Improve Question Classification Genetic Algorithm Based Feature Selection and Convolution Neural Network

Asmaa Ahmed Shama[1,*], Hadi Saboohi[2]

1. Information Technology Center, Mustansiriyah University, Baghdad, Iraq.

2. Computer Engineering, Islamic Azad University, Isfahan, Iran.

*Corresponding author:Asmaa4a4s@gmail.com

https://doi.org/10.29072/basjs.20240111

ARTICLE INFO

Keywords

Genetic Algorithm, Convolutional Neural Network Algorithm, Natural Language Processing, Feature Selection, Feature Extraction,

ABSTRACT

Natural Language Processing (NLP) approaches play a crucial role in classifying inquiries and comprehending human language in diverse applications. A Question Answering System (QAS) consists of three components are question processing, information retrieval, and answer selection. Question Answering Systems (QASs) are a distinct form of information retrieval. The most crucial aspect of QAS is deciding on the question type since it influences the other sections following. However, an important question-answering system requires a prominent question classification system. In the past, there are different methods to solve this problem, such as rule-based learning, and hybrid approaches. However, the problem with these methods is that the rules require a lot of effort to create and are very limited. In this study, the utilization of genetic algorithm and deep neural network techniques enhances the quality control problem-solving process. This research utilizes the UIUC dataset. This collection comprises 5452 questions designed for learning purposes and an additional 500 questions specifically intended for assessment. The suggested solution involves converting each query into a matrix, with each row representing the Word2vec of a word. Subsequently, a Genetic Algorithm (GA) is employed to identify the most optimal features. Ultimately, a Convolutional Neural Network is utilized for classification, yielding a remarkable accuracy of 98.2% in our experimentation with the question dataset.

Received 7 Apr 2024; Received in revised form 21 Apr 2024; Accepted 28 Apr 2024, Published 30 Apr 2024

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

The intelligent and accurate in dealing with applications than ever before, as the user or person searching for a specific question only needs a specific piece of information. To be provided instead of searching for it in many documents and thus wasting more time [1], so most of these users prefer to get a short and concise answer at the same time. The main goal of classifying questions is to learn to assign and identify questions for the purpose of answering them, and some may think at first glance that this process is easy and simple, but it is more complex if it depends on many factors and specifications that determine the quality of the system's performance and its ability to answer the questions posed. Question classification systems are not limited to quality assurance only, but also include data recovery [2]. The primary objective of developing an answering system is to enhance question classification. Proposing an automated system that automatically and the classification of questions by determining the type of question and its classification, as well as achieving the highest accuracy rate in improving the classification of expected questions by relying on clever algorithms and the data set used [3]. Providing accurate and clear information when asking any question, enabling users of these applications to obtain accurate and concise results in a short amount of time. There are two types of traditional question classification methods: rule-based methods and statistical machine learning methods. Early rule-based methods mainly used artificial analysis of syntactic structure to derive rules and then judge the question type [4]. Our method has many features. For example, it is relatively easy to implement and does not require much training data, so the classification speed is fast. But the disadvantage is that these methods rely more on experts and are subjective. In addition, the experts' classification decision is very easy to be influenced by the classification system, which makes it less flexible. Subsequently, statistical learning-based methods have shown good classification performance, which have the advantages of strong adaptability, easy integration, and extension [5]. Machine learning models based on statistical methods commonly used in question classification include Bayes [6], SVM [7], KNN [8], ME [9], etc. However, the disadvantage of the statistical learning method is that its classification accuracy is still easily affected by the syntactic analysis accuracy. Deep learning technology has gained attention in Natural Language Processing (NLP) due to its ability to extract natural language feature information without complex feature engineering. Researchers have started using deep learning methods for question classification, with Deep Nural Network (DNN) models offering advantages in query representation and feature extraction. CNN, a deep spatial

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

neural network, is particularly effective in feature extraction, reducing difficulty and improving classification accuracy. Numerous CNN-based methods have been proposed, leading to numerous research results [10]. Therefore, further research. Feature extraction is done through a genetic algorithm and CNN algorithm to improve response classification. In this paper, feature selection based on genetic algorithm and CNN classification are utilized in responding systems to increase categorization accuracy while determining relevant responses to queries. In fact, genetic algorithms are used to identify the correct features during the implementation period of the classification procedure.

2. Mythology

A novel approach to enhancing question classification via deep learning and genetic algorithms is introduced. The goal is accomplished through an accuracy indicator and attribute extraction at pre-processing steps. The system involves preprocessing, feature extraction through the BOW method, and genetic algorithm selection with the neural network [20]. The results are checked against accuracy metrics to complete the target. A schematic visualization of the system presented in the work may be found in this paper. Figure 1. The schematic layout of the proposed approach.

