Journal of Theoretical and Applied Information Technology
31st October 2022. Vol.100. No 20
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ARABIC QUESTIONS CLASSIFICATION MACHINE LEARNING ALGORITIMS

RAMI MALKAWI[1], SAJA ALSRAHAN[2], AHMAD A. SAIFAN[3]

[1]Information Systems Department, Yarmouk University, Irbid, Jordan
[2]Information Systems Department, Yarmouk University, Irbid, Jordan
[3]Information Systems Department, Yarmouk University, Irbid, Jordan
E-mail: [1]rmalkawi@yu.edu.jo, [2]2018930004@ses.yu.edu.jo, [3]ahmads@yu.edu.jo

ABSTRACT

With the increasing volume of data on the web, the user needs a fast, accurate and directly access to information and quick responses to his/her questions. This led to motivating the authors of this paper to adopt the idea of Question Answering Systems; these systems depend widely on the language. There are many research studies have been conducted focusing on the English and Latin languages. However, number of studies related to the Arabic language is very limited and it still needs to more enhancements. Question Answering Systems consist of three main modules (Question Pre-processing, Document Processing and Answer Processing). Question preprocessing form is considered as an important phase on the overall Question Answering Systems because it plays a main role in identifying the question class which effects on the detection of a candidate answer. Within this research study, the focus was on identifying the Arabic question class based on an Arabic taxonomy and building a model for classifying the Arabic question using machine learning algorithms such as Support Vector Machine, Naive Bayes, and Logistic Regression., with an attempt to build a Named Entity Recognition model for Arabic language. Although, the results were not as good as expected, however there is a chance to improve the results in future. The outcomes of the experiment were as follows: logistic regression achieved the highest accuracy 82%, linear SVM 81% and Multinomial 79%.

Keywords: *Question Classification, Arabic QA Systems, Arabic Question Classification, Machine Learning.*

1. INTRODUCTION

The web has become the main source of the information for most people, and the web contains a large amount of data that increases continuously. The user has a specific question in his/her mind and hopes to reach a specific and satisfied accurate answer. Moreover, the user needs to express his/her question in the natural language without the need to adhere to certain rules or use a specific query formulation. With this large amount of data, the user needs a system that meets his/her needs of information and answers his/her question effectively, efficiently, and expeditiously.

The behavior of these systems is supposed to simulate the style of the human being by analyzing the question and searching for answers that meet his/her needs. Search engines such as Yahoo, Google and others provide information as of set of documents. However, these documents may be larger than expected; it may take time from the user to find the answer he/she is looking for. The user always needs precise information at a real time, and this is considered as a challenge and sometimes it can be difficult. The Question Answer system (QAS) is the best alternative for search the engines, so it gives precise information.

The volume of Arabic data on the web is very large, and every day is increasing. There were many research studies related to the QAS for the English language and other Latin and European languages [1] [2] [3] [4] [42]. However, there were a few limited number of researches related to the QAS in the Arabic language due to several reasons such as Difficulty of the language itself in addition to the lack of tools that may assist the researchers in this field. Most components of the QA system are language-dependent, and there is difficulty in designing and evaluating such systems.

Arabic language is one of six official languages in the United Nations; the number of native speakers of the Arabic language is more than 450 million speakers [5]. One of the Arabic language features is that its morphological analysis is very complex such as derivational morphology.

The science of derivation is to find a fit between the two words in the origin of the meaning and the structure, so the reversion of one to the other and taking it from it is known. It is synonymous with etymology in other languages. All Arabic words have three to four characters root verbs.

Moreover, Inflectional morphology, word formation, is the science by which the conditions of the structure of a word are known, and its morphology in various ways with different meanings. Each word consists of a root and zero or prefix, infix, suffix [6]. In addition, the Arabic language contains diacritics, according to which the meaning of the word changes such as “بُر” which means wheat and “بَر” which means land. Therefore, we need to use complex morphological rules to identify tokens and parsing the Arabic text, in addition, the direction of writing in the Arabic language starts from left to right, some letters have different shape depending on where they are in the word.

The Arabic language is divided into three categories [6]:

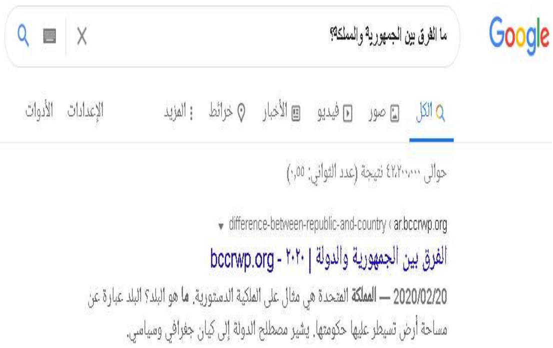
1- Classical Arabic (It is the language of Holy Quran, difficult to understand for a simple listener).

2- Modern Arabic (It follows the rules of Classical Arabic but includes simpler terms and it is the official language of the Arab world).

3- Colloquial Arabic (Informal language based on the regional dialect).

Overall aim of any QA system was to retrieve small pieces of text that contain the precise information to the question rather than the traditional retrieval of the search engine as shown in Figure 1 below. As we can see there were 42,200,000 responses for that query.

QA System combines Information Retrieval IR with Natural Language Processing NLP. Any QA System contains three modules (Question Preprocessing, Document Processing and Answer Processing). These modules are the phases of the QA System. Starting from the first module which is question preprocessing, it is question analysis that includes question classification, keyword extraction, query expansion and derivation of expected answer type. The question classification step is considered the most important phase in the QA System, this is because their results are used as an input to the next phases in the system to arrive at the correct answer the question must be classified correctly. Document Processing and Answer Processing follow the information retrieval process.



ما الفرق بين الجمهورية والمملكة؟

حوالي ٤٢٬٢٠٠٬٠٠٠ نتيجة (عدد الثواني: ٠٫٥٥)

ar.bccrwp.org › difference-between-republic-and-country

الفرق بين الجمهورية والدولة | ٢٠٢٠ - bccrwp.org

2020/02/20 — المملكة المتحدة هي مثال على الملكية الدستورية. ما هو البلد؟ البلد عبارة عن
مساحة أرض تسيطر عليها حكومتها. يشير مصطلح الدولة إلى كيان جغرافي وسياسي.

