# Chapter 1

## Introduction

The project is a specific area of research in the domain of machine learning which deals with evaluation and comparison of various algorithms such as decision trees, naïve Bayesian network, K nearest neighbours, logistic regression and random forest ensemble for binary classification of data [1]. The type of data dealt with is extracted from web pages and is hence unstructured in nature. The project involves research into the process of automatically identifying valid online course web pages after web crawling, pre-processing and filtering of this data to extract features from the page source and its contents [2]. Previous efforts in this direction deal with smaller datasets of varying sources and use dissimilar metrics for comparison of the success of the various classification algorithms mentioned above [3]. This project seeks to successfully evaluate all of them across a single large dataset and a uniform metric, namely the precision of the algorithm which can be measured in terms of the number of correctly classified instances in the evaluation test set, as well as the area under the receiver operating characteristic curve for various threshold values of the positive example rate of classification of data [4].

Classification ultimately labels whether the page represents a valid online course and aims to create an offline repository of all such course pages. This enables a user to find and select the online training courses according to interest. An offline interface exists between the user and the system where the user enters his field of interest. The results obtained are a list of all the online courses available pertaining to that field. Thus, the aim is to help users across the world to select the courses of their interest.

The software has the following major architectural units which guide the working of the whole system: A fast multi-threaded java crawler to crawl web pages, a text parsing and regular expression based parser for extraction of features, efficient machine learning framework for classification of web pages using various machine learning algorithms, visualization tools for comparison of the machine learning algorithm results and finally a front end which allows the user to view any online course related information.

Figure 1.1 depicts the general architecture. The online course web pages are crawled from which data is collected for pre-processing. The pre-processing involves

mainly crawling, featurizing and classification, each of which will be dealt in detail in chapter 4. After the pre-processing a training set is created to train the classifier model. The trained model is then given test dataset as input which was extracted from new web pages [5]. The model then classifies the web pages into relevant online course web pages and stores the data in the repository.

