Advances in Management & Applied Economics, Vol. 15, No. 1, 2025, 47-60
ISSN: 1792-7544 (print version), 1792-7552(online)
https://doi.org/10.47260/amae/1513
Scientific Press International Limited

A Classification Intelligent Question Answering Model for Retrieval-Based Chatbots

Chihli Hung[1] and Ming-Hsuan Wu[1]

Abstract

Intelligent question answering (QA) models or chatbots automatically provide appropriate responses to questions posed by users. In terms of generating continuous responses, they are divided into generative and retrieval-based approaches. For retrieval-based QA models, the key issue is how to reduce the search space. This research focuses on a retrieval-based approach and proposes a classification intelligent question answering (CIQA) model. The CIQA model contains two stages, namely a question classification stage and an answer prediction stage. The first stage consists of building a classification ensemble based on a training set. The second stage uses the first stage classification ensemble to determine the appropriate categories for a test set and selects an appropriate deep learning QA model based on a chosen category. A new benchmark dataset for chatbot, SQuAD (Stanford question answering dataset) 2.0, is used to evaluate performance. Based on the outcome of our experiments, the proposed CIQA model outperforms the baseline model and demonstrates the feasibility of the proposed approach.

JEL classification numbers: M15, O35.
Keywords: Question answering, Ensemble learning, Deep learning, Retrieval-based QA models.

[1] Department of Information Management, Chung Yuan Christian University, Taiwan.

Article Info: *Received:* October 29, 2024. *Revised*: November 13, 2024.
Published online: November 20, 2024.

1. Introduction

This research proposes a classification intelligent question answering (CIQA) model, which combines deep learning and ensemble learning to improve the performance of retrieval-based question answering models. A question answering (QA) model is one that automatically provides appropriate answers to user questions (Zhang et al., 2024; Zheng et al., 2023). It is considered one of the most challenging tasks in the field of natural language processing (Wu et al., 2018). With the development of social networks, question answering models have been extended to community question answering models (Zhou et al., 2018) and chatbots (Huang et al., 2007). A community QA model is a discussion board for an online community site. When a poster initializes with a question or topic, members of the community follow up to answer or continue to ask questions. A chatbot is an application that provides interactive questions and answers between users and the chatbot itself through text or voice. The prevalence of social networks and third-party payment technologies have ushered in the era of conversational economy. Intelligent QA models, such as Line@, Siri, Google, and Alexa chatbots, have become a necessity on social networks.

For QA or chatbot models, the generation of continued conversations can be divided mainly into generative approaches (Casheekar et al., 2024; Kim et al., 2023; Pandey and Sharma, 2023; Pathak et al., 2025) and retrieval-based approaches (Wu et al., 2018; Ma et al., 2022; Moore et al., 2023). The generative approach uses a large volume of conversations, language models and deep learning algorithms to generate a suitable continued dialogue. The retrieval-based approach selects the appropriate response by retrieving historical conversations, so it can be viewed as an extended form of information retrieval system (Abdi et al., 2016). The main difference between them is that the generative approach is able to generate a new response while the retrieval-based approach can only respond with one that has been used previously. In terms of comparison between these two approaches, Pandey and Sharma (2023) evaluate the performance between six retrieval-based and generative chatbots. Their experimental results demonstrate that the generative approaches outperform those based on retrieval. Some researchers argue that retrieval-based chatbots are superior to generative ones in response fluency, informativeness, easy construction, and evaluation (Wu et al., 2018, Ma et al., 2022, Tao et al., 2021). In comparison with generative approaches, retrieval-based approaches can be constructed at a lower cost and evaluated more objectively. We therefore focus on retrieval-based approaches and propose the CIQA model. Some retrieval-based approaches focus on selecting a proper response from historical conversations (Wu et al., 2018, Ma et al., 2022). The framework of a retrieval-based QA model usually consists of four parts, which are question analysis, semantic understanding, information retrieval, and answer extraction (Yu et al., 2018). Of these, the main task of question analysis is to find related topics and narrow down the search space of historical conversations. Therefore, this research proposes a method of question classification to effectively reduce the number of answer candidates by classifying

questions into categories. After reducing the search space of the answers, the speed and accuracy of answer extraction can be improved.

