Chapter 1

Introduction

1.1 Introduction

The technique of Information Retrieval is very successful in locating a large collection of documents those relevant to a user’s query. Habitually, however, the user doesn’t need entire document but brief answers to specific questions: How old is the President? Who was the second person on the moon? Recently, a number of research projects have investigated the computational techniques needed for effective performance.

So basically in all above questions we just want some information like how old is the president? We want some numeric answer which will categorise the question as Numeric. The second question is: who was the second person on the moon? We want some answer in the form of name and it will categorise question as Human. During this research work our main attention is on question classification, which is an imperative constituent of question answering systems. The most important job of question classification is to forecast the entity type or the type of the answer. This can be performed through different approaches. The majority of the early studies makes use of handcrafted rules to categorize questions. Though, flourishing approaches are supported by statistical learning methods. During these approaches diverse category of features from the lexical, syntactical and semantic construction of questions are extorted. During the supervised methods training is provided to the classifier with a training set and for checking the accuracy of the classifier testing is performed on an independent test set.

1.2 History of Question Answering

The main objective of question answering (QA) is to provide a laconic answer, certain question posed in natural language [1]. A brief history of the question answering field is described here, it presents a more detailed characterization of what
a question answering actually is, or can be, and also present the theme of this thesis, that is question classification.

Even though, in recent years, a great deal of attention has been given to question answering by the research community, QA is by no means a new field of research. In 1965, a survey article describing no less than fifteen question answering systems for the English language that were built in the preceding five years [2]. Among the reviewed systems, there was BASEBALL, which handled questions about baseball games played in the American League over a period of one year, after few years sponsored by NASA – developed LUNAR, which can reply to the questions concerning lunar rock and soil samples that were collected by the Apollo Moon missions [3] [4].

Both LUNAR and BASEBALL were essentially natural language interfaces to databases where a user’s question is translated into a database query, and the query’s output has come again as the answer. These systems were very restricted, in the intelligence that they merely worked with a very restricted domain (closed-domain question answering), whose facts and data’s knowledge was accumulated in a database. Additionally, these systems were very hard to port to diverse domains, and usually required a substantial human effort to build a knowledge base that comprised all the relevant information about the domain.

Previously the QA system was designed to extract answers from unstructured text, somewhat not working for a structured database. The system exploited a method that is currently referred to as full-text search, in which an access to the file was formed for every expression in the encyclopedia, excluding the general words such as the and a, which were ignored. Another unique feature of the system was its learning component, in which a human assisted to disambiguate some questions or sentences, by which outcome being stored for advance use. Open-domain question answering can be seen as a first step towards a generic question answering system that could work in an unrestricted domain – commonly referred to as an in the QA parlance. In the following years, many more question answering systems were developed, although no
significant improvements were made, since the majority of these systems was still limited to restricted-domains.

In the early 1990s, with the arrival of the WWW and the resultant explosion of electronic media, lots of assemblies started to utilize the Web as a huge text corpus, producing the so called **web-based question answering systems**, such as START [5]. Although during these preliminary efforts, QA was in some manner forgotten for few years, and it wasn’t until 1999. With the launch of the QA track for renowned Text Retrieval Conference (TREC) 2, QA became a very hot research area in the Natural Language Processing (NLP), Information Retrieval (IR) and Information Extraction (IE) communities[6]. In other words for IR, the goal was to promote a document retrieval to very short passage retrieval. Finally, these passages may possibly be condensed to a small answer for a particular question. The IE society had moreover some concern regarding QA due to the fact that the ending of the DARPA sponsored Message Understanding Conferences (MUCs) coincided with the beginning of the QA track at TREC[7]. Moreover, for the NLP community, the QA track revived the interest that began in the 1960s. Although Question Answering is far from being a finished research area, some research groups are now trying an innovative wave of QA: the question answering system which can interact and, where a dialogue interface facilitates follow-up and explanation facility of questions [8].

### 1.3 The Architecture of Question Answering System

Most of the systems are treated question answering as a system, which performs three different sub-tasks: first to perform question processing second one which deals with document processing, and a third one for answer processing [9]. Complete classical system architecture is represented in figure 1.1.

