AUT Journal of Electrical Engineering

AUT J. Electr. Eng., 55(1) (2023) 47-60
DOI: 10.22060/eej.2022.21287.5468

Classifying the Clarity of Questions in CQA Networks: A Topic based Approach

Alireza Khabbazan, Ahmad Ali Abin*

Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

ABSTRACT: Today, there are various sources of information in different fields that users can refer to. Generally, the presence of a question in users' minds leads to reference to these sources of information. Users can search for the answer by entering a few keywords in search engines. They can also ask their questions in more detail in the Community Question Answering (CQA) networks so that experts can give a more comprehensive answer to their questions. To get the proper answer, it is necessary to address all the required details in the question. The questions posted in these networks can be divided into clear and unclear. In this study, an attempt has been made to extract unique features from the questions through various machine learning approaches, which can be used to classify questions. To extract these features, the word vector of each question was created, and then using unsupervised algorithms, the questions with similar word vectors were placed in the same group. Afterwards, repetitive concepts were extracted from each group, and their repetition rate in each question makes its feature vector. Finally, the questions were classified based on the extracted feature vector, using ensemble classification models. The achievement of this study is an efficient classification model along with efficient high-resolution feature extraction for classifying clear and unclear questions in CQA networks. Compared to other baselines and transformer-based models on different datasets, the proposed method makes high accuracy results.

Review History:

Received: Apr. 18, 2022
Revised: Aug. 08, 2022
Accepted: Sep. 11, 2022
Available Online: Mar. 01, 2023

Keywords:

Community Question Answering
CQA networks
Clustering
Question Classification
Unclear Questions

1- Introduction

One of the most important sources of information on the Web, which has recently attracted many users, is Community Question Answering (CQA) networks. By visiting these networks, users can ask their questions in detail and get accurate answers. Every day, thousands of questions are registered in these networks, so increasing their quality has become one of their main goals. A network can increase its traffic and inbound users by increasing the quality of its questions. It could also boost the network's popularity [1]. Increasing the quality of a network's questions can be considered so that if a question comes into a network, it will receive an appropriate answer [2]. There are several reasons why a question will not receive a proper answer after being posted on the network. The shortness of a question, the question being too specific, and the lack of details are some of the reasons [3]. The lack of details is one of the most important reasons for not receiving a proper answer. When insufficient details are mentioned in the question, these details are placed in the respondent's mind as ambiguities and confuse him/her in answering. Figure 1 shows this concept.

Therefore, one of the most important methods for increasing the quality of a network is to increase the quality of unclear questions. To do so, an innovative approach can be considered that when the question is asked by the user, by examining to see if the question is unclear, required details are suggested to the user. This will increase the user's chances of getting a proper answer to their question by addressing these details. This approach can be divided into two steps. First, an intelligent model must be presented to determine whether an asked question is clear or unclear. In the second step, if an unclear question is encountered, it should find out the details that are not mentioned and are required, and then inform these to the user in the form of words or sentences. What is mentioned in this research is the first step of this intelligent approach. In this study, a binary classification model is presented to determine whether the question is clear or not. The classification model recognizes question clarity based on the extracted feature vector. This vector shows the repetition rate of a series of concepts in the text of the question.

This paper is organized as follows: Section 2 shows some relevant literature in the field alongside baselines. Additionally, the datasets used in this study are introduced. Section 3 introduces the proposed method. In Section 4, we evaluated the method with baselines. In Section 5, various hyper-parameters of the proposed method are investigated, and their effect on the quality of the classification model is shown. Finally, Section 6 concludes this work, followed by an outline of future works.

*Corresponding author's email: a_abin@sbu.ac.ir

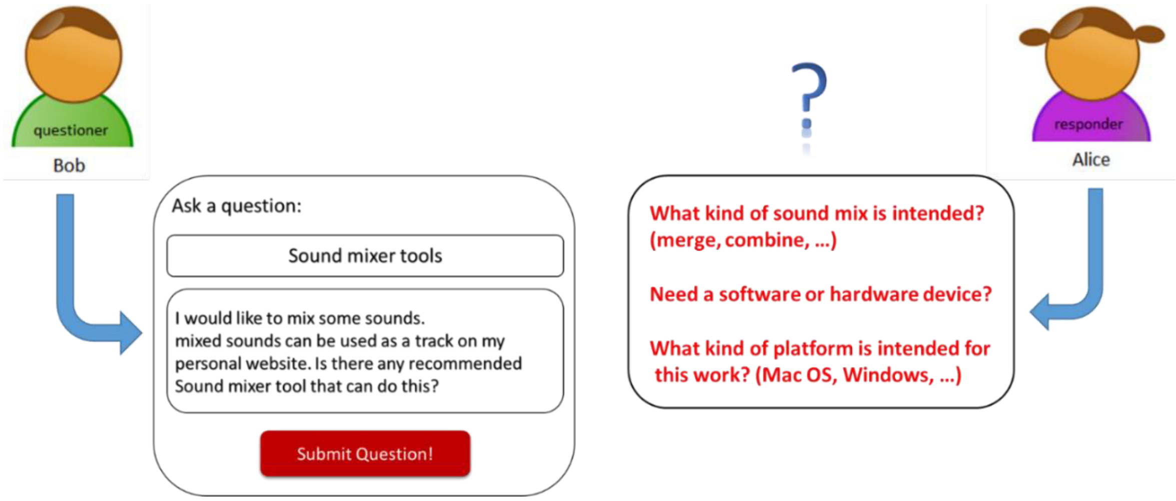

Fig. 1. questions that arise in the mind of the respondent when there are not enough details in the question

2- Related Works

This section will review all approaches to examining the quality of a question in CQAs. Additionally, various contexts in which a question may be placed are examined. Afterwards, the baselines are introduced separately. In the following, according to the areas under discussion, the position of the present research is determined. At the end, the datasets used in this research are introduced.

2- 1- Question Classification Approaches

The approaches discussed in this section are divided into the following three categories [4], and are described briefly below.

Question quality prediction: This approach tries to predict the quality of questions to find out how well the respondents understand the questions [5]. Due to the lack of a clear definition for the concept of quality and also the dependence on a specific CQA network, machine learning approaches will also have poor performance in this area.

Question review prediction: This approach tries to identify questions that need to be edited [6]. Lack of sufficient information in questions will cause them to be edited. One of the critical points in this approach is that the more the scope of the areas mentioned in the question are limited, the lower the performance of the algorithms.

Question answerability prediction: This approach tries to increase the rate of answering a question. The higher the response rate to a question, the clearer the question is to the respondents [7]. The degree of responsiveness of a question can be inferred from the behavior of the respondents in the face of the question. Respondents' comments or question scores are such behaviors.

