Research Article

COVID-19 FAQS CHATBOT USING ARTIFICIAL NEURAL NETWORK WITH BAG OF WORDS

ABDELRAHMAN R. S. ALMASSRI*[1]
NOUR AMMAR[1], UFUK FATİH KÜÇÜKALI[2]

[1] Department of Software Engineering, Istanbul Aydin University, Istanbul, Turkey
[2] Department of Architecture, Istanbul Aydin University, Istanbul, Turkey
**E-mail address: abdelrahmanalmassri@stu.aydin.edu.tr, noor101ammar@gmail.com,ufkucukali@aydin.edu.tr*
ORCID: 0000-0001-5450-7956, 0000-0002-2691-4180, 0000- 0002-2715-7046

ABSTRACT

COVID-19 is a contemporary virus with a fatal syndrome that had not been seen in the last century. The virus evolves in people of all ages in all areas around the world. As the increment of cases is growing up rapidly the worry of people is increasing, which makes it very hard for healthcare departments and governments to solve people's queries. AI solution is suggested, where a simulation of front-desk assistance. Chatbots are easy to use and simulate a human conversation through text via smartphones or personal computers. Chatbot applications can improve patient information, monitoring, or treatment adherence. The architecture is a simple neural network consisting of a single hidden layer and the sigmoid function is trained by textual data organized by multiple data organization methods. A simple GUI is provided to the classifier to be tested practically. Data used is a collection of questions and their answers about COVID-19. The approach has achieved acceptable results considering speed, and accuracy. Practical predictions were true with acceptable accuracy.

Keywords: *FAQ's, Healthcare, COVID-19, Chatbot, Neural Network, NLP.*

Received: 26.05.2022 - Accepted: 17.07.2022
DOI: 10.17932/EJEAS.2021.024/ejeas_v02i2003

1.INTRODUCTION

In the last two years, a new disease had been discovered on 31 December 2019 named COVID-19 virus. It has involved the whole world and is considered to be announced by WHO as an official pandemic on 11 March 2020, which spread anxiety between nations. Hence, people want to query about the new virus, its symptoms, and its fatality, which creates an issue of shortness in front-desk assistance employees as well as health call centers employees. To deliver the information to a bigger number of people, an Artificial Intelligence solution is suggested. A chatbot is an automated software program that interacts with humans. A chatbot is an automated computer program that fundamentally simulates human conversations such as the works of [1, 2, 3]. The evaluation of the chatbot's User Interface (UI) that was done for COVID-19 in [4] shows that the best approach is the interactive chatbot that can answer the user in conversational way and accept the free hand input, which depends on AI as this paper introduces. There is different architecture to classify text, and the word2vec embedding model [5] and GloVe [6] are embedding dictionaries, while a bag of the word (BoW) [7] and BoW TF-IDF [8] are vectorization methods that convert the textual data to numeric data in vectors shape. The artificial Neural Network (ANN) used consists of one hidden layer that uses sigmoid function and synaptic weights. The result of the architecture was fair enough to accept since the data is complex and unlike [8, 10, 11] where they used classification depending on multiple classes, in our approach, there is one class to predict which is the true question itself. The approach predicts the user's input question to its most similar true question in the dataset, then prints its answer in the Graphic User Interface (GUI) chatbot. The experiment compared with three models BERT, TF-IDF, and GloVe in section 5.

2.THEORY

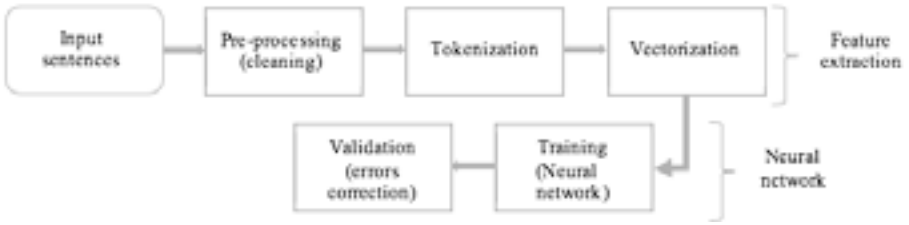

Figure 1. the algorithm flowchart of our approach

Text vectorization method used is Bag of Word [7]. The occurrence calculation used is the binary occurrences where 1 refers to the existence of a word and 0 refers to non-existence in the whole dataset, which makes it a dictionary consisting of 1688 vectors referring to several tokens, as shown in Table 1. Bert model [11], TF-IDF Representation [8], and glove embedding dictionary [6] are other models used to organize textual data. In this paper, they are applied for comparison.

