Chatbot: An automated conversation system for the educational domain

1st Anupam Mondal
Computer Science and Engineering
Jadavpur University
Kolkata, India
link.anupam@gmail.com

2nd Monalisa Dey
Computer Science and Engineering
Jadavpur University
Kolkata, India
monalisa.dey.21@gmail.com

3rd Dipankar Das
Computer Science and Engineering
Jadavpur University
Kolkata, India
dipankar.dipnil2005@gmail.com

4th Sachit Nagpal
Big Data and Analytics
S P Jain School of Global Management
Mumbai, India
sachitnagpal@gmail.com

5th Kevin Garda
Big Data and Analytics
S P Jain School of Global Management
Mumbai, India
kevingarda7@gmail.com

***Abstract*—Speech and textual information play a crucial role in communicating between humans. An article in "The New York Times" published that now-a-days the adults are spending more than 8 hours a day on screens of computers or mobiles. So the major communication between humans is conducted through web applications such as WhatsApp, Facebook, and Twitter etc as a form of speech and textual conversation. In the present paper, we have focused on designing a textual communication application namely chatbot in the educational domain. The proposed chatbot assists in answering questions provided by the users. To develop the system, we have employed an ensemble learning method as random forest in the presence of extracted features from our prepared dataset. Besides, the validation system offers an average F-measure 0.870 score on various K-values under random forest for the proposed chatbot. Finally, we have deployed the proposed system in a form of telegram bot.**

***Index Terms*—Chatbot, Educational Domain, Question answering, Machine Learning**

I. Introduction

Chatbot or Chatterbot term was introduced by Michael Mauldin (creator of the first Verbot, Julia) in 1994 to describe the conversational programs. The conversational programs provide support in designing various messenger-based applications such as Google, Facebook, and WhatsApp. Besides, the chatbot could help to improve responsiveness, increase availability, and reduce dependence man-power in today's world of automation. Responsiveness presents the quality of reacting quickly and positively at the time of multiple conversations in a particular time. So it is quite possible that a person may not be able to give an immediate response. Hence, to improve the responsiveness, the chatbot has been introduced by the researchers [1], [2].

We have observed that the interested persons as learners and students etc are always trying to communicate with a person of an educational organization via web. But everytime is not possible to answer or reply their queries due to lack of man-power and time difference between countries. So, we are motivated to design an automated conversational system namely chatbot in the domain of education. In order to design the chatbot, we have observed the following research questions:

A. Data Collection: How to collect a relevant dataset to design an automated chatbot? We have answered this question by collecting around 1500 of questions and their corresponding answers from an educational organization in unstructured form.

B. Data Preprocessing: How to convert a structured data from the crawled unstructured data? The crawled data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. To overcome this problem, we have gone through a series of steps during preprocessing such as data cleaning, integration, transformation, reduction, and discretization [3]. Thereafter, we have prepared around 1000 pairs of questions and answers as our experiment.

C. Conversation Response Selection: How to decide the response type of the conversations? Primarily, we have observed the conversations are distinguishing according to the length, coherent personality, and intention and diversity. The length of the conversation helps to decide the response of the system like linguistic or physical context. Coherent personality assists in producing consistent answers to semantically identical inputs. For example, you want to get the same reply to "How old are you?" and "What is your age?". The intention and diversity are taking an important role to produce a generic response viz. "That's great!" or "I don't know". In order to address this challenge, we have considered all the above-mentioned responses for the proposed chatbot.

D. Chatbot Building: How to design an automated conversational system namely chatbot? To design the chatbot, we have employed an ensemble learning method known as random forest on our identified features. The features are the number of words in a question, question type (e.g. why, what, and where), nouns, number of nouns, verbs, number of verbs,

Term Frequency (TF), Inverse Document Frequency (IDF), and TF-IDF. These features and random forest model helped in developing a retrieval and generative chatbot system.

E. Validation: How to evaluate the proposed chatbot? In order to validate the output of the system, we have applied random forest approach with the weighted and macro model, which provides the precision, recall, F-measure, and an accuracy score for various combinations of the dataset.

