Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

A Chatbot for Answering Frequently Asked Questions by University Students Using Natural Language Processing and Multinomial Naïvebayes Algorithm

[1]Ede Ifesinachi Chizzy; [2]Aminu Muhd Bui & [3]Hassan Suru
[1]Department of Computer, Federal University, Bennin Kebbi
[2]Department of Computer, Usman Danfodio University, Sokoto
[3]Department of Computer, Kebbi State University of Science and Technology, Aliero
Email: edechizzy@yahoo.co.uk

ABSTRACT
Chatbots are programs that impersonate human discussion and their plan should be possible utilizing different techniques. Be that as it may, little work has been finished in the use of chatbots in the instructive area, thus; this undertaking is centered on making a chatbot to be utilized by understudies to respond to their habitually posed inquiries from the school's web-based entertainment stage and regulatory workplaces. This KSUSTA Chatbot has the ability to make discussions; answer the course and workforce subtleties; answer the regularly posed inquiries instead of looking at a considerable rundown of FAQ's searching for replies. Issues engaged with making Chatbots are information assortment which winds up giving not exactly needed measure of information for preparing and retraining, utilization of Programming interface's which decreases adaptability in the bot. Be that as it may, these issues were handled by involving BeautifulSoup for information assortment, Pandas for information handling and Multinomial Naïve Bayes model with a superior presentation. To develop the Chatbot, Python Language was utilized as the fundamental language. Furthermore, its AI and Natural Language Processing Libraries, web scrapping and document handling instruments were utilized, the front end graphical UI (GUI) was designed utilizing Flask (python), Html, CSS and JavaScript, Data set was taken care of with PostgreSQL, for recovery and retraining. The system has achieved 84% accuracy of correctly classifying the questions. One of the major drawbacks was the imbalanced state of the data set. Below are the various metrics that were used to evaluate its performance.

Keywords— Frequently Asked Questions (FAQs), Natural language processing (NLP), Machine learning (ML), Multinomial Naive Bayes, Supervised Learning, Reinforced Learning

INTRODUCTION

The word chatbot originates from the two words chat and robot and describes a relatively new computer application designed to simulate conversations with users via a chat. Artificial Intelligence is best described as the development of a computer system's ability to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making and translation between languages. In general terms, AI refers to computational tools that are able to substitute for human intelligence in the performance of certain tasks. This technology is currently advancing at a breakneck pace, much like the exponential growth experienced by database technology in the late twentieth century. AI programming is an elevation of technology that has brought efficiency and optimum benefits to different companies' operations and people lives. AI has brought another level of smart technology to different industries and the prospects of its potential still grows with the expectation that it would reach the human intelligence. Below are the list of programming languages that can be used in artificial Intelligence. Python, C++, Java, LISP, Prolog. In this project, we will be making use of python programming language.

Integrating Machine Learning into Chat bot can be achieved using classification algorithms. Classification provides predictive measures which make a prediction about the outcome of data using known results found from different data. This outcome is from a fixed set of outputs. Data whose outputs are not fixed are called regression. Predictive modeling can be understood of as a learning function of mapping from an input set of vector measurements to a scalar output (Bhardwaj & Pal, 2011). Text based Machine Learning problems are usually

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

classification problems, one of the popular examples is sentiment analysis. In building Chatbot for businesses, most times those bots are there to answer questions that are commonly asked by users. Hence, the dataset needed will not be quite as much as it is in other cases. The patterns are always similar, many times the difference is the grammatical construction. A simple classifier system will be adequate to handle all use cases.

In this case, the KSUSTA BOT is built to use Machine Learning and rule based approach to adequately answer questions asked by users, which in many cases will be similar. Over time as students keep using the bot, the learning can be reinforced and make accuracy as high as possible. However, the initial dataset was from an online platform for aspiring students and entry level students and it was trained using Naïve Bayes Classifier. This classifier uses probabilities.

Statement of the Problem

Most of the existing Chatbots don't have an adequate data collection process, eventually they are left with too little data which is not enough to train the model, hence, accuracy is reduced. Furthermore, they are built using existing Chatbots assistant API's and as such flexibility of functionality is reduced due to inherent limitations on the platform themselves, also matching users' questions to stored FAQ is quite problematic. Another prevalent problem with Chatbots is processing of Natural Language, though many solutions have been proffered but the problem is not completely solved. In order to address these problems, KSUSTA Chatbot is built to be a self-updating chatbot which extends the implementation of the current Chatbots by using Web scrapping method for

data collection which can collect 10,000 questions in less than an hour, Natural Language Processing for preprocessing the data and also making use of bigrams with many range so as to allow more context which will give room for better accuracy, Pandas for cleaning and uses python machine learning library Multinomial Naïve Bayes Algorithm to train the chat bot model. Multinomial Naïve Baye's probabilistic classification will make it much more accurate and not require as much data as the former, reducing training time.

Aim and Objectives

Continuing with the problem introduced earlier, the aim of this dissertation is to create a Chatbot using Natural Language Processing and Multinomial Naïve Bayes Algorithm to get the intent of the user and adequately answer students FAQs as related to Kebbi State University of Science and Technology, Aliero. To achieve our aim, the following set objectives will be followed;

i. To develop a chatbot for university FAQ that incorporates Machine Learning Algorithm and Graphical User Interface to make interaction easier
ii. To Create a Naïve Bayes model which is a probabilistic model to determine question type
iii. To Create a rule-based response to make chat response more specific

REVIEW OF LITERATURE

Classification, a data mining technique, is the process of classifying and predicting the value of class attribute based on the values of predictors (Romero, Ventura, Espejo & Hervás, 2008). There are two main categories of classification model used in prediction: *Descriptive* and *predictive* classification

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

model. Descriptive model detects relationships or pattern in data and explore even the properties of the scrutinized data. Example of such technique which support this includes summarization, clustering, association rule etc. While, predictive model conducts prediction of unknown data values by using supervised learning function applied on known values (Jothi, Rashid & Husain, 2015). The known data is historical in nature. Example of such techniques includes Time series analysis, Prediction, Classification, Regression etc. Our interest in this study lies in predictive classification model, where the model is constructed based on a feature of historical data and are used to predict future trend (Al-radaideh & Nagi, 2012). Many classification algorithms are used for classifying categorical data e.g. *Decision tree, K-Nearest Neighbor, Naïve Bayes, SVM, J48, Random Forest* etc. In this study, we dwell on Naïve Bayes classification technique. Naïve Bayes classifier provides an analysis tool that defined a set of pattern rule which categorizes data into different classes using probabilistic approach. Initially, it would first construct a model for each of the class attributes as a function of other remaining attributes in datasets. Then, tries to correlate the class of every record using a previously designed model on unseen and even new data set (Manjusha *et al.*, 2015). This analysis aids with a good understanding of the data set and predicting future trend (Ameta & Jain, 2017).