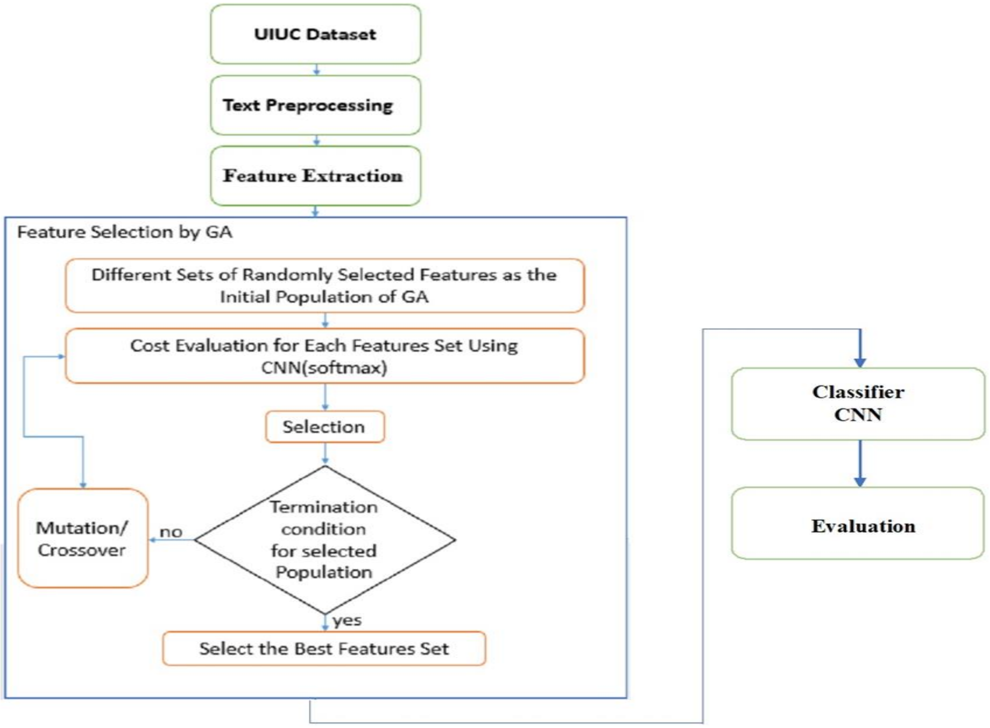
UIUC Dataset
Text Preprocessing
Feature Extraction
Feature Selection by GA
Different Sets of Randomly Selected Features as the Initial Population of GA
Cost Evaluation for Each Features Set Using CNN(softmax)
Selection
Termination condition for selected Population
no
Mutation/ Crossover
yes
Select the Best Features Set
Classifier CNN
Evaluation

Figure 1: Flowchart of the suggested approach for classifying questions.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

1. Dataset

This part provides an overview of the UIUC dataset containing 5,452 training queries and 500 assessment queries. The dataset is segmented into fifty subgroups and six primary categories [21].

2. Text Preprocessing

Data preprocessing implies the transformation of data into friendly information for machine learning. Written documents can also require text cleaning, such as removing irrelevant information, defining the unavailable value, and normalizing information, before text categorization. A couple of preparatory activities were called to have the textual materials ready for the Resume Classification task. The above method is shown in Figure 2.

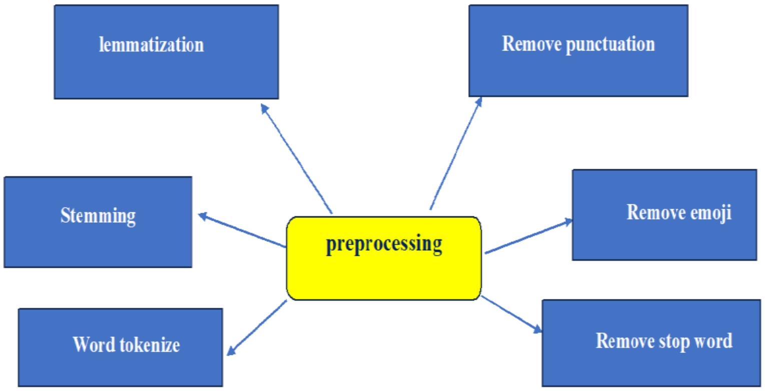

Figure 2: Text preprocessing process

2.1 Removal of punctuation

When categorizing questions the step of eliminating punctuation, from a question involves getting rid of all commas, periods, question marks, exclamation points, quotation marks and other symbols used in writing. The purpose of this action is to enhance readability and simplify the analysis for a classification model by reducing the number of tokens or distinct features it needs to process. By doing variations, in punctuation won't interfere with the model's capacity to concentrate on the core content of the questions.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

2.2 Tokenization

Tokenization involves breaking down the entries in a document, into words or tokens to better understand the text. This step is crucial for text analysis as it breaks down passages into parts by removing punctuation and spaces. These tokens are then used for analyses such as word counting and examining phrase frequency. Tokenization is essential for text processing tasks like removing stop words, stemming, and lemmatization.

