International Journal of Application on Sciences, Technology and Engineering
(IJASTE)
Volume 1, Issue 2,2023.ISSN:2987-2499

Modified K-Nearest Neighbor Web Based Chatbot as Service and Academic Information School

Natanael[1, a)] and Viny Christanti Mawardi[2, b)]

[1,2]Informatics Engineering Department Faculty of Information Technology, Tarumanagara University, Jl. Lt. Jend. S. Parman No. 1 West Jakarta 11440, Indonesia

[a)] Corresponding author: *natanael.535180006@stu.untar.ac.id*

[b)]*vinym@fti.untar.ac.id*

Submitted: January-February 2023, Revised: March 22 2023, Accepted: May 23, 2023

Abstract.

At a time when everything is done on an Online basis starting from office activities, teaching and learning activities especially administrative activities school. Where in the activity, an informant needs to be can serve academic and non-academic information needs for para parents of students and other parties. Chatbot is one of the means to overcome these needs by presenting features that allows for someone to ask when and where. Therefore, in the conclusion that chatbot can be a public means especially on schools - schools that can implement their uses as one school information sources other than the school's business. This Chatbot made using the Modified K-Nearest Neighbor method or can be briefed MKNN. MKNN is the latest version of the previous method, K-Nearest Neighbor where in the application MKNN method does the process additional after Euclidean Distance calculation, Weight Voting. Premise from the weight voting method itself is the biggest weight calculation, so chatbot will calculate the value of the shortest distance to the database from the chatbot then after completion, the chatbot will calculate the value weight or information from that distance so that the results that can be from this are the shortest distance to the data class that is being addressed by carrying the weight of information biggest or most. This is used to facilitate the deep chatbot determine the category of a database so that it can provide answers that more certain and accurate.

Keywords: Chatbot, Modified K-Nearest Neighbor, Euclidean Distance, Weight Voting, information

INTRODUCTION

Information is something that very important for most people. At this day, information is very easy to be access or obtain depends on what platform that we use to get it. Some of those platforms called Chatbot. Chatbot is a chatting system where the system itself is the one who will be the interlocutor. Unlike any other chatting system, chatbot usually only giving answer based on data that has been prepared for them. There are two method that usually used in chatbot development, one is called Retrieval Based Chatbot is the example of the statement above, Retrieval Based is a chatbot method that allow chatbot to giving answer based on the pattern that has been set on the database. And the other method called Generative Based, where the chatbot can automatically generate there on answer pattern [1].

But because of the current situation, some of that information become a little hard to get. The spread of the Covid-19 virus in the world so far has given many impacts on human survival such as economics, health especially education [2]. Some schools in Indonesia require their students and educators to carry out their school activities like classes, exam from home or in other words online learning [3]. This learning system has been applied to all schools in Indonesia as one of the one way for the Indonesian

government to continue to carry out educational activities in each country while breaking the chain of the spread of the Covid-19 virus in the school environment, but that's where new problems came out. Because some of school activities are forcedly to go online by their school, activity like Dissemination of information such as school events, administrative activities are also required to go fully online where before the pandemic this information can be required by seeing school billboard or ask directly ask the administration but because of the pandemic, these school billboards became unused because the students itself is not there to use it or even see it. On the other hand, the students and their parents usually ask the administration directly to the school office or ask them on the phone, these way of getting information is not affected by the pandemic because even before the pandemic itself most of them already did that. But even after doing it for a long time, there are still problems on it.

These problems very relate to one of the most impactful things to human lives, Time. Time can be used to control most of human activity, coordinate it so that these activities can run regularly or right on the schedule. But the main problem is the time that is owned by one individual is not necessarily the same with other individuals in the sense that everyone is believed to have their respective affairs or interests that make the time that they owned is very much limited by it. For example, there are condition where some administration can't answer or didn't have time to check their phone to see the question that given by students or their parents. Because of that, the person who ask for that school information must wait until the administration answer it, maybe a few hours or maybe tomorrow but on the other hand, that information is needed ASAP.

