*

*Jakarta, Indonesia 11480*

steffi003@binus.ac.id

*Abstract*—*Enrolling in a degree program with the hopes of gaining relevant knowledge in certain subjects of study that could later be applied in the workplace is a typical way to boost one’s chances of success. However, because there are so many different variables to ponder over, the process of selecting the right major can be tedious and frustrating. To address this issue, we propose the development of a chatbot that uses Natural Language Processing as a center of the approach. This chatbot is called Augma Bot, serving as a higher degree program consultant, that uses the multinomial naive Bayes classifier and the law of large numbers as the baseline. The dataset utilized is compilation of  responses to a survey that included questions on the respondents' favorite subjects and subjects they excel in. This experiment began with managing datasets utilizing the law of large numbers to optimize the dataset's potential and grouping existing data using Multinomial Naive Bayes. After the dataset has been processed and the model has been trained, the model will examine the accuracy and precision scores to determine the model's classification accuracy*

Keywords—Chatbot, AI, Degree Program Consultant, Multinomial Naive Bayes, Law of Large Number, Natural Language Processing

# Introduction

The advancement of artificial intelligence is exceeding people's expectations, and chatbots are increasingly expressing their importance in many sectors of life. Virtual assistants, often known as chatbots, are software agents that communicate with users using natural language processing [2]. Chatbots are frequently used to either settle common minor issues through automated messages responses that are normally handled by customer service and consultants ranging from educational advisors to medical professionals in a speedy manner or for entertainments purposes. All chatbots have their own area of expertise, but they all work in a similar manner: one human input is matched against the chatbots’s cache data [3]. Knowing the significance of artificial intelligence which chatbots are based on, we chose to contribute to education informatics by utilizing this potential.

Deciding on an appropriate major for oneself is often the source of dilemma for those who are considering a degree program to pursue. Aside from likes and dislikes, other considerations such as family background or work opportunities can have a significant impact on a persons’s final decision. If they feel as if they have chosen the wrong major for whatever reason, it won’t be as simple as quitting or switching majors as there are other serious ramifications, including squandering their time and money [4], it will result in a 68% increase in student loans for both private and public university graduates [5].

The key purpose of our project is to assist in determining which majors are appropriate for the individuals that are concerned in selecting a major by focusing on learning what areas they are interested  and excel in in order to reduce the likelihood of them switching degrees later on [5].Furthermore, users will be given a reality check connected to the majors in order to avoid misconceptions, as polls suggest that a lot of undergraduates were taken aback by materials presented in the majors they had chosen [6]. For example, we are frequently informed that computer science is a major that does not require much mathematical talents because we simply need to focus on codes. While in reality, mathematics is actually the foundation of computer science [7].

The classification method used for constructing Augma Bot, a chatbot designed to recommend relevant degree programs based on the correlation between subjects that the user enjoys and excels in, in this project is Multinomial Naive Bayes. The Naive Bayes learning algorithm is commonly used to solve text categorization problems [8]. Multinomial Naive Bayes (MNB) thinks of the document as a jumble of words and include word frequency as well as information into account [9]. After calculating the word frequencies, MNB will calculate the probability of the word appearing, which is what causes issues in the present dataset. Therefore, the dataset will be upgraded in accordance with the law of large numbers to increase the accuracy of the model that has been constructed. The fundamental notion of the law of large numbers will then aid this endeavor in resolving the dataset’s limitations [10].

The rest of the paper is organized as follows: Section II examines at previous works related to Augma Bot, which may not be in the same field as the degree program chatbot consultant we constructed, but includes the most pertinent method instances. Section III illustrates the methods used to build Augma Bot; and preparations to integrate the datasets. Section IV will cover the implementation and the end result to the experiments. And finally Section V concludes the paper.