2- 2- Question Classification in Various Contexts

In general, there are two main contexts of asking questions, in which the clarity of questions has been extensively investigated. [8]. These contexts are described briefly below.

Synchronous: Questions and answers are generated synchronously and consecutively in this context. Live question-and-answer systems, and text-based social media are manifestations of this category. According to [9], clear questions and requests will significantly affect the quality of conversation in a synchronous question and answer network.

Asynchronous: In this context, questions and answers are not synchronized, and a time interval can be considered between questions and answers. Additionally, there is not necessarily a sequence between the asked questions and the answers [10]. Various CQAs networks, such as all Stack Exchange communities, are examples of these networks.

2- 3- Baseline Models

Random: In this model, the labels of each question are generated uniformly in the dataset.

Majority: In this model, the labels are always the majority class.

Bag of words logistic regression [11]: Each question is a weighted n-gram vector that unigrams combined with phrases of length up to n = 3. The training method on the training data set was 5-fold. Ridge regularization with the strength of C = 1, which means the effectiveness of the regularization term in the loss function, is also used in this model.

Convolutional neural network [12, 13]: The architecture presented in [13] has been used. Additionally, the hyper-parameters of this architecture, including the number of filters and filter size, are set separately for each data set. The input of this model is the 300-dimensional vector created by the word2vec model [13]. This input enters the model in 64 batches to train the model.

BERT [14]: In this model, an original pre-trained version BERT by using Bidirectional Encoder representations in the transformer model on a large corpus comprising the Toronto Book Corpus and Wikipedia is used [14]. In the following,

Table 1. Dataset statistics

Community	Number of samples	Clear	Unclear
Stack Overflow	5859667	35%	65%
Super User	121998	33%	67%
Ask Ubuntu	77712	27%	73%
Cross Validated	38488	18%	82%

first this model tokenizes each question, and then by adding a linear layer to the pre-trained model and re-training it, the generated vector is classified into clear and unclear categories. This model has 12 hidden layers with a size of 768 and a contextual embedding vector for each question containing 118 dimensions [14].

SimQ [11]: In this model, a question enters the model, and several similar questions are found with clear and unclear labels. Next, some features of the input question and similar clear and unclear questions are generated. By combining these features, a feature vector for each question is generated. Finally, a classification model based on this feature vector specifies the question category.

Topic Model [15]: In this model, a topic model is made up of several topics with a medium number of middle of word distribution in all the questions. When a question enters the classification model, a feature vector is generated based on the weight of extracted topics, and then the category of input is specified by this feature vector.

In the following, the position of this research in the approaches and areas mentioned above will be investigated. Among the approaches to examining the question's quality, this study follows the "question answerability prediction" approach. Additionally, the asynchronous context is considered as the main context in this study among the two mentioned contexts.

2- 4- Datasets

The Stack Exchange platform, which includes some CQA forums in various fields, has published information from its forums for various experiments. These datasets include questions, answers, tags, and user comments, and each question is categorized into clear and unclear labels by [16]. According to the heuristics mentioned in [11, 17], A question is categorized as unclear when a question contains a comment with a question mark phrase. A question is also considered clear when there is no edit or comment, and it has an approved answer. Table 1 shows the statistical information from the datasets used in this study, including the number of questions and clear and unclear questions distribution.

3- Proposed Method

This section describes the method presented in this study for question classification. The main idea behind this method is to extract several semantic concepts from all the questions in the dataset. These extracted concepts are used to construct the feature vector for each question. Each dimension of this vector is a property related to a semantic concept, and its value is equal to the repetition of that concept in question. By plotting this feature vector, a histogram of the question can be implicitly displayed, and each bin shows the frequency of a semantic concept in the question. Figure 2 depicts an illustration of a feature vector for a question as an example.

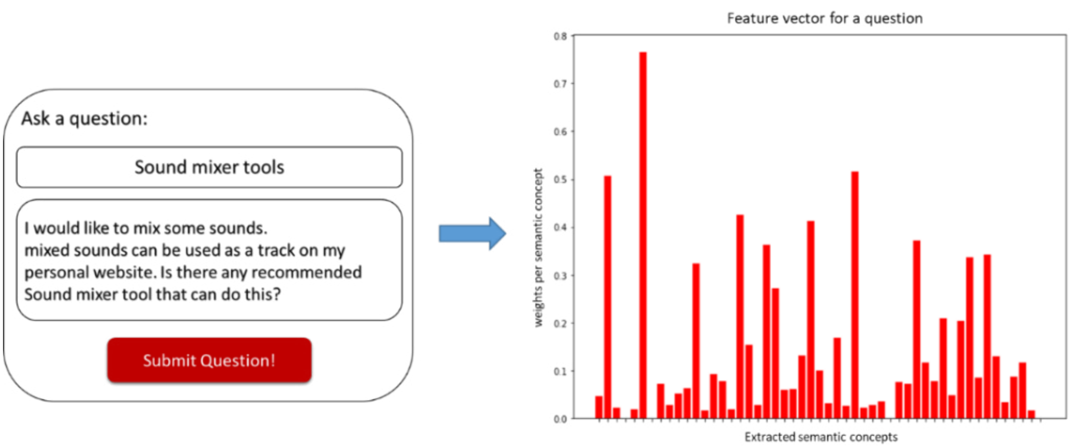

Fig. 2. an illustration of a feature vector for a question

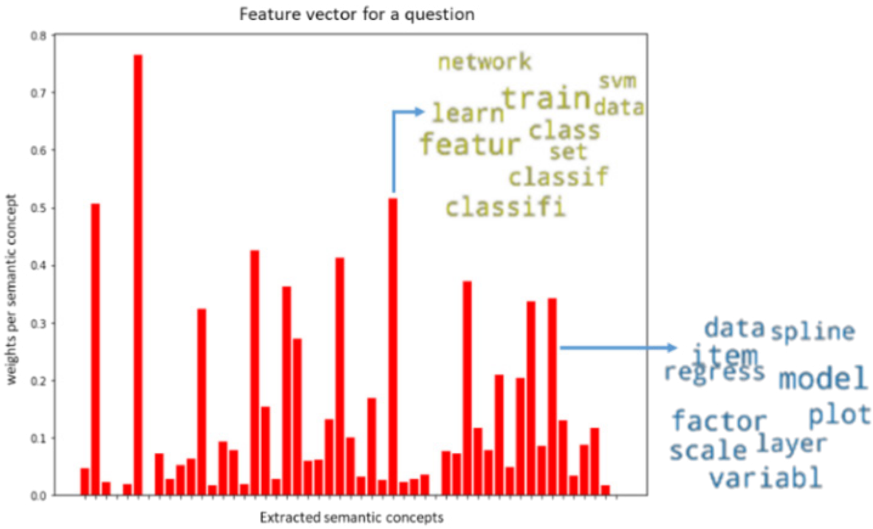

Fig. 3. semantic concepts extracted from the feature vector of a question

Fig. 4. an illustration of the idea of the proposed method

Figure 3 shows an image of some of the semantic concepts extracted in the feature vector.