Decision Tree Classifier

Decision tree is another classification algorithm that uses an organized hierarchical structure of a set of conditions to classify an instance. Decision tree is associated with a number of drawbacks such as the inability to have a representative object in real world, lack of quality measuring mechanism for

attributes cost and value, ability to fail to classify in some instances e.g. when a class is dependent on a high number of attributes and such set of conditions are not set. Decision tree supports a predictive approach in machine learning and data mining. The leaves represent the classes while branches represent junctions leading to class label (Zorman, Štiglic, Kokol, & Malčić, 1997). The initial challenge in decision tree is the lack of a distinct method for selecting attributes to be used for constructing the tree from root node to a leave. This provides means for a various construct in which some construct can provide the opportunity for failing a given instance. Decision tree classifies data by following some path of stated satisfied conditions that start from the root of a tree to the leaf called class label (Romero *et al.*, 2008)

Support Vector Machine (SVM)

SVM is a learning method that used the concept of computer science and statistics to analyze data and support pattern recognition. This approach is used in classification problem and nonlinear regression analysis. SVM is a non-probabilistic linear classifier which makes a prediction based on the set of accepted input, in which for every given input, there are two feasible classes that form the inputs (Raj & Prasanna, 2013). SVM was designed based on the principle of "Structural Risk Minimization principle" with the basic idea of finding hypothesis with the lowest minimum error e.g. error rate of a learner on data say training data set is restricted by the summation of the training-error rate (Ghumbre, Patil, & Ghatol, 2011). However, the drawback of this learner is that its computation is highly expensive thereby running slow on high data set and the classifier is also a binary classifier, therefore performing multi-class classification is done pair-wise

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

(Madzarov, Gjorgjevikj & Chorbev, 2009), and similarly, SVM provides inability to present result in a transparent manner on high dimensional data (Auria & Moro, 2008; Karamizadeh, Abdullah, Halimi, Shayan & Rajabi, 2014).

NAIVEBAYES ALGORITHM

Naïve Bayes classifier works as both supervised learning and statistical based technique for classification (Hemanth *et al.*, 2011). It works based on Bayes' theorem through finding the probability of an event occurring given the probability of another event that has already occurred. It assumes a model which relies on probability to calculate uncertainty of future events in such a principled mechanized way through estimating the probabilities of the events. Such mechanism has been widely used in prediction and diagnosis of diseases (Medhekar *et al.*, 2013). Naïve Bayes classification is simple and particularly suited when the dimensionality of the input is high. Despite its simplicity, it can outperform more sophisticated classification method. It provides perspective for understanding many learner algorithms and works on the assumptions that: is easy to construct, classifying categorical data, occurrences of an event (attributes) are independent and can be trained in a supervised manner(Patil *et al.*, 2016). The major advantage of Naïve Bayes in classification is its simplicity and its ability to approximate probabilities for a class on any given instance (Kononenko, 1991)

RELATED WORKS

University Chat bot using Artificial Intelligence Markup Language

In modern development of Chat bots, most of the chat bots utilize the algorithms of artificial intelligence (AI) in order to

get the required responses. However, Niang, N.K. and Khin M.S in 2020 designed a chatbot using Artificial Intelligence Markup Language, they designed a University Chatbot that provides an efficient and accurate answer for any user questions about university information. This is the first University Chatbot for inquiring about school information in Myanmar Language based on Artificial Intelligence Markup Language (AIML) and uses Pandora bots as the interpreter. AIML is an XML based markup language for specifying chatbot content. An AIML has an interpreter which is able to load and run the bot, then, it provides the bot's responses in a chat session with a user. AIML consists of data objects called AIML objects, which are made up of units called topics and categories. The topic is called an optional top-level element, it has a name and a set of categories related to that topic. Categories are the basic units of knowledge in AIML. One category is a rule for matching an input and converting to an output, and consists of a pattern, which represents the user input, and a template, which responses the answer. The AIML pattern is simple and consists of words, spaces, and the wildcard symbols _ and *.

This chatbot is one of rule-based chatbot and rule-based system will not function effectively to answer questions that are outside the range of patterns used. Hence, part of the proposed improvements is to include machine Learning algorithm in place of the markup language. This will reduce significantly the time it takes to get the data prepared for the Chat bot.

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

FAQ (Frequently Asked Questions) Chat Bot for Conversation

In the works of Farhana Sethi in 2020, they designed an FAQ chat Bot, Users can easily type their query in their natural language and retrieve information. In their proposed solution they have used Rule-NLP Hybrid FAQ Bot. Internally, it uses any NLP (Natural Language Processing) system to interpret the human interactions and reply back with meaningful information. This FAQ ChatBot uses simple pattern matching to represent the input and output whereas other ChatBots uses input rules, keyword patterns and output rules to generate a response. If the input is not found in the database, a default response is generated. The input and output can be customized according to the user. Based on the developer or the user, the required requests and responses can be stored in the database. The major limitations to this work were inflexibility and difficulty in understanding the intent and nuance of the language. User experience demands a consistent, clear and focused personality that mimics human interaction and makes them feel at ease. Secondly, the data collection for the Natural Language Processing have no consistent personality because the dialogue answers are all amalgamated text fragments from different sources, besides, there is a great limitation in getting sufficient training data. Adding new training data to existing models prove a herculean task. They proposed to add machine learning algorithms to handle off script questions.