In the field of question classification tasks for QA models, most studies use a single classifier for questions (Wu et al., 2018, Qian et al., 2021). Ensemble learning can improve the classification performance of a single classifier (Sagi and Rokach, 2018), but it is rarely used in retrieval-based QA models. This research integrates ensemble learning with deep learning to improve the answer accuracy for retrieval-based QA models. More specifically, this research uses SQuAD 2.0 (Rajpurkar et al., 2018) as the dataset. As the recurrent neural network (RNN) suffers from vanishing and exploding gradient issues, its extended versions, i.e., long short-term memory (LSTM) and gated recurrent unit (GRU) are used. Based on these deep learning methods, a bagging ensemble is also used. Thus, we evaluate four classification methods, which are two single classifiers and two ensemble classifiers, for a more robust composition. Finally, we combine the best three classifiers using a weighted average strategy to obtain better answers for the proposed retrieval-based QA model. The main contributions of this research are three-fold:

1) We propose a classification intelligent question answering (CIQA) model, which uses a 2-stage strategy to narrow down the search space for retrieval-based chatbots.
2) We integrate deep learning and ensemble learning in order to achieve more generalized results, and even improve the performance of a single classifier.
3) We evaluate the proposed CIQA by a new benchmark dataset, i.e. SQuAD (Stanford question answering dataset) 2.0, and demonstrate that the CIQA outperforms the baseline model.

The rest of the paper is structured as follows. Section 2 briefly reviews related work. Chapter 3 presents the methodology used in this paper. The proposed CIQA model consists of two stages, namely a question classification stage and an answer prediction stage. Chapter 4 presents the experimental design and results. The final chapter provides conclusions and some possible future research works.

2. Related Work

A question answering (QA) model is a downstream application of natural language processing, which allows users to ask questions in the form of natural language and provides proper responses. Depending on their purpose, question answering models can be divided into three types, which are task-based, chitchat, and hybrid QA models. A task-based QA model uses keyword comparison methods for specific tasks. According to pre-designed processes, some QA models, for example, Line@, Facebook, and financial management chatbots, use various components of user interfaces, such as radio buttons, checkboxes, menus, and so on to achieve their specific goals. A chitchat QA model uses rules or natural language processing techniques and combines these techniques with artificial intelligence methods to

simulate human chat behavior. For intelligent QA or chatbot systems, most research studies focus on generation or selection of continued dialogues (Kim et al., 2023). These systems can be divided into generative models (Kim et al., 2023; Pandey and Sharma, 2023) and retrieval-based models (Wu et al., 2018, Ma et al., 2022; Moore et al., 2023). The generative model creates a new continued dialogue using some deep learning techniques. A retrieval-based QA model treats questions and answers as an information retrieval task. It uses information retrieval related techniques to look for an appropriate answer or a continued dialogue from historical conversions. The techniques for selection of continued dialogues are quite diverse. Traditional techniques include using the artificial intelligence markup language (AIML), natural language processing (NLP), Markov chain model, and ontology. For example, Weizenbaum (1996) uses natural language processing and rule-based comparison methods to develop the first chatbot, ELIZA. Wallace (2009) uses AIML and advanced comparison methods to propose the ALICE chatbot. Al-Zubaide and Ayman (2011) combine WordNet ontology, natural language processing and AIML to design chatbots. Deep learning has recently extended the architecture and capabilities of traditional neural networks for the development of new solutions in many fields (Géron, 2017; Goodfellow, 2016). A recurrent neural network (RNN) with the ability to deal with words sequentially can be used for question answering or chatbot models. For example, Lowe et al. (Lowe and Pow, 2016) propose a parallel RNN architecture, which uses two RNNs to train the previous and the latter words respectively, to build an Ubuntu chatbot. A convolutional neural network (CNN) is a deep learning architecture for content-based image retrieval (CBIR) tasks and is extended to deal with text (Zhou et al., 2022). For example, Qian et al. (2021) and Wang et al. (2019) use the CNN for retrieval-based chatbots. Zhou et al. (2018) propose the recurrent convolutional neural network (RCNN), which is an integration of the CNN and RNN, to handle the question answering task.