Figure 1: How The Question Results/Answers Show On The Search Engine

CLEF (Cross Lingual Evaluation Forum) organisers provide an evaluation platform for different languages except for the Arabic language due to the absence of an Arabic text-bed used in the evaluation platform. In this work, the focus will be on identifying the Arabic question class based on an Arabic taxonomy and building a model for classifying the Arabic question using machine learning algorithms which are Support Vector Machine, Naive Bayes, and Logistic Regression. There are many studies that used the SVM algorithm in classifying Arabic and English texts [7] [8] [9] [41], in addition to SVM algorithms use hyperplane to split the classes and finding the maximum margin between the classes to decrease error rate [10], Logistic Regression and naïve bayes are used in statistics to predict the probability of a particular result and is widely popular in classification tasks, especially texts classification in the world of Natural Language Processing (NLP) [11] [12]. In addition, this research is trying to build a Named Entity Recognition model for Arabic language and use it in the Features extraction phase in attempt to improve the Arabic question classification results.

The selection of these three classifiers is that because they were used previously in several studies in English and Arabic text classification, and they all performed very well [6] [12] [13] [14] [42].

Table 1 shows the usage of the three selected classification algorithms in English and Arabic language.

Table 1 The Usage Of The Classification Algorithms In English And Arabic Language

	SVM	**Naïve Bayes**	**Logistic Regression**
Was it used to classify the English question?	Yes, [38] an average accuracy of 74.6% was recorded.	Yes, [39] Naïve Bayes classifier is found to be an effective classifier for most data, but the SVM more appropriate to the complex data.	NA
Was it used to classify the Arabic question?	Yes, [6] an average accuracy of 85% was recorded.	Yes, [13] an average accuracy of the proposed method (0.787295).	Yes, [13] an average accuracy of the proposed method (0.779508).
Are there tools that support this algorithm ?	Yes, almost 7 Free SVM Tools. “KNIME, LIBSVM …”	Yes, such as: KNIME and others.	Yes, such as: RapidMiner and others.

In This research we are going to answer the following questions:

How much it is possible to build a classification model that support the Arabic Question-Answering Systems by using machine learning algorithms to classify Arabic questions and build a Named Entity Recognition model to improve the feature extraction phase?

When the user creates a query on the search engines, he aspires to find an answer to what is going on in his/her mind quickly and accurately, and it is known that two-thirds of the answer understanding the question so in this research the focus will be on the task of classifying the Arabic question by adopting an Arabic taxonomy that was introduced in 2019 and using learning-based approaches to keep abreast of research developments in this field.

Questions Answering Systems are considered one of the best alternatives to search engines in the task of reaching the answer quickly, accurately and concisely, and since the Arabic language is considered one of the important languages in the world and spoken by millions of people, the Arabic language is distinguished by its difficulty and uniqueness of its characteristics from other languages, there are many researches Which was conducted on the English language and some of the Latin languages, but the number of researches for the Arabic language is considered limited and needs to contribute to its development and encourage researchers to delve into this field.

The rest of the paper is organized as follows: section 2 Background about the Information Retrieval and NLP, section 3 Problem Statement of the research, section 4 over views the related work on Arabic question classification, section 5 Methodology of the classification task, section 6 Results and finally section 7 the conclusion.

2. BACKGROUND

Question Answering Systems are considered as field of information retrieval (IR), and natural language processing (NLP), which is concerned with building systems that automatically answer questions that humans ask in natural language [15].

2.1 Information Retrieval

Information Retrieval (IR) is concerned with storage, representation, finding of information and searching of the information that desired by the user, IR is the core filed in information science [16]. It is used to search for specific documents, metadata that describe data, and some other type of data such as (text, images, sounds ...). Information Retrieval operations of the various types of data are now not limited to only keyword matching, IR science is a multidisciplinary field that encompasses data science, data mining, and NLP in addition to artificial intelligence and others.

2.2 Natural Language processing

Natural Language processing (NLP) is a branch of computer science and artificial intelligence (AI) that enables computers to extract meaning from unstructured text. It includes lexical analysis, syntactic analysis which include preprocessing of the text to identify name entity as example and semantic analysis which focuses on the concepts, relations, and predicates.

As mentioned above, Arabic language is divided into three categories: Classical Arabic, Modern Arabic and Colloquial Arabic. NLP in the Arabic language suffers from many challenges one of them is Modern Arabic and Colloquial Arabic may be written by the Roman script (Arabizi) which is an Arabic letter combined with numbers and punctuation marks.

Abdelali and other researchers proposed a fast and accurate Arabic segmenter, named Farasa which is known for its fast execution [17]. MADAMIRA [18] is a segmenter offers many tasks for basic language analysis such as:

1- Feature Modeling: It is a dimensionality reduction technique which is used to remove irrelevant and redundant data which increases accuracy [19].

2- Tokenization: Breaking the text into small chunks called tokens such as words, sentence, paragraph. Tokens is a basic unit contains a textual information of text [20]. The main benefit of tokens is to understand the context and meaning of the text [21].

3- Phrase Chunking: Detection of boundaries between phrases it is segmentation of the sentence to noun phrases verb phrases [22]. It is a process to identify parts of speech and short phrases present in each sentence.

4- Named Entity Recognition: A process of classifying the words or essential pieces of information to a particular category such as people, organizations etc. [23].

These tasks are not included in Farasa. However, Farasa outperformed both MADAMIRA and the Stanford Arabic Segmenter [24].

ARLST is an Arabic Stemmer presented by Abainia and other researchers to be used to remove the prefixes, infixes, and suffixes from the Arabic words [25].

Also, Khalifa and others [26] presented a morphological analyzer named YAMAMA which inspired by the rich output of MADAMIRA and the fast execution of the Farasa, the results obtained from YAMAMA was better with the context of machine translation.