There also a condition where some people didn't have a time to contact the school administration and just want all that information can be access by themselves. In this age, everything is connected through an internet. People can access or search whatever they need by using the internet. There is also a platform that can be used to access that different kind of information like Google or the platform that has a combination between giving information's and enjoyment like TikTok, Instagram, Facebook and many more. For example, something like school website where students, their parents or any other people can access to see the information about that specific school like news about some school activity that has been going on right now, or something like new student admission that kind of information can be obtained. That kind of platform is still rare, by it mean rare is not every school has it or not every school has a chance to create their own website so some of the information can be obtained by that platform.

Based on that explanation, it can be concluded that it is necessary to have a solution for this kind of problem so the students and their parents can access or ask a question about the things that they want too and is not limited to get it like the information can only can be get from one person or from a specific one place and the most important thing it has no limit of time. Which why based on what is needed from the previous sentence the right solution for this is by using a Chatbot. Chatbot is a service system where an individual human can interact with the computer via a chat interface. Chatbot will respond conversation in a conversational style and can do something the action of the conversation [4]. Chatbot is also not limited by time or place, anyone can access or use it anytime anywhere. But for this can be used for public the chatbot that we build here are Web Based Chatbot, which is a chatbot that integrated on a website and that chatbot will be one of the features in that website including the other feature of menu that is necessary for a school website.

Modified K-Nearest Neighbor or what can be shortened as MKNN is a modification version of K-Nearest Neighbor. Overall, the way MKNN works is the same as its previous version where the point is to find a shortest path from the data that has been inputted to all existing data classes on the datasets but also calculate their weight of information using Weight Voting. Weight Voting is a method to calculate how much weight of information that carry out by a single data to one another

or to the one that closes to that specific data [5]. In the process, the designed Chatbot will do several process before entering into the use of the modified K-NN algorithm which starting from the tokenization process, removing stop words, and solving affixes from the sentence that was typed to him from the user's side and then did the calculation of each occurrence of a word or term from the sentence input it to get the weight of each word in the sentence then after the above processes will be calculated using the Modified K-Nearest Neighbor Algorithm.

There are some themes about chatbot that has been raised and created like for one a paper about "Application of Technology in Virtual Conversation as Learning facilities at Immanuel Elementary School" that were made by one of the Tarumanagara University students named Jesslyn along with two lecturers from the Information Technology faculty of the Informatics Engineering study program named Viny Christanti Mawardi, M.Kom. and Janson Hendryli, S. Kom., M. Kom. The chatbot was built by implementing artificial intelligence for the process of conversation between humans and the computers. By implementing artificial intelligence, chatbot can serve and answer the curiosity of students about elementary school subject matter at any time without any limitations time by implementing artificial intelligence (Artificial Intelligence) which can simulate the process of conversation between man with computer via text chat [6].

Another one is a paper from Informatics Study Program from Faculty of Engineering, University of Muhammadiyah where they build a chatbot for school information center using Artificial Intelligence Markup Language (AIML). AIML is used in simplification of forms complex grammar into simpler forms, division sentences into sub-sentences, word equations, spelling corrections, and grammar language [7]. There also a chatbot that build by one of the Tarumanagara University students named Andrew Ciayandi with the same lectures as Jesslyn before. In here the chatbot created by applying the Multilayer Perceptron on Retrieval Based Chatbot Method to build their chatbot. The Retrieval Based method is used so that the chatbot can give the right answer to the user so it doesn't a condition occurs where the answer given is irrelevant to why that was asked. In addition, the use of Multilayer Perceptron in use it to build a chatbot that can also study the architecture of a word or sentence to provide greater accuracy better at taking answers to user questions [8].

METHODS

Chatbot

Chatbot is a virtual conversation that is carried out between humans and robots with the aim of conveying information by questions and complaints to get the respond from the robots quickly and accurately [9]. When a user uses the chatbot, the conversation that occurs is a conversation between human as a user and a computer or robot. That said the one who respond or replies the user is not a fellow human anymore but a system, a computer at this point. There are two types of Chatbots.