Table 1. simulation of Bag of Words vectorization process applied on dataset

		Word and its series							
		1	2	3	4	5	6	…	1688
Sentence and its series		"have"	"question"	"corona"	"disease"	"headache"	"…"	"symptom"	
	"I"								
1	"I have a question about corona … symptoms"	1	1	1	1	0	0	…	1
2	"is headache beinga symptom of corona disease?"	0	0	0	1	1	1	…	1
…	…	…	…	…	…	…	…	…	…
642	"what is corona?"	0	0	0	1	0	0	…	0

BoW most used method in text classification method because of its simple approach to solving classification problems [10]. BoW representation takes the sentence $x_i = \{x_1, x_2, x_3\}$ and converts it to a vector of ones and zeros $v_i = \{x_1, x_2, x_3\}$ then includes it into a matrix $v_i = \{v_1, v_2, v_3\}$.
NLTK (Natural Language Tool Kit) is a high functional NLP platform from TensorFlow that process textual data. This project is used in phases 1 and 2. Methods used are: stem() to stem the text, lower() to apply lower case, and word tokenize() to apply the tokenization process.

First phase as shown in Figure 1 is preprocessing the data by cleaning it of any punctuation marks such as (?,!,'," ", .,,), after the cleaning stemming process is applied to data then apply lower case on the whole text. Stemming is defined as reducing inflection in words to their root forms such as (saw → see, looks → look, sking → ask) to reduce the number of words that enter the tokenization process. The second phase is tokenizing the clean data which means taking the meaningful words and meaningless words such as (a, the, at). The third phase is vectorization after Features are selected in phase 2, the tokens in each sentence are numeric values in vectors as simulated in figure (BoW).

The model consists of one hidden layer multiplied by the sigmoid function, as well as input and output layers with a total of 3 layers. 30 hidden neurons in the hidden layer and 0.01 learning rate ratio (Figure 2).

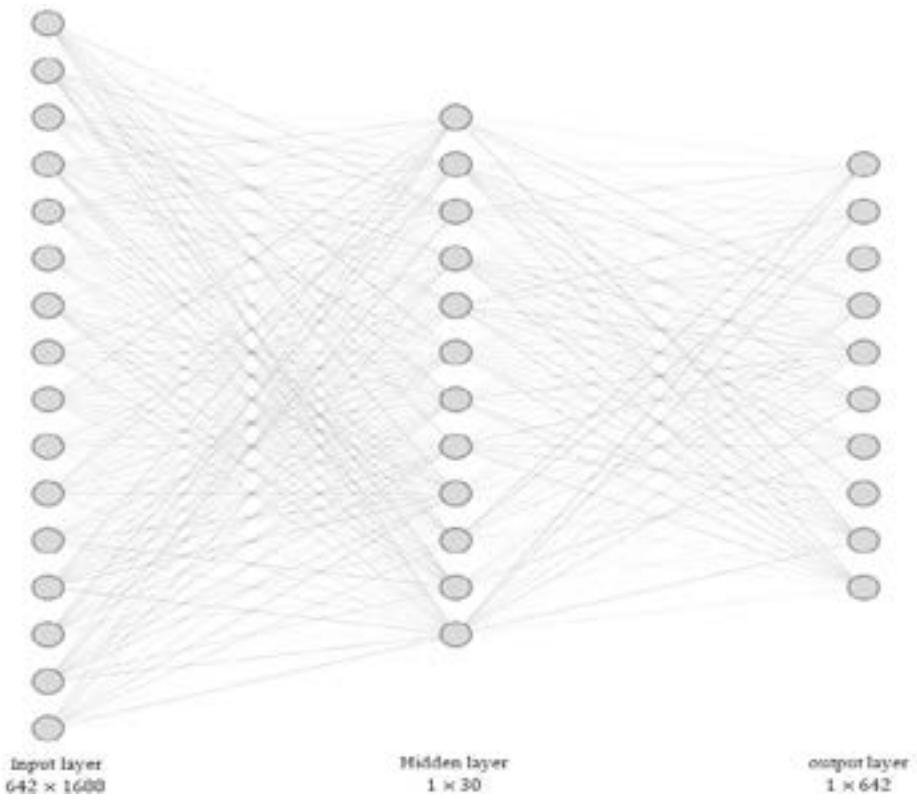

Figure 2. Simulation of ANN of our approach

Synaptic weight Refers to the measure of amplitude-change in a single iteration of the learning batch-of connection between nodes [14]. Where y_i is the output of one layer, w_i refers to weights and x_i refers to the binary BoW vectors.

$$y_i = \sum_i w_i x_i$$

Backpropagation [15] is used in this approach to learn Gradient decent, a is the learning rate which $a = 0.01$,where Δw_i refers to weights (y_i-y_i) is the evaluation of error as specified in (4).

$$\Delta w_i = (y_i - y_i)_i$$

Layers as simulated in Figure 2. The input layer is the bag of words sequences number of dimensions is 642 × 1688, where 642 refers to the number of sentences and 1688 refers to the length of BoWs sequences (each sentence converted to a numeric vector). It is multiplied by the sigmoid function (3).