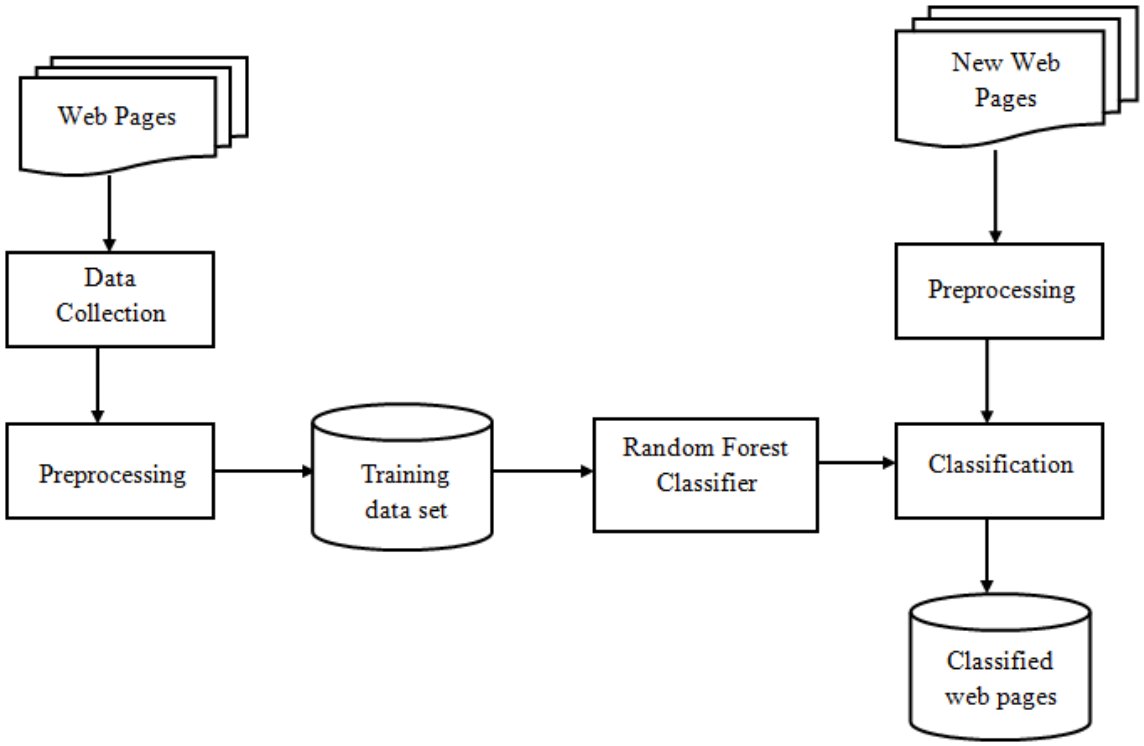


**Figure 1.1: General architecture of implementation of machine learning algorithms for efficient classification of unstructured data from online course web pages**

## 1.1 State of art developments

The basis for all research in this project is found in [6], where authors describe one of the most novel and unique techniques used in webpage classification, as it is based on natural evolutionary concepts. The motivation of this paper is to perform feature selection based on the selectivity seen in fireflies. It describes a meta heuristic algorithm developed to classify web pages that is based on the evolutionary and biological behaviour of fireflies. This seeks to find the best 'n' features out of the hundreds that can be used for web classification. They observed improvements in speed and time by using this algorithm to reduce features. However, the process was long and cumbersome, involving a large amount of computational resources and therefore future work in this direction requires the assessment of the use of this algorithm and the various

optimizations which can be applied on it. It is merely a way of feature selection to reduce the dimensionality of algorithms used.

In order to carry out any such work, it is necessary to understand the applications of data mining tools like Weka [7] by applying K means clustering to find clusters from huge data sets and find the attributes that govern optimization of search engines. The motivation of this paper is to compare and evaluate various tools and visualizations used in classification in order to select the best one for use in this project. It is important for search engine to maintain a high quality websites. This will improve the optimization. Then, a database in which following attributes -length of title, keywords in title, domain length and number of back links and Top rank website is created. After applying K means using Weka, the result window shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. The future work in this direction involves various manipulations of existing visualization algorithms in order to better understand results, and this was one of the main research gaps in this paper that the authors have used only one algorithm, therefore there was need for comparative study.

Various approaches have been tried over the years, like the efforts of [8] which introduce a shift from keyword-based representation to other perspective on representation of document's focus in form of key-concepts. This paper motivated a study of algorithms in the direction of words, concepts and intent based analysis, which requires a large amount of domain knowledge. The method is based on disambiguating the word senses using PageRank algorithm. The principle of this proposed approach is to do a two-pass ranking. So it is more accurate approach than keyword approach and its space efficient. However, there is scope for further research into graph processing algorithms which optimize storage and representation concepts without causing wastage of computational resources. There is also the research area of limited nodes in a graph, which needs to be explored to suit expanding dictionaries and vocabularies.

It is also necessary to have a quality and quantity metric such as in [9], where the author explains the way to create the website quality prediction models and nine quantitative web measures from different categories of Pixel Awards website are computed by the Web Metrics Analyzer tool. This helps to understand the various metrics used in predicting validity of a page the methodology employs the quantitative web page attributes (number of links, words etc.) to compare the goodness of the web pages and to

construct a quality prediction model utilizing Subtractive and FCM clustering model for predicting the class of website as good or bad. The methodology has 5 sections - Empirical Data Collection, web metrics analyzer, data processing, model building and data result analysis. One of the major results of this study is that ANFIS clustering techniques can be used to predict the quality of the website. This provides substantial improvement and the basic outline and structure followed in the methodology has been emulated in this project with the future improvement of dynamically completing the same, which this paper listed as a limitation.

A specific example of explicitly identifying webpage content can be seen in [10], where the authors describe the method to detect malicious web pages and to identify the specific threat types. This is a common application of web page classification and provides scope for common techniques and approaches used to be studied. The study suggests that KNN is better than SVM with 95.4% accuracy than SVM which has 92.3% accuracy. However, the method of comparison here uses a limited data set of only 4 types of web pages which is small in size and statically put together. This approach is extended to include dynamic crawling and classification, and compare multiple algorithms, one of each type in order to provide a better picture of the domain overview of webpage classification and machine learning algorithms.