Natural Language Toolkit (NLTK) is a software library in Python language specialized in natural languages processing and computing linguistics that contains a set of NLP algorithms, it is a powerful tool that contains a various text processing libraries [27].

3. RELATED WORK

This section presents the most important studies related to the QA system for the Arabic language.

Research related to the QA system began in 1972 (i.e., SHRDLU) [28], but was limited to some restricted domain with the structural data, research at the present time is aimed at finding possible answers from unstructured data and adopting open-ended questions to provide a modern framework of information retrieval related with QA problem.

At the beginning, the review of the most prominent research related to the question-answering systems in the Arabic language from the earliest to the most recent, one of the first research that worked on this topic is the QARAB system, which is a question-answering system to support the Arabic language. The research relied on a set of Arabic texts extracted from Al-Raya Qatar newspaper. They adopted the keyword matching with matching simple structure to define the possible answer from the test collection that was previously addressed; the results were promising as an initial launch in this area [15]. This research adopted the task of question classification for assist later on define the possible answer.

The Question Answering system is one of the difficult systems in design and evaluation process, Arabia-QA is an Arabic-QA system, the focus was placed on carefully designing the components of this system, starting with the use of Named Entity Recognition (NER) to identify the question class, to get to answer validation module, manually they built a special test-set that contains 200 questions and 11,000 documents extracted from Wikipedia for evaluation task of their architecture, and they have reached a precision up to 83.3% in extracting a list of possible answers of the questions [28]. This research attempts to use the NER in the feature extraction phase for attempt to improve the classifier results.

DefArabicQA system, which is a definitional Arabic question answering system, the focus of the research was on the question type "What is X?", about persons or organizations, in the question analysis module they used the lexical patterns matching approach, the results were very satisfactory about 90% of the questions used have complete definitions in the top-five answers [29]. In this research the focus on the open questions any type of Arabic questions that classified into 7 categories such as Location, Yes/No etc.

Some research conducted on restricted domains such as: Al-Bayan which is Question Answering system that specializes in questions related to the Holy Quran the system retrieves the verses most relevant to the question and then extracts the part that contains the answer from the Holy Quran and interpretation books, they focused on solving the problem of quality of data and the fear of publishing untrusted information or false information, they used the SVM to classify the questions, the results showed an accuracy of 85% of the system using the top 3 results [6], in this research other techniques from machine learning techniques have been used to classify the Arabic questions.

Recently, the trend towards solving the problems of the QA system modules and work to improve it, Al Chalabi and others [30] focused, in their research, on the first module of the Arabic QA system, which is the question classification, they proposed a method to accurately classify the questions for Arabic QA system, using regular expressions of Nooj tool and context free grammars by analyzing the interrogative particles, they used Stanford's parser to design the WH questions patterns, the results showed that out of 200 questions 186 question match with context free grammar, this research will adopting the auto model for classifying the questions, also Lahbari and other researchers [5] proposed a rule-based method to classify Arabic questions, depending on Arabic taxonomy and Li & Roth taxonomy based on a semantic interpretation of the answer type, the results showed an accuracy of 78% of the system. This research used an Arabic taxonomy for question classification task and applying set of machine learning algorithms to build a model for Arabic question classification task.

Also, from the research that was conducted on restricted domain and focused on question classification problem, Hasan and others [31] proposed a question answering system for Islamic Hadith questions used the combination of SVM with pattern-based matching techniques, three type of question were adopted in the study namely "Where", "What" and "Who", the dataset they have used consist of 200 question derived from Sahih Al-Bukhari, the results show that the F-Measure values of the "Where", "What" and "Who" are: 87.66%, 87.93% and 88.39%. This research adopted questions from different domains such as religion, sport, art, history, geography, etc.

Arabic online forums provide a website to allow the Arabic users interact and discuss different topics and solving different issues that they interest, most of the users post a questions on these forums and other users are answering these questions, Romeo and others [32] they were interested in question ranking in community Question and Answering systems in Arabic forums, this is in order to avoid posting previously answered questions and to improve the user experience by re ranking the answers most relevant to the question asked, they addressed the task by using a machine learning algorithms using advanced Arabic text representations, the main contributions were 1-proposed an Arabic language processing pipeline based on UIMA (is a framework that allows systems integration to analyze text documents intended to extract new information relevant to the specific application context) from segmentation to constituency parsing built on top of Farasa Arabic toolkit. 2 –identify the best text fragments in questions by using neural networks. In this work the machine learning algorithms have used to solving the question classification task.

As for the medical field, Dardour and other researchers [33] proposed a new method specialized in Arabic Medical Questions, and they focused on solving the problem of ambiguity in medical questions in the Arabic language they used dictionaries in addition to transducers using Nooj platform to answering the complex medical questions, they gathered medical texts and the medical questions to determine ambiguity patterns by studying the contextual features, their results showed their disambiguation task enhances the F-Measure by 28%. This work is not interested in specific domain.

From recent research related to question classification, Ouatik and Said El Alaoui [34] investigated a combination of continuous distributed word representation with deep neural networks to classify the question label; first they computed high quality vector representations that identify the syntactic and semantic features of the questions words. Then these vector representations used to feed in the deep NN to get Arabic question classification. The results show that the F1 measure value 92% in the term of micro average. This research used the Bag of Word with TF-IDF to represent the features that feed to the machine learning models.

Automatic tracking of the Arabic text on the web is considered one of the important things in the development of text mining operations in the

Arabic language, social media sites contain mostly texts in the form of questions, and these questions are some of which are looking for an answer and some not, like advertisements. Ramzy and others [35] proposed machine-learning approach to build a binary classifier which classify the social media contents which contain questions to questions that seeking of an answer or not, the main contribution of their work is using the emotional based features with the lexical features in the classification process, the results show that the emotional features had an impact in enhancing the results. In future adopting the emotional features in the feature extraction process may enhance the classifier accuracy.