# Related Work

## Multinomial naive bayes for text categorization revisited. (A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, 2004)

The MNB Classifier, along with the four existing problems alongside it, are the main topics of study in this paper. There is also a discussion about the distinctions between regular MNB and MNB which has been given weighted learning, also known as Transformed Weight-normalized Complement Naive Bayes (TWCNB) classifier included. TWCNB will not always be more accurate than the traditional MNB, according to the text. Furthermore, Support Vector Machine (SVM) could even outperform TWCNB when it comes to classifying data in a variety of datasets [8].

## Multinomial Naive Bayes Classification Model for Sentiment Analysis (M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, 2019)

In this research study, the MNB classifier is utilized to categorize articles, particularly in Sentiment Analysis (SA). The experiments were conducted on a dataset of movie reviews that yielded a satisfaction scalar ranging from 1-100. The methods utilized in this study attained a classification accuracy of 90%, and they are ready to be used as well as offered to firms or individuals who are interested in employing them [9].

## Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot (M. Y. H. Setyawan, R. M. Awangga, and S. R. Efendi, 2018)

In this research study, the MNB classifier is utilized to categorize articles, particularly in Sentiment Analysis (SA). The experiments were conducted on a dataset of movie reviews that yielded a satisfaction scalar ranging from 1-100. The methods utilized in this study attained a classification accuracy of 90%, and they are ready to be used as well as offered to firms or individuals who are interested in employing them [9].

## Health Care Counselling Via Voicebot Using Multinomial Naive Bayes Algorithm. (S. Revathy, 2020)

A chatbot was created in this study to serve as a consultant for common health conditions like fever and other complications. The chatbot will converse with the user via text and speech, allowing the user to type inputs in text format that will be voiced out by the chatbot. To improve the chatbot's accuracy, these incoming inputs will be parsed using MNB. The voicebot will perform the following steps: import the yml file (Ain't Markup Language Format), ask the user to input and validate the input, tokenize and stem the input, predict using previously generated MNB models, compare the output from the models to the yml file, and output the result [12].

## Nonlinear Mathematics for Uncertainty and its Applications ( S. Li, X. Wang, Y. Okazaki, J. Kawabe, T. Murofushi, and L. Guan, eds., 2011)

This book is a compilation of papers that have been combined to provide a comprehensive explanation of uncertainty, or probability, with in-depth explanations and extensive examples. However, because the law of large numbers is the topic of this study chatbot consultation, we will primarily focus on the paper inside this book titled "Strong Laws of Large Numbers for Bernoulli Experiments under Ambiguity." The laws of large numbers focus on the theorem which states that the more a simulation is performed, the closer the result of the simulation is to what is expected [10].

# Method

## System Design

Our dataset was obtained through a survey of more than 115 people and the list of questions are as follows:

TABLE I

SURVEY QUESTION

|  |  |  |
| --- | --- | --- |
| **Questions given** | **Questions translated to English** | **Reasons of why the questions were given** |
| Nama | Name | Identity |
| Umur | Age | Identity |
| Mata pelajaran apa yang anda sukai semasa SMA/SMK? | What subjects during highschool did you like? | Used in the MNB Model |
| Mata pelajaran apa yang anda kuasai semasa SMA/SMK? | What subjects during highschool did you excel in? | Used in the MNB Model |
| Jurusan manakah yang menarik perhatian anda atau anda minati? (Kuliah) | Which degree program interests you or you enjoy? (University/College) | Used in the MNB Model |
| Jurusan apa yang anda putuskan untuk masuk? (Kuliah) | Which degree program did you end up pursuing? (University/College) | Used for data analyzation |
| Sesuai dengan mata pelajaran yang anda kuasai dan sukai, jurusan apa yang menurut anda cocok dengan anda? (Kuliah) | According to the subjects you mastered and liked, which major do you think is best suited for you? (University/College) | Used for data analyzation |
| Sesuai dengan mata pelajaran yang dikuasai, Apakah anda merasa puas dengan pilihan jurusan kalian (saat ini atau yang akan kalian ikuti)? | According to the subjects you mastered, are you satisfied with you choice of major (undergoing or will be entering)? | Used for data analyzation |
| Bolehkah anda memberikan kami alasan atas jawaban diatas? | Can you explain the reason for your answer to the question given above? | Used for data analyzation |