1. Chatbot that can only give answer based on the output answer that their sentence pattern has been set before. This type of chatbot called Retrieval Based Chatbot.
2. Chatbot that can generate their own sentence for the output answer by doing learning process of grammar and how to make its own sentence pattern. This type of chatbot called Generative Based Chatbot.

Preprocessing

Text preprocessing is a process for filtering a raw text data to become more structured by going through a series of stages like,

1. Case Folding

   Case Folding is a process of changing capital letters into lowercase. An Example of this method can be seen on **TABLE 1** below.

TABLE 1. CASE FOLDING

Before	After
Siapa wali kelas 1A?	siapa wali kelas 1a?

2. Tokenization

Tokenization is the process of dividing a sentence of text or a document into an individual words. An Example of this method can be seen on **TABLE 2** below.

TABLE 2. TOKENIZATION

Before	After
siapa wali kelas 1a?	‘siapa’, ‘wali’, ‘kelas’, ‘1’, ‘a’,‘?’

3. Stemming

Stemming is the process to know if there’s an affixed words in a sentence and then turn it into its basic words. An Example of this method can be seen on **TABLE 3** below.

TABLE 3. STEMMING

Before	After
‘siapa’, ‘wali’, ‘kelas’, ‘1’, ‘a’,‘?’	‘siapa’, ‘wali’, ‘kelas’, ‘1’, ‘a’,‘?’

4. Stop words Removal

Stop words Removal is the most commonly process that been used in a sentence, text or document to remove their stop words out of them. This process uses a documentation Indonesian type of stop words such as ‘Siapa’, ‘?’, ‘apa’ and many more. An Example of this method can be seen on **TABLE 4** below.

TABLE 4. STOP WORDS REMOVAL

Before	After
‘siapa’, ‘wali’, ‘kelas’, ‘1’, ‘a’,‘?’	‘wali’, ‘kelas’, ‘1’, ‘a’

Confusion Matrix

Confusion matrix is one of classification methods where it measures accuracy of the model based on the table field of predicted and actual values [10]. Thera are four different fields of confusion matrix,

1. True Positive (TP), The Predicted is positive and its positive
2. False Positive (FP), The Predicted is positive and its false
3. False Negative (FN), The Predicted is false and its positive
4. True Negative (TN), The Predicted is false and its false

K-Fold Cross Validation

K-fold Cross Validation is a method to evaluate a model of dataset to finding the parameter that on this case is the K value that giving an accuracy close to the ideal accuracy of all folds that have been done [11]. The description can be seen **FIGURE 1. b**elow.