A distinction needs to be made between the identification explained above and the general working of a search engine. This is seen in [11], the author focuses on estimating semantic similarity measures as an alternative to normal search engines, and the motivation of this paper is to study search hits of various patterns. The method is proposed via exploiting the page counts of two biomedical concepts returned by Google web search engine. The similarity scores of different patterns are evaluated, by adapting support vector machines for classification, to leverage the robustness of semantic similarity measures. The gaps in this paper include use of a single algorithm and hence the paper was studied with a view of understanding semantic search which can be used as an improvement in further iterations of the current system and provide multilingual support.

One of the applications of machine learning in webpage navigation is seen in [12] which deals with the application of Neural Networks for the Classification of Chinese web pages to develop a Web Information Navigation System. The motivation in studying

this paper is to understand how neural networks can be used in multilingual classification, which can be further explored as a research area. The quantum neural network has three layers: input layer, hidden layer, and output layer. The keywords of Web document are chosen as inputs of quantum neural network classifier. The quantum neural network classifying result and the manual work classifying result are compared, and there are 867 web documents rightly classified into correct subject class by quantum neural network classifier. The test result shows that the average system performance of the quantum neural network classifier is about 86.7%. The scope for future work includes use of a larger dataset with different types of samples which can be improved in precision over the current output.

Another application specific to forums is seen in [13] where the author presents Forum Crawler Under Supervision (FoCUS), a supervised web-scale forum crawler. The motivation of FoCUS is to crawl relevant content, i.e., user posts, from forums with minimal overhead. Forums exist in many different layouts or styles and are powered by a variety of forum software packages, but they always have implicit navigation paths to lead users from entry pages to thread pages and these can be exploited for efficient crawling. Forum threads contain information content that is the target of forum crawlers. This paper also shows how to learn accurate and effective regular expression patterns of implicit navigation paths from automatically created training sets using aggregated results from weak page type classifiers. Robust page type classifiers can be trained from as few as five annotated forums and applied to a large set of unseen forums. This can be applied in multiple domains with numerous applications and can also be extended to improve the precision of classifiers by including more data in the automatically created training sets.

Advances in dynamic HTML and webpage scripting requires advanced parsing such as in [14] which describes a system which uses vision and DOM based methods to apply page segmentation in obtaining academic data. This methodology consists of an algorithm which combines vision-based and DOM based segmentation methods, Bayesian network classification and post-processing. It can improve initial classification results by repairing wrong results and adding unclear results, this improves the result greatly: the precision is improved from 0.90 to 0.96, the recall is improved from 0.89 to 0.98, and the F1 is improved from 0.90 to 0.97. There is scope for further improvement since this system relies heavily on post processing which involves manual effort. Also,

the system is offline in nature because it relies on a static dataset of conference pages, which is a major research gap.

Along with HTML content parsing, it is required to extract intent related data as well. The work in [15] presents an algorithm is presented to extract keywords, which can be used to classify pages based on intent. The algorithm generates and extracts keywords from a poetry book. A matrix is formed in which words are number of columns and distiches are number of rows. The elements of matrix are filled by zero and one. The results indicate that “Love”, “Heart”, and “Eyes” are the most important keywords of the selected book with frequency 5%, 14%, and 9% respectively. This is able to tell us about the nature of content on the page and will be useful in large scale classification which involves text mining instead of conventional algorithms. Sparse representation of matrix and efficient processing of high dimensionality matrices need to be explored.

Other methods involving structured data include schemas as shown in [16] where the author explores the use of formal source code structure for classifying a large collection of the web content. The motivation was to focus on use of schemas collection Schema.org to classify web pages and categorize them unambiguously. Future work included implementation of this concept in larger domains and using multiple keywords.

In the absence of structured data, it is required to have structural units. An effort in [17] incorporated named entities as feature for web page classification. Named Entities (NEs) are phrases that contain names of persons, organizations, locations, numeric expressions including time, date, money, percent expressions and names of other miscellaneous things. The motivation to research this topic is that it gives better results for narrow domains. Some relevant classification algorithms include HMMs, Maximum Entropy (ME), Transformation Based error-driven Learning (TBL), SVMs and Conditional Random Fields (CRF). The technique used is information gain (IG) feature selection to reduce the dimensionality. Future work in this domain involves distinction between Unicode characters and extended character sets so that this can be applied in websites which do not have plain text.

