Different Approaches for Multi-Class Classification using Machine Learning Techniques

Prashant Y Niranjan, Vijay S Rajpurohit

Department of Computer Science & Engineering, KLS Gogte Institute of Technology, Belagavi.

ABSTRACT

Web content increasing in every sector has become challenging task to find the useful information. Question Answering system in agriculture domain help farmers to provide the accurate answer. Farmer asking the queries relevant to pomegranate fruit in which question classification plays an important role for various questions asked by farmer to categorize or to identify the type of question. Different question classification methods have been proposed to provide solutions for classification of question. In this research, we are considering the different pomegranate questions that can be asked by farmer to classify the questions correctly using different machine learning methods. A proposed framework for question classification having multiple classes i.e. name, descriptive, location, numeric and entity: other which enables machine learning algorithms to categorize the type of question. This paper compares various machine learning algorithms and result shows K-Nearest Neighbor, SVM & Decision tree performed well with good accuracy.

Keywords: Question classification; Machine learning, Natural Language Processing, Text Mining.

1. Introduction

Agriculture is the need of every country and it increases the economy of a country. Providing the needs of our farmer is one of the important tasks to carry out their work effectively and smoothly. In this research pomegranate fruit has been considered, Pomegranate is a commercial fruit plant which belonging to Punicaceae family [1]. It has good source of protein, minerals, antioxidants, carbohydrate, vitamins A, B and C, also it's used for restrain or to control heart diseases, cancer, fever, leprosy, abdominal pain etc. Pomegranate is in more demand; its production is increasing day by day and exporting over the past years and earning higher profit from the production.

Huge amount of data flow becomes very difficult to get the accurate content which the user is looking for. Current search engine do not have eliminating capability to reduce the content and providing the exact answer to users. Normally it gives number of URL for the user asked questions which makes tedious and time consuming task for the user to open the correct link to find the required or exact answer.

Considering these drawbacks we have proposed the development of question answering system. Farmer growing the pomegranate in his field has lot of queries about various parameters which may involve water, growing methods, soil, disease etc., hence QA system has been introduced to provide solution to every user according to their requirement and can be used in all sectors.

Question answering system is used to provide the intercommunication between human languages and computer, it programs the machine to understand and analyze the large amounts of human language data. QA is a slice of NLP, wherein developing the farmer Question answering system that has been trained to provide automatic responses to the questions asked by farmer in natural language to receive the exact answers.

QA system has become one of the booming areas of research, for identification of questions types question classification module plays a major role in the classification process.

QA system has become user friendly, improving the user needs and methods/techniques used in finding the associated results, because of drastic growth in the amount of web content and getting lesser responses to the asked queries. For the expected answer type, the task of classification process is to assign the accurate class outcomes to the queries asked by the user.

Based on class outcome of a question, it displays the reply to the user questions. So performing classification of questions directly affects the results or answers, miss-classification of questions in QA system leads to the erroneousresults.

Now a day most of the farmers are literate, they have good ideas about how to generate the revenue from their growing field, based on the climate, soil and water conditions most of the people choose some items like fruits, vegetables or some grains to grow in their field. To grow any kind of item we need support of technology to get some help in online mode and the same kind of techniques/tools/procedures can be applied by the farmer to grow in his/her field.

Farmer growing the pomegranate has number of queries like water, soil, new techniques/technologies/tools and disease related issues. He/she typing the question for his problems and the type of question will be classified as Descriptive, Name, Location, Numeric & Entity: other, this will help our system to extract the exact answer to the farmer. Here the task of proving accurate answers to farmers is more dependent on the question type; hence question classification task directly affects the answer. Based on expected answer type, classification task is used to allot the appropriate class outcome to questions.

Most used algorithm for multi-class classification is Support Vector Machine & Logistic regression; even other ML algorithms used are Nearest Neighbor, Decision Tree, Naive Bayes & Neural networks.

In this research, we propose the pomegranate data set consisting of 300 questions built from authentic NRCP file [1]. Here comparison of different ML algorithms for question classification process has been performed for the asked questions.

This paper is organized as: Introduction of the paper is presented in Section 1; Previous work carried out in QA system and machine learning techniques used for question classification approaches outlined in section 2. Section 3 discusses the research objectives. Different question classification approaches are described in section 4. The experimental setup and results are shown in Section 5; results are discussed in Section 6. Finally, Section 7 concludes the paper and outlines directions for future work.