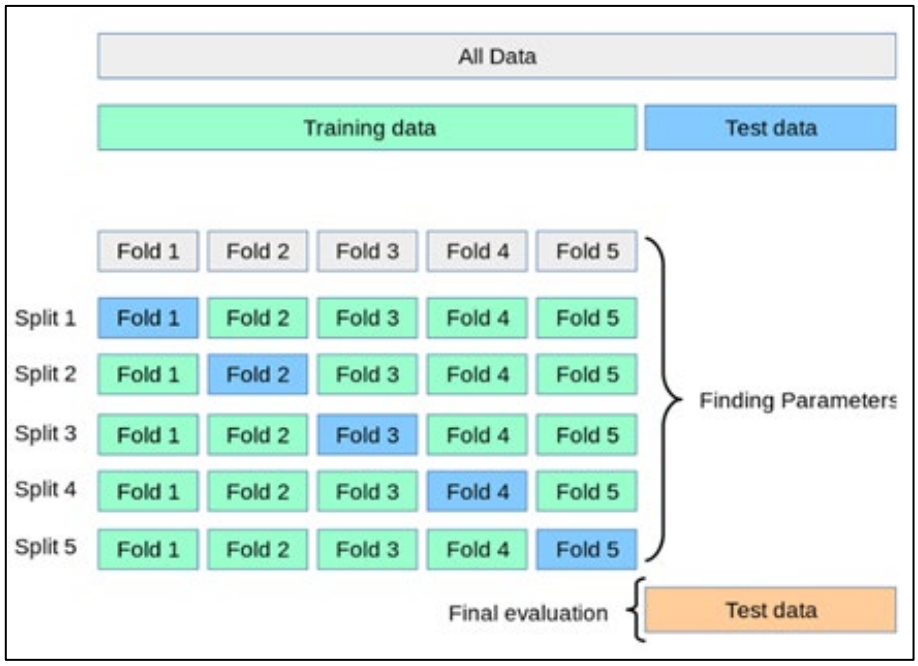

FIGURE 1. K-FOLD CROSS VALIDATION

Splitting the dataset into training and testing data is the first step that had to be done, then fit those training data into the model and calculate throughout the cross validation and the folds that had been decide before. For example, if the folds that decide is in range of 1 until 10 then K-Fold will loop through that range in index start from 0 to 9.

For the data that tested in each fold had to be split depending on how many it wants. An ideal number that can be used in this are 3, 5 and 10. It means that if it decided that each fold will have three data that that's how many that fold will calculate, if five than five data on each fold then if its ten than ten data on each fold.

Modified K-Nearest Neighbor

Modified K-Nearest Neighbor or can be shortened as MKNN is the method that we use in this project development. This method is used to predict and find the similarity between the data input and the available data class by combining Euclidean Distance to find it shortest path and using Weight Voting to calculate the weight of the data. we used K-Fold Cross Validation the find the best K value for MKNN use in the calculation [12].

1. The first step is to define the K value. This K value earn by calculating the data using K-Fold Cross Validation.
2. After getting the best K value, calculate the distance between training data as x and testing data as y using Euclidean distance and sort it to find the shortest path or value between them.

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

After calculating all of the data, sort it from the smallest distance value to largest distance value. The ideal distance is the distance that has smallest value than any other data.

3. And then calculate the validation value of the training data to find the similarity of data classes to one another.

$$validation(x) = \frac{1}{K}\sum_{i=1}^{n} S(label(x), label(Ni(x))) \quad (2)$$

4. The final step is calculating a process called Weight Voting. This process is done by calculate the weight of all data throughout Euclidean distance that has been calculated alongside the value from validation process.

$$W_{(x)} = validation(x) * \frac{1}{d_e + 0.5} \quad (3)$$

RESULT AND DISCUSSION

Scheme

Based on the methods above, the schematic of the process can be seen in **FIGURE 2**.