The contribution of the paper is to design an automated chatbot in the educational domain, which satisfies the above-mentioned research questions. Besides, we have also deployed the proposed chatbot under telegram bot environment.

The overall structure of the paper is as follows. Section II presents the related work carried out in chatbot. Section III and Section IV describe the model building and its evaluation process. Section V describes the concluding remarks and future scope of the research.

II. Related Work

Chatbot aims to make communication between a human and machine such as computer and mobile [4]. Recently a considerable amount of promising work has been conducted in the area of chatbot design. O. V. Deryugina [2] presented a detailed survey on the history of the chatbot, their applications, and the first designs of such systems. Bordes et. al [5], presented an intelligent question answering system which achieved competitive results. They trained their model using low-dimensional embedding of words and knowledge base constituents and used these representations to score natural language questions against candidate answers. Pareira et.al [1] presented an overview of the chatbot early contributions and tried to map those with the current works in the human-machine interaction research.

We have noticed that the chatbot related research is mainly distributed in the following areas, (i) different approaches (e.g. retrieval and generative), (ii) length of the conversation, and (iii) according to the domain (e.g. open and closed).

Retrieval-based models use a repository of predefined responses and a heuristic to pick an appropriate response based on the input and context [6], [7]. The heuristic could be as simple as a rule-based expression match or as complex as an ensemble of machine learning classifiers [8]. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns.

Besides, generative models don't rely on predefined responses. They generate new responses from the scratch. Generative models are typically based on Machine Translation techniques where we "translate" from an input to an output (response) [9]. Both approaches have some obvious pros and cons. Due to the repository of handcrafted responses, retrieval-based methods don't make grammatical mistakes. However, they may be unable to handle unseen cases for which no appropriate predefined response exists. For the same reasons, these models can't refer back to contextual entity information like names mentioned earlier in the conversation. Generative models are "smarter", however, these models are hard to train, are quite likely to make grammatical mistakes (especially on longer sentences), and typically require huge amounts of training data. The longer the conversation, the more difficult to automate it [10]. On one side of the spectrum are Short-Text Conversations (easier) where the goal is to create a single response to a single input [7].

In an open domain setting, the user could take the conversation anywhere. There isn't necessarily have a well-defined goal or intention [11]. The infinite number of topics and the fact that a certain amount of world knowledge is required to create reasonable responses makes this hard problem. On the other-side, a closed domain setting, the space of possible inputs and outputs is somewhat limited because the system is trying to achieve a very specific goal [12]. Technical customer support or shopping assistants are examples of closed domain problems.

III. Methodology

Chatbot refers a computer program which conducts a textual or audio based conversation between humans via web [13], [14]. Such programs are typically used in dialog systems for various applications including information acquisitions, customer services, and questions answering etc. In order to design this chatbot, we have used sophisticated natural language processing approaches over simple keyword or similar word pattern matching from a predefined database. The following subsections discuss the data preparation, feature extraction, and model building in details.

A. Data Preparation:

A label data is essential in building an automated chatbot application, which we could present a question answering system. We have initially collected around 1500 number of educational conversations from an educational organization. For maintaining privacy, we will not disclose the name of the organization. Thereafter, we have preprocessed the crawled data and converted to structured data.

The preprocessing steps are data cleaning, integration, transformation, reduction, and discretization [3]. So we have written python [1] scripts to execute these steps. Data cleaning helps to remove the noise and inconsistencies from the crawled data, whereas integration step assists in combining various pairs of questions and answers from multiple sources. We have processed the data transformation and reduction steps, which normalize and eliminate the redundant data from the initial dataset. Our experimental dataset contains around 1000 pairs of unique questions and answers.

Afterwards, we have extracted the features from the experimental dataset to build the proposed model. Finally, we have applied discretization step of preprocessing on the extracted features to discover the knowledge to improve the quality of the data.