Supporting Creation of FAQ Dataset for E-Learning Chatbot

Yasunobu S. et.al in 2020 set out to create a dynamic support system that is lacking in many of the e-learning systems. This was implemented using a chat bot, however, they noted the difficulty in collecting the large number of Q&A data or high-

quality datasets required to train the chat bot model to obtain high accuracy. A proposal for data collection using a novel framework for supporting dataset creation was made. This framework provides two recommendation algorithms: creating new questions and aggregating semantically similar answers.
The core contribution of this study is to provide recommendations that are applicable to small-sized datasets. Compared with previous studies on dataset creation, their framework uses two unsupervised learning algorithms: supporting creation of new questions and finding semantically similar answers. The following assumptions were made in the process:

- It is difficult to automatically create FAQ datasets from small Q&A datasets.
- We can manually create FAQ datasets from small Q&A datasets.
- Supporting manual creation is beneficial to decrease the costs even though we can create the dataset without any tools.

They first collected raw data from logs of users of the e-learning system introduced in Tokyo Metropolitan University and recorded the questions they asked and answers provided by system engineers who managed the e-learning system in practice. They created a dataset which includes 200 Q&A pairs in total. They introduced our categorization scheme for the collected raw Q&A based on features of the e-learning system. The objective was to organize answers; which was useful for analyzing the kinds of features users often have difficulties with and understanding the feature that should be focused on when preparing FAQ dataset. For Future Improvements qualitative evaluation will be done as this paper focused on quantitative evaluations; however, analyzing what users feel

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

and think about using chatbots is also important for practical usage.

Dexter the College FAQ Chatbot

In this paper by Ajinkya H. et al, They designed of chatbot which comprised of two isolated parts in particular "chatbot engine" and "language model" which offers the chance to effectively execute a chatbot in a recently created knowledge model in addition, a graphical user interface was added to the inquiry chatbot. This, unlike the previous bots which use command line, makes interaction much more effective and simulate a typical conversation with any user. The chatbot will be the Generative model because of which it will be able to generate new responses from scratch. It will be done by using RNN (Recurrent Neural Network) and LSTM (Long Short Term Memory so that the chatbot will be able to frame its answer if the answer for that particular question is not available in the database. One major limitation in this work is inadequate dataset for prediction.

Implementing a College Enquiry Chatbot

In a paper published by Ujaliben, K in 2019, it focuses on creating a College Enquiry Chatbot that has the capacity to make friendly conversations; respond to course and faculty details; give the link for the academic calendar; answer the frequently asked questions; calculate the fees based on the student's input; and give the timings, address, contacts, and events information of the departments like Union, Library, IPGE, and AIRC. To build the chatbot, Microsoft Azure bot service as well as Microsoft cognitive services, namely, Text Analytics, LUIS, and QnA Maker are used. Most of the existing chatbots lack empathy and fail to accommodate anything

outside of the script. In order to address these problems, the College Enquiry Chatbot extends the implementation of the current chatbots by adding sentiment analysis and active learning. Although, sentimental analysis correctly recognizes the user's query as positive, negative and neutral, the system was partially successful in adding empathy to the chatbot. It is because the system requires more rigorous training data to handle all queries which are off-script. However, for such queries, active learning helps to improve the chatbot. A major setback in the work is the collation of varied data and limited scope due to the dataset, social media integration and speech recognition.

Chatbot for University Related FAQs

In this paper Bhavika, R. R, Nidhi R and Sanjay S. Y. 2017 designed a chatbot using Artificial Intelligence Markup Language, they designed a University Chatbot that provides an efficient and accurate answer for any user questions about university information. This is the first University Chatbot for inquiring about school information in Myanmar Language based on Artificial Intelligence Markup Language (AIML) and uses Pandora bots as the interpreter. AIML is an XML based markup language for specifying chatbot content. An AIML has an interpreter which is able to load and run the bot, then, it provides the bot's responses in a chat session with a user. AIML consists of data objects called AIML objects, which are made up of units called topics and categories. As a future work we can make a chatbot which is blend of AIML and LSA. This will enable a client to interact with chatbot in a more natural fashion. We can enhance the discussion by including and changing patterns and templates for general client queries

using AIML and right response are given more often than not utilizing LSA.

Recommending Moodle Resources Using Chat bots

Kamal S. in 2019 proposed designing a Recommendation system using the most suitable educational resources in the field of E-Learning, which has been a huge challenge. This challenge has pushed educators and researchers to implement new ideas to help learners improve their learning and their knowledge. New solutions are using Artificial Intelligence (AI) techniques such as Machine Learning (ML) and Natural Language Processing (NLP), however, the approach is centered primarily on the use of a custom chatbot which can be linked to Moodle's platform using a web configuration. The major setback for this work is that it cannot support other applications and platforms.

Developing an AI-Powered Chatbot to Support the Administration of Middle and High School Cybersecurity Camps

Jonathan He and ChunSheng Xin in 2021 published a paper which distinguished the different types of approach used in developing a chatbot occluding the rule based approach and the AI-based chatbots. The rule-based chatbots primarily If-Else statements to filter the question-answer pairs meanwhile the AI-based chatbots use a repository of predefined responses and some methods to select the appropriate response. They again further classified chatbots retrieval-based and generative-based (Maroengsit et al., 2019). Retrieval-based approach searches a user-issued query from a database and returns a reply that best matches the query (Song et al., 2016). When the database is small, it may not find a reply. In contrast, generative-based approach often uses recurrent neural

networks to generate new responses, but it could generate meaningless responses (Song et al., 2016). In their case, they adopted a retrieval-based approach with an added advantage of being best suited for use in consumer support, lead generation, and feedback. They however noted the following drawback

- Chatbots can become outdated later if no further questions/responses are added to the chatbot.

As a way to solve this problem, they suggested Chatbots should be trained regularly as new training phrases are being included in the chatbot (Herriman et al., 2020) to achieve better accuracy. They proposed as a future work, to add more knowledge base to the chat bot for better response and make it more versatile.