From the research that adopted the learning-based approach, Hamza and his team [36] have investigated the use of Embeddings from Language Models from Language Models ELMo which is contextual continuous word representation, then they applied the neural network to classify the Arabic questions by study the behavior of the words representation, they conducted comparison between the context-free words representation with ELMo, the ELMo get a better performance the results show that the percentage of the accuracy they achieved is 94%. The nature of the dataset in this research are questions in the modern Arabic language, so the focus was on extracting the most important features that could affect the question classification process. In future there is an attempt to use the dataset containing question written in colloquial Arabic and in this case, the need to study the context of the questions is very important.

4. METHODOLOGY

In this section, the Arabic-Question classification model is presented with its components, starting from data collection phase to the question classification phase that classifies the Arabic-Question based on the novel Arabic taxonomy proposed by Hamza and his researchers team [37] using machine learning algorithms: Support Vector Machine, Naïve Bayes and Logistic Regression.

4.1 Data Collection

Due to the limited numbers of Arabic-QA systems research, there is no Arabic-Question test-bed available on the internet, for this task the dataset available online from Kaggle was used that was published for the Semantic Question Similarity task that conducted in the NSURL 2019 Kaggle competition (NSURL 2019: Task8).

4.2 Data Description

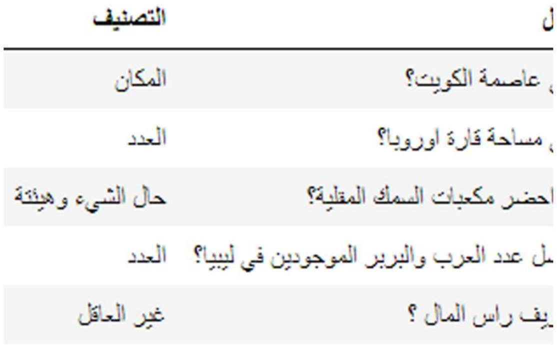
The dataset consists of two columns which represent a pairs of Arabic questions (Question 1, Question 2) that have the same meaning for the Semantic Question Similarity task, but for the Arabic Question classification task the question of the dataset was labeled manually based on the Arabic taxonomy that classify the Arabic question into seven categories which is ("Human, Group", "Entity", "Animals", "Status, Structure", "Location", "Time", "Numbers" and "Yes/No") Table 1 shows the interpretation for these categories [37].

Table 1: Table 1: Arabic Question Taxonomy

Class	Explanation
Human, Group العاقل	who are they, who is she, who is he…? العاقل: الأسئلة التي تحتوي على "من هو، من هم، من هي، من التي ...
Entity, Animals غير العاقل	Questions that ask about animals, organizations, about any entity. غير العاقل: الأسئلة التي تسأل عن حيوانات، منظمات عن أي شيء غير عاقل.
Status, Structure حال الشيء وهيئته	Questions that are looking for a way, how to prepare something or the steps of a certain thing. حال الشيء وهيئته: الأسلة التي تبحث عن طريقة، عن كيفية إعداد شيء أو خطوات شيء معين.
Location المكان	Questions looking for names of capitals, continents, cities المكان: الأسئلة التي تبحث عن أسماء عواصم، قارات، مدن...
Time الزمان	Questions that answer a specific date or time, such as: “When did the Battle of Karameh begin?” الزمان: الأسئلة التي تكون إجابتها تاريخ أو وقت معين مثل: "متى بدأت معركة الكرامة؟"
Numbers العدد	Questions that contain "how many?", "how many times?” العدد: الاسئلة التي تحتوي على "كم عدد"، "كم مره... "
Yes/No التصديق	Questions that begin with “DO, Does, is, are…” that answer “yes” or “no”. التصديق: الأسلة التي تبدأ ب "هل ..." والتي يكون جوابها عبارة عن "نعم" أو "لا."

https://www.kaggle.com/c/nsurl-2019-task8

Figure 2 and Table 2 show the First 5 records from the data, and Figure 3 shows random sample from the data.

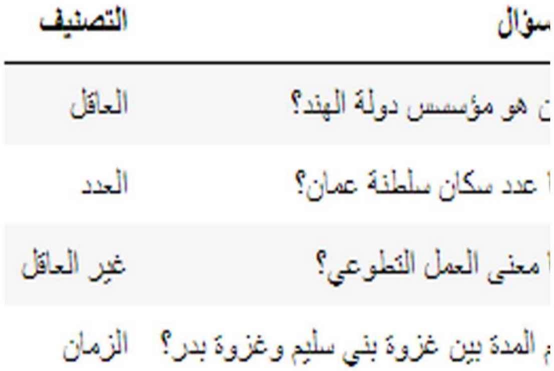


التصنيف	ل
المكان	عاصمة الكويت؟
العدد	مساحة قارة اوروبا؟
حال الشيء وهيئتة	احضر مكعبات السمك المقلية؟
العدد	ل عدد العرب والبربر الموجودين في ليبيا؟
غير العاقل	يف راس المال ؟

Figure 2: The First 5 Records From The Dataset

Table 2: The First 5 Records From The Dataset

Class	Question
Location	What is the capital of Kuwait?
Numbers	What is the area of Europe?
Status, Structure	How do you get crisp fried fish?
Numbers	How many Arabs and Berbers are in Libya?
Entity, Animals	What is the definition of capital?