[1] https://www.python.org/

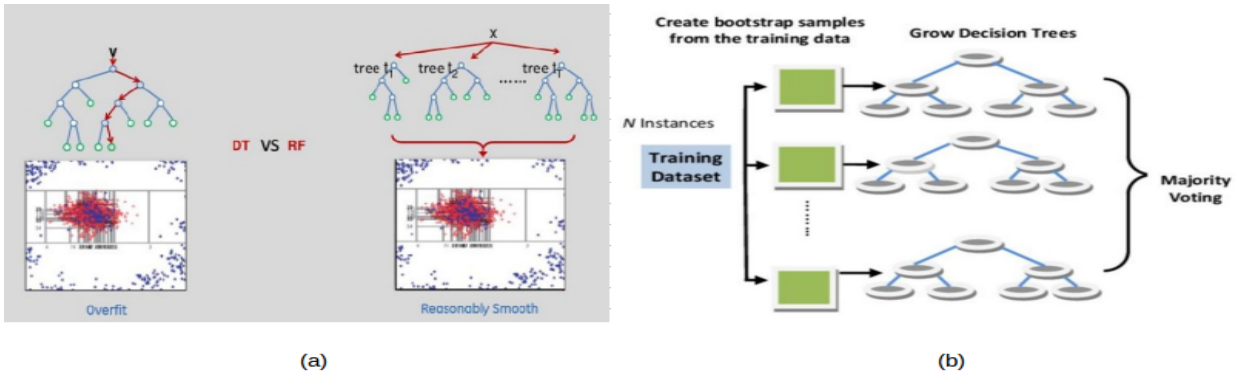

Fig. 1. The above figure depicts (a) difference between decision tree (DT) and random forest (RF) and (b) Standard steps of any random forest model.

B. Feature Extraction:

The chatbot is presented as a classification problem according to the response of chatbot such as retrieval and generative. A bag-of-words (BoW) model is essential to design a retrieval-based chatbot, its response is primarily generated from the predefined BoW. On the other hand, the contextual and semantic features help to develop a generative chatbot. The generative chatbot is not directly copied the response from the BoW. So the feature extractions are taking a crucial role to build the retrieval and generative both combined chatbot.

Hence, we have extracted various features from 1000 pairs of questions and answers, which refer our training dataset. The features are the number of words in a question, question type (e.g. why, what, and where), nouns, number of nouns, verbs, number of verbs, Term Frequency (TF), Inverse Document Frequency (IDF), and TF-IDF. The extracted nouns and verbs assist in preparing a BoW model. Besides, rest of the mentioned features help to design the proposed chatbot system.

These features have been extracted through our written python scripts (python 2.7 version) in the presence of nltk package [2] and WordNet lexicon [3]. The WordNet lexicon supports to identify the synonyms and hyponyms for the extracted words or phrases (nouns and verbs), which enrich the BoW.

For example, the following question is asked by the visitor, from where we identified the features 1. length as ”12”, 2. question type as ”may”, 3. nouns as ”fees, BDAP, and S_P_Jain”, 4. number of nouns as ”3”, 5. verbs as ”know”, and 6. number of verbs as ”1”. Thereafter, we have applied these features in an ensemble learning to build the model, which provides the following chatbot output.

Visitor: ”May I know the fees of the BDAP at S P Jain?”
Chatbot: ”The fees of the BDAP is INR 5,00,000.”

The following subsection discusses, how we have employed the extracted features to design an educational chatbot.

C. Model Building:

According to the response classification such as retrieval and generative of the chatbot, we have distributed the above-mentioned features as semantic and contextual. The semantic features assist in identifying the response keywords, whereas contextual features help to carry the contextual knowledge from questions to answers. Hence, we have used an ensemble learning approach known as random forest or random decision forest. The learning approach is constructing a multitude of decision trees at the time of training and predicting the classes of response. Random forest uses decision trees but follows a different approach. The decision trees are growing as a ”single” very deep tree, whereas random forest relies on aggregating the output from many ”shallow” trees. Figure 1 shows the difference between decision tree (DT) and random forest (RF). Hence, we have selected random forest, which additionally helps to overcome the over-fitting problem of each decision tree.