Figure 4 depicts an illustration of the idea behind the proposed method, which consists of four steps, as described below.

3- 1- Preprocessing

Preprocessing is done to extract keywords from different parts of the questions, consisting of three steps. The first step combines the title, body, and tags of the questions, which are then tokenized. In the next step, each word in question enters the word stemming and lemmatization process. Finally, each question is converted into a set of words. Figure 5 depicts the various step of the preprocessing section.

3- 2- Question Clustering

After preprocessing, we cluster the questions based on their semantic similarity. To this end, in the first step, a Word2Vec model has trained on all the questions words. The output of this step is a 300-dimensional vector for each word of all questions. Next, for each question, the vector of all the words in that question is averaged a 300-dimensional vector is generated for each question. Figure 6 depicts the construction of the question vector from Word2Vec vectors.

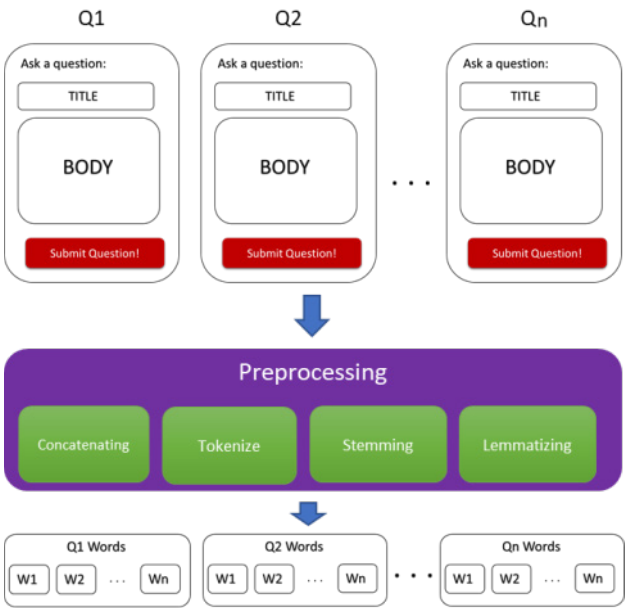

Fig. 5. an illustration of the preprocessing steps

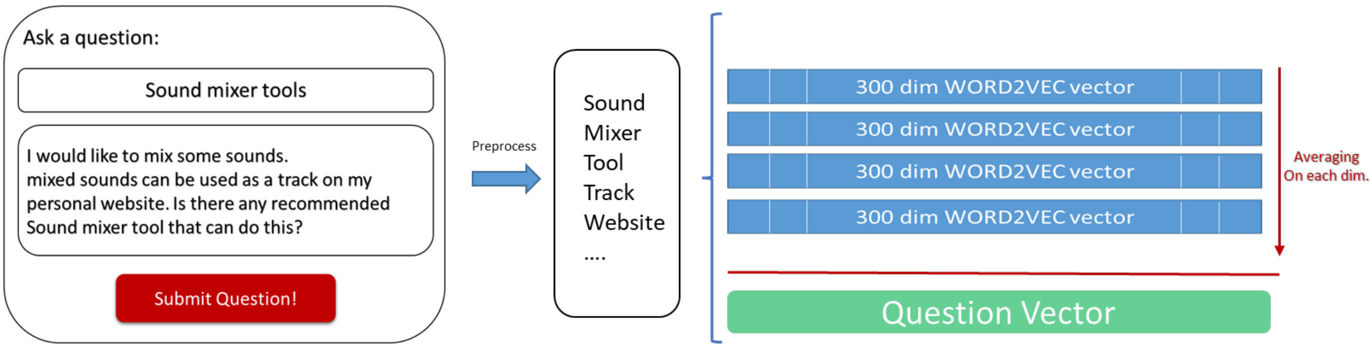

Fig. 6. An illustration of making a question vector by Word2Vec vectors in words

Afterwards, the K-Means clustering is applied to the questions vectors, and all questions are grouped into k clusters based on their 300-dimensional word vectors.

3- 3- Feature Extraction

After clustering the questions in section 3-2, the feature vector of each question is created based on the obtained clusters. This vector has dimensions equal to the number of obtained clusters, and each dimension represents the repetition rate of similar semantic concepts extracted from the corresponding cluster. Figure 7 illustrates this concept.

To construct a feature vector, it is first necessary to extract the similar semantic concepts of each cluster. Each cluster contains several similar questions. Additionally, each question contains some words. Therefore, some words can be extracted from each cluster.

Each word is repeated several times in the questions of a cluster. Therefore, the repetition rate for each word can be considered. This repetition rate can be normalized and the words sorted accordingly. A few most repeated words in each cluster can be considered as the semantic concepts of each cluster. Figure 8 illustrates this step.