2. Background

Previous work on question classification based on user intent has been discussed in this section. Different class labels of user intent are described in Section 2.1, previous work on question classification techniques/methods are described in section 2.2., and user asked questions the figure 1. shows the different steps applied for QA system to get the accurate answer. Different algorithms or techniques have been provided for the different modules of QA system.

A. Question Categories

Different class labels for questions are defined, which is shown in below table 1. Major question types are: Descriptive type, Name, Numeric, Location and Entity: other.

Table 1. Categories of question.

Authors	Categories
[2]	definition, casual, relationship, Factoids, list, procedural, hypothetical, and confirmation questions
[3]	Solution, Definition, Navigation, Fact, List, Reason,
[4]	Cause and effect, Definition, Rationale, significance Advantage/Disadvantage, Comparison, Explanation, Identification, List, Opinion,
[5]	Entity. Human, Location and Numeric as coarse classes; colour, city, Abbreviation, Description

Different question labels are available like factoid, definition, procedural, List, description, opinion, solution, explanation & Reason all these classes are available under the Descriptive type question. A descriptive type question usually starts with Wh-type questions where it normally uses what and How keyword.

For Name type label, it is used to find the answer for name type question. It uses the WHICH keyword to ask the questions and needs to find the name that can be fruit name, vitamin name, disease name or seed name etc.

For Numeric label, it requires to find the numeric relevant answers like quantity, value, measurement, feet, inch, weight and profit/loss. It uses how much, What, When, How many keywords are used to ask the questions to find the numerical related answer to the user.

For Location label, it requires to find the location relevant answers like place, country, state and area. It uses WHERE keyword to ask the questions to find the location related answer to the user.

For Entity: other label, it requires to find the kind of object information except from the given labels. User is asking the question relevant to any real time entity or other questions will be classified as Entity: other to find the entity related or any other answer to user.

2.2 Question classification methods

This section discusses the previous work carried out on question classification methods and used different ML algorithms.

Authors in Alaa Mohaseeb et al [7]. have proposed using grammatical features for Classification of factoid questions intent, they have used techniques of SVM and J48 and it contains the Grammatical features the word like what, which, who, when. Domain specific features contain the features like religious terms, health terms, and events. Grammatical patterns have structure for type of question. It has been proposed for type of question to be classified and have limitation where rule based and grammatical patterns are used and it is not accessing the recent and authorized links.

Muhammad Wasim et al. proposed the Classification of Factoid and List type Questions in Biomedical Question Answering for Multi-Label Question [8]. Authors have highlighted the *technique of binary* relevance transformation method and with consideration of extracting features (Question Focus). It's been used to classify list and factoid type questions but using this technique it's a time consuming process.

Bastian Haarmann et al. have proposed the [9] Mighty Data set for stress testing QA system, which uses SPARQL Queries. It's used to extract data from fixed set of NLP questions. Its limitations is that it requires separate SPARQL queries to be written for all different types of questions and works only for fixed set of questions.

Hasangi Kahaduwa et al [10] proposed the Question Answering system for the travel domain. Techniques used are linear SVM algorithm, rule based approach and SPARQL queries are used. It has been designed to provide the answer for travel related queries and works only for fixed set of NLP questions.

M.R. Sumalatha et al [11] "*Analyzing and predicting knowledge of contributors in community question answering services" Here the* knowledge contributors are involved and its based on the number of likes and votes. This works good for all the domain related queries and its open for all the people question and answers. This system is very time consuming and it cannot assure that all responded answers are accurate.

Sharvari Gaikwad et al [12] ***"AGRI-QAS question-answering system for agriculture domain" [12]*** have used the NER and POS tagging , system works only for factoid questions such as 'which', 'what', 'where', 'who' and for semantic type questions it does not perform well.

Payal Biswas et al [13] "A framework for restricted domain Question Answering System" have used the Alchemy content extractor, Paragraph extractor, standford coreNLP toolkit, WH-type questions, head word. Here answer extraction works good when the keywords, Headword is present and the answer statement is in correct format. Here rule based approach is used, if answer statement is not in the proper format answer extraction fails to work.

Haarmann et al. (2018) have proposed [14] the QA system which has mighty dataset of stress-testing and it contains large number of NLP questions and SPARQL queries are used.NLP questions are in fixed format and it uses SPARQL queries. Here a given natural language question is converted into a suitable SPARQL query that displays the correct answer.