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is an extended model of RNN, which uses an input gate, an output gate, and a forget gate to improve the performance of RNN. The Google research team uses LSTM to propose an automatic email reply system (Kannan et al., 2016). Some studies demonstrate that LSTM is more suitable than RNN for designing chatbot systems (Kadlec et al., 2015). A single LSTM is extended to be bidirectional, namely biLSTM (Graves and Schmidhuber, 2005). Gu et al. (2019) use biLSTM in the sentence encoding and aggregation layers for their retrieval-based chatbots. Based on biLSTM, Ma et al. (2022) propose a global and local interaction matching model (GLIMM) to match a response to the context-knowledge pair. The gated recurrent unit (GRU) is also an extension of RNN (Cho et al., 2014) and has been also used for building retrieval-based chatbots (Pandey and Sharma, 2023; Tao et al., 2019). The difference with LSTM is that GRU uses two parameters, namely an update gate and a reset gate, for recall and forgetting respectively.

On the other hand, ensemble learning uses the integration results of multiple classifiers to present the final decision, also known as a multi-classifier system,

which can achieve more generalized results, and even improve the performance of a single classifier (Sagi and Rokach, 2018). According to the classification algorithm and data sampling approach, ensemble learning is divided into three methods, which are stacking, bagging and boosting. Stacking combines a variety of different classification algorithms. Each classification algorithm uses the same training data and adopts a majority decision method to determine integration results. It is generally recognized that although this method has the advantages of individual classification algorithms, it also includes their disadvantages. Hung and Chen (2009) propose a selective stacking ensemble learning method for bankruptcy prediction, which attempts to inherit the advantages of individual algorithms and reduce the influence of the algorithms' shortcomings. Bagging uses the same classification algorithm for each individual classifier, and uses a bootstrap sampling method to randomly take samples from the dataset as the training set for individual classifiers. Similar to bagging, boosting uses the same classification algorithm for each individual classifier. Unlike bagging, the weights of individual classifiers vary based upon their performance. Specifically, boosting increases the weight of data with higher error rates to provide more opportunities for re-learning. Ensemble learning has been widely used in the field of machine learning and has achieved strong results, but it is used less in the QA models (Bühlmann, 2012). Banerjee and Bandyopadhyay (Banerjee and Bandyopadhyay, 2013) use bagging, boosting, and stacking in a Bengali classification task. The experimental results show that stacking is slightly better than bagging, and bagging is slightly better than boosting. However, Lei et al. (2009) compare bagging with boosting in the tourist question answering system and find that bagging is better than boosting.

In the fields of question answering and ensemble learning, there is a lack of research on ensemble learning for QA, and there are very few studies which use an ensemble of deep neural network classifiers. Therefore, this research integrates bagging with LSTM and GRU to improve the accuracy of responses for retrieval-based QA and proposes a classification intelligent question answering (CIQA) model.

3. Methodology

The overall methods are divided into two stages, namely a question classification stage and an answer prediction stage, as shown in Figure 1. The purpose of the first stage is to look for a suitable ensemble decision for use in the second stage. Based on the question part of the dataset, the first stage selects the best combination of classification decisions from two deep learning techniques (i.e., LSTM and GRU) and their bagging strategies. In the second stage, the training set is divided into seven training subsets according to the topics. Each training subset is used to build a deep learning QA model for a specific topic. The ensemble of trained classifiers built in the first stage is used for the test set to determine the proper deep learning QA model. Thus, we use several smaller deep learning QA models instead of a larger deep learning model for the QA task. The detailed steps are as follows.