Structural units are easier to categorize with link based information that can simplify parsing within a domain. For example, in [18] the author presents a web page classification algorithm, Link Information Categorization (LIC) based on the K nearest

neighbor method, it combines information on the website features, to implement the Web page link to information classification. It was seen LIC is more suitable for Web information classification, especially for professional websites, it has better classification results than other traditional classification algorithms. Future work involves implementing LIC in systems which have completely unlabelled data.

A complete project based on multiple previous efforts can be seen in [19] the author focuses on categorizing product pages on the Web depending on their information. Naive Bayes and the complement naive Bayes classifier are used. The experiments showed that the product pages can be classified most correctly depending on only the nouns of the titles of the product pages. It was found that the complement naive Bayes classifier outperformed the naive Bayes classifier and the future work involved using pictorial product catalogues for classification as well, by combining image processing related concepts.

The above work can be further refined by experimenting with the algorithm and mathematical parameters. An example is [20] where the author uses Naive Bayes Expectation Maximization (EM) algorithm classification method (using hierarchical clustering EM framework) that trains the Naive Bayes classifier iteratively. It is used to classify the massive unlabeled data iteratively until all labelled data (pages) has been classified completely. The proposed M-EM algorithm can produce high classification accuracy. With the increase in number of labelled pages, F I-score ( F I-score  is an important indicator to assess the performance of classification result) value in both algorithms also became larger. However, the algorithm performance would be stable or even decline when the labelled data increased to a certain number. Therefore future work involves finding the inflection point in terms of size of dataset for which this algorithm is favourable.

Classification can also be done using a focused crawler [21], which is a web crawler that attempts to download only web pages that are relevant to a predefined topic or set of topics. In order to determine a web page is about a particular topic, focused crawlers use classification techniques. In this study the motivation is the classification of links instead of downloaded web pages to determine relevancy. Naïve Bayes classifier is combined for classification of URLs with a simple URL scoring optimization to improve the system performance. Focused crawlers use page scoring and link scoring techniques.

The results demonstrate that the link scoring method based on the Naïve Bayes (NB) classifier increases accuracy of the system considerably. Based on these results, the modified NB link classifier to incrementally update its training set during crawling will improve the system performance further.

The best features for web page classification problem, accuracy and run time performance of the classifiers can be improved by using Genetic Algorithm (GA) [22]. GA is important because it apparently attempts to decrease the feature space and determines the best features of the given set of web pages. It is found that when features selected by their genetic algorithm are used and a KNN classifier is employed, the accuracy improves up to 96%. However, since this is time consuming and computationally intense, future work aims to make it more efficient.

Malicious web pages can be detected using a two-stage classification model [23] which uses something called static and run-time features which are defined and used efficiently in each stage based on their values. It happens in 2 classification stages. The operation flow basically starts with a list of URLs which are fed into the static feature extractor which extracts the properties of the web page. This is then fed into the first stage of classification to estimate the maliciousness or basically categorize the web pages into potential malicious and benign one. Research gap includes work on both static or runtime features for classifying.

The unstructured data from web pages can be transformed into structured relations using a relation predictor to filter out irrelevant pages so that speed of the overall information extraction process is increased substantially. SVM [24] approach is used for this purpose. The evaluation pages on a sentence level, where each sentence is transformed into a token representation of shallow text features. The concept of prediction + extraction speeds up the process of extraction by a factor of 2 and research gaps include text extraction to remove advertisements, spam and other unwanted data.

One of the application of Neural Networks is the Classification of Chinese web pages to develop a Web Information Navigation System [25]. The quantum neural network has three layers: input layer, hidden layer, and output layer. The keywords of Web document are chosen as inputs of quantum neural network classifier. The test result shows that the average system performance of our quantum neural network classifier is

about 86.7%. Future work consists of six main parts: information collecting part, web document indexing part, web document classifying part, hyperlink depositing part, information inquiring part, and hyperlink updating part.