Gautre et al. (2018) have proposed [15] a QA system which has the submitted question and produces it to the speaker, crowd sourcing application designed to produce the suitable answer to the questioner.

Devi and Dua (2017) et al. [16] developed a QA system for agriculture domain using ontology. ontology web language and Resource description form of data available. SPARQL protocol and RDF query language is used to extract this data.

3. Research Objective

In order to provide valuable support to the farmers in decision making by providing the precise answers to the farmers when they don't have knowledge about the asked query.

In this system firstly it finds out the type of question, expected type of answer, getting keywords of the question and focus of the question then it collects or retrieves the data from the WWW, news, articles etc and then finally it extracts the exact answer using the trained model which is developed using machine learning techniques.

QA system it performs the following functions:

i) Question processing module
ii) Document processing module
iii) Answer extraction module

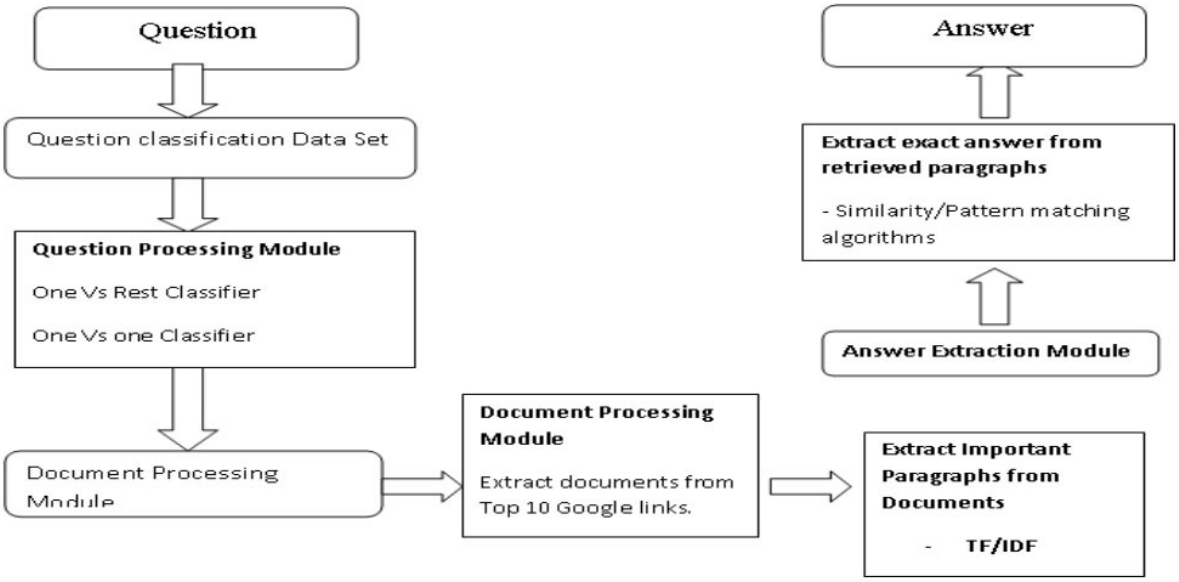

Fig. 1 Question Answering System Architecture

Our research is on QA system for particular or closed domain. Mainly the QA system is used for providing the accurate answer to the user and it has mainly three steps to provide the exact answer. In this paper we are working on question classification task to classify the given question accurately for the expected answer type.

Our research focuses on QA system for pomegranate fruit, Farmers those who are into agriculture and growing specifically the pomegranate fruit has number of queries, especially those are new into farming of this fruit. To provide solution for this problem we are designing and developing a QA system which helps all the famers.

In this paper, we are working on question classification task of pomegranate QA system which is used to provide the expected answer type according to the categorization of question.

The goal of this research paper is to:

1. Investigating the different types of pomegranate relevant questions under multi-class to check the performance ofquestion classification system.
2. Evaluating and comparing the different ML algorithms used for classification task of a question to check theperformance of different algorithms.
3. Evaluating the result of each algorithm to conclude the best algorithm of machine learning can be used for multiclass classification.

In the first objective some sample question figure/questions will be taken and illustrations of data set with different categories are discussed.

Second Objective is to compare and discuss the various ML algorithms for classification of question.

Third objective shows the result of each algorithm and discussing the same with code implementation. Afterdiscussing the result part of each algorithm and it concludes with the highest accuracy giving algorithm.