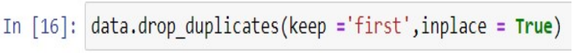


التصنيف	سؤال
العاقل	هو مؤسسس دولة الهند؟
العدد	عدد سكان سلطنة عمان؟
غير العاقل	معنى العمل التطوعي؟
الزمان	المدة بين غزوة بني سليم وغزوة بدر؟

Figure 3: Random Sample From The Dataset

4.3 Data Exploratory

During the data exploration phase, it was noticed that the questions are from different domains and are not related to a specific domain. It is open questions written in modern Arabic. In addition to that the number of questions was 2709 questions classified according to the Arabic taxonomy, but it was noticed that some questions were repeated. 30 questions contain missing information so we removed them. Moreover, 98 duplicate questions were removed from the dataset using statement shown in Figure 4, and the repetition of each category showed it is possible that some kinds of bias may occur, which may affect the accuracy of the model. Table 3 shows the number of questions presented and the dataset for each class label.



```
In [16]: data.drop_duplicates(keep ='first',inplace = True)
```

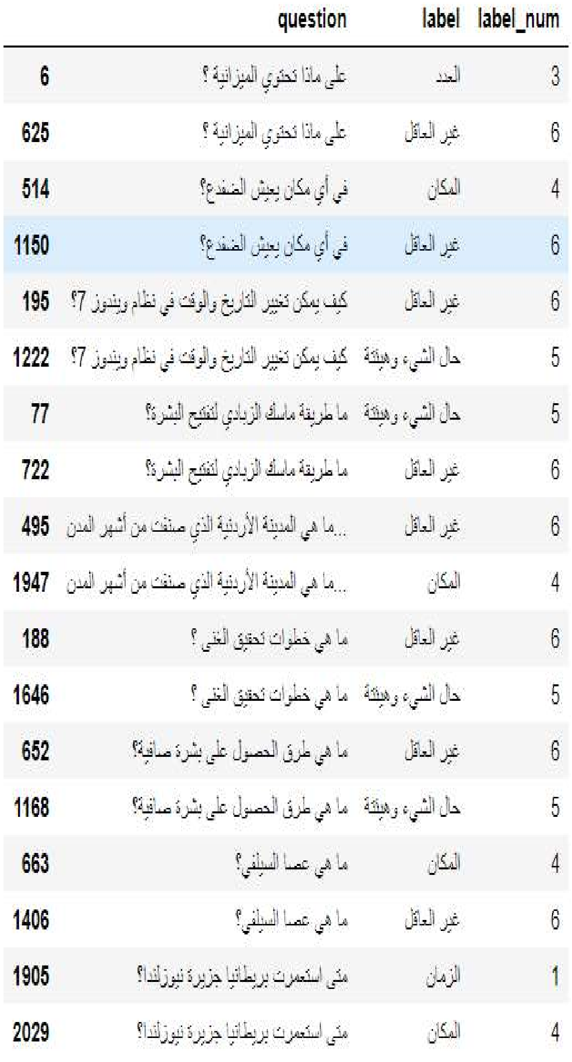
Figure 4: Remove Duplicate Data

What is meant by bias? Is that we do not want the classified model to be biased to the lesser category such as "العاقل" with 45 question count, as an example. So, we need to balance the number of questions associated with each category. Several questions have been manually added to each category reaches to almost 250 questions count.

Table 3: Shown The Number Of Questions For Each Class

Class Label	Class	Question Count
0	العاقل	45
1	غير العاقل	867
2	حال الشيء وهيئته	465
3	المكان	192
4	الزمان	129
5	العدد	179
6	التصديق	45

It is noticeable that there are many questions that belong to more than one category at the same time, such as "What is the capital of Jordan?" " ما هي عاصمة الاردن؟" this question may be classified into "المكان" ,"غير العاقل" or both "المكان وغير العاقل", as is shows in Figure 5.



	question	label	label_num
6	على ماذا تحتوي الميزانية ؟	العدد	3
625	على ماذا تحتوي الميزانية ؟	غير العاقل	6
514	في أي مكان يعيش الضفدع؟	المكان	4
1150	في أي مكان يعيش الضفدع؟	غير العاقل	6
195	كيف يمكن تغيير التاريخ والوقت في نظام ويندوز 7؟	غير العاقل	6
1222	كيف يمكن تغيير التاريخ والوقت في نظام ويندوز 7؟	حال الشيء وهيئته	5
77	ما طريقة ماسك الزبادي لتفتيح البشرة؟	حال الشيء وهيئته	5
722	ما طريقة ماسك الزبادي لتفتيح البشرة؟	غير العاقل	6
495	ما هي المدينة الأردنية الذي صنفت من أشهر المدن...	غير العاقل	6
1947	ما هي المدينة الأردنية الذي صنفت من أشهر المدن...	المكان	4
188	ما هي خطوات تحقيق الغنى ؟	غير العاقل	6
1646	ما هي خطوات تحقيق الغنى ؟	حال الشيء وهيئته	5
652	ما هي طرق الحصول على بشرة صافية؟	غير العاقل	6
1168	ما هي طرق الحصول على بشرة صافية؟	حال الشيء وهيئته	5
663	ما هي عصا السيلفي؟	المكان	4
1406	ما هي عصا السيلفي؟	غير العاقل	6
1905	متى استعمرت بريطانيا جزيرة نيوزلندا؟	الزمان	1
2029	متى استعمرت بريطانيا جزيرة نيوزلندا؟	المكان	4

Figure 5: Multiclass Question Issues

This research has relied on adopting the case of single classification, noting that the proposed model is able to deal with the Arabic question with multiclass classification, if it is fed by a dataset that is designed into multiple categories/classes.

4.4 Question Analysis Model

The proposed model built on three machine learning algorithms which are Support Vector Machine, Naïve Bayes and Logistic Regression, Figure 6 shows the overall process of the Arabic Question Classification model. Below we described each process in detailed.

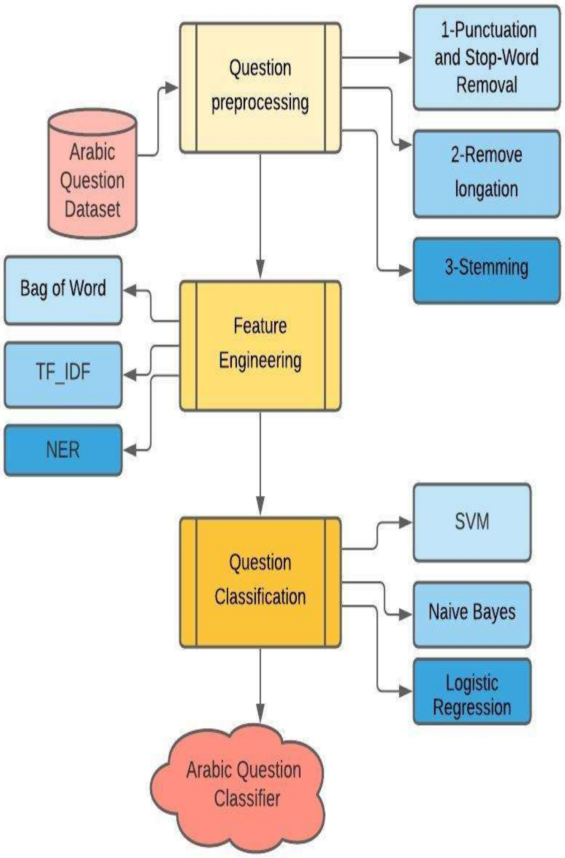
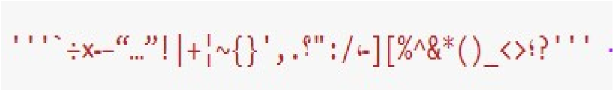