**Question processing:** Question processing consists of an assembly of question representation that is the classification of question, derivation of expected answer type, and keyword extraction. But in some systems, there is also a provision of questionable reformulation, where a question is transformed into a number of declarative equivalents. For structural representation of a question, parsing is performed; the structure is a syntactic tree or a dependency tree. This structure can afterward be used
to place and substantiate answers within retrieved documents and passages [10]. The keywords which are extracted are used by retrieval engines to fetch relevant documents. The expected answer type is also derived for this purpose.

**Document processing:** Document processing usually includes keyword expansion, document retrieval, and passage identification. Keyword expansion typically involves taking the keywords which are extracted in the question processing stage and looking them up in a thesaurus, or in other resources, and adding similar search terms in order to fetch as many relevant documents as possible. A term such as ‘kill’ might be expanded to ‘murder’ and ‘assassinate’ for instance. Document retrieval is restricted to passing the expanded keywords to a standard search engine (e.g. Google) and retrieving the documents with higher ranks. During passage retrieval, within each document the paragraph or section containing the possible answer is identified.

**Answer processing:** Answer processing consists of candidate answer recognition, answer grading, and answer formulation. Identifying answers mean taking the results from the passage identification phase and further processing it, this means doing a full parse of the passage and comparing it to a full parse of the question. The result of this is a set of candidate answers that are then ranked according to an algorithm or set of heuristics. In many cases the formulation of an answer is omitted completely and the answer is presented as it was found in the document.

---

**Figure 1.1** Architecture of Question Answering System
1.4 Research Issues

The focus of this thesis is mainly on question classification. More specifically, classification of questions with respect to expected answer types. The research is motivated by the fact that previous research has shown that correctly predicting the expected answer type is imperative for a question answering system as a whole to be successful [11]. It is very important and well understood that if the categorization of question will be wrong then it will lead to find out the wrong answer so it is very significant for question answering system that it will classify the exact problem domain of a question. So it is implicit that the proper classification of a question will also provide us a reduction of steps in two ways. For example, suppose if a question is classified as a category of CITY then the system will find out only cities no need to find out anything else and if a question belongs to category DEFINITION then the QA system will find out the answer for explanation part using or connecting Wikipedia. The data for this research work is taken from the Cognitive Computing Group at University of Illinois at Urbana-Champaign UIUC. Which consists of a training set of 5500 questions, and a test set with 500 questions (http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/).

Research issue 1 Classification of categorization of questions and to achieve that what different types of techniques can be used?

In order to establish this different machine learning algorithms has been used and comparison has been performed on them, the various algorithms are: Naïve Bayes, Back-Propagation-Feed-Forward Artificial Neural Network and SVM. The first work of this research is to train the machine with TREC 10 dataset for all three above mentioned algorithm then to check the performance by using a testing dataset for the same corpus after then compare the result for the same dataset. A second aim of the thesis is to make an analysis of the performance of machine learning algorithms with the combination of the different question feature set. More specifically, looking for the cases where the different combinations will increase the accuracy of accurate classification.
Research issue 2 Are there rich sets of question features and what are the contributions of each feature in question classification?

The main challenge is to develop a supervised classifier with the rich set of features. So the thesis will be going to present the different features along with their contributions to the classification accuracy of a question.

Research issue 3 Can different combinations of features increase the performance of question categorization?

The influence of different features and their combinational influence is also one of the motivational issue of this research work, along with how they are extracted from questions.

Research issue 4 How the optimization of the classifier will be performed for question classification?

The main influence of any classification algorithm is based on their parameters so one of the main important research issue is to optimization of the parameter of the classifier so that the optimized accuracy can be found. In order to do this the thesis introduces the use of the particle swarm optimization technique.

1.5 The Role of Question Classification

For any question answering system, Question classification part plays an important role because on the basis of it the answer depends for any question. Even if diverse categories of Question Answering System have a different structural design, the majority of them pursue a structure in which question classification plays a significant responsibility [12]. Additionally, it has been revealed that the presentation of question classification has important influence on the overall presentation of a QA system [13], [14], [10].

Essentially, there are two most important inspirations for question classification that are as follows:
1. For establishment of the answer

2. For deciding the exploration policy

- **For establishment of the answer:** Meaningful question class cannot simply diminish the exploration gap which needs to find the answer; it can also discover the accurate answer in a given set of applicant answers. Suppose, for example, if the question is “where was the birth place of Lord Ram” the system know that the question is related to “location” type category. Means the answer should be a place so the answering system does not require to examination of all phrase within a passage for answering the question.