When combining the datasets, there are numerous stages that must be completed. Filtering the data is the first step that has to be taken. The intention of filtering data here is to get eliminate of any existing survey spam. There are four spam surveys with the following contents: empty data, data with unpleasant phrases, and data with periods (Punctuation). This data filtering is done so that the data that is eventually entered into the model is clean, in the sense that it contains no odd data.

After the data has been filtered, the next step is to cluster the data manually. Because the obtainable dataset is somewhat minimal, this is done. The law of large numbers will be used here to enhance the existing dataset. The law can be applied to this dataset since the constructed model generates probability values, allowing for the application of probability theory [10]. The three columns used in the model will be retrieved, and the column side will be disregarded, as shown in Table I. Unimportant aspects of the column linked to the subject will be eliminated, and the words of the typo will be rectified, such as changing "saya sangat meyukai bahasa Ingeris (non-standard word)" to "bahasa Inggris (standard word)." This is done to maximize the MNB model's probability calculation. Because of the dataset's shortcomings, Augma Bot can only suggest six majors: "Science," "Socials," "Art," "Religion," and "Computer." which is the direct translation of  ”MIPA”, “IPS”, “Seni”, “Agama”, and “Komputer”. Then, to improve the accuracy of the model outcomes, merge the columns of these subjects from two to one.

## Model

The model used in Augma Bot is the Multinomial Naive Bayes Classifier. MNB operates by measuring the probability of each word, and the model’s formula is as shown in:



As explained in both section I and the previous paragraphs, MNB is a classifier oriented towards the probability of a word c appearing in document d with c∈C where c must be in a certain class and c^ states the estimation of the class that is correct. From the above, it is found that P(c|d) is a conditional probability and that probability can be derived as follows [1].

2

The calculation of the class estimation is very contingent on the probability of the word appearing in a given class divided by the probability of the word appearing in each existing document [1], as shown by the preceding formula.

However, there are significant flaws in the formula as now. If a new word is added that isn't in the model, for example, the result will change to 0, which isn't feasible in the MNB calculation. As an outcome, the smoothing method will be used in the MNB. The term Frequency (TF), or the amount of words in each document, is multiplied by 1 to ensure that the final probability does not equal 0 [8].

# Results

The Augma Bot experiment was conducted on the same three datasets, with the only difference being how the datasets were handled. By combining the columns of subjects that the respondents liked and mastered in one row, the first dataset is created. The second dataset combines the columns of subjects they liked and mastered by combining the two columns into one, resulting in a longer first column. The third dataset is processed in a similar way to the second dataset, with the exception that it adds 6 rows, which are added to provide improved accuracy using the Law of Large Numbers concept. The dataset will be vectorized and separated into training sets and test sets in order to be included in the model.

The model will be trained with a training set and a test set after the datasets indicated in the techniques section are created, as well as the existence of three datasets. Due to the scarcity of data, the model is trained with training sets and test sets. Here is the dataset's accuracy and precision after the model has been trained.

TABLE II

DATASET ACCURACY

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dataset 1** | **Dataset 2** | **Dataset 3** |
| **Training Set** | 31.11% | 32.95% | 31.52% |
| **Test Set** | 54.35% | 63.64% | 75% |
| **Average Accuracy** | 42.73 | 48.295 | 53.26 |

TABLE III

DATASET PRECISION ON TRAINING SET

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dataset 1** | **Dataset 2** | **Dataset 3** |
| **Agama** | 0% | 0% | 100% |
| **IPS** | 14.75% | 25% | 21.21% |
| **Komputer** | 5.13% | 0% | 0% |
| **Literatur** | 0% | 0% | 0% |
| **MIPA** | 95.65% | 75% | 79.17% |
| **Seni** | 4.35% | 50% | 25% |