4. System Architecture

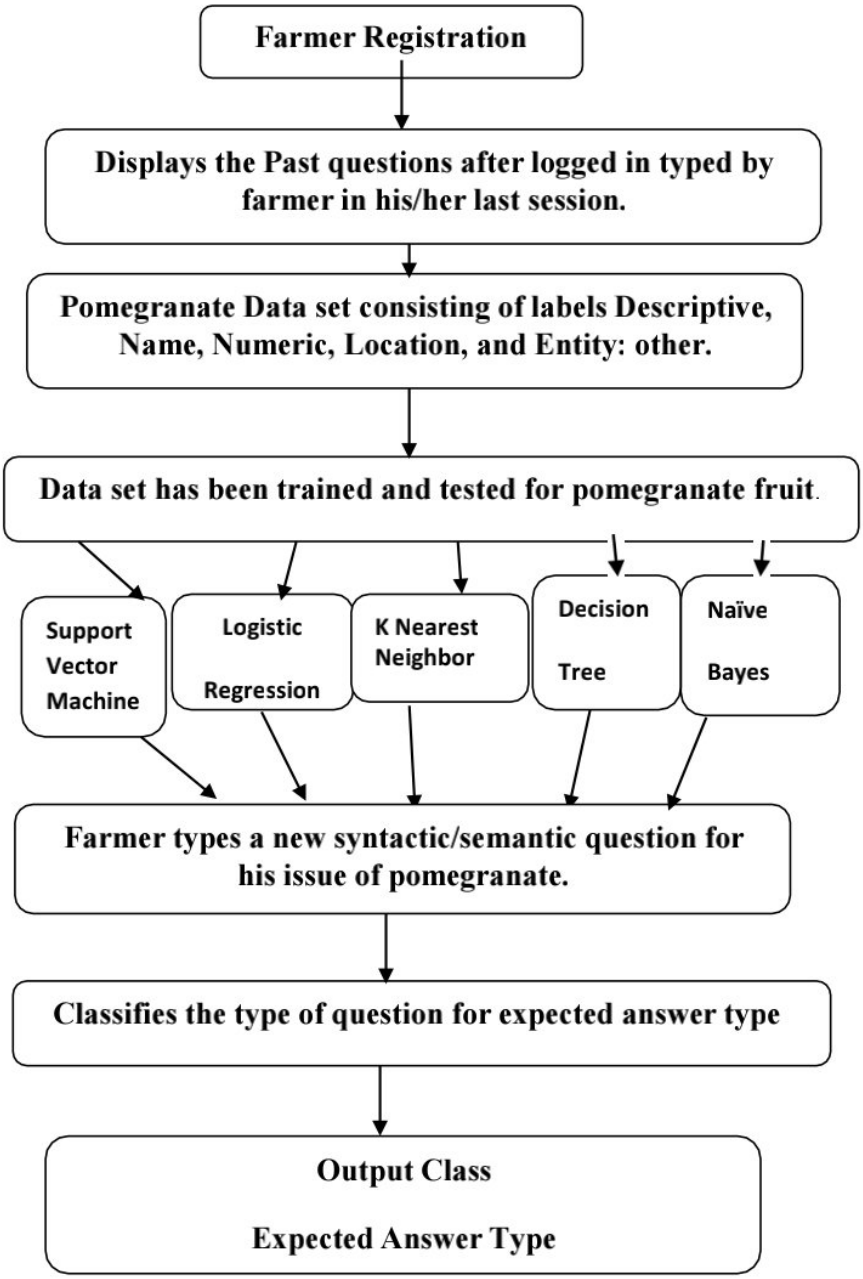

Fig. 2 System Architecture of Question Classification

4.1. Farmer Registration

Farmer growing the pomegranate has number of queries in his mind to ask the question to QA system. First the farmer needs to register to QA system with his details and mobile number can be taken as key for registration. Next time when he logs in he can his username as mobile number and the entered password. Once he login into QA system he can ask any kind of queries relevant to pomegranate.

4.2. Displays the Past questions after logged in

It displays the past questions which are entered by the user. After registration user has entered many questions in QA system, but after every login it displays the past or some previous questions entered by the user.

4.3. Pomegranate Data set consisting of classification labels

Pomegranate data set consisting of multi class labels specified for the number of questions to produce the output. It has labels like Descriptive, Name, Location, Numeric and Entity: other labels.

