**AI CHATBOT**

Name: **KISEMBO BANGIRANA RODGERS**

Course name:***Bachelors of Science in Software Engineering***

Reg no.: **20/U/2036/EVE**

Student no.: **2000702036**

Email: [**Kisemborodgers100@gmail.com**](mailto:Kisemborodgers100@gmail.com)

***Abstract***—**In the last few years, there have been a rapid growth in artificial intelligence chatting systems** **example the growing chatGPT, quill Bot, writely.ai, that are capable of answering questions in various fields such as engineering, agriculture, marketing, healthcare and others. This paper develops a Question Answering System in educational domain. The question answering system is a system that can analyze and understand the questions expressed by users in natural language, and has a complete information retrieval function, and answer the questions. [1] Using Word Embedding and Text summarizing method build 3 knowledge bases to provide necessary answers to students on their student systems. Chatbot is a program that processes and simulates human-to-human dialogue through natural language [2]**

***Index Terms*—chatbot, student, information system, artificial intelligence**

1. **INTRODUCTION**

This document is a model for the student Chatbot artificial intelligence system. It will mainly be used by students at the college or university level of education. It would be implemented in the software information systems that these institutions use to provide information and procedures done while at the campus premises. (1) The artificial intelligence will provide answers to the questions provided by the students on the information systems. (2) This will reduce the rate of queues made by students at offices to get information and procedures to follow to solve problem in the university yet they could get these instructions on thus system especially for the new students in the university that are not well informed about the requirements and procedures done in university. Chatbot capabilities can be expanded by including them in other systems such as e-learning systems and library systems. The Chatbot built aims to provide self-learning services as if communicating with the teacher. This study uses a qualitative approach and by conducting a literature review, so that the authors acknowledge that there may be bias in the results of the analysis.[3]

**II. BACKGROUND AND MOTIVATION**

***A.*** ***Selecting a Voice And Requirements; The*** chatbot would be user friendly and in a voice that is preferred by the user in the information system and in case of failure to get needed replies, it would refer the students to the recommended offices for further help from the university staff. The student should already be registered in the institution in order to access the artificial intelligence.

***B.*** ***Maintaining Access to the Chatbot;*** The artificial intelligence systems would be available as long as the person is still a student in the institution and is still in need of the university information systems, and he or she could turn the artificial intelligence chatbot on or off depending on their preferences at different situations.

1. **PROJECT REVIEW**

Before you begin to use the artificial intelligence, it has to be trained by supervised learning training techniques by providing possible questions (input data/ training data) and providing possible answers to the questions given and then using testing data to test whether the artificial intelligence can predict the possible outputs while using linear regression algorithms and also being able to classify and group answers that are related or belong to a certain sector or college. Then it would be able to use algorithms do determine the offices to forward the students to in the case of failure to provide answers to the questions asked. Many previous works discussed the structure of dialogue management and how a computer system can communicate with users with emotion. [4][4][5]

1. *Research Gaps in Project;*

The gaps and weaknesses of these techniques used to implement the artificial intelligence would be limited training data since students may ask questions out of range of the trained data and also failure to understand the different languages that may be used when asking the questions. Another gap would be the occurrence of the information systems blackout used by the university which will sequentially lead to in access to the artificial intelligence chatbot. There would be an occurrence of students using the chatbot for malicious purposes such as cheating during online quizzes and examinations to help answer questions in the examinations.

*B. Summary of term paper contributions*

• The use of unsupervised learning techniques like clustering and association to determine possible answers to the questions and also group the relationships between questions and offices or departments that may handle the same issues to students.

• Avoiding the blackouts of the information systems by the university information and technology staff in the university.

• Limiting the use of the artificial intelligence chatbot by implementing a feature that will automatically disable the artificial intelligence chatbot temporarily as they undertake tests and quizzes on the online information systems.

• Use of a centralized language to ask and answer questions in the information system so as to eradicate the confusion brought about by using different languages that may cause inefficiency of the artificial intelligence.

*C. Equations*

The equations being used here are an example of linear progression equations. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting line (also known as the regression line) that describes the relationship between the variables[6][7].

The regression line is represented by an equation of the form:

y = 0 + 1x1 + 2x2 + ... + nxn

where: y is the dependent variable, x1, x2, ..., xn are the independent variables, 0, 1, 2, ..., n are the coefficients that represent the slope of the line for each independent variable, and n is the number of independent variables.