- **For deciding the exploration policy:** In addition question class can be also utilized for deciding the exploration policy. It is required when the question is transformed into a query over the search engine. For example, if any question is related to definition means in any question description or definition has asked so in this type of situation the system can search for possible answers in encyclopedic sources e.g., Wikipedia, which are more likely to contain an accurate answer. And it will just narrow down the search policy.

The growth of a question answering system which will take the question class and generate the appropriate SQL query according to question class has been mentioned [15].

### 1.6 Literature Survey

For Question Answering System, question classification plays a vital role. In this chapter different methodology has been covered that employed for implementation and study of question classification systems. The difficulties encountering during Question classification problem have already considered in a lot of previous work. The present chapter discusses the evaluation of various question classification works along with their outcomes.

#### 1.6.1 Comparison of Learning Approaches
The categorization of question has immense consequence for searching the kind of class of question in the question classification and it also decreases the attempt of exploration. The majority of the occasion it happens that public wasted excessively a large amount of time for exploring on the web and as well as do blunder in the detection of accurate response.

History shows that a variety of classification techniques and algorithm has been proposed and are also used for an extremely huge region of classification and web mining. Naïve Bayes (NB), $k$ Nearest Neighbors ($k$-NN), and Support Vector Machine (SVM) are the well known classifiers and generally used for the pattern recognition or for the classification purpose, the above mentioned algorithm also proved their presence in probabilistic learning method. Literature Survey says that out of NB, k-NN and SVM, SVM classifier is most flourishing classed for learning to classify text documents. The hybrid recommender system employs neural network classifier and joint filtering, the basic concern of Hybrid system are machine learning and data mining method for sorting undetected information and can calculate about the need of source by the user means whether the user would like a given resource. Nowadays there is also a huge amount of use of neural network in the online application and this is again an easy illustration of machine learning and successful calculation[16]. The various types of survey regarding question classification have been discussed below:

The outcome of the mechanistic classification of the question in the support of a machine learning algorithm and description of different types of machine learning algorithm, their evaluation, comparison of all algorithms has been discussed [17]. The literature states that the evaluation of the SVM has an optimistic result in content [18]. The Naïve Bayes classifier is an extremely admired classifier the reason behind that is their straightforwardness and computational effectiveness and various times employed for real world problem. The problem associated with Naïve Bayes is that for most of the time it is not capable of solving the complicated classification problem the comparison have been performed between the Naïve Bayes and SVM and found that the performance of SVM is much better than Naïve Bayes in many cases. Observation for the diverse question set has been performed on the basis of various measurements and the different result has been calculated from a different aspect. Different
applications regarding SVM classifier and Naïve Bayes classifier have been also discussed [18].

The hypothesis for primarily re-examining the earlier effort on different machine learning method and their result with the comparison has also discussed. The evaluation performed for the question classification using Naïve Bayes, k-nearest neighbors, decision tree learning, sparse network of windows and support vector machine and furthermore communicated the earlier effort concerns [19].

The use of words related features like stop-word and about the stemming procedure and about the tools desirable for building-up these features in the French language database and some retrieval research is also mentioned [20]. The work has been performed for two medium-sized French language test collections. The concluding research for carrying out the estimation of the efficiency of the planned work explained in detail and also provided a concise proposal regarding the common Stopword List, Stemming Procedure [21].

The establishment of an innovative functioning for using sub tree-mining in favor of the question Classification trouble has also discussed. For this purpose the utilization of a forest tree with the setting of some definite levels has also mentioned. Basically the sub tree is used for the forest created tree classification [21]. For providing the training to the classifier the utilization of boosting model and maximum entropy has mentioned. The exercise of boosting model and maximum entropy specify that this approach accomplished improved performance than kernel methods and in addition develop testing effectiveness [22].