Class Label	Answer
Descriptive	Provides two or more lines of answer, definition types of answer.
Name	Provides the name of the fruit, disease, seed, pesticide etc.
Numeric	Provides quantity, pounds, rupees, time and date etc.
Location	Provides city, state, country and area.
Entity: Other	Provides other types of answer.

4.4. Data set has been trained and tested of pomegranate fruit

Pomegranate data set have many or number of questions considering the several parameters like its growth, disease, water, pesticides etc. Here some questions of data set have been trained and some questions of data set have been included for testing.

Numbers of classification algorithms have been used like logistic regression, K-Nearest Neighbor, Decision Tree, support vector machine, and Naïve Bayes algorithm.

4.5 Farmer types a new syntactic/semantic question.

Farmer who grows the pomegranate has number of issues, once the data has been trained and tested or model has been built using number of classification algorithms, farmer asking the query relevant to pomegranate for his issue. Farmer has number of queries in his mind while growing the pomegranate specially when his new into farming or into growing of pomegranate fruit.

4.5 Classifies the type of question for expected answer type

Farmer typed question it classifies the type of question to produce the expected answer type, it involves the number of labels/category like Descriptive type, Name, Location, Numeric and Entity:other.

4.6 Expected Answer Type

Finally it displays the result that is expected answer type to the farmer.

5. Experimental Study and Results

This section examines the different machine learning classifiers to categorize on different question types. Pomegranate data set for question classification is important component to have one of the predefined categories for the farmer entered question. Data set for question classification task has been prepared by using NRCP pomegranate file [1]. For our research work 300 pomegranate questions has been prepared along with outcome labels that is Descriptive, Name, Numeric, Location and Entity: Other.

Slno	Question	Classlabel
1	what are the ailment symptoms of bacterial leaf spot?	DESC:desc
2	What are the sickness signs and symptoms of baterial fuit spot?	DESC:desc
3	what is the predominant source of inoculum in contaminated cuttings?	ENTITY:other
4	what is the optimum temperature for the initiaion of fruit spot in pomegranate?	Numeric
5	what is the beneficial climate situation for the inititation of leaf spot in pomegranate?	DESC:desc
6	what is the favourable temperature for the initiation of leaf spot in pomegranate?	Numeric
7	which disease is the most frequent in pomegranate?	Name
8	what are the disease signs of anthracnose?	DESC:desc
9	what is the source of foremost spread of leaf spot in pomegranate?	DESC:desc
10	how can i forestall leaf spot from spreading?	DESC:desc
11	how can i forestall the fruit spot from spreading?	DESC:desc
12	what are the favourable prerequisites for anthracnose?	DESC:desc
13	what are the sickness symptoms of Fusarium wilt?	DESC:desc
14	what are the signs and symptoms of Alternaria fruit spot?	DESC:desc
15	what are the beneficial weather condition for bacterial blight?	DESC:desc
16	what are the reasons for bacterial blight?	DESC:desc
17	what is the beneficial weather for the unfold of anthrecnose?	DESC:desc
18	what are the signs of cercospora fruit spot?	DESC:desc
19	what are the signs of cercospora leaf spot?	DESC:desc
20	What are the signs and symptoms of stem-canker?	DESC:desc

Fig. 4 Pomegranate Questions with Output Labels (Supervised Data set).

To convert your words into numbers. To process machine learning algorithm on a sentence. These words cannot interpret by machine learning algorithm. So we need to convert these words to numbers.

► Word2Vec

► Count Vectorizer

► TF-IDF Vectorizer

Count Vectorizer

Text= [“hello my name is aman and I am a data scientist”]

Text1=[“you are watching unfold data science aman”]

0	1	2	3	4	5	6	7	8
Am	Aman	And	Data	Hello	Is	my	Name	scientist

Fig 5 Representation of word embedding.

After preprocessing the data in background the above diagram shows the representation of word embedding. With reference to text1 below is the output

[0 1 0 1 0 0 0 0 0 0]

If there are common words between these it will be 1. This will not be helpful when documents or domain is different. If text1 is about cricket and text2 is about football. This will not be helpful. So we have TF-IDF vectorizer.

TF-IDF Vectorizer

Text = [“ Aman is data scientist in india” , This is unfold data science”, “Data science is promising carrer”]Above text has 3 documents, After preprocessing the data word to vec is created.