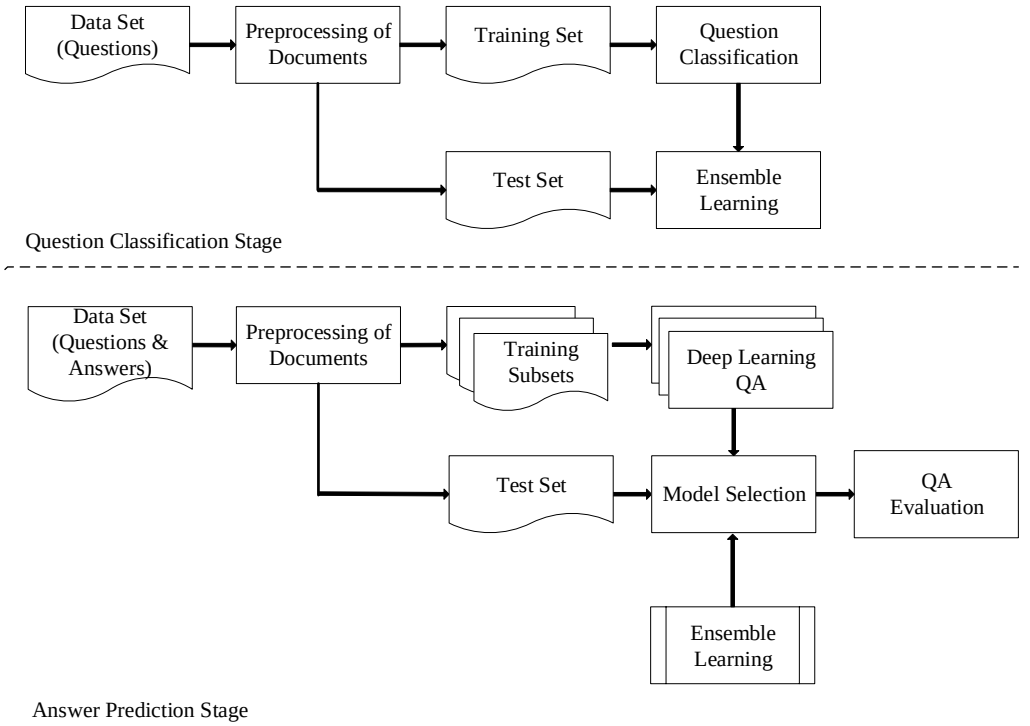

Figure 1: The conceptual framework of the classification intelligent question answering

3.1 Question Classification Stage

3.1.1 Preprocessing of Documents

SQuAD (Stanford question answering dataset) is a new benchmark dataset for question answering tasks (Rajpurkar et al., 2018; Guven and Unalir, 2022; Konrad, 2018; Rajpurkar et al., 2016). It consists of more than 10 million questions posed in Wikipedia articles, where the answer to each question is a paragraph of text from the corresponding article. The newest version of SQuAD is SQuAD 2.0, which is used in this research. Each data sample in the SQuAD 2.0 dataset contains four fields including article, title, question, and answer. SQuAD 2.0 has a total of 442 titles, which are divided into 130,319 questions. Among these questions, 43,500 have no corresponding answers. Since the main purpose of this research is to design a retrieval-based question answering model, only 86,819 questions with answers are used. The preprocessing of documents includes tokenization and removal of punctuation.

3.1.2 Question Classification

The purpose of the question classification stage is to provide suitable categories for questions of the test set in the answer prediction stage, so that the appropriate deep learning QA models can be selected. Konrad (2018) divides SQuAD 1.0 into seven question categories according to the title of the article, such as organization,

personality, geographic location, abstract concept, art, sport, and other. Thus, following the approach of Konrad (2018), we divide SQuAD 2.0 into these seven question categories according to the title of the article. Table 1 shows the distribution of seven categories for the dataset. In the question classification stage, only the question and classification label are used. We randomly select 80% of the dataset as the training set and the rest as the test set.