Named Entity is a very important feature of the sentence which is extremely employed in natural language processing and basically was created for Message Understanding Conference (MUC). In this research the main concern was the configuration of facts and information from the unstructured information. The main concern and conversation was regarding Language factor, domain factor, Entity type factor and also diverse Learning methods like supervised learning, Semi-supervised learning, etc., with different feature spaces. Conclusion offered as a summary of the procedure working to build up NERC Systems, also offered a summary of the
estimation technique that were in utilization of the key structure of the NERC research community [23].

The utilization of language modeling skill is specially designed for solving the problem of question classification by the way of entity-tagged language model. This procedure is completely mechanical and not involved automatic convention. The outcome of language modeling skill is evaluated against the result of the SVM classification. The important feature of the sentence that is named entity is also conversed and the name in this present literature is given to the named entity is an identifier. The conversation about the question classification and categorization approaches has also explained. Relative study for discovering which algorithm is superior and discussion about the investigation procedure, baseline and result have discussed [24].

One hybrid Turkish passage summarization classification which joins the construction of semantic features is presented. In this hybrid classification primarily the utilization of 5 constitutional features has performed out of which two features were recently projected features and 3 were semantic feature whose standards was taken from the Turkish Wikipedia link. Merging of The features is performed by using the weight calculation and it employed by 2 new approaches. Subsequent to feature mining the artificial BEE algorithm is employed for the automatic formation of the weight in the future. The development of each feature independently presented the better outcome and also presented the exploring performance of the AHP and the ABC algorithm [25].

The suggestion for head word feature has proposed and also presented for exploring the headword the most prominent semantics features through the word net. The proposal has also given for the use of a linear SVM classifier for enhancing the precision of the classifier [26].

The introduction and the utilization of the extractable classes which does not require human intervention and which is the part of the syntactic features of the text classification is given. Syntactic features are the foundation and its forms the sub tree of parse tree. The recommendation has provided which advised that producing a huge
special set of these features mentioned above may provide the facility for enhancing the accuracy. Including the various types of features related to syntactic category the dimension is also given for bags-of-words (unigram) demonstration uniquely [27].

The assessment and proposal of a text classification technique which supported on skilled replica has presented. The replica employed for text classification is mainly dealing with the theory of divide and conquer. The replica also makes the use of abstracting less significant categorization problems based on predefined hierarchical structure. For text classification the learning method required and which is based on machine learning. The demonstration of outcome of hierarchical structure indicates that it is much better than the performance of the plain structure for the classification of text. The initiative for neural network and about the Rocchio algorithm and their presentation is evaluated by the consecutive level of neural networks [28].

The essential conversation concerning the neural network and their diverse property and meaning has mentioned. Explanation is given about the multilayered networks and their ability for calculating a huge variety of Boolean functions. Various types of conversation regarding neural network like activation functions, Local minima of the error function and outline of common feed-forward networks, imitative of network functions, and other Artificial Neural Network functions have presented. Discussion about the back propagation algorithm their approaches and their best suitability for a graphical problem which diminishes the labeling problem in graph have also presented [29].

1.6.2 Machine Learning Approaches for Question Classification

Basically what is the main fundamental or aim of machine learning is that it deals with the problem to make such type of computer system which can learn itself with the time or experience [30]. So in this context a computer program learns from the experience of the past history, which is performed by the system for the same task and in the learning processs it is going to improve its performance.

Mainly the machine learning started in the 1950s, Arthur Samuel ŕ who was the founder of machine learning, has created a program for playing checkers the
main quality of this program was learning it learns how to play checkers, by identifying different patterns of game by which player can win or lose the game [31].

The major working of this program was to learn and examine the result of each shifting frequently by playing the game against a replica of itself and by the way identified the pattern [31]. This checker playing game which was created by Samuel was the landmark for machine learning, and it was the possibly the first program written in a computer that in fact learned from past data and occurrence [31]. The different types of machine learning are:

i) Supervised Learning

In this type of machine learning, learning is basically performed through a tagged training data with the help of training examples the construction of training information is performed. In supervised learning, every illustration consists of two kind representations one is the input objective and second is a preferred output assessment. The main task of supervised learning algorithms is to examine the training input set and create a conditional function, which can be further employed for planning the new examples. A most favorable situation will permit and used for the algorithm to appropriately conclude the category tags for unobserved illustrations. There is a need to the learning algorithm to simplify and make a general view from the training input set of unobserved conditions for making a "logical" technique [32].