Figure 6: The Overall Process Of The Arabic Question Classification Model

4.4.1 Question preprocessing

4.4.1.1 Punctuation and Stop-Word Removal

First, we define a list of Arabic and English punctuation that will get rid in text. List of these marks represents in figure 7.



```
'''`÷×-"…"!|+¦~{}',.؟":/-][%^&*()_<>؛?'''
```

Figure 7: List Of Punctuation Marks

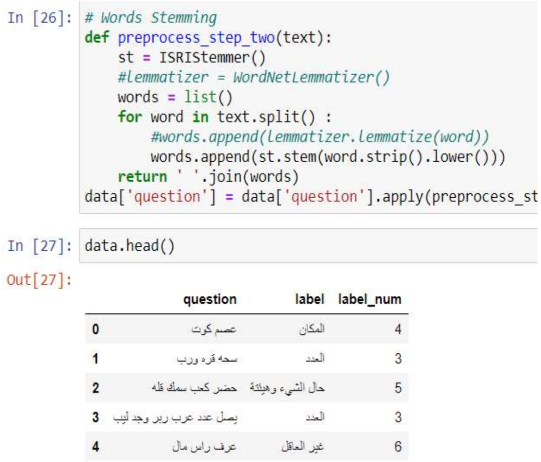
After that we removed the Arabic stop words using **nltk library** in python, such as “بكم” and “على” etc. In addition to remove a diacritic such as “◌ُ◌َ◌ْ~◌ً◌ٍ◌ٌ”.

4.4.1.2 Remove longation

Then we removed longation, to delete repeated letters such as transform the “إأآى” to “ا”, “ك” to “گ”, “ء” to “ى ؤ ئ”.

4.4.1.3 Stemming

Stemming is used to reduce variant word forms to common roots, Figure 8 shows the choices for transform the Arabic words to its root by using "stemmer" uses the stem of the word or "lemmatizer" uses the context in which the word is being used, both are methods supported by **nltk library** in python.



```
In [26]: # Words Stemming
         def preprocess_step_two(text):
             st = ISRIStemmer()
             #lemmatizer = WordNetLemmatizer()
             words = list()
             for word in text.split() :
                 #words.append(lemmatizer.lemmatize(word))
                 words.append(st.stem(word.strip().lower()))
             return ' '.join(words)
         data['question'] = data['question'].apply(preprocess_st

In [27]: data.head()

Out[27]:
```

	question	label	label_num
0	عصم كوت	المكان	4
1	سحه قره ورب	العدد	3
2	حضر كعب سمك قله	حال الشيء وهيئتة	5
3	يصل عدد عرب ربر وجد لبب	العدد	3
4	عرف راس مال	غير العاقل	6

Figure 8: Stemming Step

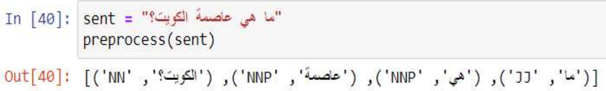
We can see from Figure 8, in our study we adopted the stemmer rather than lemmatizer, because it reduced the number of features by almost 1000 feature in comparison with the lemmatizer results, so we want to maintain a large number of words represented with few features.

4.4.2 Feature Engineering

We treat a question in any QA system as a query, so this phase is very important from a semantic and syntactic point of view to assist to find the expected answer by understanding the question first.

This research applied the Bag of Word technique with TF-IDF in addition to the NER which is considered a part of research contribution, to identify the most important features or question words that help into find the proper class to any Arabic question. Bag of Word with TF-IDF gave almost the same results; the number of extracted features was 1921 features in both.

Figure 9 shows an example of applying the part of speech tagging to the Arabic question, which the "JJ" refers to adjective, "NNP" refer to proper noun and "NN" refer to noun singular or mass.



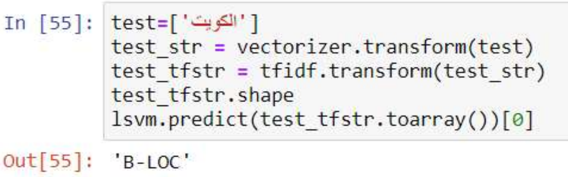
```
In [40]: sent = "ما هي عاصمة الكويت؟"
         preprocess(sent)

Out[40]: [('ما', 'JJ'), ('هي', 'NNP'), ('عاصمة', 'NNP'), ('الكويت؟', 'NN')]
```

Figure 9: The Result Of Applying Pos_Tag On The Arabic Question

But when trying to apply NER, there is no suitable model available for the Arabic language, which prompted to build a new model as an attempt to test the results of applying the NER to the Arabic questions.