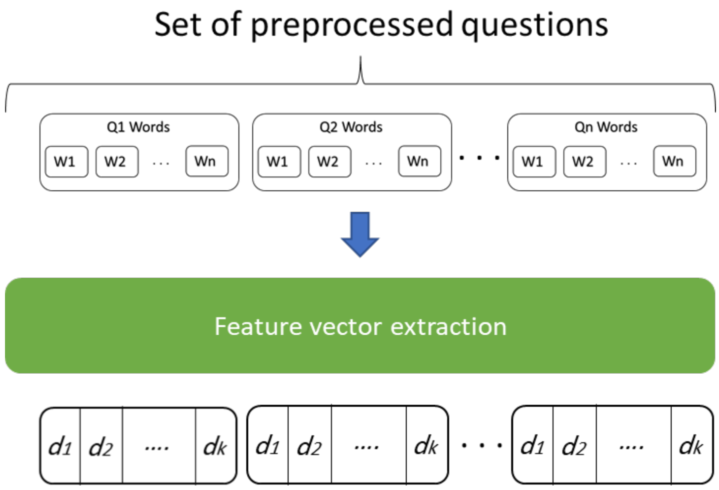

Fig. 7. an illustration of feature vectors made up of several questions

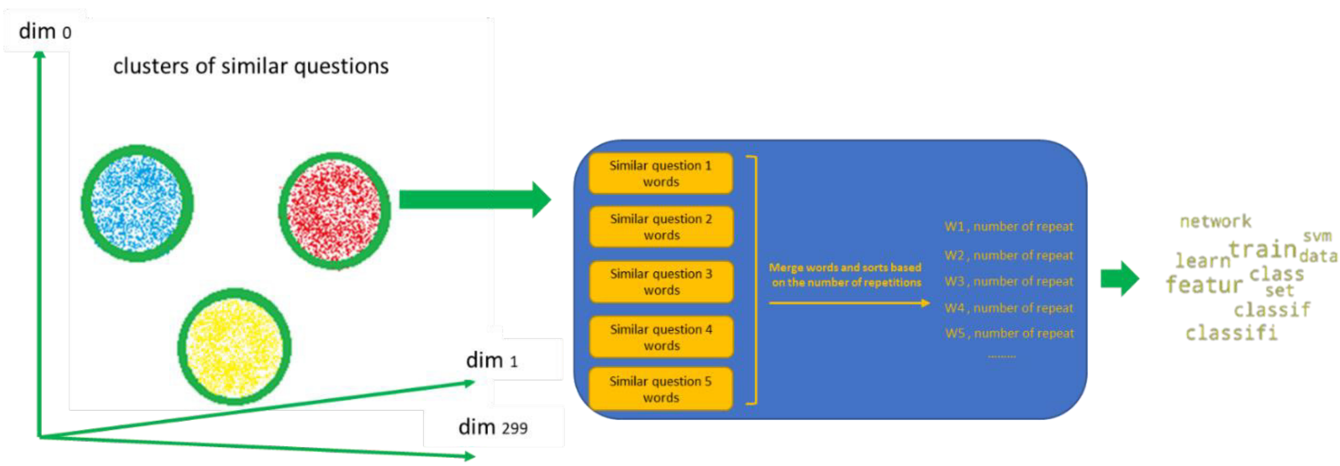

Fig. 8. an illustration of the semantic concepts extracted from each cluster

After obtaining similar semantic concepts from each cluster, it is time to construct a feature vector of a question. In this step, the question is preprocessed, and then a vector equal to the number of clusters is created for it. In each dimension of this vector, the weight of the words in the cluster, which is also in the question, is added. Figure 9 illustrates this step.

After constructing the feature vector, in the next section, these vectors are classified into two categories, clear and unclear, by the classification model.

3- 4- Question Classification

After feature extraction, the questions are classified into clear and unclear classes based on the feature vector extracted for each question. In this study, an ensemble-based classification model called AdaBoost [18, 19] was used for classifying the clear and unclear questions.

4- Experiments and Results

In this section, the performance of the presented model is compared with other baselines. These models are explained in detail in section 2-3. The datasets used in this comparison are Cross Validated, Ask Ubuntu, Super User and Stack Overflow, whose details are mentioned in section 2-4. Model training is done on 70% of the dataset and the rest is used for model evaluation. Additionally, the Google Colab platform has been used to train the model. All hyper-parameters selected for the model and training procedure are

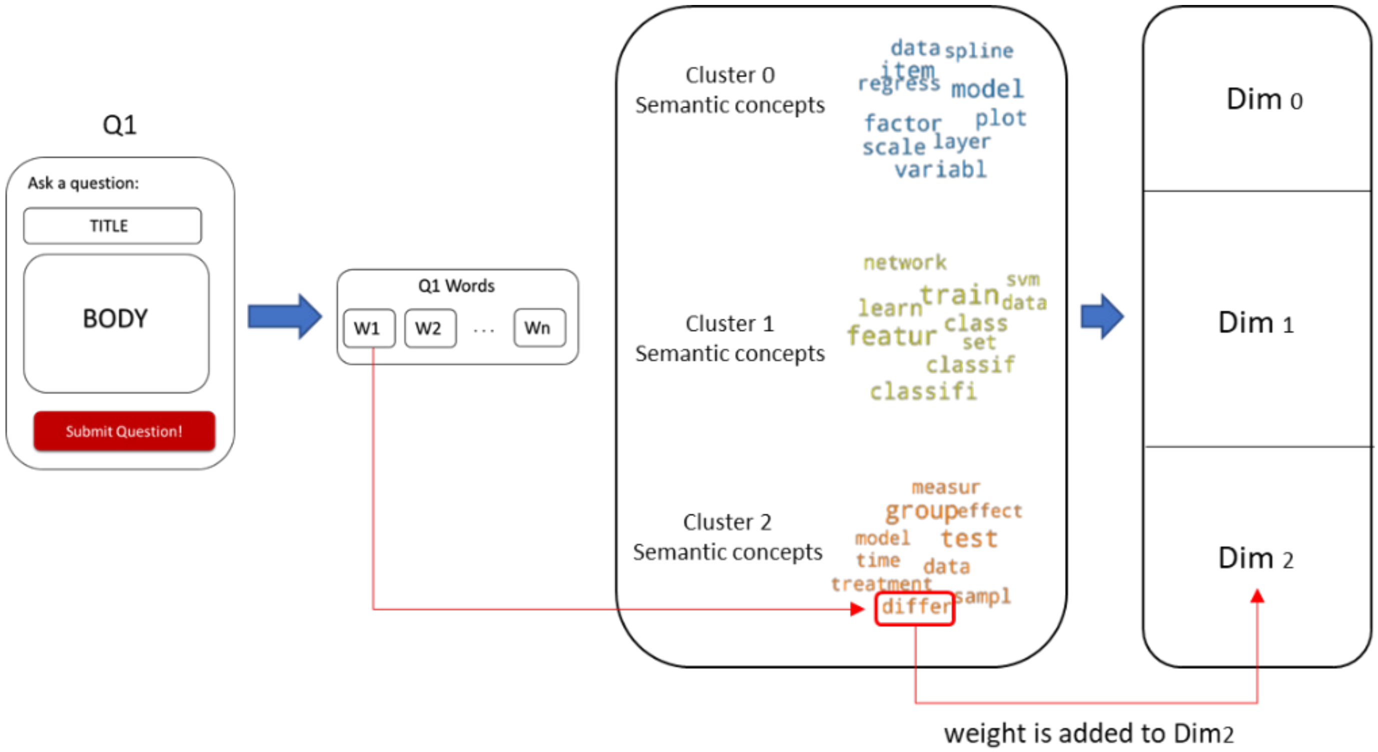

Fig. 9. an illustration of making a feature vector for a question

Table 2. result for unclear question detection

Models \ Datasets	Cross Validated		Ask Ubuntu		Super User		Stack Overflow	
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
Random	49%	61%	48%	56%	50%	57%	49%	56%
Majority	81%	90%	65%	77%	66%	80%	64%	78%
BoW Logistic Regression	81%	90%	65%	78%	70%	80%	69%	78%
CNN	81%	89%	68%	80%	70%	80%	69%	79%
BERT	70%	80%	69%	68%	70%	69%	69%	69%
SimQ ML	81%	90%	65%	77%	68%	80%	67%	78%
Topic ML	95%	96%	88%	89%	85%	87%	NA	NA
Proposed Method	**96%**	**97%**	**90%**	**90%**	**86%**	**88%**	**75%**	**81%**