ii) Unsupervised Learning

The main trouble or the difficulty associated with unsupervised machine learning category is the demanding to discover an unseen organization in the data that have not tagged. Because the data which is provided to the learner is not tagged it is hard to calculate possible and this is the main difference associated with the unsupervised learning and make it differentiate to others that are supervised learning and reinforcement learning. This type of learning is mainly associated with the problem of the calculation of density is statistics. But many times unsupervised learning taken into the account of numerous additional methods that look forward the review and give the details of key features of the data. A lot of techniques which is
working for unsupervised learning uses the foundation of data mining methods and used to preprocess the input data stream. Approaches to unsupervised learning includes: Clustering, hierarchical clustering, and Hidden Markov models. Along with neural network models, usually the utilization of unsupervised learning algorithms are: self-organizing map (SOM) and adaptive resonance theory (ART). Topographic association which is SOM in which close position in the map corresponds to contribution of related belongings. The ART model permits the amount of clusters to fluctuate by the way of difficulty dimension and give permissions to the client to manage the amount of connections among constituents of the identical clusters by means of a client-stated invariable called the vigilance restriction. The utilization of ART networks is also in various fields likes for numerous sample identification responsibilities, such as automatic target recognition and signal processing. In 1988 the earliest edition of ART which was "ART1", developed by Carpenter and Grossberg [32].

iii) Reinforcement Learning

It is the field of machine learning which is enthused by the performance psychology, related to how software, representative have to obtain accomplishment in a situation so as to make best use of some concept of increasing returns. The difficulty level regarding its simplification is considered in other numerous regulations, such as for example the theory of games, in genetic algorithms, optimized simulation and in operation research. Especially the use of reinforcement learning in the field of operations research is known as dynamic programming. In game theory the use of reinforcement learning may be viewed as the use to give details about how symmetry may take place in boundary rationality. The use of Markov decision process that is MDP in the field of machine learning surroundings is prominent and many reinforcement learning is highly incorporated with the influence of dynamic programming methods or procedures. The major and the most important dissimilarity among the traditional systems and reinforcement learning algorithms are that the reinforcement learning do not require information regarding the MDP and they aim huge MDPs where correct process turn out to be infeasible [33].
The difference between the Reinforcement learning and the typical supervised learning is that the exact contribution/productivity pairs not at all offered, nor sub-optimal events unambiguously approved. Additionally, there is main concern on the presentation of on line performance, which concern discovery of equilibrium among investigation (of unexplored region) and utilization (of existing information). The utilization vs. investigation transaction in reinforcement learning has been mainly concerned systematically and considered throughout the multi-armed-bandit difficulty and in restricted MDPs.

iv) Semi-Supervised Learning

It is a special category of supervised learning which utilizes the untagged or unmarked data for training purpose with the combination of small amount of data which are tagged or marked. Semi-supervised learning is a learning mechanism which is in between of unsupervised learning which works with excluding the tagged or marked training data and supervised learning which uses entirely tagged or marked training data. Numerous researchers associated with the field of machine-learning has originated that untagged or unmarked data, when employed in combination of a little amount of tagged or marked data, can construct significant development in learning precision. The achievement of marked or tagged data for a learning difficulty frequently involves an expert human being as an agent for example, calculating about the organization of protein structure which is 3D or investigating about the oil location which one is exact. The charges connected with the classification development, thus may provide a fully tagged training set is infeasible, while acquirement of untagged data is comparatively low-priced. In such circumstances, semi-supervised learning can provide an immense useful assessment. This kind of learning is also in interest of researchers as well as learners [34].

1.6.3 Machine Learning for Question Classification

The first approach for the question classification is to explore the classifier for the optimal classification and there are so many classifiers and if the searching is performed for it, then the most of the successful work regarding the question classification is based on the SVM (support vector machine) approach [11], [35], [36].
Especially when the data stream is large it is very successful because they are judicious well-organized particularly when the feature vectors are sparse. Another method has been done by using Maximum Entropy models for Question classification like the Sparse Network of Winnows (snow), and language modeling [11], [37-39].