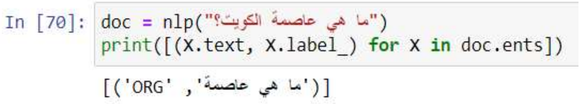
So we built another model for the NER using the dataset available on the internet that classified according to list of named entity types (O, PERSON, LOCATION, ORGANIZATION, MISCELLANEOUS), we applied liner SVM in the classification task the accuracy was 95%, figure 10 below show a sample of the NER results for the "الكويت" word which classified as B-Loc (location).



```
In [55]: test=['الكويت']
         test_str = vectorizer.transform(test)
         test_tfstr = tfidf.transform(test_str)
         test_tfstr.shape
         lsvm.predict(test_tfstr.toarray())[0]

Out[55]: 'B-LOC'
```

Figure 10: A Sample Of NER Classification Test



```
In [70]: doc = nlp("ما هي عاصمة الكويت؟")
         print([(X.text, X.label_) for X in doc.ents])

         [('ما هي عاصمة', 'ORG')]
```

Figure 11: A sample of NER classification test (2)

In Figure 11 the question of "what is the capital of Kuwait?" classifies as "organization" which is not true which means that the NER model is not accurate enough. NER model has been dispensed due to limited time of this work but in future there is an attempt to improve the NER results for Arabic language, this model is known to have a strong effect on improving question classification results for Arabic QA systems.

4.4.3 Question Classification

Machine learning algorithms have the best performance in text classification and CNN algorithms are a good candidate for classifying Arabic texts due to the difficulty of the language itself [10,40], we applied the CNN algorithm to classify the Arabic Questions, but the result was not good because the deep learning algorithms need a big dataset to get more efficient results. We used the Support Vector Machine, Naive Bayes, and Logistic Regression to classify the Arabic Questions.

5. RESULTS

To answer the question presented in the introduction which is How much it is possible to build a classification model that support the Arabic Question-Answering Systems by using machine learning algorithms? We examine evaluation results that were achieved by the three popular machine learning classifiers we were used in our research: SVM, Naive Bayes and Logistic Regression. We compared the performance of different classification algorithms to classify the Arabic question class.

Determining the type of data and the type of problem helps in choosing what is the appropriate classification model. It is always preferable to compare more than one of classification models on a given data set, and in our case, there is no testbed for Arabic Questions that used as a reference to other studies in this filed, it is important to compare more than one classification model on this dataset, this research relied on the use of a dataset that containing 2581 questions in the Arabic language classified according to an Arabic question taxonomy proposed above [37], for training the model 2064 questions and 517 for testing are used, as mentioned earlier

The extracted features using Bag of Words and TF-IDF were (1921) feature fed into the classifiers: two types of SVM which are Linear SVM and RBF SVM, in addition to three types of Naïve Bayes which are GaussianNB, MultinomialNB, BernoulliNB, and the Logistic Regression. Then we computed the overall accuracy that measure the goodness of classification as a ratio of correctly predicted instances from the dataset. Figure 12 represent the histogram of models accuracy and Table 4 show the accuracy percentage to each of them.

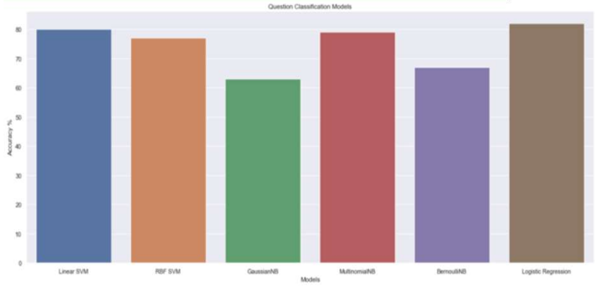


Figure 12: Arabic Questionclassification Models Accuracy

Table 4: Models Accuracy

Models	Accuracy %
Linear SVM	81%
RBF SVM	77%
GaussianNB	63%
MultinomialNB	79%
BernoulliNB	67%
Logistic Regression	82%

In most of the previous research, the SVM and the Logistic Regression achieved the highest accuracy, and as noted from the above table that the accuracy values of Linear SVM, MultinomialNB and Logistic Regression are somewhat close and considered good, however, there is a need to improve the model accuracy by working to increase the dataset size to train the model more to give better results.

6. THREAT TO VALIDITY

There are two kinds of threat to the validity of the experiment presented in this paper. The primary and most pressing one comes from the dataset used in the experiment. Due to the limited numbers of Arabic-QA systems research and because there is no Arabic-Question test-bed available on the internet we used only one dataset. Clearly, in order to make more general statements about the effectiveness of building a model for classifying the Arabic question using machine learning algorithms, it would be necessary to apply the approach introduced in this paper to more datasets.

The second threat to the validity of this work comes from the machine learning algorithms used in the experiment. It would be necessary to apply the approach introduced in this paper to all machine learning algorithms.

7. CONCLUSION

This research has attempted design a new approach for the Arabic Question Classification model for the Arabic Question Answering systems, the questions from different domains classified based on novel Arabic taxonomy proposed by Hamza and his team researchers [37]. The new approach began experimenting with more than one of machine learning algorithms which are SVM, Naive Bayes, and Logistic Regression. This research attempted to contribute in building a NER model for Arabic language but the results was poor and the researcher looking for improve the results in future. In this paper we showed that the machine learning algorithms used are The outcomes of the experiment are good, logistic regression achieved the highest accuracy 82%, linear SVM 81% and MultinomialNB 79%.

REFERENCES:

[1] Olvera-Lobo, M. D., & Gutiérrez-Artacho, J. (2011, September). Multilingual question-answering system in biomedical domain on the web: an evaluation. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 83-88). Springer, Berlin, Heidelberg.

[2] Waltz, D. L. (1978). An English language question answering system for a large relational database. *Communications of the ACM*, *21*(7), 526-539.

[3] Laurent, D., Séguéla, P., & Nègre, S. (2006, September). Cross lingual question answering using qristal for clef 2006. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 339-350). Springer, Berlin, Heidelberg.

[4] Quaresma, P., Quintano, L., Rodrigues, I., Saias, J., & Salgueiro, P. (2004, September). University of Évora in QA@ CLEF-2004. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 534-543). Springer, Berlin, Heidelberg.

[5] Lahbari, I., Ouatik, S. E. A., & Zidani, K. A. (2017, November). A rule-based method for Arabic question classification. In *2017 international conference on wireless networks and mobile communications (WINCOM)* (pp. 1-6). IEEE.