The sigmoid function that first suggested by [15] and approved its functionality in backpropagation learning networks. In our method, it is used to normalize values since it is the most useful activation function for the textual prediction that applies feature selection on one probability between 0 and one as shown in Figure 3.

$$a(x) = \frac{1}{1 + e^{-x}}$$

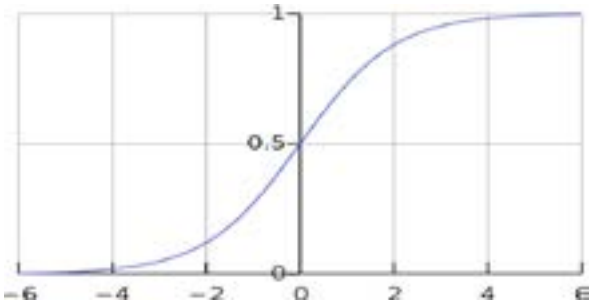

Figure 3. Sigmoid function curve

Hidden layer: the input layer multiplied by sigmoid activation function 1 × 30, where 30 refers to the number of features (neurons) and 1 refers to a BoW vector. It is multiplied by the sigmoid function then MSE (4) is applied in this layer. Output layer: 1 × 642, where 1 means that there is one labeled true question as an output of prediction.MSE is an appropriate choice of error measuring in single hidden layer neural networks, where n is several BoW vectors, y_i is the actual output and y_i is the predicted output. Each actual vector is subtracted from the predicted vector and then squared $(y_i - y_i)^2$. The result is the mean which evaluates the ratio of error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.RESULT AND DISCUSSION

Figure 4 shows a sample of data that the model trained on. Data used in this approach is a collection of questions about the new disease COVID-19 [16]. The dataset contains 642 questions about COVID-19. The question title is a short

main question with questions about the same subject in different phrasing ways appended to each question title in the question field, in our approach, the question title is trained.

question_id	title	question	answer_id	answer	answer_type	wrong_answer
14057	Can pets catch the cold?	Last night I was drying my cat with a towel af...	14083	Yes they can. The viruses that cause a cold in...	Accepted	That is a Priapulid worm, also known as a "pen...
89709	Is the Common Cold an Immune Overreaction?	It's my understanding that the majority of sym...	89712	Can someone die of the common cold? \n\nNo. \nT...	Accepted	The dash ("-") does not represent a negative c...
89886	Air purifier against bacteria and viruses?	We would buy a mobile air purifier in our home...	89887	The aforementioned filter will filter microbes...	Accepted	It's a bleu ray jellyfish, don't touch is becau...
89929	Why are bats the source of dangerous coronavir...	Why do coronaviruses come from bats?\n\nI mean...	89944	\n The preponderance of links between bat and...	Accepted	First of, depending on your definition of life...
89938	How do bats survive their own coronaviruses?	How do bats survive their own coronaviruses (w...	89975	It's common for the reservoir host of a zoonot...	Accepted	I think that "career in synthetic biology" and...

Figure 4. Sample of dataset visualization

Training is done on the Number of epochs=100 000, 10 000 in each iteration. A large number of epochs makes the error correction higher, time consumed in the training is about 30 minutes on SSD, 12 RAM, i7 core processor PC. Experiments Figure 5 shows the simulation of process flow in the training phase, first the input data "covid1" inputs the model the data converted from nominal to numerical vectors by BoW process then vectors multiplied by the hidden layer input test data in "apply model" state, finally, the output comes out from "performance" phase which is the prediction of a true question then its label-answer-is printed in chatbot as shown in Figure 6:

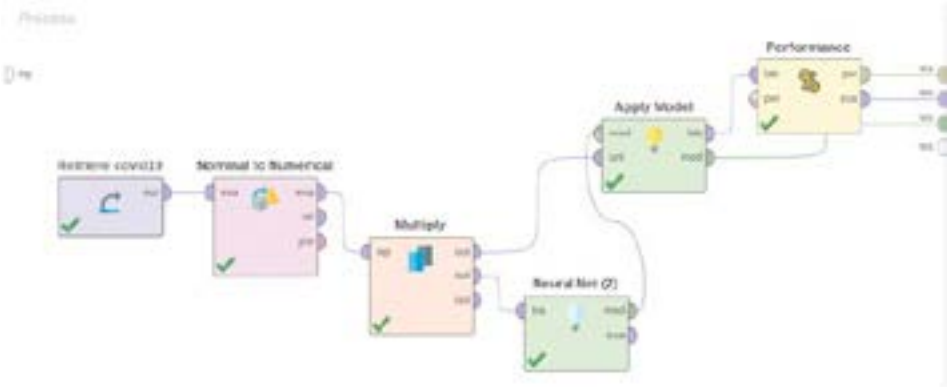

Figure 5. Process visualization of trained model

As shown in Figure 6 the process of NLP must be done before inserting the question of the user into the classifier, after vectorization of the user's input, it is fitted into the classifier, and based on similarity the nearest question from the dataset is chosen to be answered, then the labeled answer is printed in chatbot UI.