0	1	2	3	4	5	6	7	8	9	10
aman	carrer	Data	In	India	Is	Promising	science	scientist	This	unfold

Fig 6 Representation of word embedding

Word vec will be compared with the given input sentence if its present it will be 1 otherwise it will be 0.Text as input = text[0]

Text [0] = “aman is data scientist in India”[1 , 0 , 1, 1, 1, 1]

Text [2] = “Data science is promising career”[0 ,1 , 1, 0, 0 , 1 , 1 , 1]

Different machine learning algorithms have been used for question classification. These are briefly described below. Multi-class classification is not supported by all classification models; for binary classification designed algorithms such as the Perceptron, Logistic Regression, and Support Vector Machines and for more than two classes these algorithms do not support.

Multi-classification problems can be solved by using binary classification algorithms which splits the multi-class classification dataset into multiple binary classification datasets to fit a binary classification model on each.

Binary Classification: For two classes Classification.

Multi-class Classification: For more than two classes Classification.

For Binary classification problems some of the designed algorithms are which includes: Support Vector Machines, Logistic Regression, Perceptron, but they cannot be used for multi-class classification tasks that too not directly.

Two different approaches used are

- One Vs Rest and
- One Vs One

5.1 Logistic Regression

For multi-class classification, logistic regression uses the technique of One-vs-rest which uses binary classification algorithms that splits the multi-class dataset into multiple binary classification problems. For each binary classification problem, a binary classifier is trained and predictions are made using the model that is the most confident. It uses one-versus-rest approach, in which we train C binary classifiers, fc(x), where the data of class c is treated as positive, and data from all the other classes is treated as negative.

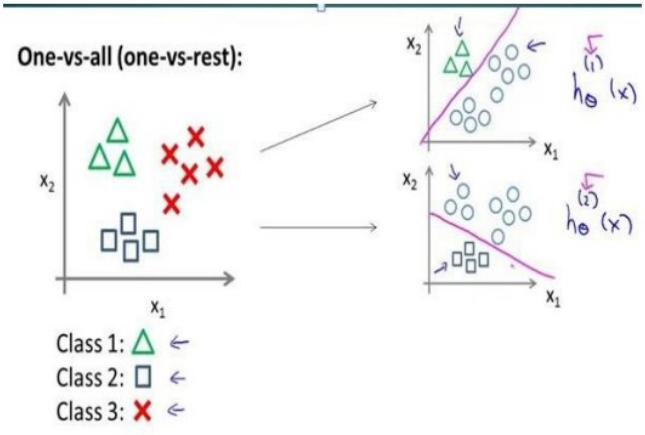

Fig.7 One Vs Rest Classifier

For 3 categories one group will be positive and other 2 group will be negative (M1 model is created), For M2 model other side one group will be positive and other group will be negative and continues.
For M1 model it gives one probability value 0.20, M2 model gives probability value 0.25 and M3 model gives 0.55. Then it will check which has given the highest probability value, M3 model has given highest probability value so it means given new data belongs to O3 category.

Logistic regression algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For 300 questions supervised data set has been applied with logistic regression and it is giving the accuracy of 68.85%. For K-fold cross validation the test accuracy score mean is 57.38% and it shows the below confusion matrix.

Table. 1 Example of One Vs Rest Classifier

f1	f2	f3	O/P	O1	O2	O3
I1	I2	I3	O1	+1	-1	-1
I4	I5	I6	O2	-1	+1	-1
I7	I8	I9	O3	-1	-1	+1
I10	I11	I12	O1	+1	-1	-1
I13	I14	I15	O2	-1	+1	-1
I16	I17	I18	O3	-1	-1	+1

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[32  0  0  2  0]
 [ 5  5  0  0  0]
 [ 0  1  0  0  0]
 [ 4  1  0  1  0]
 [ 5  0  0  1  4]]
                precision    recall  f1-score   support

  DESC:desc        0.696     0.941     0.800        34
ENTITY:other       0.714     0.500     0.588        10
    LOCATION       0.000     0.000     0.000         1
        Name       0.250     0.167     0.200         6
     Numeric       1.000     0.400     0.571        10