Table 1: Distributions of dataset in the question classification stage

Title Category	Dataset #	Dataset %
Organization	6,206	7.15
Personality	12,977	14.95
Geographical Location	24,290	27.98
Abstract Concept	8,978	10.34
Art	3,625	4.17
Sport	2,749	3.17
Other	27,994	32.24
Total	86,819	100.00

This research uses two renowned deep learning classification models, i.e. LSTM and GRU, as our based classification models. We use a deep learning structure with six layers (Figure 2). The first layer is an input layer. Among the 86,819 data samples, 20% of the training set is the validation set. The second layer is a Keras embedding layer. The parameters of input_dim, output_dim, and input_length are 20,000, 200, and 1000, respectively. The Keras embedding layer uses sequences of words from the corpus without the training of an artificial neural network. It is provided from the Python Keras package and has been used for many deep learning models, such as RNN, LSTM, GRU (Pandey and Sharma, 2023). The third layer is a recurrent layer. Long short-term memory (LSTM) and gated recurrent unit (GRU) are used for this layer. The fourth layer is a dense layer, which is a fully connected layer of 256 units. The fifth layer is a dropout layer. Its function is to discard some neurons in the neural network to avoid overfitting. The dropout rate is 0.5. The final layer is an output layer of 7 units. The Softmax activation function is used for this layer. It should be noted that the above parameters are obtained from suitable results in different parameter combinations in preliminary experiments.

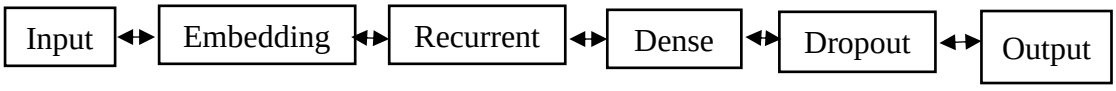

Figure 2: A deep learning structure for the question classification stage

3.1.3 Ensemble Learning

The bagging strategy is used to further improve the classification performance for the two best deep learning models. Thus, we have four deep learning models, which are LSTM, GRU, LSTM-bagging and GRU-bagging. In terms of the bagging models, the number of bags is three. We randomly select three subsets of training set for each bag, which contains around 55,600 samples. All four deep learning models use the same test set, which contains 17,186 samples. The average accuracy of the test set represents the performance of the specific bagging model. Finally, an average weighted approach is used to obtain the three best deep learning models as the final result in this stage. According to Table 2, LSTM, LSTM-bagging and GRU-bagging are the winners, which will be used in the second stage. Thus, the first stage can assess the appropriate topic category for the unseen question.

Table 2: Performance of models in the question classification stage evaluated by accuracy (%)

Model	Training Set	Test Set
LSTM	74.31	66.54
GRU	73.44	66.07
LSTM-bagging	74.43	67.10
GRU-bagging	75.51	67.18

3.2 Answer Prediction Stage

3.2.1 Preprocessing of Documents

At this stage, an answer is derived in response to a user question. As a QA task is a supervised learning task, a correct answer is labeled 1 and an incorrect answer 0. For the entire dataset, we have 86,819 question-answer pairs whose labels are 1. We duplicate the 86,819 questions and randomly choose their wrong answers from different topic categories. Thus, we have 173,638 question-answer pairs as the dataset. Based on seven topic categories, the dataset is divided into seven data subsets. Like the document preprocessing in the question classification stage, the answer prediction stage includes tokenization and punctuation removal. Each data subset is divided into training set and test set in the ratio of 8:2. Table 3 shows the distribution of seven categories for the test set. Unlike the question classification stage which uses question and classification labels, this stage uses question, answer and answer labels.