Linear regression is used to make predictions about the dependent variable based on the values of the independent variables[14]. For example, if we have a set of data that shows the relationship between the amount of time a student studies and their exam score, we can use linear regression to predict the exam score for a student who studies for a given amount of time.

Linear regression is a simple yet powerful technique that is widely used in various fields, including economics, finance, biology, and engineering, to name a few. It is a useful starting point for modeling relationships between variables and can be extended to more complex models as needed [8]

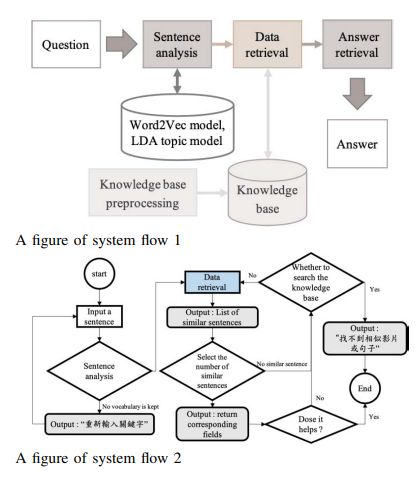
*D. Some Common Mistakes*

• Please use short and precise sentences or questions in the input section of the artificial intelligence example” How to pay functional fees on the portal” which shows exactly what the student needs and is clear.

• Please don’t use the other languages to ask the artificial intelligence as it will fail to interpret the word structure and sequentially will cause inefficiency due to lag in solving problems. • Please note that the use of abbreviations and slang /colloquialisms /informal language will not be understood by the artificial intelligence chatbot for example using” u” instead of” you” or ”hw” instead of” how” • In your questions, it is important to add” how to” at the start of the questions so as to have easy access to the answer with relative examples, links to follow and offices to visit in case of further assistance is required. • The artificial intelligence can’t read your mind and also cannot determine the required need of help without asking it questions.

1. **METHODOLOGY**

The problem being solved it the provision of quick solutions to students without the need of making queues in offices for procedures. The automatic conversation system is an intelligent human machine interaction using natural language. This chatbot receives questions from students, tries to understand the question and provides appropriate answers. The chatbot will be fed with data to enable it answer questions and have a stateful service, remembering previous commands(conversations) in order to provide functionality. The significance of this project is the quick availability of resources to students, including the ones that maybe physically impaired or disabled by making it effortless thus reducing the ignorance levels significantly in university especially amongst new students. The project scope is defined by the objectives, deliverables, milestones, team, and time-frame. Therefore, its objectives, as seen has been cleared stated and is to provide swift access to procedures and information; deliverables are the capabilities of the artificial intelligence such as proving answers with link and examples. The methodology used here is the supervised learning method that will integrate the linear regression algorithms to predict solutions to users by having a set of training data and their corresponding outputs and then using their relationships to predict other various answers that may be related or not. Artificial intelligence evaluation frameworks used in the project may include ScikitLearn to be able to create visual graphics and illustrations that show analysis and reports graphically.



1. ***Dataset Description;***

The dataset used to implement this machine learning model is known as ”Open Book by Bot press” that provides input data also known as training data that are in form of questions that maybe asked in the system and coming up with possible answers or output.

Some factors considered while choosing the dataset to work with.

• Size of dataset: The dataset is fairly large with a lot of training data that can be used in the training of the data model while implementing the linear regression algorithms in the system. The number of questions to be used are very countless and thus being able to answer at least 90 percent of the questions asked.

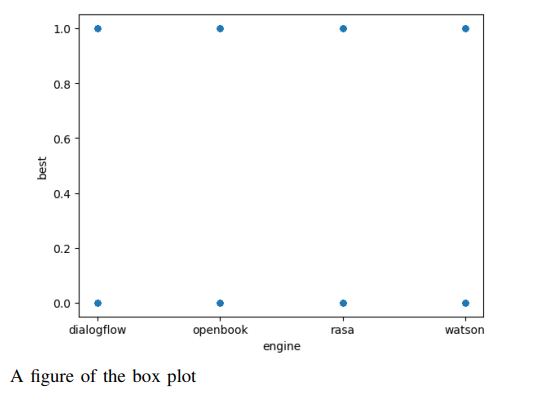
• Representation of real-world problems: We are training machine learning models to solve real world problems. So, the data should be real too thus use of this dataset that helps collect data from the real world and use it to train the models.

• Dataset matches context: Always we must make sure the characteristics of the dataset used for training the model matches with the conditions we have when the model goes live in production. Therefore, it indeed matches the context of our problem here.