2.3 Removing stop words

An important step in data preparation is eliminating stop words. While commonly found in text data words like "'s" "each " and "and" hold significance, for classification models. As a result, removing stop words from the corpus improves the classification model's performance.

2.4 Stemming and Lemmatization

Stemming and Lemmatization, categorized under Text Normalization or Word Normalization techniques in NLP, aim to reduce word inflection in classification text by mapping words to their root stem. Both methods remove prefixes and suffixes from words, including affixes like -es, -s, -ed, in-, un-, -ing, etc., which alter the meaning of words.

Table 2. shows the text preprocessing, which will remove the HTML tags, punctuation, numbers, stop words, and extra characters. It will also perform encoding, mode Stemming and Lemmatization.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

Table 1: Text preprocessing

Questions	Pre-process	Category1	Category2
How did serfdom develop in and then leave Russia?	serfdom develop leave Russia	DESC	Manner
What films featured the character Popeye Doyle ?	film featured character popeye doyle	ENTY	Cremat
How can I find a list of celebrities ' real names	find list celebrity real name	DESC	Manner
What fowl grabs the spotlight after the Chinese Year of the Monkey ?	fowl grab spotlight chinese year monkey	ENTY	Animal

3. Splitting the data set

A standard practice for dividing a dataset involves allocating 70-80% for training and reserving the remaining 30 – 20 % for testing. Additionally, a smaller portion of the training set, typically 20 – 10 % is designated for validation. The objective is to ensure sufficient data in each subset for robust model training and evaluation, while also maintaining a balanced distribution of data across all subsets

4. Feature extraction

Following the preprocessing phase, the dataset now comprises crucial words essential for classification purposes. To showcase their importance, various methods of feature extraction, including Bag-of-Words (BOW) with different n-gram ranges, were assessed.

5. Feature selection with genetic algorithm and classification with CNN

In this part, the integrate and install genetic algorithm and CNN algorithm is explained. In short, a group of people have (solutions) [22]. And each person represents a group of characteristics

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

selected randomly, so that the characteristic selected from the data has a value of 1 or unselected, so its value is zero, and then a number is entered, a certain number of people (solutions) such as 10 people and fit is calculated for each of these solutions [23]. closest to the intended solution. From here, it should be defined that the process of calculating the fit is done by entering the features or the person in the CNN algorithm, so that the process of calculating a fit through (MSE) is done. And after performing the process of calculating the value for each individual, depending on the lowest error coefficient, it is sorted in ascending order [24], where the values with the lowest error are at the top and are sorted in this way, then the genetic algorithm works with the crossover process. Slow and crossover means creating new solutions to the problem, through mating between individuals and some of these solutions, they are exposed to a mutation, i.e. a small number of characteristics randomly selected by the person changes [25]. After that, we select the best individuals produced by parents and children, 10 people are selected to complete the next stage, and note that the number of people who will enter must be equal to the number of people in the next generation. This process is repeated until the best solution or the best subset of features is obtained or until the number of iterations is reached. And then feed it back into the CNN algorithm to classify the questions and identify each question for each class. Confusion matrix in order to evaluate the results by calculating the value of precision, accuracy, recall and F1 score for project evaluation [26]. Algorithm 1 shows the combination of these two GA-CNNs for feature selection and question classification. The objective function One of the most important functions that must be correctly determined for optimization is the objective function used in the optimization algorithm. This function should be adjusted according to the optimization problem and the objective of the problem. Since our goal in this optimization work is to reduce the classification error, so the fitness function used in this optimizer algorithm is defined based on the classification error function, this means that our goal in this work is to reduce the classification error or increasing the accuracy of the convolution classifier is also. Next, the relationship given in (3) shows the objective function or the fitness function used in this algorithm [27].

$$costfun = MSE = 1/ n \sum \mathrm{ni}\text{-}1\ (Y\mathrm{i} - Y\ ‘\mathrm{i}\ )2 \qquad \ldots 1$$

```
// Initialize population
FOR i = 1 TO n DO
    Xi = InitializeChromosomeWithRandomFeatures([0, 1])
END FOR
```

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

```

// Main loop for evolutionary process
WHILE (CurrentIteration < MaxIterations) DO

    // Fitness evaluation
    FOR EACH Chromosome Xi IN Population DO
        Network = CreateCNN(Xi)
        TrainCNN(Network)
        Fitness(Xi) = EvaluateAccuracy(Network)  // Using MSE as accuracy metric
    END FOR

    // Genetic operations
    Parents = SelectParents(Population, Fitness)
    Offspring = Crossover(Parents)
    Mutate(Offspring)

    // Update population with new generation
    Population = Offspring

    INCREMENT CurrentIteration
END WHILE

// Select the best performing chromosome
BestChromosome = SelectBestChromosome(Population, Fitness)

// Classify using the selected features
FinalClassification = ClassifyUsingCNN(BestChromosome)
RETURN FinalClassification
```

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

6. Evaluation metrics

We assessed the performance of the classification models using various evaluation metrics. These metrics included Overall Accuracy, Precision, Recall, and F-Score[1].