Moreover, the random forest is a bagging algorithm that aims to reduce the complexity of models that over-fit the training data. Random Forests train a number of decision trees from bootstrap samples from the training set with replacement. In addition to the bootstrap samples, the algorithm also draws a random subset of features for training the individual trees in contrast to regular bagging where each tree is given the full set of mentioned features. The following steps are discussed the working principle of the random forest.

[2]http://www.nltk.org/
[3]https://wordnet.princeton.edu/

Fig. 2. A sample output of the proposed chatbot.

K-fold values	Weighted			Macro		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
K=10	0.864	0.882	0.867	0.869	0.888	0.874
K=20	0.867	0.884	0.868	0.860	0.877	0.861
K=30	0.873	0.889	0.875	0.869	0.886	0.867
K=40	0.863	0.886	0.867	0.871	0.888	0.872
K=50	0.869	0.888	0.870	0.862	0.883	0.865

TABLE I
AN EVALUATION OF THE PROPOSED CHATBOT USING PRECISION, RECALL, AND F-MEASURE FOR WEIGHTED AND MACRO RANDOM FOREST.

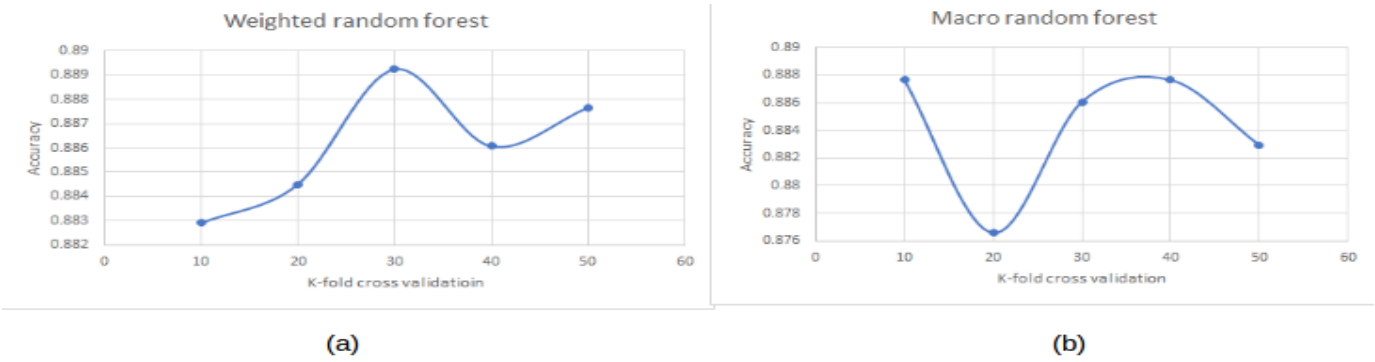

Fig. 3. Accuracy verses K-fold cross validation value (a) for weighted random forest (b) for macro random forest.

Step-1: Each tree is trained on roughly 2/3 of training data, which randomly replace the original data. This part of the data is used as a training data for growing the tree.
Step-2: Thereafter, randomly selected some prediction variables as m, which is used to split the node. The value of m is held constant during the forest growing.
Step-3: The rest 1/3 of training data helps to calculate the misclassification rate or Out Of Bag (OOB) error rate. Aggregate error from all trees to determine overall OOB error rate for the classification.
Step-4: Each tree provides a classification on 1/3 of training data and its OOB, which helps to decide the class after voting between them. Finally, the forest selects the classification output with most votes over all the trees in the forest.

The model provides better predictive classes and faster due to better variance-bias trade-offs and each tree learns only from a subset of features respectively. Finally, the random forest and the extracted features help to extract the chatbot response for the corresponding question of the visitors in a single environment as shown in Figure 2.

IV. VALIDATION AND DEPLOYMENT

To validate the proposed chatbot, we have used weighted and macro based random forest approaches with K-fold cross-validation. We have varying K value from 10 to 50, which showing a saturated F-measure score for both weighted and macro model. The F-measure has been calculated with the help of standard evaluation matrices such as precision and recall score. Table I presents the distribution of precision, recall,

and F-measure score for the proposed system under both of the models.