Table 3: Distributions of test set in the answer prediction stage

Title Category	Test Set #	Test Set %
Organization	2,666	7.76
Personality	5,122	14.90
Geographical Location	8,708	25.33
Abstract Concept	3,352	9.75
Art	1,294	3.76
Sport	718	2.09
Other	12,512	36.40
Total	34,372	100.00

3.2.2 Deep Learning QA

For each training subset, we build a deep learning QA structure of seven layers (Figure 3). This structure is similar to the deep learning structure in the first stage. For the deep learning QA structure, questions and answers are imported separately. They have their own input layer, embedding layer, and recurrent layer. A Keras embedding approach is used for the embedding layer. Long short-term memory (LSTM) and gated recurrent unit (GRU) techniques are used for the recurrent layer. The fourth layer is a concatenated layer, which concatenates two recurrent layers from questions and answers. The fifth and sixth layers are dense and dropout layers, respectively. The final layer is an output layer of two units so the Sigmoid activation function is used. All parameters are the same as those in the first stage.

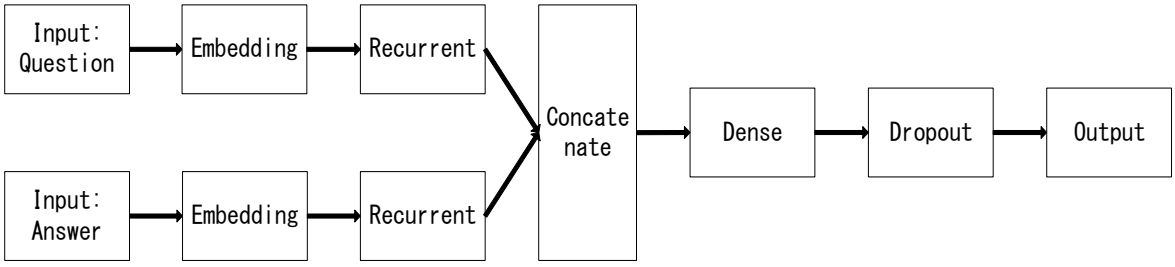

Figure 3: A deep learning QA structure for the answer prediction stage

3.2.3 Model Selection

In the previous step seven deep learning QA models are built, and each model is for one specific topic category. For the questions in the test set, the first stage ensemble learning model is used to choose a suitable question category. Based on this category, the specific suitable deep learning QA model of the second stage is selected to determine the appropriate answer.

3.2.4 QA Evaluation

This research uses accuracy (ACC) as an evaluation metric (1), which is very common in the field. TP (true positive) indicates the number of data samples for which the answer to the question is correct and for which the model also predicts it as a correct answer. TN (true negative) represents the number of data samples for which the answer to the question is incorrect and for which the model also predicts to be an incorrect answer. FP (false positive) indicates the number of data samples where the answer to the question is incorrect and the model predicts the correct answer. FN (false negative) represents the number of data samples where the answer to the question is correct but the model predicts a wrong answer.

$$\text{ACC}=\frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

4. Experiments

The traditional retrieval-based QA model generally uses the training set to train the classification model, and uses the test set to verify the retrieval performance. Thus, this model tries to retrieve a proper answer from a whole knowledge resource. We treat this model as a baseline in this research. The baseline model uses the same document preprocessing as our proposed classification intelligent question answering (CIQA) model. After document preprocessing, the same deep learning QA structure in Figure 3 is also used. Our proposed CIQA model uses a two-stage approach to narrow down the search space. We present the performance evaluated by the measure of accuracy in Table 4. Specifically, a *t*-test is used to test if there is a significant difference between our proposed model and the baseline model. P-value is presented in brackets. For results shown in Table 4, the deep learning model with a GRU recurrent layer slightly outperforms the model with an LSTM recurrent layer. All of the proposed models perform significantly better than the baseline models at a significance level, α, of 0.01.