7.Results and discussion

In the previous section, the method that improves the accuracy of the Convolutional Neural Networks using GA algorithm by selecting the best features is proposed. This system should be able to classify questions. To prove this, evaluations and tests should be performed on the system. In this section, we will evaluate the performance of the proposed classifier on the UIUC dataset. Performance analysis is done according to the evaluation scales in the test set. We used various evaluation measures including classification accuracy, coverage, accuracy, and F1 score. The results of the test without feature selection using the ML classifier are shown in Table 2.

Table 2: ML classification with Non-selection of features

Classifier	Type-Class	Precision	Recall	F1-score	Accuracy
MLP	coarse-grained	0.66	0.66	0.66	0.66
Random Forest	coarse-grained	0.68	0.66	0.66	0.67
SVM	coarse-grained	0.68	0.65	0.43	0.65
KNN	coarse-grained	0.67	0.44	0.43	0.44
NB	coarse-grained	0.43	0.41	0.41	0.41
MLP	fine-grained	0.59	0.58	0.57	0.58

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

Random Forest	fine-grained	0.60	0.55	0.55	0.55
SVM	fine-grained	0.61	0.61	0.51	0.51
KNN	fine-grained	0.50	0.27	0.29	0.27
NB	fine-grained	0.33	0.31	0.30	0.31

this one has nearly the lowest values. We can draw the conclusion that the suggested approaches have a positive impact on the way questions are categorized. The above table shows the implementation of machine learning classification algorithms in two types of data, primary and secondary type, without using genetic algorithm. The classification accuracy of the MLP algorithm in the main type was 66%, while the classification leading to the subgroup was 58%, and this shows that the results in the main type are better than the. Figure 3.a and Figure 3.b show the comparison of these algorithms on the data set

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

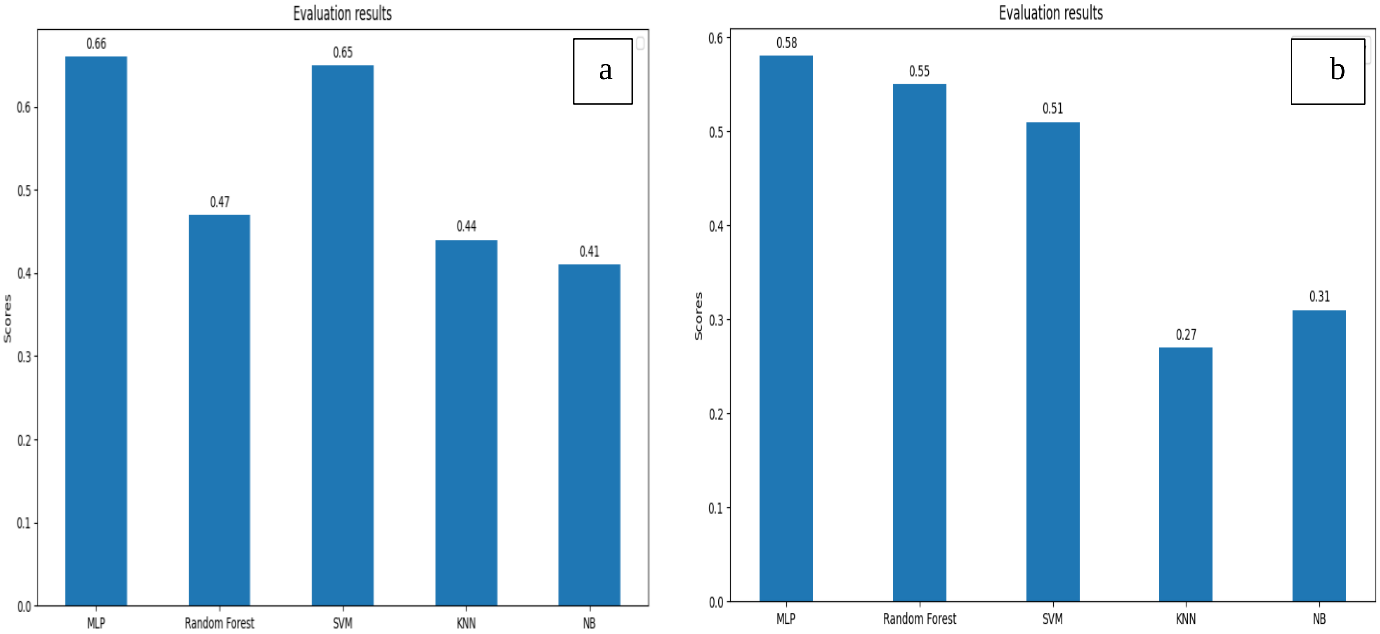

Figure 3: (a) and (b) algorithms comparison results on the original sticky

7.1.1 Feature selection with ML classification

Table 3 listed the outcomes of feature selection using ML algorithms since both GA techniques can be used for feature selection. As the values of these scales rise, Table 3's values demonstrate that feature selection has enhanced performance in the majority of the suggested comparison scales. This is because enhanced feature selection contributes to improved classification.