    accuracy                           0.689        61
   macro avg       0.532     0.402     0.432        61
weighted avg       0.693     0.689     0.656        61
```

Fig. 8 Logistic Regression classification result.

5.2. Support Vector Machine

Support Vector Machine uses the technique of One-vs-One strategy of binary classification algorithms for multi-class classification. One-vs-One splits a multi-class classification dataset into binary classification problems which splits the dataset into one dataset for each class versus every other class.

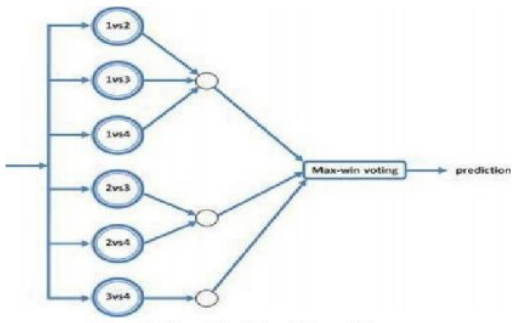

Fig. 9 One Vs One Classifier

It predicts one class label for each binary classification model and the model with the most predictions or votes is predicted by the one-vs-one strategy. This approach is used by support vector machines (SVM).

SVM algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with SVM and it is giving the accuracy of 70.49%. For K-fold cross validation the test accuracy score mean is 59.04% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[33  0  0  1  0]
 [ 4  6  0  0  0]
 [ 0  1  0  0  0]
 [ 4  1  0  1  0]
 [ 6  0  0  1  3]]
              precision    recall  f1-score   support

  DESC:desc       0.702     0.971     0.815        34
ENTITY:other      0.750     0.600     0.667        10
    LOCATION      0.000     0.000     0.000         1
        Name      0.333     0.167     0.222         6
     Numeric      1.000     0.300     0.462        10

    accuracy                          0.705        61
   macro avg      0.557     0.407     0.433        61
weighted avg      0.711     0.705     0.661        61
```

Fig. 10 SVM classification result.

5.3. K-Nearest Neighbor

Supervised Learning of KNN which is one of the ML algorithms that assumes similarity between new data and available or existing data and classify the new data into the category that is similar to the available category. Here new data is classified into appropriate category by using K-NN algorithm.

It's not learning from training set so it's called as lazy learner algorithm, it performs an action on the dataset. Training phase in KNN algorithm stores the dataset and when it has new data it classifies that given data into an appropriate class/label that is much similar to the new data.

KNN method has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with SVM and it is giving the accuracy of 71.42%. For K-fold cross validation the test accuracy score mean is 67.22% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[44  3  0  3  1]
 [ 5  9  0  0  0]
 [ 1  0  0  0  0]
 [ 5  1  0  4  1]
 [ 4  2  0  0  8]]
              precision    recall  f1-score   support

  DESC:desc       0.746     0.863     0.800        51
ENTITY:other      0.600     0.643     0.621        14
    LOCATION      0.000     0.000     0.000         1
        Name      0.571     0.364     0.444        11
     Numeric      0.800     0.571     0.667        14

    accuracy                          0.714        91
   macro avg      0.543     0.488     0.506        91
weighted avg      0.702     0.714     0.700        91
```

Fig. 11 KNN classification result

5.4 Decision Tree

Decision Tree algorithm is an supervised learning algorithms used for solving classification problems. It creates a training model that predicts the class by learning simple decision rules taking from training phase.

It starts with a root of tree for predicting a class label, it compare the values of the root attribute with the record's attribute and after comparison we follow the branch with the value and jumps to the next node.

Decision tree algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with DT and it is giving the accuracy of 70.32%. For K-fold cross validation the test accuracy score mean is 61.66% and it shows the below confusion matrix.

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[44  2  0  5  0]
 [ 2 12  0  0  0]
 [ 0  1  0  0  0]
 [ 6  1  0  2  2]
 [ 5  0  0  3  6]]
                precision    recall  f1-score   support

   DESC:desc        0.772     0.863     0.815        51
ENTITY:other        0.750     0.857     0.800        14
    LOCATION        0.000     0.000     0.000         1
        Name        0.200     0.182     0.190        11
     Numeric        0.750     0.429     0.545        14

    accuracy                           0.703        91
   macro avg        0.494     0.466     0.470        91
weighted avg        0.688     0.703     0.687        91
```

Fig.12. Decision tree classification result.

5.5 Naive Bayes

Naive Bayes classifier used for multiclass learning. Trained Classification Naïve Bayes classifiers that store the training data, prior probabilities, parameter values. It uses these classifiers to perform tasks such as estimating resubstitution predictions and predicting labels or posterior probabilities for the new given data.