Table 4: The performance for the baseline and BIQA models evaluated by accuracy (%)

	LSTM	GRU
Baseline	71.34	71.84
Abstract concept	79.29	77.56
Art	75.19	74.03
Geographical location	76.15	76.03
Organization	73.89	76.18
Other	74.18	75.76
Personality	72.58	72.99
Sport	76.74	76.74
Average	75.43 (0.0027**)	75.61 (0.0007**)

**denotes significance level $\alpha < 0.01$

5. Conclusion and Possible Future Work

Generally speaking, intelligent question answering or chatbot models are divided into generative and retrieval-based approaches for continued conversations. This research focuses on a retrieval-based approach and proposes the classification intelligent question answering (CIQA) model. The CIQA model uses a two-stage approach, i.e. the question classification stage and answer prediction stage. The first stage builds an ensemble of classifiers based on the training set. The second stage uses the classifier ensemble from the first stage to determine a suitable category for the questions of the test set and selects a suitable deep learning QA model. The SQuAD (Stanford question answering dataset) 2.0 is used to evaluate the performance of the CIQA model. Based on the experiments, the proposed CIQA model effectively reduces the search space and significantly outperforms the baseline model.

For future work, there are several possible directions. For example, this research only uses the bagging ensemble strategy as a method for improving the performance of basic deep learning. Other ensemble strategies, such as boosting and stacking, may be investigated in further work. In the deep learning architecture proposed in this research, Keras embedding is used in the embedding layer. Some other embedding techniques, such as Word2Vec, Doc2Vec, and GloVe, may also be used for possible further work. Finally, an investigation into the category imbalance of the QA dataset is another possible research direction.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2410-H-033-040-MY2.

References

[1] Zhang, J., Wang, X., Lu, J., Liu, Li. and Feng, Y. (2024). The impact of emotional expression by artificial intelligence recommendation chatbots on perceived humanness ad social interactivity. Decision Support Systems, vol. 187, no. 114347, pp. 1-11.

[2] Zheng, S., Yahya, Z., Wang, L., Zhang, R. and Hoshyar, A.N. (2023). Multiheaded deep learning chatbot for increasing production and marketing. Information Processing and Management, vol. 60, no. 130446, 2023, pp. 1-14.

[3] Wu, Y., Li, Z., Wu, W. and Zhou, M. (2018). Response selection with topic clues for retrieval-based chatbots. Neurocomputing, vol. 316, pp. 251-261.

[4] Zhou, X., Hu, B., Chen, Q. and Wang, X. (2018). Recurrent convolutional neural network for answer selection in community question answering. Neurocomputing, vol. 274, pp. 8-18.

[5] Huang, J., Zhou, M. and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, pp. 423-428.

[6] Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K.S. and Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. Computer Science Review, vol. 52, no. 100632, 2024, pp. 1-24.

[7] Kim, J.K., Chua, M., Rickard, M. and Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. Journal of Pediatric Urology, vol. 19, no. 5, 2023, pp. 598-604.

[8] Pandey, S. and Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. Healthcare Analytics, vol. 3, no. 100198, pp. 1-12.

[9] Pathak, K., Prakash, G., Samadhiya, A., Kumar, A. and Luthra, S. (2025). Impact of Gen-AI chatbots on consumer services experiences and behaviors: Focusing on the sensation of awe and usage intentions through a cybernetic lens. Journal of Retailing and Consumer Services, vol. 82, no. 104120, 2025, pp. 1-13.

[10] Ma, H., Wang, J., Lin, H. and Yang, L. (2022). Global and local interaction matching model for knowledge-grounded response selection in retrieval-based chatbot. Neurocomputing, vol. 497, pp. 39-49.

[11] Moore, K., Zhong, S., He, Z., Rudolf, T., Fisher, N., Victor, B. and Jindal, N. (2023). A comprehensive solution to retrieval-based chatbot construction. Computer Speech and Language, vol. 83, no. 101522, pp. 1-21.

[12] Abdi, A., Idris, N. and Ahmad, Z. (2016). QAPD: an ontology-based question answering system in the physics domain. Soft Computing, vol. 22, no. 1, pp. 213-230.

[13] Tao, C., Feng, J., Yan, R., Wu, W. and Jiang, D. (2021). A survey on response selection for retrieval-based dialogues. Proceedings of the 15th International Joint Conference on Artificial Intelligence. pp. 4619-4626.