Table 3: feature selection with ML classification

Classifier	Type-Class	Precision	Recall	F1-score	Accuracy
GA-MLP	coarse-grained	92.2	91.9	92.3	91
GA-Random Forest	coarse-grained	79.1	83.1	81.4	81.7

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

GA-SVM	coarse-grained	93.1	92.2	92.5	91
GA-KNN	coarse-grained	81.71	82.03	81.86	75
GA-MLP	fine-grained	89.4	89.03	89.23	89.1
GA-Random Forest	fine-grained	76.5	80.2	77.6	77.3
GA-SVM	fine-grained	90.1	88.9	96.5	87.4
GA-KNN	fine-grained	75.18	75.63	75.52	69

The above table shows the implementation of ML classification algorithms on the data that has two sticky types, the main type and the secondary type using the genetic algorithm, in which the classification accuracy is obtained using the genetic algorithm with the ML algorithms. The accuracy of the MLP algorithm has reached 91% in the first type and 89.1% in the second type. From Table 3 and Table (4) we find that there is an improvement in performance in most of the comparison criteria, which confirms the importance and effectiveness of using the genetic algorithm and its effective role in improving the classification in different criteria. Figure (a,b) and Figure (4.a) (4.b) show the comparison of these algorithms with the genetic algorithm for selecting features on the data set.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

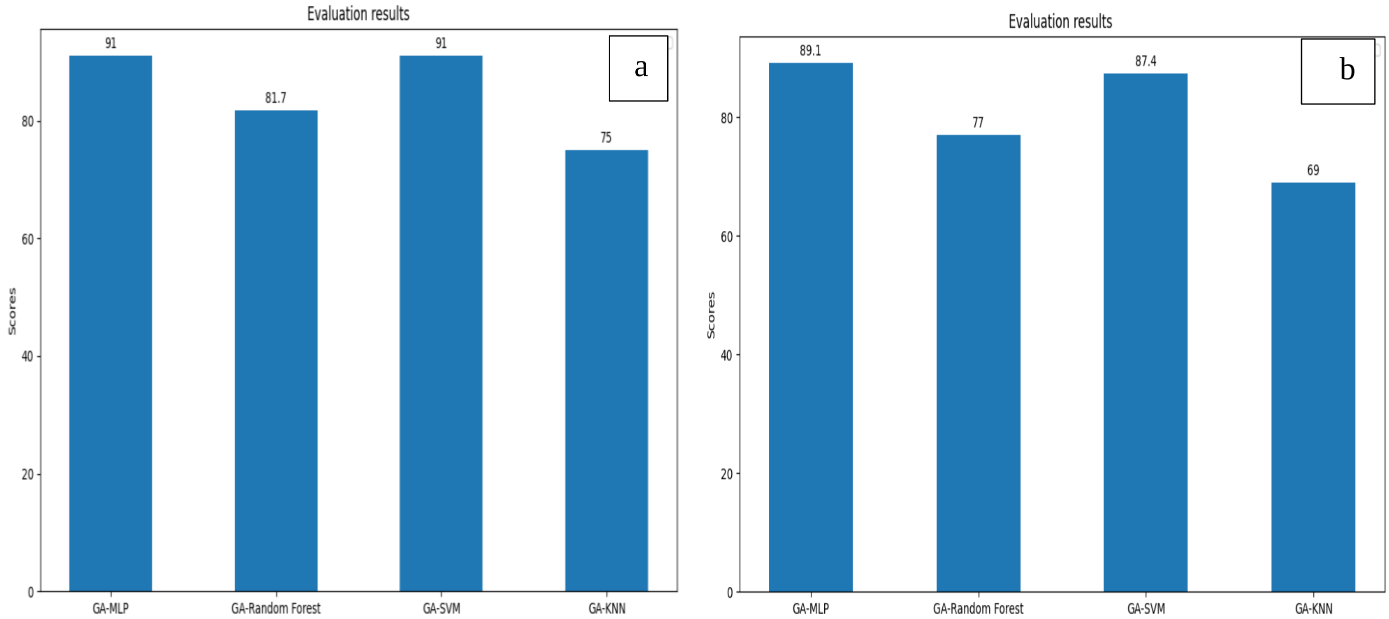

Figure 4: (a) Algorithm comparison results on the original sticky data using genetics and (b) Algorithm comparison results on sub-labeled data using genetics.

7.1.2 Suggested classification without feature selection

Table 4 presents the outcomes of the evaluation of collective classification using CNN algorithms without the use of GA feature selection.