Table 3. result for clear question detection

Models / Datasets	Cross Validated		Ask Ubuntu		Super User		Stack Overflow	
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score
Random	49%	49%	48%	52%	50%	51%	49%	52%
Majority	81%	88%	65%	78%	66%	85%	64%	79%
BoW Logistic Regression	81%	88%	65%	75%	70%	79%	69%	72%
CNN	81%	79%	68%	74%	70%	79%	69%	77%
BERT	70%	28%	69%	71%	70%	69%	69%	70%
Topic ML	95%	95%	88%	89%	85%	86%	NA	NA
Proposed Method	**96%**	**96%**	**90%**	**90%**	**86%**	**87%**	**75%**	**79%**

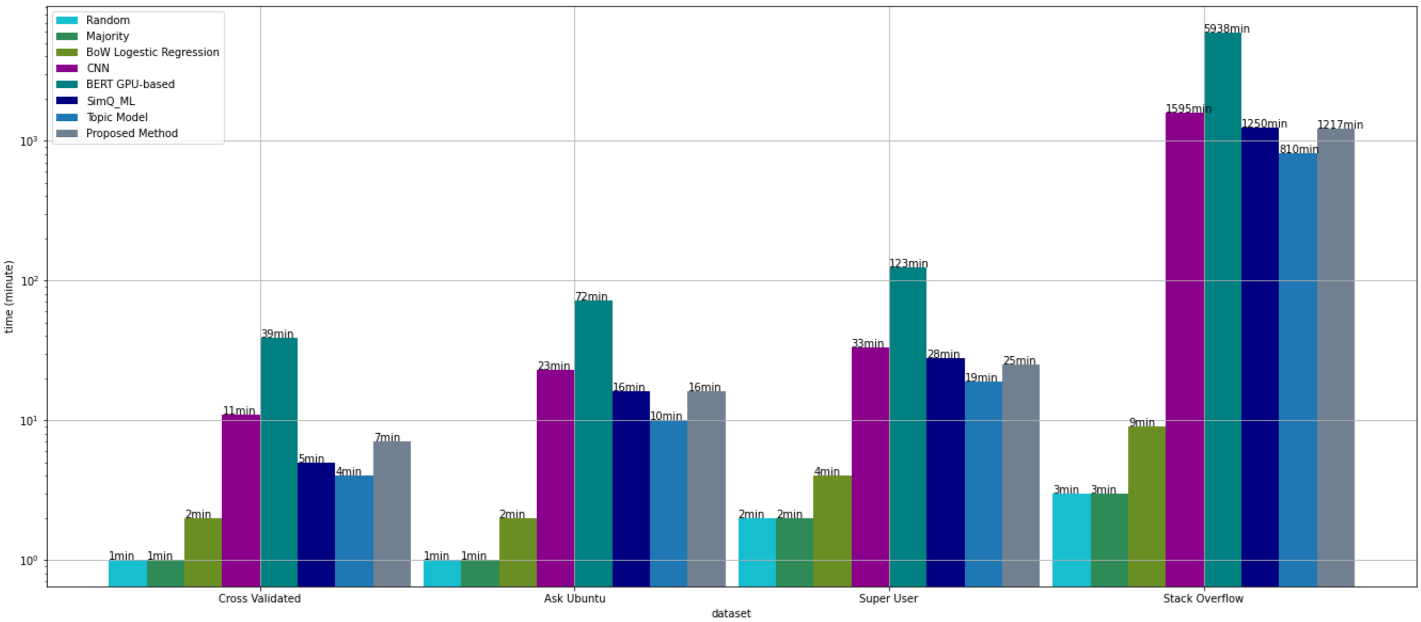

Fig. 10. Comparison of the training time of different models in the introduced dataset

also mentioned in section 5. This comparison is made from two aspects. First, the performance of the presented model is examined to identify unclear questions. This means that unclear questions are positively labelled. The second aspect is that clear questions will be labelled positive. The evaluation criteria used are accuracy and F1-score. The results of this comparison are shown in Table 2 and Table 3. At the end of these comparisons, Figure 10 depicts a comparison of the training time of the presented model and baselines on different datasets.

5- Discussion

In this section, the effect of various hyper-parameters is examined. According to the approach in the first section, the proposed model includes a four-step approach. The parameters of the question clustering method and the parameters of the feature vector extraction method are examined. Additionally, the classification model's hyper-parameters and how to search them are declared at the end of this section