[14] Yu, B., Xu, Q. and Zhang, P. (2018). Question classification based on MAC-LSTM. Proceedings of 2018 IEEE Third International Conference on Data Science in Cyberspace, pp. 69-75.

[15] Qian, H., Dou, Z., Zhu, Y., Ma, Y. and Wen J.R. (2021). Learning implicit user profiles for personalized retrieval-based chatbot. Proceedings of 30th ACM International Conference on Information and Knowledge Management, pp. 1467-1477.

[16] Sagi, O. and Rokach, L. (2018). Ensemble leaning: a survey. WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, e1249.

[17] Rajpurkar, P., Jia, R. and Liang, P. (2018). Know what you don't know: unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 784-789.

[18] Weizenbaum, I. (1966). ELIZA: a computer program for the study of natural language communication between man and machine. Communications of ACM, vol. 9, no. 1, pp. 36-45.

[19] Wallace, R.S. (2009). The anatomy of ALICE. Parsing the Turing Test, pp. 181-210, Springer, New York.
[20] Al-Zubaide, H. and Ayman, A.I. (2011). Ontology based chatbot. Proceedings of the 2011 Fourth International Symposium on Innovation in Information and Communication Technology.
[21] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly, USA.
[22] Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning, MIT Press.
[23] Lowe, R., Pow, N., Serban, I. and Pineau, J. (2015). The Ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 285-294.
[24] Zhou, Y., Li, J., Chi, J., Tang, W. and Zheng, Y. (2022). Set-CNN: a text convolutional neural network based on semantic extension for short text classification. Knowledge-Based Systems, vol. 257, no. 109948, pp. 1-11.
[25] Wang, H., Wu, Z. and Chen, J. (2019). Multi-turn response selection in retrieval-based chatbots with iterated attentive convolution matching network. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1081-1090.
[26] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, vol. 9, pp. 1735-1780.
[27] Kannan, A., Kurach, K. Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., Young, P. and Ramavajjala, V. (2016). Smart reply: automated response suggestion for Email. Proceedings of 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
[28] Kadlec, R., Schmid, M. and Kleindienst, J. (2015). Improved deep learning baselines for Ubuntu corpus dialogs. Proceedings of Machine Learning for SLU & Interaction NIPS 2015 Workshop.
[29] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, vol. 18, pp. 602-610.
[30] Gu, J.-C., Ling, Z.-H. and Liu, Q. (2019). Interactive matching network for multi-turn response selection in retrieval-based chatbots. Proceedings of 28th ACM International Conference on Information and Knowledge Management, pp. 2321-2324.
[31] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1724-1734.

[32] Tao, C., Wu, W., Xu, C., Hu, W., Zhao, D. and Yan, R. (2019). Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. Proceedings of the 12th ACM International Conference on Web Search and Data Mining, pp. 267-275.
[33] Hung, C. and Chen, J.H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. Expert Systems with Applications, vol. 36, no. 3, pp. 5297-5303.
[34] Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In: Gentle, J., Härdle, W., and Mori, Y. (eds.) Handbook of Computational Statistics. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg.
[35] Banerjee, S. and Bandyopadhyay, S. (2013). An empirical study of combining multiple models in Bengali question classification. Proceedings of International Joint Conference on Natural Language Processing, pp. 892-896.
[36] Lei, S., Hongzhi, L., Zhengtao, Y. and Quan, Z. (2009). Ensemble learning for question classification. Proceedings of 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, vol. 3, pp. 501-505.
[37] Guven, Z.A. and Unalir, M.Q. (2022). Natural language based analysis of SQuAD: an analytical approach for BERT. Expert Systems with Applications, vol. 195, no. 116592, pp. 1-16.
[38] Konrad, J. (2018). Transfer Learning for Question Answering on SQuAD. Department of Cybernetics, Czech Technical University, Technicka 2, 166 27 Praha, Czech Republic.
[39] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383-2392.