• Dataset is unbiased and not redundant: A bias dataset never produces an unbiased trained model. Always the dataset we choose should be balanced and not bias to certain cases. It is also not data redundant or has less data duplication.

1. ***Data preparation and Exploratory Data Analysis***

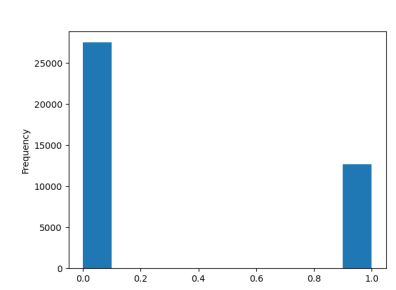
During data exploration and exploratory data analysis, the training data was cleaned by first removing the irrelevant data in the dataset such as the rows of p1, p2, p3, until the very end of the dataset. The duplicate data in the dataset was also then eradicated since there was a great sum of it in the dataset. Duplicate records slow down analysis and require more storage. Even more importantly, however, training a machine learning model on a dataset with duplicate results, the model will likely give more weight to the duplicates. Structural errors include things like misspellings, incongruent naming conventions were then dealt with as these can affect analysis because, while they may be obvious to humans, most machine learning applications would not recognize the mistakes and analysis would be skewed. The dataset had no missing data in it which made it favorable in the data cleaning process. Additionally, the data had various data outliers that had to be worked on in order to easily integrate the data visualization process of the dataset being used.



• Dataset training: This was implemented by using the questions columns as the training data of the model.

• Dataset testing: This was then implemented using a few testing data that provides the model with data to use so as to provide a possible output.

• Dataset validation: After training and testing the data, it was then brought to attention that the dataset is valuable and useful in training the machine learning model.



A figure of the bar graph

1. ***Artificial Intelligence Models***

In simple terms, an AI model is a tool or algorithm which is based on a certain data set through which it can arrive at a decision – all without the need for human interference in the decision-making process. Below are the models used in implementation of the project.

• Linear Regression: Linear Regression is a model that is based on supervised learning. The main task of this model is to find the relationships between the input and output variables. In simpler words, it predicts the value of a dependent variable based on a given independent variable. The parameters used here where the predictor values which were the questions and then the response variables/ independent variables which are the possible answers to the questions provided. Hyper-parameters used to tune the model would be noise term/ random error of the design matrix.

• Logistic Regression: Logistic regression estimates the probability of an event occurring, such as voted or did not vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds.

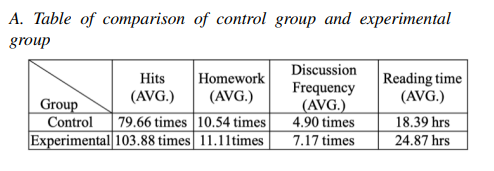
• Decision Trees: Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Here, the trees take in two inputs as arrays, an array X, sparse or dense, of shape (questions) holding the training samples, and an array Y of integer values, shape (possible answers), holding the class labels for the training samples.

1. ***Artificial Intelligence Models Accountability***

Accountability of artificial intelligence models refers to the ability to explain and justify the decisions made by an AI system. This is important because AI systems are increasingly being used in decision-making processes and we need to ensure that they are fair, transparent, and free of bias. An artificial intelligence model should be able to have the following aspects of explain ability, transparency, auditability and responsibility. In the project, linear progression model accountability was implemented by clearly documenting the steps taken to clean and process the data, the feature selection process, and any data transformations that were performed (transparency) Additionally, the model interpretability was implemented by providing information on the magnitude and direction of the relationship between the variables, so as to understand which independent variables have the most impact on the prediction and whether the relationship is positive or negative. The auditability and responsibility of the model was done by having the inputs, outputs, and decisions made by the linear regression model logged and stored, so that they can be reviewed and audited later. This can help ensure that the model’s decisions are transparent and can be corrected if necessary.