[6] Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N. M., & Torki, M. (2014, October). Al-Bayan: an Arabic question answering system for the Holy Quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 57-64).

[7] Liao, W. (2008). An implicit fourth-order compact finite difference scheme for one-dimensional Burgers' equation. *Applied Mathematics and Computation*, *206*(2), 755-764.

[8] Young, D. (1954). Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society*, *76*(1), 92-111.

[9] Young, D. M. (1970). Convergence properties of the symmetric and unsymmetric successive overrelaxation methods and related methods. *Mathematics of Computation*, *24*(112), 793-807.

[10] Mazlin, I., Rawi, I. M., & Zakaria, Z. (2021). Hadith arabic text classification using convolutional neural network and support vector machine. In *Computational Science and Technology* (pp. 371-382). Springer, Singapore.

[11] Gao, H., Zeng, X., & Yao, C. (2019). Application of improved distributed naive Bayesian algorithms in text classification. *The Journal of Supercomputing*, *75*(9), 5831-5847.

[12] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1-16.

[13] Alammary, A. S. (2021). Arabic questions classification using modified TF-IDF. *IEEE Access*, *9*, 95109-95122.

[14] Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *Int. Arab. J. e Technol.*, *2*(2), 124-128.

[15] Hammo, B., Abu-Salem, H., Lytinen, S. L., & Evens, M. (2002, July). QARAB: A: Question answering system to support the Arabic language. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*.

[16] Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing.

[17] Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016, June). Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations* (pp. 11-16).

[18] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014, May). Madamira: A fast,

comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 1094-1101).

[19] Anoual, H., Aboutajdine, D., Elfkihi, S., & Jilbab, A. (2010, September). Features extraction for text detection and localization. In *2010 5th International Symposium On I/V Communications and Mobile Network* (pp. 1-4). IEEE.

[20] Park, K., Lee, J., Jang, S., & Jung, D. (2020). An empirical study of tokenization strategies for various Korean NLP tasks. *arXiv preprint arXiv:2010.02534*.

[21] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *COLING 1992 volume 4: The 14th international conference on computational linguistics*.

[22] Daelemans, W., Van den Bosch, A., Zavrel, J., Veenstra, J., Buchholz, S., & Busser, B. (1998). Rapid development of NLP modules with memory-based learning. *Proceedings of ELSNET in Wonderland*, 105-113.

[23] Mikheev, A., Moens, M., & Grover, C. (1999, June). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 1-8).

[24] Monroe, W., Green, S., & Manning, C. D. (2014, June). Word segmentation of informal Arabic with domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 206-211).

[25] Abainia, K., Ouamour, S., & Sayoud, H. (2017). A novel robust Arabic light stemmer. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(3), 557-573.

[26] Khalifa, S., Zalmout, N., & Habash, N. (2016, December). Yamama: Yet another multi-dialect Arabic morphological analyzer. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations* (pp. 223-227).

[27] Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

[28] Benajiba, Y., Rosso, P., & Lyhyaoui, A. (2007, April). Implementation of the ArabiQA question answering system's components. In *Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April* (pp. 3-5).

[29] Trigui, O., Belguith, L. H., & Rosso, P. (2010). DefArabicQA: Arabic definition question answering system. In *Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta* (pp. 40-45).

[30] Al Chalabi, H. M., Ray, S. K., & Shaalan, K. (2015, May). Question classification for Arabic question answering systems. In *2015 International Conference on Information and Communication Technology Research (ICTRC)* (pp. 310-313). IEEE.

[31] Hasan, A. M., & Rassem, T. H. (2018, June). Combined support vector machine and pattern matching for arabic islamic hadith question classification system. In *International Conference of Reliable Information and Communication Technology* (pp. 278-290). Springer, Cham.

[32] Romeo, S., Da San Martino, G., Belinkov, Y., Barrón-Cedeño, A., Eldesouki, M., Darwish, K., ... & Moschitti, A. (2019). Language processing and learning models for community question answering in arabic. *Information Processing & Management*, *56*(2), 274-290.

[33] Dardour, S., Fehri, H., & Haddar, K. (2019, June). Disambiguation for Arabic Question-Answering System. In *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ* (pp. 101-111). Springer, Cham.

[34] Ouatik, S. E. A. (2020). Exploring Convolutional Neural Networks and Recurrent Neural Networks for Arabic Question Classification. *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019): Volume 4-Advanced Intelligent Systems for Applied Computing Sciences*, *1105*, 233.

[35] Ramzy, A., & Elazab, A. (2020). Question Identification in Arabic Language Using Emotional Based Features. *arXiv preprint arXiv:2008.03843*.

[36] Hamza, A., En-Nahnahi, N., & Ouatik, S. E. A. (2020, April). Exploring contextual word representation for Arabic question classification. In *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1-5). IEEE.

[37] Hamza, A., En-Nahnahi, N., Zidani, K. A., & Ouatik, S. E. A. (2021). An arabic question classification method based on new taxonomy and continuous distributed representation of

words. *Journal of King Saud University-Computer and Information Sciences*, *33*(2), 218-224.

[38] Sangodiah, A., Muniandy, M., & Heng, L. E. (2015). Question Classification Using Statistical Approach: A Complete Review. *Journal of Theoretical & Applied Information Technology*, *71*(3).

[39] Panicker, Arun D., U. Athira, and Sreesha Venkitakrishnan. "Question classification using machine learning approaches." *International Journal of Computer Applications* 48, no. 13 (2012): 1-4.

[40] Saifan, A.A.; and Al Smadi, N. (2019). Source code-based defect prediction using deep learning and transfer learning. Intelligent Data Analysis, 23(6), 1243-1269.

[41] Saifan A. A., “Test case reduction using data mining classifier techniques”, Journal of Software, Vol. 11, No. 7, pp. 656-663, 2016.

[42] Nahar, K. M., Ra'ed, M., Moy'awiah Al-Shannaq, Daradkeh, M., & Malkawi, R. (2020). Direct text classifier for thematic arabic discourse documents. Int. Arab J. Inf. Technol., 17(3), 394-403.