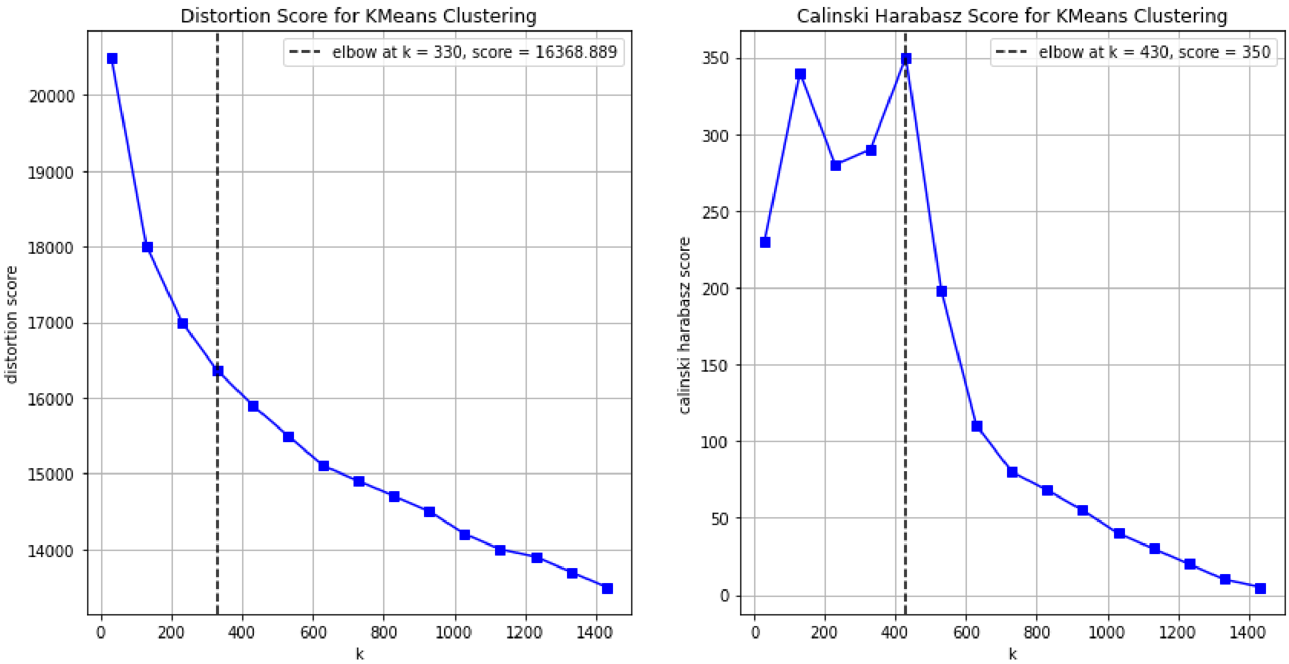

Fig. 11. The effect of the number of clusters on clustering quality

5- 1- Discussion about Question Clustering Method

The hyper-parameters studied in this section are the number of clusters, the number of initializations, and the maximum number of iterations. These parameters directly affect the quality of clusters, so their impact is examined through clustering evaluation criteria.

The number of clusters: This parameter specifies the number of clusters in the clustering model of section 3-2. To find the optimal value of this parameter on a specific dataset, separate clustering models are trained with different values of that. In the following, the quality of clusters produced by each of these models is evaluated through evaluation criteria Calinski-Harabasz [20] and Distortion score, and the obtained result is displayed separately in the form of a curved line. Figure 11 depicts the curves generated from the values of the evaluation criteria for the Cross Validated dataset. The Distortion score is calculated by the average of the Euclidean squared distance from the centroid of the respective clusters. The Calinski-Harabasz score is calculated by dividing the variance of the sums of squares of the distances of individual objects to their cluster center by the sum of squares of the distance between the cluster centers. The higher the Calinski-Harabasz Index value, the better the clustering model. According to Figure 11, the Distortion curve has significantly decreased with the increase in the number of clusters up to around 400 clusters. However, with the increase in the number of clusters from 500 onwards, the decrease is not very high. On the other hand, in the range of 400 to 500 clusters, the Calinski-Harabasz score with the number of clusters of 430 has the highest value. By combining the results obtained from two curved lines, it can be concluded that the clustering model has shown good quality for the number of clusters equal to 430 on the Cross Validated dataset.

The number of initializing: Since the initial point of the clustering algorithm is selected randomly, this parameter shows the number of initializing of the algorithm to get the optimal result. Figure 12 shows the inertia and silhouette score [21] curves for different parameter values for the Cross Validated dataset. The inertia score is calculated by the sum of squared distances of samples to their closest cluster center. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. In the inertia curve of Figure 12 for the values of 12, 21 and 26, it can be seen that the decrease is not significant. Further, by referring to the silhouette score curve, we can see that the value of this curve is at its highest value at points 11, 14 and 26. By combining the results obtained from two curved lines, it can be concluded that the clustering model has shown good quality for the number of initializing equal to 26 on the Cross Validated dataset.

The maximum number of iterations: This parameter specifies the maximum number of iterations in the clustering model. Figure 13 shows the inertia and silhouette scores curves for different parameter values. The clustering model has shown good quality for the maximum number of iterations between 250 to 300.