Examples of different types of classifiers are:

**a) Naive Bayesian Classifier**

The most common method in supervised learning technique is a naive Bayesian classifier that has been widely used in many functions concerning classification tasks, like spam filtering in e-mails [40]. It reaches to high-tech outcome, regardless of its simplified *naive* hypothesis. If so many inputs are given to naïve classifier and there is having some predefined classes like $\text{CL} = \{\text{CL}_1, \ldots, \text{CL}_m\}$ and suppose the inputs are represented by $\text{I} = \{I_1, \ldots, I_n\}$ then the work of Bayesian classifier is to find out the most appropriate class for that particular input instance. For performing this activity probabilistic model is used that selects the class $\text{CL}$, from the different set of probable classes, that make the most of the posterior probability $P(\text{CL}/\text{I})$ that mean the probability of $\text{CL}$ find out the feature vector $\text{I}$. Inside information, this is frequently called a *maximum a posteriori* (MAP) assessment, and is specified by [33][41].

$$\text{cl} = \text{MAP}(\text{I}) = \text{arg } \max_{\text{CL}} P(\text{cl}/\text{I}), \text{ cl} \in \text{CL}$$  \hspace{1cm} (1.1)

where $\text{CL} =$ Set of Predefined Classes

$\text{cl} =$ Any one class which belongs to $\text{CL}$ with maximum posterior probability.

The detail explanation of eq. (1.1) is given in chapter 4.

**b) Support Vector Machine**

Support Vector Machine (SVM), is very commonly used method for classification task like text classification [42], [43]. It is a supervised learning technique developed by Vladimir Vapnik and his colleagues in the early 1990s [44], [45]. SVM works with hyper plane separation which is also known as linear decision
boundary finding specially for two classes and it is performed with the training data for every class additionally, this design can be extended to nonlinear decision boundaries using kernel functions.

The critical idea is to come across a hyper plane, which separate the d-dimensional data entirely into its two classes if SVM will be using for Binary classification. Conversely, many example data are not linearly separable. SVM introduces the notion of a kernel induce feature which direct the data into an advanced dimensional breach where the data is distinguishable. Typically, directing in such a space would start problems. SVM in the higher-dimensional space doesn't necessitate to be dealt with directly, which can eliminate the above disquiet. Overall, SVM's are discriminating tentatively well-founded, and have made to be basically unbeatable. SVM's have also been extended to solve degeneration tasks.

When scheming for SVM, the main problem is the selection of kernel to use, and for a given kernel, how to position the restriction. The novice strength is tempted to strive all types of kernel/parameters, and choose the one which perform best among others. So the performance or the result of any work based on SVM is also depends upon the type of kernel is selected and for which conditions. So the user who is using the SVM classifier should collect the data and knowledge about the problem for which SVM classifier is in use. For example, if one researcher were training an SVM classifier for face acknowledgment, then they should have an idea about the previous work related to the SVM and face detection. As this research is dedicated for classification accuracy of questions, linear kernel is best among all others for classification task [46]. The basic working algorithm of SVM for question classification is explained in detail in chapter 4.

c) Sparse Network of Winnows

This type of multi class design is specially used for the problem having a huge amount of data feature which may not be known a priori [47].

In SNoW's most basic configuration, a two-layer network is created, with an input layer comprising a set of nodes, each representing a distinct feature in the
training set; and an output layer consisting of a set of target nodes, one for each possible class label. This setting resembles a single-layer neural network. The objective is then to find a set of weighted edges connecting the target nodes to the input features, with the restriction that an edge $w_{t;i}$ is allocated and linked, if and only if the $i$-th feature is seen associated with target $t$ in the training set (hence the name **sparse network**). Conceptually, $w_{t;i}$ represent the importance of the $m_e$-th feature for the classification of the class associated with $t$ [21].

d) *K*-Nearest Neighbors

The *K*-Nearest Neighbor algorithm (or *k*-NN for short) is a technique basically used for pattern recognition, classification. For using this technique *k* training example, has to take into account for input data set. Also for regression technique *k*-NN is used, how the *k*-NN is working is totally depends on the purpose for which it is used like the classification or regression.

If *k*-NN is used for classification purpose, then the output will be the class association. Any entity will be classified by the membership of neighbors where will be the maximum number of neighbors, the entity belongs to that particular class mainly frequent among its *k* nearest neighbors where *k* is a positive integer. If the value of *k* = 1, in this case the entity is simply assigned to the class of that solitary nearest neighbor.