Another web page classification algorithm which can be used in the E-government system combines the Support Vector Machine (SVM) and the Unsupervised Clustering (UC) methods, called Combined UC and SVM algorithm (CUCS) [26]. The CUCS improves the speed and accuracy of the web page classification through pruning the training set. The accuracy rate, CUCS is much higher than UC, slightly higher than SVM. At the point of the time cost, UC is shortest, SVM is longest, and CUCS is in between, far less than SVM because the pruned training set reduces the training time complexity. This algorithm is implemented in E-government system theoretically and future work includes practical implementation.

The analogies of students' behaviour and internet usage pattern related to academic issues can be surveyed. Based on the classification scheme [27], out of 7000 visited websites by students during one year's data collection period, 36% of websites are classified as ACademic (AC) and remaining (64%) are classified as Non-ACademic websites (NAC). In the category of NAC visited websites, results presented that, the most visited websites belongs to Entertainment (43%) and the next most favourite websites were mentioned Social Networks (37%). The survey results of BTECH students, declared that department of computer science having majority of users and department of mathematic and civil having minimum number of users among all departments.

The three techniques generally used in classification are Naïve Bayes (NB), Decision Trees (DT) and Neural Networks (NN). NB [28] models are popular in machine learning applications, due to their simplicity in allowing each attribute to contribute towards the final decision equally and independently from the other attributes. This enhanced NB classifier has 3 stages - Crawler, Trainer and NB classifier with a common Indexer. The results showed that the enhanced NB classifier was comfortably in the lead by over 7% in both accuracy and F-Measure value. The NB classifier achieved the highest accuracy (97.89%), precision (99.20%), recall (98.61%) and F-Measure (98.90%) values, however, the DT classifier achieved the fastest execution time. However, the NB classifier achieves higher accuracy, precision and most importantly, overall F-measure value.

Social tags from social networking sites like Facebook, Twitter, MySpace can be used in classification of web pages. Social tagging [29] is a process in which many users add metadata to a shared content. Through the past few years, the popularity of social tagging has grown on the web and this motivates further research. In this paper the use of social tags for web page classification: adding new web pages to an existing web directory, has been investigated. The future work is based on applying different automatic approaches of using social tags for extending web directories with new URLs.

Another highly efficient classifier called Random Forest (RF) [30] classifier that can classify web pages efficiently according to their corresponding class without using other feature selection methods is generally used for large data sets. Random Forest classifier grows many classification trees. Each tree is trained on a bootstrapped sample of training data and at each node the algorithm only search across a random subset of the variables to determine a split. To classify an input vector in random forests, the vector is submitted as an input to each of the trees in the forest. Each tree gives classification, and it is said that the tree votes for that class. In the classification, the forest chooses the class having the most votes. The experiments have shown that the proposed approach is suitable for the multi-category web page classification.

A web-crawled academic video search engine, named as LeeDeo [31] that can search, crawl, archive, index, and browse academic videos from the Web, provides an alternative for other popular academic search engines like CiteSeer and Google Scholar. They built a proof-of-concept system and demonstrated how their classification technique worked. There were multiple gaps in this paper which provide scope for future research and work, such as extension of techniques used to multiple forums and across various languages, domains and providers. A focus on images and video grabs instead of complete videos is also suggested.

Another algorithm for web page classification called CUCS (Combined UC and SVM) [32] is used when the training set is large. The motivation is that it provides an automatic web page classification method to help users locate the required information on the internet in a convenient way. CUCS combines the advantages of SVM (Support Vector Machine) and UC (Unsupervised Clustering), achieving high precision and fast speed. In precision, CUCS is far higher than UC and a little better than SVM. As to time

efficiency, CUCS costs more time than UC, which forma the research gap of this paper. The future work is to improve the time efficiency of CUCS algorithm.

The web pages can be classified automatically using Rough Set Theory [33]. The motivation provided is to overcome the difficulty in mining high dimensional data because of the curse of dimensionality. The definition of a set of attributes consists of one conditional attribute which is evaluated on the basis of the others. The rules for classification are extracted iteratively by choosing different attributes as the conditional so that we can find which ones are superfluous and eliminate them. Experimental results using a dataset from www.sohu.com show that the precision and recall of this method is much higher. The way this is achieved is by noting the decision weight age given to different conditional variables in the VSM (Vector Space Model). The research gap is that this classification effect is not very good for the web page classifications in which class partitions are not obvious.