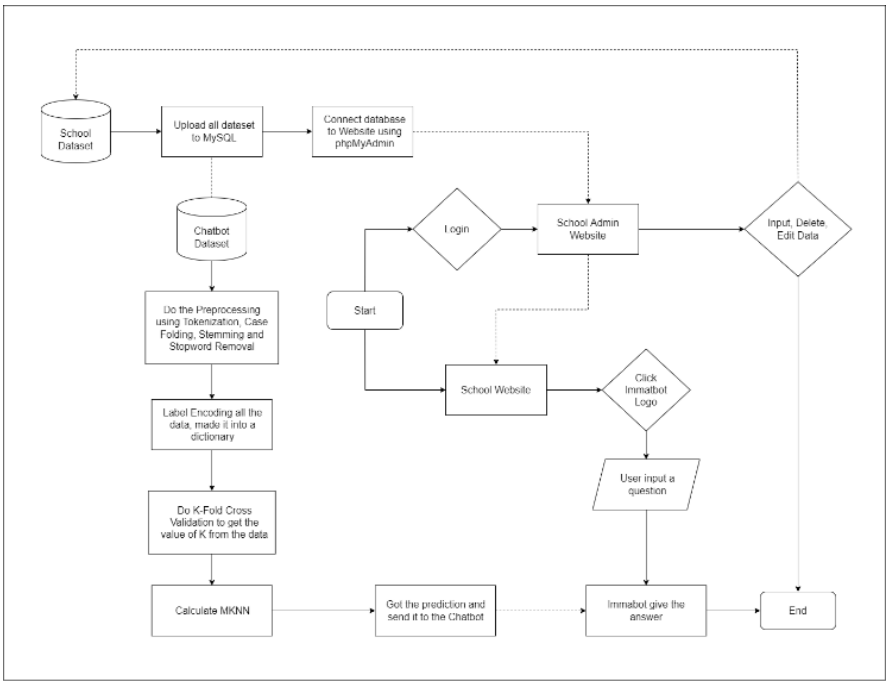

FIGURE 2. System Flowchart

Normalization Result

The dataset that used for the chatbot are based on the database that created beforehand on MySQL. The dataset has three data on it, first data of questions, second data of answers and lastly categories. For the data can be use on the website backend itself, the data is imported to phpMyAdmin server. After that the database will be preprocessed first then do the normalization using Label Encoding to all data on dataset. The example of dataset can be seen on **TABLE 5** below.

TABLE 5. EXAMPLES OF CHATBOT DATASETS

No	Questions	Answers	Categories
1	Siapa wali kelas 1A?	Wali kelas 1A adalah ibu Kartika Irawati	trivia
2	Siapa nama ketua yayasan SDS Immanuel?	Ketua yayasan SDS Immanuel adalah bapak Julius Prakasa Chandra	trivia
3	Apa seragam sekolah untuk hari senin	Seragam hari senin itu Putih Putih	kelas
4	Berapa harga SPP untuk kelas 2 SD?	Untuk Kelas 2 SD, SPP-nya sebesar Rp. 705.000 yang sudah termasuk di dalamnya uang kegiatan Tari, Lukis dan Komputer	administrasi
5	Berapa biaya mutasi untuk kelas SD?	Biaya mutasinya sebesar Rp. 200.000	administrasi
6	Apa seragam sekolah untuk hari jumat	Seragam hari jumat itu Batik Merah	kelas

Category 1 is for “trivia”, category 2 is for “administrasi” and category 3 for “kelas”. From the real dataset we convert the data for normalization.

K-Fold Calculation

After the normalization, the data will be calculated using K-Fold Cross Validation. At first the calculation starts by calculate the ideal mean of the data which is 0.733 then start a looping through the data using how many fold that been set, on this one the data will loop through 10-fold, or it can be said the data will loop 10 times from 1_{st} – 10_{th} loop. After the calculation complete, the result will show what K is the best to use by program and the other result of K in this **TABLE 6.** below.

TABLE 6. K-FOLD CALCULATION

Folds	min	max	mean
1	0.600	0.800	0.733
2	0.400	0.800	0.667
3	0.600	1.000	0.800
4	0.400	1.000	0.733
5	0.600	0.800	0.733
6	0.400	0.800	0.667
7	0.600	0.800	0.667
8	0.400	1.000	0.667
9	0.800	0.800	0.800
10	0.600	0.800	0.733

Calculate Modified K-Nearest Neighbor

After getting the K value, the dataset that has been normalize will be calculated with the input data by user on the chatbot. Then, those input will be normalized using Label Encoding and transform it into a point like (x, y). the x are the input data that has been normalize while the y is a random number that define an unknown category for the input data. this unknown category will be changed right after the MKNN determines the class data that matches for the input data. after the normalization, the input will be calculated with the dataset using Euclidean distance and sort it to find to shortest value or shortest path between the input and the dataset. For Example, the input data is “seragam hari rabu”. that sentence will be normalized into (0, 5) and the result can be seen on **TABLE 7**.

$$dist(x, y) = \sqrt{(0-6)^2 + (5-60)^2} = 55{,}33 \quad (4)$$

"TABLE 7. EUCLIDEAN DISTANCE VALUES"

X, y	Distance
4	4.47
3	5.1
2	6.71
8	7.81
5	9.29
6	15.52
10	46.1
1	55.33
9	88.32
7	114.02

After getting all the distance value, the last step is to calculate the weight of every distance value to search the highest weight between those data then sort it from the highest value to the lowest one after that return the correct category of the input data with a condition like this:

1. if category == 'Trivia' then category += (1/distance value('Trivia') + 0.5) Return 'Trivia'
2. if category == 'Administrasi' then category += (1/distance value('Administrasi') + 0.5) Return 'Administrasi'
3. if category == 'Kelas' then category += (1/distance value('Kelas') + 0.5) Return 'Kelas'

The calculation of weight voting can be seen on Table 8 below.

$$W_{('Kelas')} = \frac{1}{4.47+0.5} = 0.201 \quad (5)$$

TABLE 8. WEIGHT VOTING VALUES

X, y	Weight	Category
4	0.201	3
3	0.178	3
2	0.139	1
8	0.120	2
5	0.102	3
6	0.062	2
10	0.021	3
1	0.018	1
9	0.011	1
7	0.009	2

After the calculation, we need to define which is the highest of all of them by using K value from K-Fold calculation before, the ideal K is 3. So, by that means we take three of the highest weight value to be compare one another to find the highest of those three. The three highest weight value can be seen on TABLE 9. below.

TABLE 9. HIGHEST WEIGHT VALUES

X, y	Weight	Category
4	0.201	3
3	0.178	3
2	0.139	1

From that table, it concluded that the highest weight value is 0.201 and this concluded that the input data before are categorized into data class category "Kelas". With the category of the input data has been found, now the program can continue to find the answer of that input by searching in the database with the condition that chatbot only searching the answer on the class data from the result of MKNN process. The result are "Kelas" then chatbot will only search throughout the "Kelas" category on the database.

Beside the method itself, there are some conditions where the chatbot cannot answer the input even it already being calculated using the method that has been mention before for example, on **FIGURE 3**. This happen because some input has their own sentence structure, and this sentence is less precise than what the chatbot can understand.

FIGURE 3. EXAMPLE OF INCORRECT SENTENCE INPUT

Some of incorrect sentence that can be inputted are:

1. Sentence that using English Language. Because the can only accepted Indonesian Language at this point.
2. Sentence that using slang words on it or the sentence are passive sentence. For now, chatbot can only accept input that use standard pattern.

The correct sentence that can be inputted are:

1. Sentence that using Indonesian Language.
2. Sentence that using standard sentence pattern.
3. Sentence where all the words on it either it's all uppercase or lowercase words
4. Sentence where it just the keyword of the question like on **FIGURE 3.** before.

An example for the correct sentence that chatbot can accept can be seen on **FIGURE 4.**

FIGURE 4. EXAMPLE OF THE CORRECT SENTENCE INPUT

Confusion Matrix Result

The testing process that carried out using Confusion Matrix are using 70% of data training and 30% of data testing from 474 data consisting of Questions data, Answers and Categories. On this term, the model that tested here using that 30% data which around 142 data and the remaining 70% use to train the model. The testing result of this process can be seen **FIGURE 5** and **FIGURE 6.**

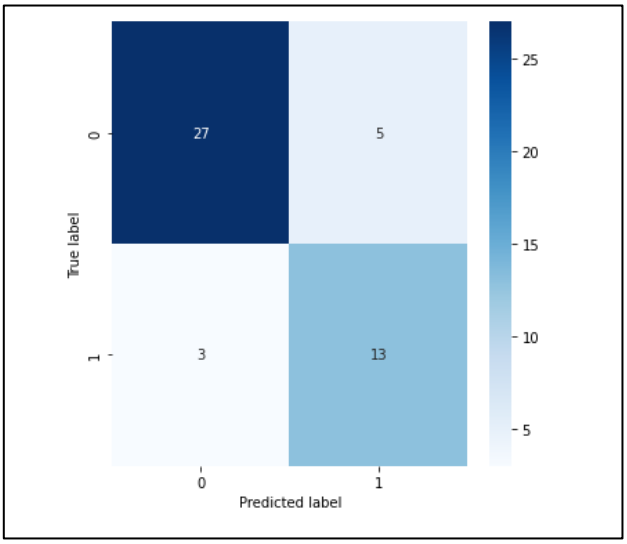

FIGURE 5 CONFUSION MATRIX RESULT (TESTING)