Naïve Bayes algorithm has been applied on the supervised data set and for the given new data; it's been classified into the appropriate category. For given 300 questions supervised data set has been applied with DT and it is giving the accuracy of 65.93%. For K-fold cross validation the test accuracy score mean is 58.22% and it shows the below confusion matrix.

perspective. Information Sciences, 181(24),

```
#mythreshold=0.5
from sklearn.metrics import confusion_matrix
#predicted= text_clf.predict(X_test)>= mythreshold).astype(int)
cm = confusion_matrix(y_test, predicted)
print(cm)

print(metrics.classification_report(y_test, predicted, digits=3))
```

```
[[50  0  0  1  0]
 [ 7  7  0  0  0]
 [ 0  1  0  0  0]
 [ 9  1  0  1  0]
 [12  0  0  0  2]]
                precision    recall  f1-score   support

  DESC:desc        0.641     0.980     0.775        51
ENTITY:other       0.778     0.500     0.609        14
   LOCATION        0.000     0.000     0.000         1
       Name        0.500     0.091     0.154        11
    Numeric        1.000     0.143     0.250        14

   accuracy                           0.659        91
  macro avg        0.584     0.343     0.358        91
weighted avg       0.693     0.659     0.585        91
```

Fig.13 Naïve Bayes classification result.

6. Discussion

We have applied the several classification algorithms for the question data set and the above figure classification result shows there is an increase in the results which indicate SVM, KNN & Decision tree have performed well and showing the good accuracy levels compared to logistic regression, decision tree, and naïve bayes.

After comparing the several algorithms, the accuracy plays a major role to judge the quality of an algorithm to meet the standard value.

$$\text{Accuracy} = \frac{\text{Number of classified questions correctly}}{\text{Total Number of questions}}$$

7. Conclusion

In this paper we have proposed and compared the several classification algorithms to classify the questions by considering the domain specific data set i.e. pomegranate fruit. In future classification task can be done on general data set and on multimedia data sets.

References

[1] https://krishi.icar.gov.in/jspui/bitstream/123456789/4319/2/Bulletin%20English-2.pdf

[2] Kolomiyets, O., & Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412–5434.

[3] Bu, F., Zhu, X., Hao, Y., & Zhu, X. (2010). Function-based question classification for general QA. Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics 1119–1128.

[4] Bullington, J., Endres, I., & Rahman, M. (2007). Open ended question classification using support vector machines. MAICS 2007.

[5] Nguyen, T. T., Nguyen, L. M., & Shimazu, A. (2007). Improving the accuracy of question classification with machine learning. Research, innovation and vision for the future, 2007 IEEE international conference on. IEEE234–241.

[6] Li, X., & Roth, D. (2006). Learning question classifiers: The role of semantic information. Natural Language Engineering, 12(03), 229–249.

[7] Alaa mohasseb, mohamed bader-el-den and mihaela cocea "Question categorization and classification using grammar based approach" Information Processing and management https://doi.org/10.1016/j.ipm.2018.05.00.

[8] Muhammad Wasim et al. – "Multi-Label Question Classification for Factoid and List type Questions in Biomedical Question Answering" in 2018 IEEE transactions and content mining volume 7, 2019.

[9] Bastian Haarmann "Mighty Data set for stress testing QA system" in

[10] Hasangi Kahaduwa ; Dilshan Pathirana ; Pathum Liyana Arachchi ; Vishma Dias ; Surangika Ranathunga ; Upali Kohomban

" Question Answering system for the travel domain " 2017 Moratuwa Engineering Research Conference (MERCon).

[11] M.R. Sumalatha ; N.Ahana Priyanka "Analyzing and predicting knowledge of contributors in community question answering services " 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET).

[12] Sharvari Gaikwad; Rohan Asodekar; Sunny Gadia; Vahida Z. Attar "AGRI-QAS question-answering system for agriculture domain" 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) pp. 1474 - 1478, 2015.

[13] Payal Biswas; Aditi Sharan; Nidhi Malik "A framework for restricted domain Question Answering System" International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),pp. 613 – 620, 2014.

[14] B. Haarmann, C. Martens, H. Petzka and G. Napolitano, "A Mighty Dataset for Stress-Testing Question Answering Systems," *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018, pp. 278-281, doi: 10.1109/ICSC.2018.00054.

[15] Gautre, T.A. et al. (2018) 'An analysis of question answering system for education empowered by crowdsourcing', Proceedings of the Second International Conference onInventive Systems and Control (ICISC 2018).

[16] M. Devi and M. Dua, "ADANS: An agriculture domain question answering system using ontologies," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 122-127, doi: 10.1109/CCAA.2017.8229784.