The identification of web pages, by making use of Intensive Explosive Devices or IEDs [34] to detect the malicious use by terrorists and extremists is the prime need of the hour by various security organizations. The motivation is that it provides an improvised explosive device web pages representing a significant source of knowledge. Here the approach used is to make a lexicon of 100 Arab words related to IEDs and then crawl specifically those websites which are mentioned in known sources and intelligence agencies as potential targets for terrorists. In this way Focused Crawling is used, an approach to select and narrow the search space. Three types of features were used-stylistic features deal with words, vocabulary, richness, frequency etc, while structural features deal with HTML tag elements like number of paragraphs. The topical features used include bag of words model etc. and this forms an extended feature set. Classification accuracy exceeded 88%. This paper provides only a framework for genre classification of IED related web pages, which poses as the research gap. The future work is to research and implement this framework.

A graph based approach [35], for classification, that scores over other approaches and that does not depend on labelled data is gaining popularity. The motivation is that it discusses about semi-supervised learning, which has been an important research oriented

topic in recent years, because it is difficult and expensive to obtain labelled data for classification purposes. It assigns weights according to similarity to unlabelled web pages and uses similarity measures to construct a K Nearest Neighbours graph. The problem then simplifies to that of label propagation between graph nodes. The noisy links like advertisements are removed and Transducer SVMs or TSVMs are used to compare the results. The research gap is that there still exists noisy links which reduces the efficiency of this approach. Hence future work is to effectively remove noise from the link information and optimize the calculation algorithm of semi-supervised learning, so as to investigate its new applications in the future.

The use of ontology [36], or a domain based representation of knowledge, in order to provide mechanism to enable machine reasoning has increased during the last few years. The motivation is the use of ontology for document classification which expresses terminology information contained in web documents. This approach seeks to exploit the fact that typically web documents have the following characteristics - First, they are structured in a web site with the web site as a unit, and in other words, a web site normally consists of many web texts with the same or a similar subject. Second, each web document exists as a part of other web sites. Finally, individuals or organizations that have specific purposes manage typical web sites. The advantages of an ontology model has higher accuracy which has been proved later in the paper and it is also mentioned that it needs no training data unlike existing methods. It is compared with an existing Bayesian classifier and it is found that this model outperforms. The research gap is that it is required to develop more efficient and accurate ontological expressions and to document classification methods. Future work would be to conduct further studies on how to improve the efficiency of an information search using the document classification technique suggested in this paper and how to automatically determine the meaning of concepts and relations from Web documents.

In dynamic and hierarchical classification system, the hierarchical structure[37] is exploited making use of categories of web pages. The motivation of this paper is the automatic classification of web pages to overcome the difficulty of retrieving information from the internet. First the web pages are arranged in a format that supports classification. They are then classified into categories using a "bag-of-words" approach. The categories

are then hierarchically arranged to form a tree where common features are propagated from parent to child node and at the same time each node can be reached through a unique path because it has its own set of unique features. It has been tested on two aspects of their system viz unique path traversal and dynamic updation where they find 84% and 78% accuracy respectively. The research gap is that this paper provides just a starting point for classification of web pages from the internet but the internet is highly unstructured and thus future work would be to convert the whole internet to a well-structured system based on some classification standard.

All the efforts above have dealt with isolated instances and examples of algorithms used in specific domains for webpage classification. Many research gaps remain such as integration of multiple approaches, dynamic updation, improvements in algorithms accuracy and precision as well as scalability for larger and more complex datasets. Each of these problems has to be dealt with and researched upon in order to improve this domain and use the results in productive and feasible final applications.

## 1.2 Motivation

All work till date in the field of web page classification has dealt with a single algorithm on a limited dataset and a specifically selected metric for measurement. There has been no attempt to compare multiple algorithms across large complex datasets and use multiple measurement metrics. There is also a need to examine a domain of widely changing web pages such as those of online courses which can change as per date and location. This project is motivated to solve the above issue. There is also no system in place which provides access to information about these courses offline unless it has been manually composed. Such efforts are tedious and prone to error and obsolescence. Therefore it is strongly required to provide multiple facets of information to query over 30,000 courses from over 60 providers worldwide and create an offline repository that can be accessed easily. The application developed is motivated to be dynamic in updation and also provides the first of its kind, an offline automatically sourced and accurate standalone application to provide online course information.