Thereafter, we have also measured the accuracy score of classification of responses for the chatbot. Figure 3 (a) and (b) show accuracy scores against various K-fold cross-validation value for weighted and macro random forest respectively.

We have also deployed the proposed chatbot on Telegram bot to design a messaging platform. In order to deploy, we have introduced the Telegram bot API. Hence, we have created an account with Telegram and registered using Botfather [4]. The account helps to send and receive messages or commands through the Telegram bot. Thereafter, we have applied the following steps to complete the deployment process.

The deployment of the proposed chatbot assists in communicating between a visitor and a machine in the domain of education.

Step-1: Requests module of python was used, along with Telegram's API, to deploy the chatbot on Telegram Messenger.
Step-2: Every 0.5 seconds, we fetch new, unseen messages, if there are any.
Step-3: In case of new messages, the sender ID and messages are retrieved from the API.
Step-4: The new message along with sender's unique ID, is input as parameters to the chatbot function. The unique ID is input to fetch the features from past conversation with the user.
Step-5: The output of the chatbot, which is essentially the response to the query, is sent to the respective user through the requests module and Telegram API.

V. Conclusion and Future Scope

The research was primarily focused on developing a chatbot in the educational domain. To design the chatbot, we have prepared a training data from the crawled data, which contains around 1000 unique pairs of questions and answers. Thereafter, we have identified various features from the data and processed through random forest algorithm. The ensemble learning based random forest helps to enhance the classification of the response classes of the proposed chatbot. The chatbot is able to respond most of the queries with an accuracy of 88.60%. Besides, we have also deployed the proposed chatbot on Telegram bot for communicating between the visitor and machine.

In future, we will try to introduce an open domain chatbot with various additional features, which will help in designing a quality conversation system.

[4]http://telegram.me/bot

References

[1] M. J. Pereira, L. Coheur, P. Fialho, and R. Ribeiro, "Chatbots' greetings to human-computer communication," *arXiv preprint arXiv:1609.06479*, 2016.
[2] O. Deryugina, "Chatterbots," *Scientific and Technical Information Processing*, vol. 37, no. 2, pp. 143–147, 2010.
[3] J. S. Malik, P. Goyal, and A. K. Sharma, "A comprehensive approach towards data preprocessing techniques & association rules," in *Proceedings of The4th National Conference*, 2010.
[4] B. A. Shawar and E. Atwell, "Chatbots: are they really useful?" in *LDV Forum*, vol. 22, no. 1, 2007, pp. 29–49.
[5] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," *arXiv preprint arXiv:1406.3676*, 2014.
[6] B. Setiaji and F. W. Wibowo, "Chatbot using a knowledge in database: Human-to-machine conversation modeling," in *Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference on*. IEEE, 2016, pp. 72–77.
[7] H. Wang, Z. Lu, H. Li, and E. Chen, "A dataset for research on short-text conversations." in *EMNLP*, 2013, pp. 935–945.
[8] D. Britz, "Deep learning for chatbots, part 1–introduction," 2017.
[9] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models." in *AAAI*, 2016, pp. 3776–3784.
[10] J. Hill, W. R. Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations," *Computers in Human Behavior*, vol. 49, pp. 245–250, 2015.
[11] N. Asghar, P. Poupart, J. Xin, and H. Li, "Online sequence-to-sequence reinforcement learning for open-domain conversational agents," *arXiv preprint arXiv:1612.03929*, 2016.
[12] A. Kerly, R. Ellis, and S. Bull, "Calmsystem: a conversational agent for learner modelling," *Knowledge-Based Systems*, vol. 21, no. 3, pp. 238–246, 2008.
[13] I. Ahmed and S. Singh, "Aiml based voice enabled artificial intelligent chatterbot," *International Journal of u-and e-Service, Science and Technology*, vol. 8, no. 2, pp. 375–384, 2015.
[14] S. du Preez, M. Lall, and S. Sinha, "An intelligent web-based voice chat bot," in *EUROCON 2009, EUROCON'09. IEEE*. IEEE, 2009, pp. 386–391.