Table 4: Suggested categorization without feature selection

Type-Class	Accuracy	Precision	Recall	F1 score
coarse-grained	89.43	89.05	89.22	89.03
fine-grained	85.75	85	85.37	85.03

The above table shows the implementation of the CNN classification algorithm on the main sticky type without using the genetic algorithm with a classification accuracy of 89.03%, while the classification accuracy on the subtype without using the genetic algorithm was 85.3%. Note that the classification accuracy was better in the main type than in the sub-type.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

7.1.3 Feature selection with proposed method classification

Experiments were carried out utilizing CNN algorithms since the suggested classification technique for feature selection includes evolutionary algorithms (GA) for both feature selection and collective classification. Table 5 reports the outcomes found for these algorithms. demonstrate how the suggested method's classification corresponds with feature selection.

Table 5: Feature selection with the classification of the proposed method

Type-Class	Accuracy	Precision	Recall	F1 score
coarse-grained	0.982	0.97	0.98	0.977
fine-grained	0.956	0.958	0.95	0.954

The above table shows the execution of the genetic algorithm with the CNN algorithm on the data with two main and secondary types of sticky, the results for the main type are as follows: accuracy 98.2%, accuracy 97%, recall 98%, and 97.7f1-score%. While the results of applying the genetic algorithm with the CNN algorithm on the data subtype are as follows: accuracy was 95.6%, accuracy was 95.8%, recall was 95%, and f1-score was 95.4%. The result of using two algorithms together with their integration has had a significant effect on improving the classification in different categories.

8.Comparison with previous studies

To complete the analysis of this research in an integrated, scientific and systematic manner, the model presented for this study was compared with the testing of other methods and techniques presented by a number of researchers in recent scientific articles. In this section, we compare the proposed method with other methods of classifying the query Table 6. Show the comparison of the proposed method with other methods in detail and Table 7 show the proposed method with other methods in terms of accuracy.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

Table 6: Comparison of the proposed method with other methods in detail.

Reference	Methods	Proposed method	Disadvantages	Advantages	Accuracy
[11]	SVM, Albanian Collection, Machine Learning	It uses question classification, Albanian and SVM to improve the QA system.	-	Best SVM Preforming Algorithm	Good Accuracy
[12]	CNN with LSTM, SVM and TF-IDF	Analysis of question classification methods for low-resource languages	-	Experimental evaluation increases the reliability of the method	CNN with LSTM 94.6% reached SVM and TF-IDF 93% F1
[14]	،DPCNN BiLSTM	Model proposes MCDPLSTM to classify disease questions with improved performance	-	The MCDPLSTM model shows improved performance especially in terms of accuracy, recall and F1 rating, while also being computationally efficient at a lower cost of time.	Accuracy 0.18% - 5.43%, recall 0.42% - 7.42%, and F1-Score increased 0.47% - .7.23%
[15]	،GRU ،LSTM CNN and Word2Vec with CBOW	Explores deep learning techniques for categorizing questions in .Turkish	Lack of a comprehensive analysis of the generalizability of deep learning models and techniques used	Successfully uses deep learning techniques to categorize questions in challenging Turkish language	93.7%
[17]	Group	Uses set classification	-	The proposed method	Increases the accuracy of

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

Reference	**Methods**	**Proposed method**	**Disadvantages**	**Advantages**	**Accuracy**
		and feature selection to improve QAS.		increases the accuracy of classification compared to the absence of these methods.	the classification.
[18]	CNN-LSTM CNN-SVM Word2vec	Question classification is an essential aspect of automated answering systems.	Limited access to tagged Turkish question data	Deep learning techniques, including LSTM and CNN, were successfully employed in the challenging task of categorizing questions in .Turkish	Accuracy 94%

Table 7: Comparison of the proposed method with other methods in terms of accuracy.

Research	**Precision**	
Kote, Trandafili and Pelpi 2022	SVM	75.1%
	Logistic regression	72.6%
	Random forest	64.1%
Gong, Liu et al. 2023	95.59%	
Golzari, Sanei et al. 2022	91.80%	
The proposed method	98.2%	

The accuracy results of the tests are displayed in Table (6). Table (7) illustrates that the suggested strategy, which simultaneously employs feature selection and combination classification methods,

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

has higher acceptable efficiency in the majority of cases. This is a result of correctly choosing the right features and utilizing the classifiers' capabilities. With an efficiency of 98.2%, the technology used in this study has the fastest speed in terms of efficiency.