On the other hand, if *k*-NN is used for regression the output value depends on its nearest neighbor and it is the average of its *k* nearest neighbors. In *k*-NN it is hard to find the proper computation until the complete classification because it totally depends on the instance of all the occurrences despite it is very simple among all the machine learning algorithms. In both the cases (classification and regression) the weight is assigned to its neighbors and on the basis of its distance can be easily calculated because the nearer neighbor contributes more than the distant one [21].

So the decision simply depends for which condition *k*-NN is going to be used for regression purpose or for classification. If classification is performed by using *k*-NN then neighbors are selected for which the class is known and in the case of
regression the neighbors are selected such that whose object value is known by this way the training is performed, so no extra training is required but the major drawback of k-NN algorithm is that it is susceptible to the restricted structure of the data.

e) Back-Propagation Neural Networks

Feed forward terminology used in neural network basically explains that how the actual working of neural networking is done and how the progress is performed, how it remembers about the samples. Within a feed forward neural network, the connection of neurons is established in such manner that the neurons are only associated foreword. Layers are connected in such a way that if there is an input layer then it is connected with hidden layer, but here each layer of the neural network contains connections to the next layer (for example, from the input to the hidden layer), but there are no reverse linking. Despite this, another type of neural network where two way association or linking (onward direction and back to forward direction) is possible is the Hope field neural network.

The terminology back propagation is used for, providing training to the neural network, it exactly tells about the guidance of the network. Back propagation which is one kind of technique of supervised learning, in which training is provided to the network with both example inputs and expected outputs. The real output is evaluated with expected output for the example input. The anticipated outputs are compared against the actual outputs for the given input. The back propagation training algorithm finds the difference of errors between the expected outcome and the real outcome and on the basis of it changes the weights of the different layers just in the opposite direction from the output layer to the input layer [26].

Generally the both algorithms: back propagation and feed forward are working together, but it is also possible that design a neural network that apply the feed forward algorithm to resolve its output and does not utilize the back propagation training algorithm. And just its opposite it is not necessary to use feed forward for the output if back propagation is used for training. But both above described cases are limited to use any one of them (use any one of them either feed forward algorithm or back propagation training algorithm).
For carrying out the proposed research work Back-Propagation Feed-Forward Artificial Neural Network (BPFFANN) has been used and the detailed explanation of its working and role is explained in chapter 4.

1.7 Contributions of Proposed Work

There are many challenges and issues on question classification problem. The main contribution presented in this thesis is to improvement of the performance of learning-based question classifier systems with different combinations of question features to contribute to the next generation of question answering systems. Following are the main challenges (research questions) in question classification problem that are addressed in this work:

1. What category of features can be extracted from a question written in natural language?
2. What is the role of each feature in question classification?
3. How these features are extracted?
4. How efficient features are combined, which will increase the accuracy of question classification?
5. Which learning algorithm is appropriate for question classification?
6. What are the parameters which are needed for the optimization of classifier and how the parameter optimization has performed for improving the classification accuracy?

For giving the answer of all above mentioned questions, implementation of question classifier system has been performed. And the research work successfully introduced the following contributions to question classification problem:

1. Development of a proficient set of features for question classification.
2. A better combination of features for question classification.
3. Improved perceptive of mis-classification causes in question classification and:
4. An additional perfect question classifier than previous works along with the optimization.
For finding the above mentioned contribution, classifiers that has been used for this research work are: Naïve Bayes, Support Vector Machine and Back Propagation Feed Forward Artificial Neural Network. The reason behind the selection of these classifiers are so simple because they are the most common classifiers used for supervised machine learning. And for carrying out this research work training datasets has been used so for that reason supervised machine learning has been choosen for this research.

1.8 Thesis Outline

This thesis organized as follows:

1. Chapters 2 presents question classification approaches.
2. Chapter 3 describes the classification, representation of a question.
3. In chapter 4, the generative model and the enhancement in generative model for question classification is represented along with the evaluation of the overall performance of the system using a test set of questions collected from the TREC 10 dataset.
4. Chapter 5, presents the optimization of the improved generative model.
5. Finally, chapter 6 concludes this thesis by our main contributions to the question classification by discussing the direction of future work.

+++