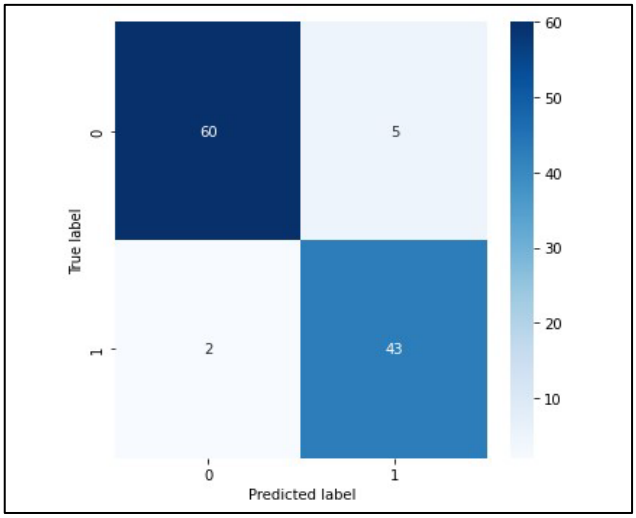

FIGURE 6. CONFUSION MATRIX RESULT (TRAINING)

The website has two designs, one is only can be access by the admin and the other can be access publicly. For the admin, first of all the user that want to access this have to input their username and password. Those username and password only belong to the school teacher and other school staff so public user can't access this site after that, the user can finally access its home screen and use all the feature to input new data for the chatbot or for the school website itself.

As for the school website, it can be access publicly as it says before. The website has 5 menus, the first one is the home menu, second one contains the history of how the school itself can be exist in the first place, the third one shows what vision that the school is chasing on and what mission that the school want to do for their student, the fourth one is where all the news can be seen or access by user and the last menu has location, email or phone number that can be used to contact the school staff if it's needed by them.

These two websites were build using HTML and CSS as their frontend based, as for the backend it used PHP to become the connector between the websites and the databases on phpMyAdmin.

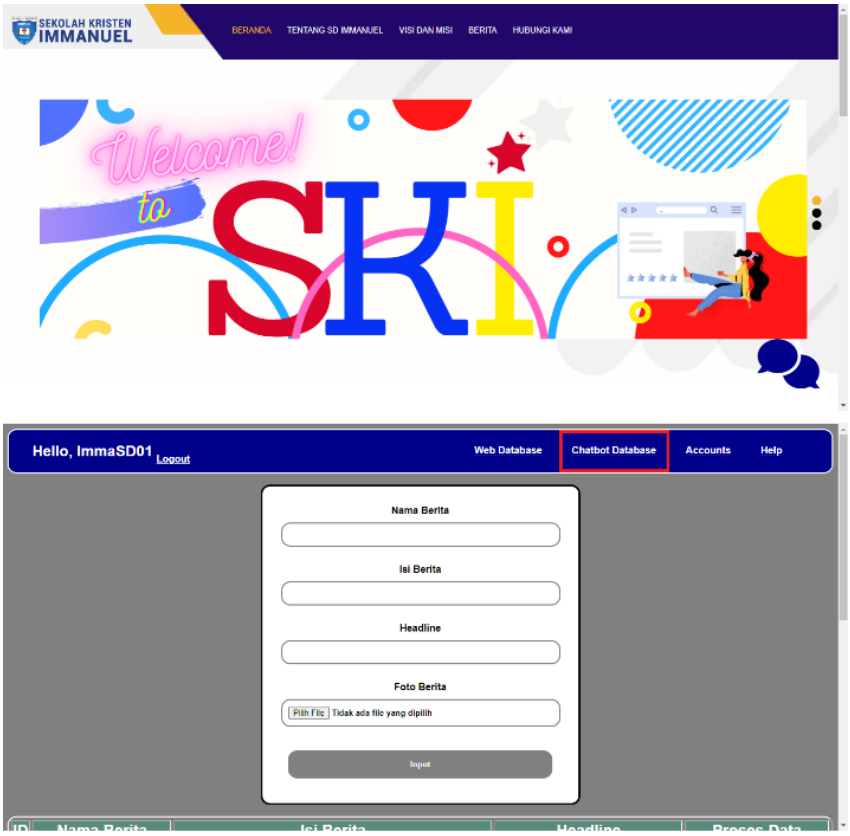

FIGURE. 6. Website Interface (Top Image: School Website Home menu, Bottom Image: Admin Website Home menu)