Building a Medical Chatbot using Support Vector Machine Learning Algorithm

Tamizharasi B., Jenila Livingston L.M.* and S. Rajkumar in 2021 published a paper to develop chat bot using Support Vector Machine, SVM is a supervised learning algorithm and is classified into two classes using "hyper plane". The hyper plane has the highest margin for separating given data into classes. SVM is a regulated AI calculation which can be utilized for both arrangement and relapse difficulties. In SVM, it is anything but difficult to have a direct hyper-plane between these two classes. SVM has a method called the Kernel. Its capacities that take up low-dimensional information space and transform it into a higher-dimensional space. It is helpful in classifying non-directional issue. The final model was tested against KNN and Naïve Bayes algorithms are also used. Mostly SVM gives accurate results and it worked better with large number of data and working faster too. The comparative analysis is

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

depicted with the column chart (Figure 1). SVM produces 92.33%, KNN with 87.66% and Naïve Bayes with 81%.

SOFTWARE DESIGN

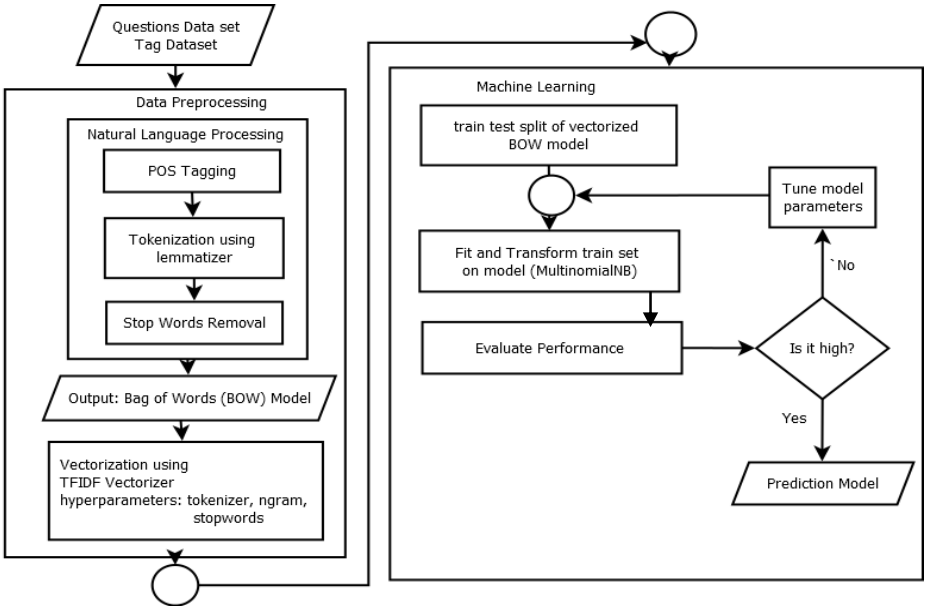


Figure 1: System Architecture

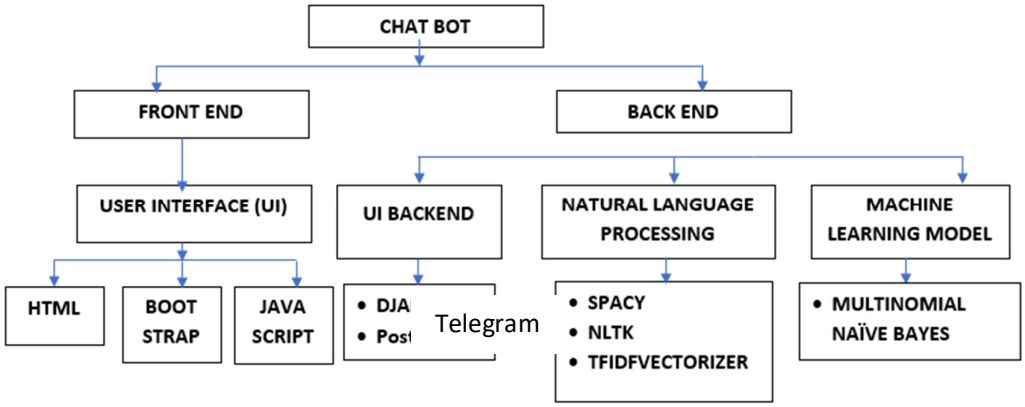


Figure 2: Development Tool

Requirement Elicitation

The system is intended to serve as an assistive tool for students to get answers quickly without having to go to school office and reading through a long list of FAQ's on the school's website. It can also be used by any other individual who wishes to inquire about the status of the school as the information is being updated. Therefore, the users are expected to supply their questions while the system would automatically answer them.

Outline of the Requirement

i. User opens the platform.
ii. A textbox is availed to them to type their question.
iii. The user gets the result displayed as a chat.
iv. The message get stored behind the scene for further learning.

Software Requirement

The software is designed using Python programming language which supports Pandas, Scikit – Learn library which includes all the modules that are needed for cleaning, preprocessing, building of the model and evaluation metrics. This language was chosen over others in order to have maximum control on interface design, better presentation of statistical result and flexibility and also the ease of programming as highlighted in the introductory section.

Hardware Requirement

The minimum hardware requirements include the following:

1. Windows 8.1 or 10, 64 bits (PC or Mac computers).
2. All CPU (Intel family, Xeon, AMD)
3. 3 GB RAM or above,
4. 50GB HDD or SDD Free Space

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

PREDICTION PROCEDURE

Data Collection

The data to be used was collected from an online student forum and divided into two portions which are the Training set and the Testing set. The training set was used to build the model while the testing set was used to test the performance of the model and based on that, retraining could therefore be carried out.