Conclusions

Advancements in global science and technology underscore the pressing demand for automated systems mirroring human cognition, adaptable across diverse scientific and practical domains. This study unveils a refined method enhancing question classification, vital for interpreting inquiries based on their formulation nuances. The proposed method streamlines the utilization and preprocessing of the UIUC dataset, culminating in a feature-rich dictionary transformed into vector representations via TF-IDF or BOW methods. Leveraging genetic algorithms, optimal feature selection precedes CNN-based feature extraction, facilitating precise question categorization across domains such as medical, sports, and scientific disciplines. The dataset comprises two categories: subtype (50 classes) and basic type (6 classes), with superior classification accuracy demonstrated in the latter, reaching an outstanding 98.2%. Future endeavors entail augmenting NLP for multilingual support, automating question answering systems, resolving contextual ambiguity, refining interpretability of classification models, genetic algorithm optimization of CNN parameters, and exploring domain-specific models for heightened industry relevance.

Acknowledgments

The authors would like to thank Mustansiriyah University, Baghdad-Iraq for its support the present work.

References

[1] S. Chotirat, P. Meesad, Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning, Heliyon, 7 (2021) 1-13, https://doi.org/10.1016/j.heliyon.2021.e08216

[2] A. Khan, M.Z Asghar, H. Ahmad, F.M Kundi, S. Ismail, A rule-based sentiment classification framework for health reviews on mobile social media, J Medical Imag Health Inf., 7 (2017) 1445-1453. https://doi.org/10.1166/jmihi.2017.2208

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

[3] W.A Qader, M.M. Ameen, B.I. Ahmed, An overview of bag of words; importance, implementation, applications, and challenges in 2019 international engineering conference (IEC). (2019) IEEE.200-204. DOI:10.1109/IEC47844.2019.8950616

[4] E. Sherkat, M. Farhoodi, A hybrid approach for question classification in Persian automatic question answering system, 4th International Conference on Computer and Knowledge Engineering (ICCKE), (2014) 279-284, http://doi.org/10.1109/ICCKE.2014.6993377

[5] Y Sarica, S.J. Luo, Stopwords in technical language processing. Plos one J, 16 (2021) e0254937.https://doi.org/10.1371/journal.pone.0254937

[6] A. Aouichat, M.S. Hadj Ameur, A. Geussoum, Arabic question classification using support vector machines and convolutional neural networks. in Natural Language Processing and Information Systems: Proceedings of 23rd International Conference on Applications of Natural Language to Information Systems, 10859 (2018) 113-125, http://dx.doi.org/10.1007/978-3-319-91947-8_12

[7] D. Han, T. Tohti, A .Hamdulla, Attention-based transformer-BiGRU for question classification Information, 13 (2022) 214-235, https://doi.org/10.3390/info13050214

[8] S.K Ray, S. Singh, B.P. Joshi, A semantic approach for question classification using WordNet and Wikipedia. Pattern recognition letters, 31 (2010) 1935-1943. http://dx.doi.org/10.1016/j.patrec.2010.06.012

[9] B. Wutzl, K. Leibnitz, F. Rattay, M. Kronbichler, M. Murata, S.M Golaszewski, Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness. PLoS ONE 14(2019) e0219683, https://doi.org/10.1371/journal.pone.0219683

[10] P. Gong, J. Liu, Y. Xie, M. Liu, X. Zhang, Enhancing context representations with part-of-speech information and neighboring signals for question classification, Complex & Intelligent Systems, 9(2023) 6191–6209, doi.org/10.1007/s40747-023-01067-7

[11] N. Kote, E. Trandafili, G. Plepi, Question Classification for Albanian Language: An Annotated Corpus and Classification Models. International Conference on P2P, Parallel, Grid, Cloud and Internet Computing,14 (2023)737- 744, http://dx.doi.org/10.14569/IJACSA.2023.0140385

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

[12] E. Cortes, V. Woloszyn, A. Binder, T. Himmelsbach, D. Barone, S. Möller., An Empirical Comparison of Question Classification Methods for Question Answering Systems. In Proceedings of the Twelfth Language Resources and Evaluation Conference,(2020) 5408–5416, Marseille, France. European Language Resources Association.

[13] J. Suzuki, T. Hirao, Y. Sasaki, E. Maeda, Hierarchical directed acyclic graph kernel: Methods for structured natural language data. in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, (2003)32-39, DOI:10.3115/1075096.1075101.

[14] Y.X. Yu, R. Gong, P. Chen, Question Classification Method in Disease Question Answering System Based on MCDPLSTM, 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), Hainan, China, (2021)381-387, http://dx.doi.org/10.1109/QRS-C55045.2021.00063

[15] M. Zulqarnain, K. Z. Alsaedi, R. Ghazali, M.G Ghouse, W. Sharif, N.A. Aida Husaini, comparative analysis on question classification task based on deep learning approaches, Peer J Comp Sci,7 (2021) 570-596, http://dx.doi.org/10.7717/peerj-cs.570

[16] L. Zhen, X. Sun, The research of convolutional neural network based on integrated classification in question classification, Scientific Programming, (2021) 1-8. https://doi.org/10.1155/2021/4176059