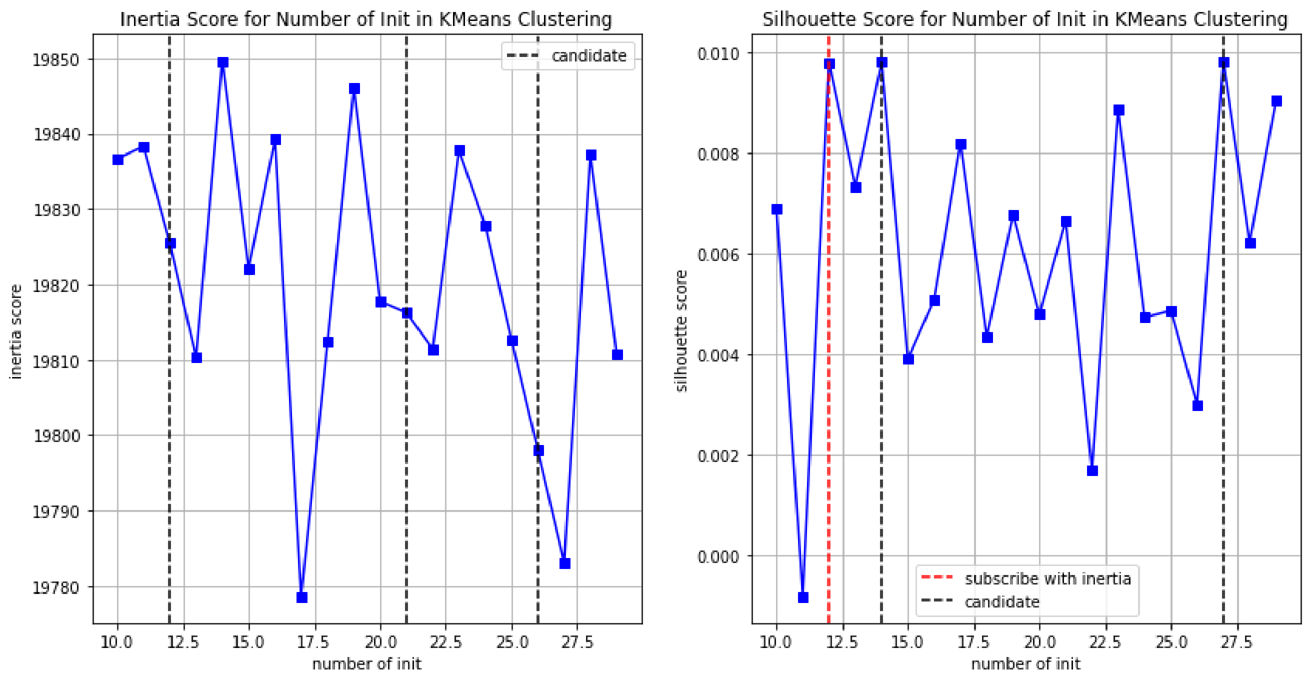

Fig. 12. The effect of the number of initializing on clustering quality

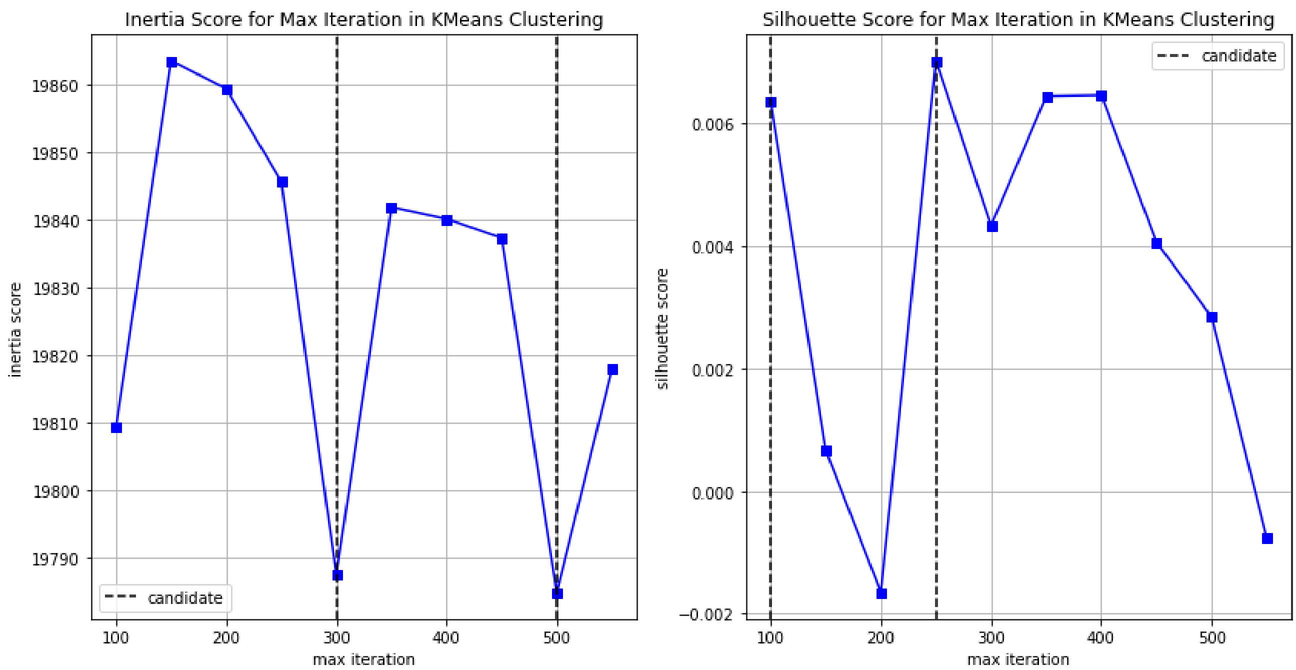

Fig. 13. The effect of the maximum number of iterations on clustering quality

5- 2- Discussion about Feature Vector Extraction Method

The hyper-parameter studied in this section is the number of the first words extracted for the feature vector extraction method. This parameter directly affects the quality of the classification model, so its impact is examined through classification model evaluation criteria.

The number of the first words extracted: This parameter specifies the number of the first word extracted in the feature vector extraction in section 3-3. To find the optimal value of this parameter on a specific dataset, separate classification models are trained with different values.

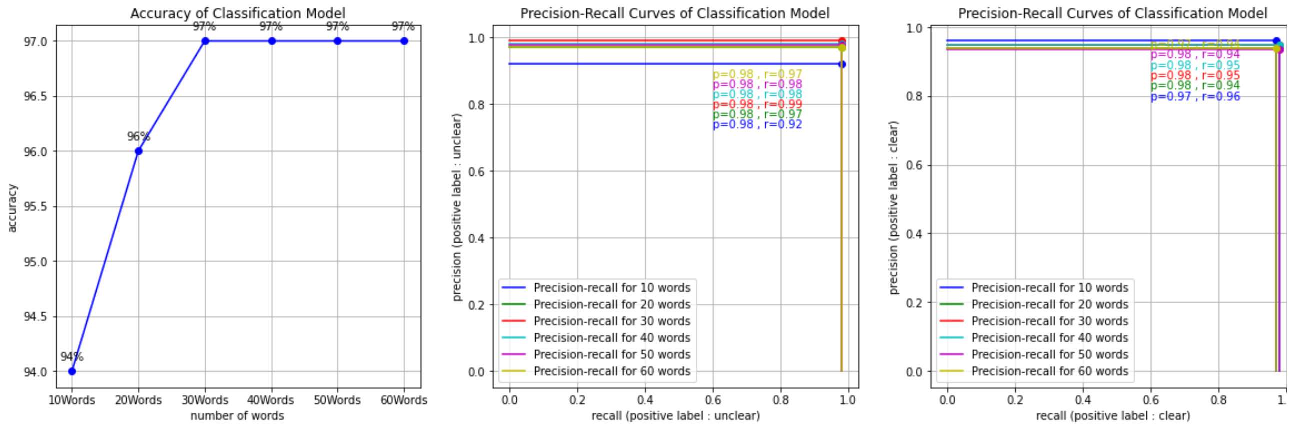

Fig. 14. The effect of the number of the first words extracted on classification model performance