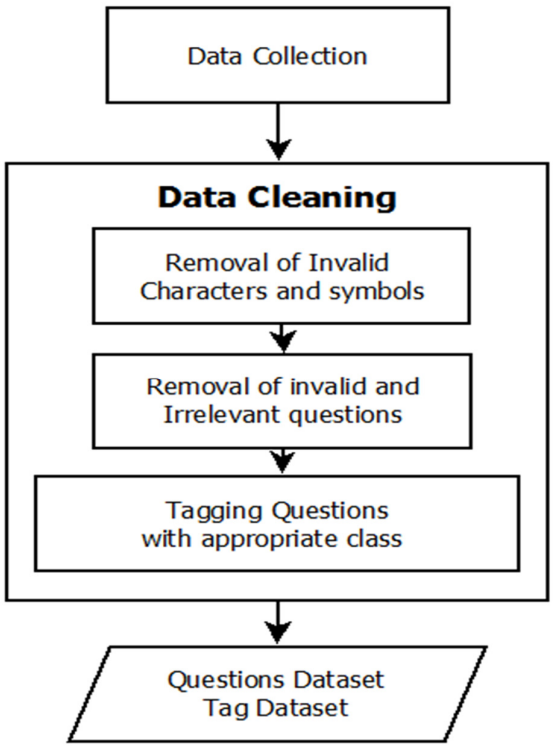


Figure 3: Data Collection Process

Data Preprocessing

Data preprocessing was carried out to clean the data from error and noise, removing outliers, and prepare the data for machine Learning.

Transformation

Before a data set can be fitted for use by a machine learning model, the data has to be processed through cleaning, encoding

(where necessary). Processing numerical data is usually quite straight forward, if the dataset has texts as part of its attributes, some encoders like one hot encoder can be used to transform it into numerical values. When the dataset is entirely composed of texts from sentences or documents, it will not be practical to use the conventional encoders as this might cause it to lose its meaning and make the machine learning model perform poorly, The same words in a different order can mean something completely different. Even splitting text into useful word-like units can be difficult but with the use of **SPACY,** it is easy to annotate the sentences being asked and help give meaning to the words. The major features of **SPACY** are its tokenization feature, lemmatization feature, and linguistic annotations feature. First the dataset is broken down to sentences, since students are going to use the bot to ask question after which the corresponding answer be sent to them. Tokenization is the task of splitting a text into meaningful segments, called *tokens*. During processing, Spacy first **tokenizes** the text.

The Spacy's Lemmatizer is a pipeline component that provides lookup and rule-based lemmatization methods in a configurable component. An individual language can extend the Lemmatizer as part of its language data.

TFIDFVECTORIZER

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. This formula has an importance consequence that a high weight of the *tf-idf* calculation is reached when we have a high term frequency (*tf*) in the given document (*local parameter*) and a low

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

document frequency of the term in the whole collection (global *parameter*)

Assuming we have a train set and a test set as seen below

```
Train Document Set:
d1: The sky is blue.
d2: The sun is bright.
Test Document Set:
d3: The sun in the sky is bright.
d4: We can see the shining sun, the bright sun.
```

Figure 4: Test data set for transformation

After being vectorized and transformed by tf-idf, it gives us overall documents of weight of words.

```
TD-IDF Vectorizer

          blue    bright       sky       sun
Doc1  0.707107  0.000000  0.707107  0.000000
Doc2  0.000000  0.707107  0.000000  0.707107
```

Figure 5: tf-idf result

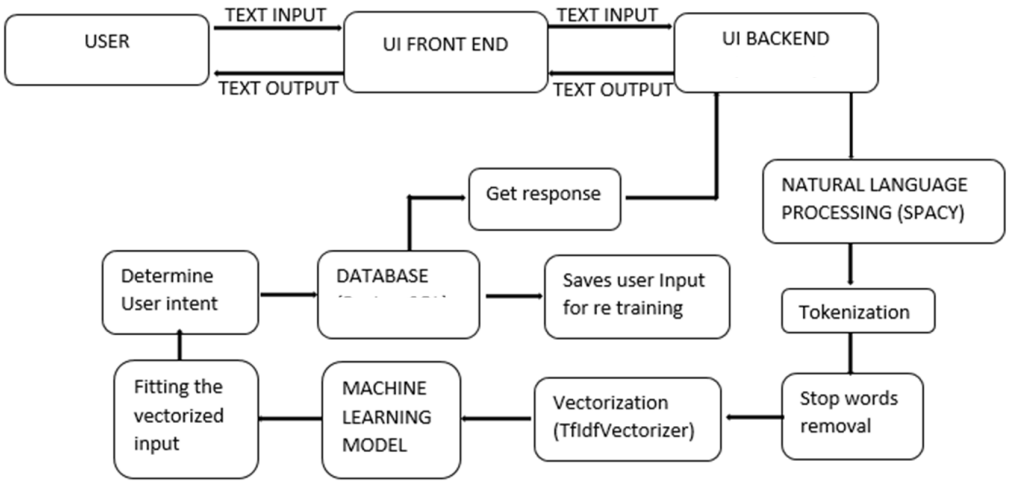


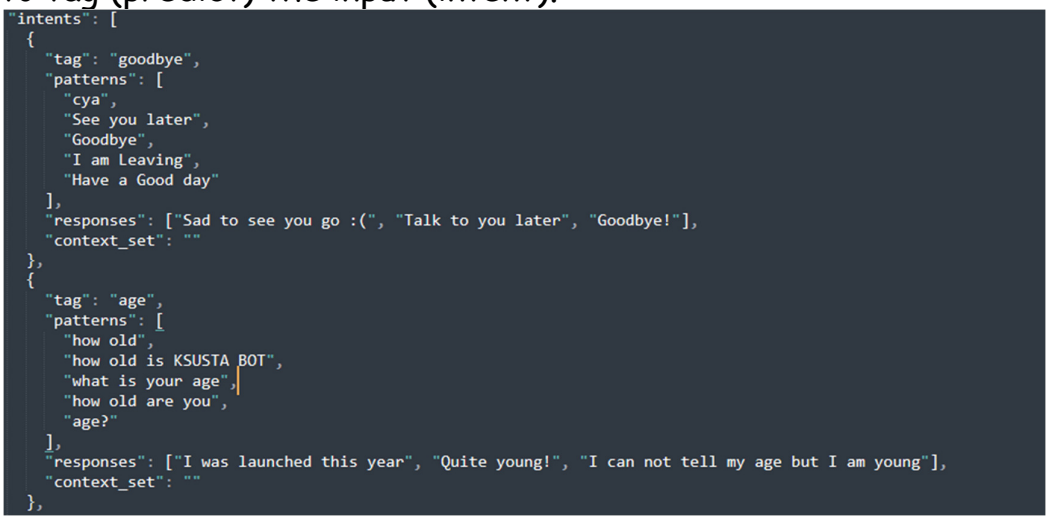
Figure 6: Prediction Architecture

CHATBOT TRAINING

Training a bot can likened to giving the bot brain and helping it to understand what functions it has to perform and how to go about them, all these are termed as the intents. Each intent has the following parameters:

1. Tag
2. Pattern
3. Response

The tag is the name or identity of the intents, while the patterns are the various texts which all mean the same thing or have the same pattern. These patterns formed the machine learning data to be trained while the tag represents the label for each data in the supervised machine learning model. The response is the output to the user after the bot has been able to tag (predict) the input (intent).



```
"intents": [
  {
    "tag": "goodbye",
    "patterns": [
      "cya",
      "See you later",
      "Goodbye",
      "I am Leaving",
      "Have a Good day"
    ],
    "responses": ["Sad to see you go :(", "Talk to you later", "Goodbye!"],
    "context_set": ""
  },
  {
    "tag": "age",
    "patterns": [
      "how old",
      "how old is KSUSTA BOT",
      "what is your age",
      "how old are you",
      "age?"
    ],
    "responses": ["I was launched this year", "Quite young!", "I can not tell my age but I am young"],
    "context_set": ""
  },
```

Figure 7: Intent Classification

Classification and Prediction

Based on the nature of variable in our dataset, we will use Multinomial Naïve Bayes classification techniques pipeline. Working of the framework is illustrated as follows:

i. Data collection

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

ii. Data Cleaning
iii. Data Classification
iv. Data preprocessing
v. Preprocessed data is stored in training and test sets.

Data Classification

Data is classified into Admission, Cut off, Post Utme, Resumption, School Fees, Subject Combination, Courses and Minor questions. A question will be given its appropriate tag as soon as its question is asked. The system would be designed in such a way that the users are only required to supply their questions.

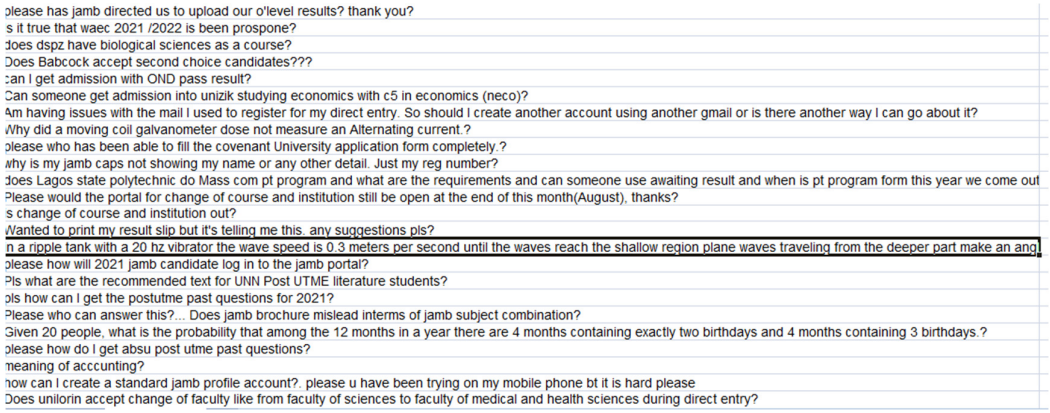


Figure 81: Raw Data Collected before tagging

DATA COLLECTION

Before collecting the data, the researcher was guided with all ethical training certification on data collection, right to confidentiality and privacy reserved called*Institutional Review Board* (IRB). Data was collected from the online archive of my school ng using webscrapping technique then transformed the data to electronic form and stored in POSTGRESQL database.

Sample Size Determination

Naïve Bayes technique requires that data should be as high as possible because its accuracy depends on how high the volume of the data (Qian, Zhou, Yan, Li, & Han, 2015). There are thousands of pages of question to scrap but not all questions were relevant and those were dropped off during data cleaning. The initial data cleaning was done manually because of the nature of the data. Albeit, We used the formula developed by Krejcie and Morgan, (1970) to understand minimum sample size required for the study.

$$s = \frac{X^2 NP\,(1-P)}{d^2(N-1)} + X^2 P(1-P) \qquad \textbf{(5)}$$

s = the required sample size.

X^2 = is the table value of chi-square for 1 degree of freedom at confidence level (3.841).

P = is population proportion (assumed to be .50since this would give the max. sample size).

N = is population size.

d = is a degree of accuracy expressed as a proportion (.05).

Sample size

(S) = (3.841) x (1000) x 0.50(1-0.50) ÷ $(0.05)^2$ x (1000-1) + (3.841) x 0.50(1-0.50) = 385.6

A total of 1325 cleaned preprocessed records were collected and stored in our database According to a popular scholar,Sordo & Zeng (2005) which states that a sample size of ~150 -~8500 should be adequate for training while testing set of 10 - 60 should be adequate for a classifier performance measure (Indira, Vasanthakumari, Jegadeeshwaran, & Sugumaran, 2015). In this study,out of 700 records collected, 80% was used for training while 20% was used for testing.

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

SETUP

A portion of real data was used for training the model. We have one training set.

Using,

$$\text{Naive Bayes; } P(t/y) = \frac{P\left(\frac{y}{t}\right)P(t)}{P(y)} \quad (6)$$

Where t: is the class

P(*t*|*y*): is a posterior prob. of *class* given *predictor*.

P(*t*): is the past (prior) prob. of *class*.

P(*y*|*t*):the prob. Of *predictor* given *class*.

P(*y*): past prob. of the *predictor*.