1. **RESULTS AND DISCUSSION**

Using evaluation metrics such as accuracy, f1-score, precision, and recall. These four metrics are often used to evaluate the performance of binary classification models, and provide different perspectives on the performance of the model. The choice of metric will depend on the specific use case and the desired outcome, and it is often useful to use a combination of metrics to get a more complete picture of the model’s performance. Accuracy evaluation metric was quite low at a percentage of 60 which was determined as the number of correct predictions divided by the total number of predictions. Precision was provided too at a lower percentage of 50 which was calculated as the number of true positive predictions divided by the number of true positive predictions plus the number of false positive predictions. Recall metric was gotten at a 50 percent rating which was calculated as the number of true positive predictions divided by the number of true positive predictions plus the number of false negative predictions. Finally, the f1-score is brought about by the formula: 2 \* (precision \* recall) / (precision + recall)



1. **CONCLUSION AND FUTURE WORKS**

In summary, the project would be more effective with a dataset whose training data is not so biased and less duplicacy in the data provided. I would like to conclude that the project would be implemented for other various institutions of learning. Main findings of this project is that its abundantly rich in data and could revolutionize artificial intelligence as we know it in many sectors not only education but also in health care systems, agriculture, industrial, and others. Integrating further improvement of the model, such as exploring different architectures, incorporating additional data sources, and incorporating unsupervised learning methods to improve performance would be another recommendation to follow up with.

1. **DATASET AND PYTHON SOURCE CODE DATASET:**

https://www.kaggle.com/datasets/talatiqbal/ botpresschatbotqnamontrealhotel

*PYTHON SOURCE CODE*: /notebook/chatbot-system

**ACKNOWLEDGMENT**

More advice would be integrating unsupervised learning techniques like clustering and association algorithms.

**REFERENCES**

[1] L. E. Chen, S. Y. Cheng and J. -S. Heh,” Chatbot: A Question Answering System for Student,” 2021 International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 2021, pp. 345-346, doi: 10.1109/ICALT52272.2021.00110.

[2] S. Reshmi and K. Balakrishnan,” Implementation of an inquisitive chatbot for database supported knowledge bases,” sadhan ¯ a, vol.41, no. ¯ 10, pp. 1173-1178, 2016

[3] D. F. Murad, M. Irsan, P. M. Akhirianto, E. Fernando, S. A. Murad and M. Hendra Wijaya,” Learning Support System using Chatbot in ”Kejar C Package” Homeschooling Program,” 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2019, pp. 32-37, doi: 10.1109/ICOIACT46704.2019.8938479.

[4] S. -H. Wu, L. -P. Chen, P. -C. Yang and T. Ku, ”Automatic Dialogue Template Synthesis for Chatbot by Story Information Extraction,” 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 485-490, doi: 10.1109/IRI.2018.00077.

[5] S. -H. Wu, L. -P. Chen, P. -C. Yang and T. Ku,” Automatic Dialogue Template Synthesis for Chatbot by Story Information Extraction,” 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 485-490, doi: 10.1109/IRI.2018.00077.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989

[8] D. Bailey and N. Almusharraf,” Investigating the Effect of Chatbotto-User Questions and Directives on Student Participation,” 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 2021, pp. 85-90, doi: 10.1109/CAIDA51941.2021.9425208.

[9] M. J. Some et al.,” Improving Covid-19 vaccine literacy among undergraduate students in Burkina Faso,” 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, 2022, pp. 522-523, doi: 10.1109/ICHI54592.2022.00096.

[10] G. K. Reddy K, A. Pal, S. K. V, R. J and S. K,” Cross Domain Answering FAQ Chatbot,” 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), Coimbatore, India, 2022, pp. 1-4, doi: 10.1109/ICACTA54488.2022.9752986.

[11] R. Al Farisi, R. Ferdiana and T. B. Adji, ”The Effect of Anthropomorphic Design Cues on Increasing Chatbot Empathy,” 2022 1st International Conference on Information System Information Technology (ICISIT), Yogyakarta, Indonesia, 2022, pp. 370-375, doi: 10.1109/ICISIT54091.2022.9873008.

[12] A. Herrera and N. Piedra, ”Ontology design proposal to formalize knowledge and manage responses in a conversational interface of the Open Campus Initiative,” 2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO), Loja, Ecuador, 2020, pp. 1-8, doi: 10.1109/LACLO50806.2020.9381131.

[13] M. R. Fonna and D. H. Widyantoro, ”Tutorial System in Learning Activities Through Machine Learning-Based Chatbot Applications in Pharmacology Education,” 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Bandung, Indonesia, 2021, pp. 1-6, doi: 10.1109/ICAICTA53211.2021.9640275.

[14] S. Alqaidi, W. Alharbi and O. Almatrafi,” A support system for formal college students: A case study of a community-based app augmented with a chatbot,” 2021 19th International Conference on Information Technology Based Higher Education and Training (ITHET), Sydney, Australia, 2021, pp. 01-05, doi: 10.1109/ITHET50392.2021.9759796. IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.