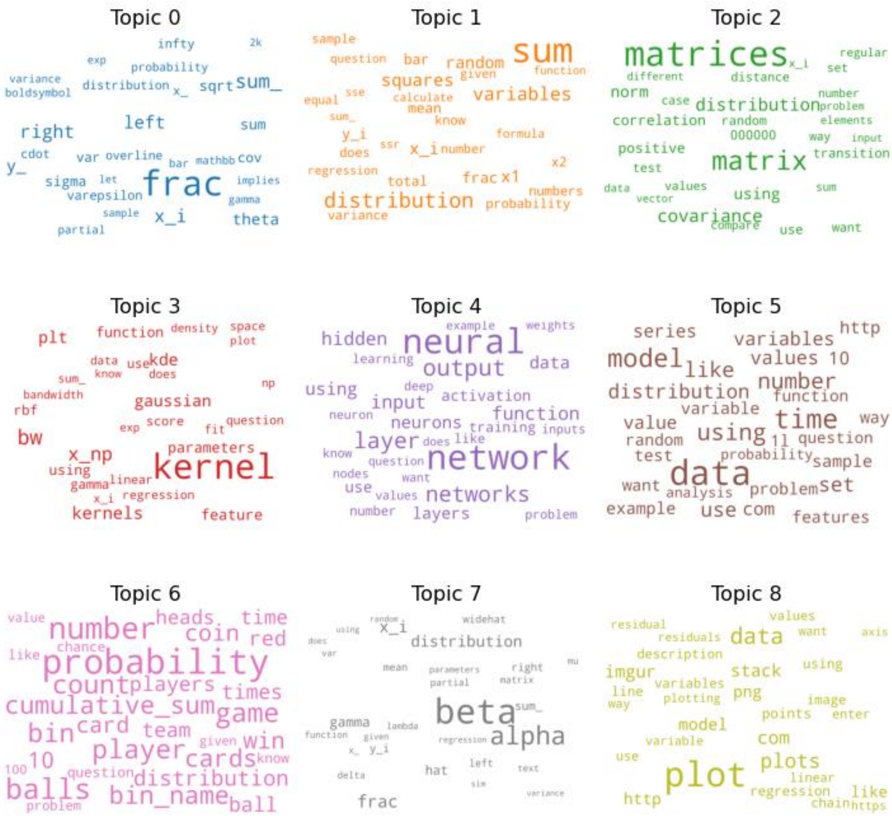

Fig. 15. Words extracted on the first nine cluster

Figure 14 shows the accuracy and precision-recall curve for different values of this parameter for the Cross Validated dataset. Considering that the feature vector has dimensions equal to the number of clusters, increasing the number of the first words extracted from each cluster only affects the value of each dimension. The extracted words in each cluster have a weight based on which they are sorted. This weight is actually the normalized number of their repetitions in the questions of that cluster. The increase in the number of the first words extracted in each cluster indicates the influence of the weight of new words on the value of each dimension. Considering that the extracted words are sorted based on weight, the new words will have less weight, and their impact on the value of each dimension is low. Therefore, a large increase in the number of the first words extracted can have a negligible effect on the performance of the classification model. Figure 14 depicts that the classification model shows good performance for the number of the first word extracted is equal to 30 on the Cross Validated dataset.

Figure 15 also shows the words extracted from the first 9 clusters for the Cross Validated data set.

5- 3- Discussion about Classification Model

The hyper-parameter studied for the AdaBoost classification model is the number of estimators and the learning rate. Additional parameters such as maximum depth and maximum leaf nodes were also studied due to the determination of the Decision Tree Classifier as a base estimator for this model. The optimal values of all these hyper-parameters were obtained and used by the grid search approach.

6- Conclusion

This study presented a classification model to identify clear and unclear questions in an asynchronous context for community question and answering networks. This model consisted of four steps and tried to cluster the questions. Afterwards, by extracting several semantic concepts from each cluster, the frequency of these concepts in each question was examined. Finally, a feature vector of the frequency of semantic concepts was constructed for each question. These feature vectors are learned by an ensemble-based model called AdaBoost. Next, the performance of the presented model with other baselines in this field were compared on different datasets and based on the evaluation criteria of accuracy and F1-score, an excellent result was obtained. The outputs obtained in this study can be used to increase the quality of unclear questions in a CQA network, which lead to an increase in its popularity. This can be done by first examining every question that enters the network by this model, and different approaches can be adopted to clarify the question by classifying it as unclear.

References

[1] B. Li, T. Jin, M.R. Lyu, I. King, B. Mak, Analyzing and predicting question quality in Community Question Answering services, in: Proceedings of the 21st International Conference on World Wide Web, Association for Computing Machinery, Lyon, France, 2012, pp. 775–782.

[2] S. Ravi, B. Pang, V. Rastogi, R. Kumar, Great Question! Question Quality in Community Q&A, in: ICWSM, 2014.

[3] M. Asaduzzaman, A.S. Mashiyat, C.K. Roy, K.A. Schneider, Answering questions about unanswered questions of Stack Overflow, in: 2013 10th Working Conference on Mining Software Repositories (MSR), 2013, pp. 97-100.

[4] L.A. Adamic, J. Zhang, E. Bakshy, M.S. Ackerman, Knowledge sharing and yahoo answers: everyone knows something, in: Proceedings of the 17th international conference on World Wide Web, Association for Computing Machinery, Beijing, China, 2008, pp. 665–674.

[5] Y.R. Tausczik, J.W. Pennebaker, Predicting the perceived quality of online mathematics contributions from users' reputations, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, Vancouver, BC, Canada, 2011, pp. 1885–1888.

[6] M. Dehghan, M. Biabani, A.A. Abin, Temporal expert profiling: With an application to T-shaped expert finding, Information Processing & Management, 56(3) (2019) 1067-1079.

[7] S.M. Nasehi, J. Sillito, F. Maurer, C. Burns, What makes a good code example?: A study of programming Q&A in StackOverflow, in: 2012 28th IEEE International Conference on Software Maintenance (ICSM), 2012, pp. 25-34.

[8] A. Bouziane, D. Bouchiha, N. Doumi, M. Malki, Question Answering Systems: Survey and Trends, Procedia Computer Science, 73 (2015) 366-375.

[9] M. Harper, D. Raban, S. Rafaeli, J.A. Konstan, Predictors of answer quality in online Q&A sites, in: 26th Annual CHI Conference on Human Factors in Computing Systems, CHI 2008, 2008, pp. 865-874.

[10] M.P. Kato, R.W. White, J. Teevan, S.T. Dumais, Clarifications and question specificity in synchronous social Q&A, in: CHI '13 Extended Abstracts on Human Factors in Computing Systems, Association for Computing Machinery, Paris, France, 2013, pp. 913–918.

[11] J. Trienes, K. Balog, Identifying Unclear Questions in Community Question Answering Websites, in: ECIR, 2019.

[12] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A Convolutional Neural Network for Modelling Sentences, in, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 655-665.

[13] Y. Kim, Convolutional Neural Networks for Sentence Classification, in, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746-1751.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171-4186.

[15] A. Khabbazan, A.A. Abin, A Topic Based Method to Classify the Question Clarity in CQA Networks, in: 2021 12th International Conference on Information and Knowledge Technology (IKT), 2021, pp. 96-101.

[16] P. Braslavski, D. Savenkov, E. Agichtein, A. Dubatovka, What Do You Mean Exactly? Analyzing Clarification Questions in CQA, in: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, Association for Computing Machinery, Oslo, Norway, 2017, pp. 345–348.

[17] S. Rao, H. Daumé, III, Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information, in, 2018, pp. 2737-2746.

[18] Y. Freund, R.E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 55(1) (1997) 119-139.

[19] T.J. Hastie, S. Rosset, J. Zhu, H. Zou, Multi-class AdaBoostB Statistics and Its Interface, 2 (2009) 349-360.

[20] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics, 3(1) (1974) 1-27.

[21] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20 (1987) 53-65.

HOW TO CITE THIS ARTICLE

A. Khabbazan, A. A. Abin, Classifying the Clarity of Questions in CQA Networks: A Topic based Approach, AUT J Electr Eng, 55(1) (2023) 47-60.

DOI: 10.22060/eej.2022.21287.5468