Admission class label

$$P(\text{Admission} \mid \text{User}) = \frac{P(\text{Admission})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (7)$$

Courses class label

$$P(\text{Courses} \mid \text{User}) = \frac{P(\text{Courses})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (8)$$

Cut Off class label

$$P(\text{Admission} \mid \text{User}) = \frac{P(\text{Cut off})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (9)$$

Combination

$$P(\text{Combination} \mid \text{User}) = \frac{P(\text{Combination})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (10)$$

Post Utme

$$P(\text{PUtme} \mid \text{User}) = \frac{P(\text{PUtme})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (11)$$

Resumption class label

$$P(\text{Resumption} \mid \text{User}) = \frac{P(\text{Resumption})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (12)$$

Fees class label

$$P(\text{Fees} \mid \text{User}) = \frac{P(\text{Fees})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (13)$$

Registration class label

$$P(\text{Registration} \mid \text{User}) = \frac{P(\text{Registration})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (14)$$

Others class label

$$P(\text{Others} \mid \text{User}) = \frac{P(user \mid P)P(\text{Others})}{\sum_{i=0}^{n} P(\text{user} \mid \text{Class}_i) * P(\text{Class}_i)} \quad (15)$$

PERFORMANCE METRICS

1. Confusion Matrix: A confusion matrix is a table that describes the performance of a classifier. Confusion matrix demonstrates the accuracy of a solution to a given classification. It contains information about the predicted and actual classifications done by a classifier system. The performance of the model is normally evaluated using the data in the confusion matrix.
2. Accuracy: is the ability of a model to appropriately predict the class label of previously unseen data or new data. It is a measure of how well the classifier makes a prediction on average. A good classification algorithm will try to minimize the number of times it makes the wrong prediction.
3. Precision: Precision is the portion of retrieved cases that are relevant. It is the measure of correctness or excellence
4. Recall: This a measure of when the outcome of a prediction is said 'P' and the classifier have actually predicted the value to be same 'P'. It is a measure of comprehensiveness or magnitude

RESULTS AND DISCUSSION

This chapter discusses the result of the testing by using screenshots to present the design of our implementation which includes the models used, training data, chat interface along with the performance evaluation of the classifier. The prediction framework has been implemented using Python programming language. A web scrapper (Beautiful Soup), jupyter notebook and Multinomial Naive Bayes model was used. A total sample record of 3000 sample questions were scrapped online and used in the training after cleaning, it took 654 seconds (approximately 10 minutes) to get them. This is a major improvement over the previous data collection methods where

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

significant amount of time (like days or weeks) are required to be able to collate questions which are not up to 1000. Unlike in existing systems where API's are used to build Chatbots, most of their preprocessed data have a fixed format which is usually JSON (JavaScript Object Notation), this format is not typically what is used in the data collection, hence they have to manually edit the collected data into this format which is usually time consuming, appending new data for retraining also follows the herculean pattern and thus increases time for data processing and formatting. However, the use of Pandas gives the flexibility of taking in data in any format and outputting same in any format.

The prediction Model (Multinomial Naïve Bayes)

The prediction Model built using our training data sets is used to predict the intention of the questions.

```
Enter your question
Is the admission list out

Probability of each class
Admission: 0.39774320964774845
Combination: 0.05596285322976162
Courses: 0.07061628515780945
CutOff: 0.10427274567536687
Fees: 0.05116616202287255
minor: 0.17828057160799615
PostUtme: 0.04754653883385442
Remedial: 0.03270265085049759
Resumption: 0.06170898297409225
The question is related to Admission
```

Figure 9: Probabilites for classification

Classification report

	precision	recall	f1-score
Admission	0.94	0.97	0.95
Combination	1.00	0.80	0.89
Courses	1.00	0.87	0.93
CutOff	0.88	1.00	0.94
Fees	1.00	0.81	0.90
PostUtme	0.61	1.00	0.76
Remedial	1.00	0.52	0.68
Resumption	0.00	0.00	0.00
minor	1.00	0.58	0.73

Chat Interface

In asking questions, the user is first required to type in the question. The system would use the designed model to learn and predict the class label of the question automatically based on the training. Once the label has be generated, the learning model now generates an appropriate response to the user based on the key words or intent of the question. For example we have different responses for the same question but with different intents as shown below

Figure 10: Bot Responses during Training

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

Figure 11: Bot Response after Retraining

RESEARCH CONTRIBUTION

This research uses a supervised and reinforced learning algorithm on large data sets of 1325 questions obtained from "***https://myschool.ng/questions/school-based-questions?page={i}***" to build a Chatbot which answer questions by students using various data processing task and Naive Bayes classification technique. This research was able to bring to light how effective web scrapping for data collection is by giving thousands of raw data in less than an hour, It further shows a high accuracy for Naïve Bayes would perform as a model for machine learning and how a flexible development can

be done without the use of Application Programming Interface (API) from existing Chatbot building platforms.

RECOMMENDATIONS

a. Naïve Bayes can suitably creating a chatbot with a higher response rate than what is normally obtainable.
b. The web scrapping method of data collection is responsible for giving high quality data collection techniques.

Future Work

This study can further be experimented on other machine and deep learning models and another classification technique as well as on more attributes for better performance and other features can be included such as auto-predict and auto complete which will help to reduce spelling errors and help improve prediction accuracy.

REFERENCES

Abu Shawar, B., & Atwell, E. S. (2009). Arabic question-answering via instance based learning from an FAQ corpus. In *Proceedings of the CL2009 International Conference on Corpus Linguistics*. UCREL, Lancaster University.

AbuShawar, B., & Atwell, E. (2015). ALICE chatbot: trials and outputs. *Computación y sistemas*, *19*(4), 625-632.

Ahmad, N. A., Hamid, M. H. C., Zainal, A., & Baharum, Z. UNISEL Bot: Designing Simple Chatbot System for University FAQs.

Al-Radaideh, Q. A., & Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employee's performance. *International Journal of Advanced Computer Science and Applications*, *3*(2), 144-

Journal of Physical Science and Innovation
Volume 14, Number 1, June 2022
ISSN: 2277-0119
http://www.cenresinjournals.com

151.

Auria, L., & Moro, R. A. (2008). Support Vector Machines (SVM) as a technique for solvency analysis. *DIW Berlin*, (August), 1–16.

Baby, M. N., & Priyanka, L. T. (2012). Customer classification and prediction based on data mining technique, *2*(12).