GUI Chatbot

Figure 6. Testing the classifier system

Figure 7 shows the chatbot tested with the user, the speed of answering is measured in milliseconds, Table 2 shows the results of the models compared with the ANN BoWs model, as shown the results depend on true or false values, the technique used is the cosine similarity that shows the similarity probability between the predicted question and the questions in the dataset.

Figure 7. Screenshot of our GUI chatbot tested with user

Table 2. Results of models

Question		Prediction result according to model			
		Binary BoW with ANN (our model)	BERT embedding	GloVe embeddingdictionary	TF-IDF BoW
X_actual	Can pets catch the cold?	true	false	true	true
X_prediction	Can dog catch the cold?				

4.CONCLUSION

Chatbots are easy to use and simulate a human conversation through text via smartphones or personal computers. Chatbot applications can improve patient information, monitoring, or treatment adherence. The architecture is a simple neural network consisting of a single hidden layer and the sigmoid function is trained by textual data that is organized by multiple data organization methods. In the final of this paper, the model used is a single layer ANN with BoW text organization method using stemming and tokenization. Results are fairly accepted where the data is not big enough to get high accuracy. The results show that 3 questions are answered true out of 5 questions as the example in Table 2. In future work, the accuracy is planned to be higher, by expanding the neural network to be a deep neural network with more than one hidden layer as well as experimenting with another text organization method such as GloVe embedding dictionary.

REFERENCES

[1] Kumar, A., Meena, P. K., Panda, D., & Sangeetha, Ms. (2019). Chatbot in Python. International Research Journal of Engineering and Technology. Volume 6, issue 11. 391-395. e-ISSN: 2395-0056. https://www.irjet.net/archives/V6/i11/IRJET-V6I1174.pdf .

[2] Park, H., Moon, G., & Kim, K. (2021). Classification of Covid-19 Symptom For Chatbot Using Bert. Advances In Mathematics: Scientific Journal. Volume 10. no.2. 1857-8438 (electronic). ISSN: 1857-8365 (printed). https://doi.org/10.37418/amsj.10.2.34 .

[3] Lei, H., Lu, W., Ji, A., Bertram, E., Gao, P., Jiang, X., & Barman, A. (2021). COVID-19 Smart Chatbot Prototype for Patient Monitoring. arXiv preprint arXiv:2103.06816. https://arxiv.org/abs/2103.06816 .

[4] Höhn S., Bongard-Blanchy K. (2021). Heuristic Evaluation of COVID-19 Chatbots. In: Følstad A. et al. (eds) Chatbot Research and Design. CONVERSATIONS 2020. Lecture Notes in Computer Science, vol 12604. Springer, Cham. https://doi.org/10.1007/978-3-030-68288-09 .

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of ord representations in vector space. https://arxiv.org/abs/1301.3781 .

[6] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162, https://doi.org/10.3115/v1/D14-1162.

[7] Joachims T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683.

[8] Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009). Improving arabic text categorization using decision trees. 110 - 115. 10.1109/NDT.2009.5272214.

[9] Van T. P., Thanh, T. M. (2017). "Vietnamese news classification based on BoW with keywords extraction and neural network," 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), pp. 43-48, doi: 10.1109/IESYS.2017.8233559.

[10] Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. arXiv preprint arXiv:1503.08895.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805 .

[12] Naik, C., Kothari, V., & Rana, Zankhana. (2015). Document Classification using Neural Networks Based on Words. International Journal of Advanced Research in Computer Science. Volume 6, No.2 ISSN: 0976-5697. https://doi.org/10.26483/ijarcs.v6i2.2429 .

[13] Iyer, R., Menon, V., Buice, M,. Koch, C., Mihalas, S. (2013). The Influence of Synaptic Weight Distribution on Neuronal Population Dynamics. PLoS Comput Biol 9(10):e1003248.https://doi.org/10.1371/journal.pcbi.1003248 .

[14] David, E., James, L. M. (1987). "Learning Internal Representations by Error Propagation," in Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations , MIT Press, pp.318-362.

[15] Han, J., Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. Lecture Notes in Computer Science, vol 930. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3- 540-59497-3_175.

[16] COVID-19 csv format dataset, https://www.kaggle.com/xhlulu/covidqa . accessed [June 16, 2021].