## 1.3 Problem Statement

The problem statement is to compare and analyse the various machine learning algorithms used for binary classification of unstructured data collected from web pages of online course providers across the internet. There are over 35,000 online courses from world's leading universities. According to a recent survey, more than 7 million students are enrolled in them. Manual updation and collection of information about these courses is tedious and static in nature. Automatic analysis of web pages in order to extract information about these courses is the need of the hour. A system is to be developed which can crawl the web, filter unnecessary pages, extract content, gather social tags, pre-process links and metadata, compute features and then classify whether each page visited is a relevant online course or not. After refining the system and evaluating the results of multiple algorithms, the project aims to use the best one to classify a large dataset and store offline information about relevant course pages. This offline information must then be presented to the users with a fresh and intuitive interface and the feature to update dynamically.

## 1.4 Objectives

The main objective of the project is to be able to collect information on web pages relevant to training courses according to various factors such as cost, curriculum, provider, university, instructor, description, webpage link and certification. The first way to achieve this is to fulfill the aim of successfully and quickly crawling multiple domains while filtering out advertisements, spam and unwanted pages. Preprocessing is required in terms of analyzing text contents as well as page metadata enriched features, date and time, social tagging and other parsing and text extraction based features. The next objective is to create an innovative, efficient and fast way to classify courses and training web pages according to user need and relevance. This could be done by evaluating various binary classification algorithms such as decision trees, logistic regression, naïve Bayes network, K nearest neighbors and random forest ensemble. Once done, this must be available in the form of a fresh and interactive offline standalone application which users can query according to multiple parameters. Automatic updating of content from the retrieved web pages according to the relevance for storage and querying for long term use is ensured.

## 1.5 Scope

The scope of the project is that the end users would have an offline system which would allow them to search for the available online courses according to the area of interest. For initial setup as well as subsequent updates over a defined time frame like a month, the system requires accessing of the internet to perform crawling and indexing of web pages. Once this data is collected, efficient classification algorithms will be applied offline to segregate the pages and fill them into a database for the user to access in future.

## 1.6 Methodology

The methodology of the project involves firstly, gathering of data from multiple sources to create a dataset large enough for training a model classifier. The next step is analysis and preprocessing of data collected, selection and creation of features with innovative improvements. After this, training of model and testing using various methods like bootstrapping and cross validation must be done. The penultimate step is comparison using multiple parameters such as F measure and ROC curves. Finally, the development a complete product with the best model and an interactive front end for user and business applications must be carried out.

## 1.7 Organization of the report

This section of report briefly describes the organization of each topic in the report. The report contains 8 chapters altogether. The following paragraphs describe the brief summary of each chapter respectively.

Chapter 2 is Software Requirement Specifications. It gives an overall description of the functionalities available in the system along with the specific requirements to achieve each of these functionalities.

Chapter 3 is High Level Design. It focuses on High level Design of the system. It gives an abstract view of the architecture of the basic building blocks of the system. It explains how the data flows between various layers of system.

Chapter 4 is Detailed Design, which focuses on the structure chart of this project and and flowchart diagrams. It explains the key modules involved in the project.

Chapter 5 is Implementation, which explains the programming language, development environment and code conventions followed during the project. This chapter puts light on the various difficulties encountered in the project and how they were overcome.

Chapter 6 is Testing, which explains the test environment and briefly explains the test cases which were executed during unit testing, functional testing and integration testing.

Chapter 7 is Experimental Analysis and Inference of results where the results of the project are listed and inferences are made about the efficiency of the editor.

Chapter 8 is conclusion, which gives the outcome of the project work carried out and also brings out the limitations of the project and future enhancements.

## 1.8 Summary

This chapter has provided an introduction to the various domains covered in this project as well as relevant information regarding the same. It has provided an extensive literature survey about the background of the field in which research is conducted, as well as introduced various objectives, aims and overview of the entire project, while setting the tone for the rest of the detailed phases conducted through the project. The remaining chapters of the report deal with further stages in the conceptualization, design and development of the idea explained herewith.