TABLE IV

DATASET PRECISION ON TEST SET

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Dataset 1** | **Dataset 2** | **Dataset 3** |
| **Agama** | N/A | N/A | 0% |
| **IPS** | 35.71% | 40% | 80% |
| **Komputer** | 12.5% | 0% | 0% |
| **Literatur** | 0% | 0% | N/A |
| **MIPA** | 100% | 87.5% | 100% |
| **Seni** | 0% | 83.33% | 100% |

Table II shows that dataset 1's accuracy is significantly lower than that of datasets 2 and 3. This is because the subjects that are liked and mastered could be very different, such as the respondent liked the subject "Sports" and excelled in  "Math," yet majoring in "social studies." Dataset 3 is superior than Dataset 2 since it has an additional 6 rows of data that can assist the model in determining the existing major. Although dataset 3's accuracy in the training set is lower than dataset 2, the test set's accuracy differs considerably enough that these flaws can be overlooked. Additionally, because the difference between datasets 3 and 2 is only 1.43 percent, it can be ignored.

Tables III and IV illustrate that the precision of dataset 1 is rather high only among mathematics majors. This is due to the fact that dataset 1 has twice as much MIPA data as datasets 2 and 3. Dataset 3 has more precision than datasets 1 and 2, yet there are some areas where other datasets stand out more than dataset 3. This is because dataset 3's accuracy is higher, thus other datasets aren't used.

# Conclusion

The average accuracy of the training set and test set with dataset 3 is 53.26 percent, indicating that MNB is sufficient in classifying these data, according to the findings of the experiment above. The important takeaway from this experiment is that the dataset is quite limited that the model's accuracy is still insufficient. However, with a larger and more thorough dataset, experiments can be carried out for further improvements.

##### References

1. Sergios Theodoridis. (2015). Machine Learning, A Bayesian and Optimization Perspective. First. Jonathan Simpson. India. ISBN: 9780128015223.
2. T. T. Nguyen, A. D. Le, and  H. T. Hoang,  NEU-chatbot: Chatbot for admission of National Economics University. Computers and Education: Artificial Intelligence 2. (2021): 100036
3. S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari, and S. Kalpana Devi. A self-diagnosis medical chatbot using artificial intelligence. Journal of Web Development and Web Designing 3, no. 1 (2018): 1-7.
4. W. T. Nelson, and R. Yazdanifard. (2021). A Study of Malaysian High School Leavers’ Dilemma in Choosing Varsity Courses.
5. Z. Friedman, "Student loan debt statistics in 2018: A $1.5 trillion crisis." *Forbes* (2018).
6. S. Carrell, L. Lusher, and D. Rury, Major Disappointment: A Large Scale Experiment on (Non-) Pecuniary Information and College Major. Working Paper, 2020.
7. T. Beaubouef, Why computer science students need math. ACM SIGCSE Bulletin 34, no. 4 (2002): 57-59.
8. A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes. Multinomial naive bayes for text categorization revisited. In Australasian Joint Conference on Artificial Intelligence, pp. 488-499. Springer, Berlin, Heidelberg, 2004.
9. M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed. Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur 19, no. 3 (2019): 62.
10. S. Li, X. Wang, Y. Okazaki, J. Kawabe, T. Murofushi, and L. Guan, eds., Nonlinear Mathematics for Uncertainty and its Applications. Vol. 100. Springer Science & Business Media, 2011.
11. M. Y. H. Setyawan, R. M. Awangga, and S. R. Efendi. Comparison of multinomial naive Bayes algorithm and logistic regression for intent classification in chatbot. In 2018 International Conference on Applied Engineering (ICAE), pp. 1-5. IEEE, 2018.
12. S. Revathy,. Health Care Counselling Via Voicebot Using Multinomial Naive Bayes Algorithm. In 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 1063-1067. IEEE, 2020.