[17] S. Golzari, F. Sanei, M.R. Saybani, M. Basir, Question classification in question answering system using combination of ensemble classification and feature selection, J AI and Data Mining, 10 (2022) 15-24, https://doi.org/10.22044/jadm.2021.10016.2142

[18] S. Yilmaz, S. Toklu, A deep learning analysis on question classification task using Word2vec representations, Neural Comp. Appl., 32 (2020) 2909-2928. DOI:10.1007/s00521-020-04725-w

[19] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Sys. Appl., 36(2009) 6527-6535. DOI:10.1016/j.eswa.2008.07.035

[20] J. Chen, H. Huang, S. Tian, Y. Qu, Feature selection for text classification with Naïve Bayes. Expert Sys. Appl., 36(2009)5432-5435, http://dx.doi.org/10.1016/j.eswa.2008.06.054

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

[21] https://www.kaggle.com/datasets/nltkdata/qc-corpus

[22] W.A Qader, M.M. Ameen, B.I. Ahmed, An overview of bag of words; importance, implementation, applications, and challenges. International Engineering Conference (IEC), 2019. IEEE, http://dx.doi.org/10.1109/IEC47844.2019.8950616

[23] E. Hovy, U. Hermjakob, D. Ravichandran, A question/answer typology with surface text patterns. in Proceedings of the Human Language Technology conference (HLT), (2002)247-251, http://dx.doi.org/10.3115/1289189.1289206

[24] Y. Ehrentraut, M. Ekholm, H. Tanushi, J. Tiedemann, H. Dalianis, Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting, Health Informatics J. 24, (2018)24-42. https://doi.org/10.1177/1460458216656471

[25] W.T Yih, X. He, C. Meek, Semantic Parsing for Single-Relation Question Answering, In Proceedings of ACL 2(2014)643–648, https://doi.org/10.3115/v1/P14-2105

[26] Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil, Learning Semantic Representations Using Convolutional Neural Networks for Web Search, 23rd International Conference on World Wide Web (2014)373–374 DOI:10.1145/2567948.2577348 ,

[27] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, P. Kuksa, Natural Language Processing (Almost) from Scratch, J. Machine Learning Res., 12 (2011) 2493-2537., https://doi.org/10.48550/arXiv.1103.0398

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

تحسين تصنيف الأسئلة القائمة على الخوارزمية الجينية اختيار الميزة والشبكة العصبية الالتفافية

اسماء احمد شمه [1] وهادى صبوحى [2]

1. مركز تكنولوجيا المعلومات ، الجامعة المستنصرية ، العراق-بغداد
2. كلية الهندسة جامعة ازاد اصفهان ، الجمهورية الاسلامية الايرانية

المستخلص

مع تطور تكنولوجيا المعلومات والذكاء الاصطناعي، هناك حاجة ملحة لاقتراح أنظمة ذكية تحاكي العقل البشري، وتلعب معالجة اللغة الطبيعية دورا مهما في تصنيف الأسئلة وفهم اللغة البشرية في التطبيقات المختلفة، حيث تساعد تقنيات البرمجة اللغوية العصبية على تحديدها. هناك ثلاثة مراحل تشكل نظام الإجابة على الأسئلة (QAS): وهي معالجة الأسئلة واسترجاع المعلومات واختيار الإجابة. QASs هي نوع فريد من استرجاع المعلومات. الجانب الأكثر أهمية في QAS هو تحديد نوع السؤال لأنه يؤثر على الأقسام الأخرى التالية. يتطلب نظام الإجابة على الأسئلة المهمة نظاما بارزا لتصنيف الأسئلة. مصنف الأسئلة هو نظام يقوم بتعيين تسمية لكل سؤال. في الماضي ، كانت هناك طرق مختلفة لحل هذه المشكلة ، مثل التعلم القائم على القواعد والنهج الهجين. ومع ذلك ، فإن المشكلة في هذه الأساليب هي أن القواعد تتطلب الكثير من الجهد لإنشائها وهي محدودة للغاية. لحل هذه المشاكل ، في هذا البحث ، تم استخدام مناهج الخوارزمية الجينية والشبكة العصبية العميقة وذلك لتوفير تحسين مراقبة الجودة. تم استخدام مجموعة بيانات UIUC. تحتوي هذه المجموعة على 5452 سؤالا للتعلم و 500 للاختبار. في الطريقة المقترحة ، يتم تحويل كل استعلام إلى مصفوفة حيث يمثل كل صف Word2vec للكلمة. بعد ذلك ، يتم استخدام الخوارزمية الجينية (GA) لتحديد أفضل الميزات ، وأخيرا ، يتم استخدام الشبكة العصبية التلافيفية لغرض التصنيف ، تم الحصول على دقة عالية بنسبة 98.2٪ وهي اعلى دقة تم الوصول اليها ولم تتمكن الابحاث السابقة في الحصول عليها في مجموعة بيانات السؤال.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).