CONCLUSION

Based on the result of several testing to the chatbot using Modified K-Nearest Neighbor, several conclusions can be taken such as:

- The Confusion Matrix result show almost the same accuracy result as K-Fold Calculation result where the confusion resulting an accuracy on 83.3% success rate to categorize an input using the testing model by calculating all prediction number on every fields.

- From the 83.3% accuracy of the model testing get from K-Fold cross validation, there is still 20% accuracy where chatbot cannot understand what the user input that is include the incorrect sentence pattern that the chatbot itself cannot understand.
- Although the K = 4 seemingly has the closest mean than K = 3 from **TABLE 6.**, but the better accuracy resulted from K-Fold Calculation, the nearest one are K = 3 at 0.800.
- From the Confusion Matrix result, authors knew that there is still a possible situation where the chatbot cannot understand the input given to them because the False Positive and False Negative fields still giving their prediction number beside 0.
- From what has been describe above, the correct sentence has to be inputted in order to get the answer of the question and the chatbot can run the prediction without an error.
- If the incorrect sentence is the one that inputted to the chatbot, the chatbot will tell the user that the question or the sentence that inserted by them is the sentence that the chatbot did not understand or cannot be classify by
- MKNN.
- Even the percentage of the testing model giving a satisfactory percentage at 83.3%, it still can be improved by using other method that matches the theme of the project.

ACKNOWLEDGMENTS

Many thanks to Faculty of Information Technology, Tarumanagara University for the guidance, feedbacks and the corrections that has been given in the preparation of this journal.

REFERENCES

1. Vamsi, G. K., Rasool, A., & Hajela, G. (2020, July). Chatbot: A deep neural network based human to machine conversation model. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

2. Miralay, F. (2020). Evaluation of distance education practice in 2020 Covid 19 pandemic process. Near East University Online Journal of Education, 3(2), 80-86.

3. Lee, Y. C., Yamashita, N., Huang, Y., & Fu, W. (2020, April). " I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

4. Duijst, D. (2017). Can we improve the user experience of chatbots with personalisation. Master's thesis. University of Amsterdam.

5. Parvin, H., Alizadeh, H., & Minaei-Bidgoli, B. (2009, May). Validation Based Modified K-Nearest Neighbor. In AIP Conference Proceedings (Vol. 1127, No. 1, pp. 153-161). American Institute of Physics.

6. Jesslyn, J., Mawardi, V. C., & Hendryli, J. (2021). Penerapan Teknologi Dalam Percakapan Virtual Sebagai Sarana Pembelajaran di Sekolah Dasar Immanuel. PROSIDING SERINA, 1(1), 1479-1488.

7. Ciayandi, A., Mawardi, V. C., & Hendryli, J. (2020, December). Retrieval based chatbot on tarumanagara university with Multilayer Perceptron. In IOP Conference Series: Materials Science and Engineering (Vol. 1007, No. 1, p. 012146). IOP Publishing.

8. Sianipar, Y. P., Mawardi, V. C., & Sutrisno, T. (2022). PENGGUNAAN APRIORI PADA REKOMENDASI PAKET MENU DAN DILENGKAPI FITUR CHATBOT. Jurnal Ilmu Komputer dan Sistem Informasi, 10(1).

9. Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. International journal of remote sensing, 22(16), 3223-3235.

10. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

11. Gazalba, I., & Reza, N. G. I. (2017, November). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. In 2017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE) (pp. 294-298). IEEE.