Bhardwaj, B. K. (2011). Data Mining : A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security*, *9*(4).

Bhirud, N., Tataale, S., Randive, S., & Nahar, S. A Literature Review On Chatbots In Healthcare Domain.

de Lacerda, A. R., & Aguiar, C. S. (2019, August). FLOSS FAQ chatbot project reuse: how to allow nonexperts to develop a chatbot. In *Proceedings of the 15th International Symposium on Open Collaboration* (pp. 1-8).

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, *217*(1), 48-58.

Ghumbre, S., Patil, C., & Ghatol, A. (2011). Heart disease diagnosis using Support Vector Machine. *International Conference on Computer Science and Information Technology (ICCSIT'2011)*, 84–88.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

Hassouna, M., Tarhini, A., Elyas, T., & AbouTrab, M. S. (2016). Customer Churn in Mobile Markets A Comparison of Techniques. *arXiv preprint arXiv:1607.07792*.

Hemanth, K. S., Vastrad, C. M., & Nagaraju, S. (2011). Data mining technique for knowledge discovery from engineering

materials data sets. *Advances in Computer Science and Information Technology*, 512-522.

Husin, N. A., Mustapha, N., Sulaiman, M. N., & Yaakob, R. (2012, September). A hybrid model using genetic algorithm and neural network for predicting dengue outbreak. In *Data Mining and Optimization (DMO), 2012 4th Conference on* (pp. 23-27). IEEE.

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data mining in healthcare - A Review. *Procedia Computer Science*, *72*, 306–313.

Kar, R., & Haldar, R. (2016). Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*.

Karamizadeh, S., Abdullah, S. M., Halimi, M., ShaJothyan, J., & javad Rajabi, M. (2014, September). Advantage and drawback of support vector machine functionality. In *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on* (pp. 63-65). IEEE.

Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Machine Learning—EWSL-91* (pp. 206-219). Springer Berlin/Heidelberg.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, *30*(3), 607-610.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning*, *30*(2-3), 271-274.

Kriegel, H. P., Borgwardt, K. M., Kröger, P., Pryakhin, A., Schubert, M., & Zimek, A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*(1), 87-97.

Laishram, D. D., Sutton, P. L., Nanda, N., Sharma, V. L., Sobti, R. C., Carlton, J. M., & Joshi, H. (2012). The complexities of malaria disease manifestations with a focus on

asymptomatic

Madzarov, G., Gjorgjevikj, D., & Chorbev, I. (2009). A multi-class SVM classifier utilizing binary decision tree. *Informatica*, *33*(2).

Manjusha, K. K., Sankaranarayanan, K., & Seena, P. (2015). Data mining in dermatological diagnosis: A method for severity prediction. *International Journal of Computer Applications*, *117*(11).

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science & Technology*, *8*(1), 13-19.

Oguntimilehin, A., Adetunmbi, A. O., & Abiola, O. B. (2013). A Machine Learning Approach to clinical diagnosis of typhoid fever. *A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever*, *2*(4), 671-676.

Patil, R. R. (2014). Heart disease prediction system using Naive Bayes and Jelinek-mercer smoothing. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(5), 2278-1021.

Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2014). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote Sensing*, *7*(1), 153-168.

Ranoliya, B. R., Raghuwanshi, N., & Singh, S. (2017, September). Chatbot for university related FAQs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1525-1530). IEEE.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of

data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

Razzak, M. I. (2015). Automatic detection and classification of malarial parasite. *International Journal of Biometrics and Bioinformatics (IJBB)*, *9*(1), 1–12.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In *Educational Data Mining 2008*.

Saa, A. A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science & Applications*, *1*, 212-220.

Sala al-Din Abdullah, A. (2016). Using Data mining techniques to identify the causes of deaths in al-gedaref hospital. *European Journal of Computer Science and Information Technology*, *4*(2), 1–8.

Santoso, H. A., Winarsih, N. A. S., Mulyanto, E., Sukmana, S. E., Rustad, S., Rohman, M. S., ... & Firdausillah, F. (2018, September). Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services. In *2018 International Seminar on Application for Technology of Information and Communication* (pp. 417-423). IEEE.

Sharma, V., Kumar, A., Lakshmi Panat, D., & Karajkhede, G.(2015). Malaria outbreak prediction model using Machine Learning. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, *4*.

Shawar, B. A. (2008). Chatbots are natural web interface to information portals. In *Proc. of INFOS 2008 the Sixth International Conference on Informatics and Systems* (Vol. 2008).

Shinde, R., Arjun, S., Patil, P., & Waghmare, P. J. (2015). An intelligent heart disease prediction system using K-Means

Clustering and Naïve Bayes Algorithm. *International Journal of Computer Science and Information Technologies*, 6(1), 637–639.

SigmaPlot. (2014). ROC Curves Analysis. *SigmaPlot*, 1–20. Retrieved from http://www.sigmaplot.com/products/sigmaplot/ROC_Curves_Analysis.pdf

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

Sordo, M., & Zeng, Q. (2005). On sample size and classification accuracy: a performance comparison. *Biological and medical data analysis*, 193-201.

Stauffer, W., & Fischer, P. R. (2003). Diagnosis and treatment of malaria in children. *Clinical infectious diseases*, *37*(10), 1340-1348.

Taneja, A. (2013). Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, 6(4), 457-466.

Tarekegn, G. B. (2016). Application of data mining techniques to predict students placement in to Departments. *International Journal of Research Studies in Computer Science and Engineering*, *3*(2), 10–14.

Tribhuvan, A. P., Tribhuvan, P. P., & Gade, J. G. (2015). Applying Naive Bayesian classifier for predicting performance of a student using Weka. *Advances in Computational Research*, *7*(1), 239.

Weißensteiner, A. A. A. (2018). Chatbots as an approach for a faster enquiry handling process in the service industry. *Signature*, *12*, 04.

Zorman, M., Štiglic, M. M., Kokol, P., & Malčić, I. (1997). The limitations of decision trees and automatic learning in real world medical decision making. *Journal of Medical Systems*, *21*(6), 403-415