Automatic Text Classification System

Vrusha U.Suryawanshi, Pallavi Bogawar, Pallavi Patil,
Priya Meshram, Komal Yadav, Prof. Nikhil S. Sakhare

Abstract- **The goal of text classification is to classify the text documents into a certain number of pre-defined classes. As per demands for text is increased with the evolution of large amount of information available in internet, news, institutes. For this large amount of data we need a text classifier which will help in classifying text documents. In classification there are major issues such as handling large number of features, unstructured text documents, and selecting a machine learning technique suitable for the text classification application. The Automatic Text Classification System solves those problems by giving the text document to a set of pre-defined classes by using machine learning techniques. The appropriate machine learning technique for text classification is k-nearest neighbour. The classification is mainly done on the basis of significant words or features extracted from the text document. It also can be used for creating training documents and for creating databases for various texts.**

Index Terms- **Automatic Text Classification, Text Mining (Mining Algorithm), K-Nearest Neighbour algorithm, Text Data.**

I. INTRODUCTION

The System explains the strategy for automatic text classification. Automatic Text Classification has a set of classes which is pre defined in the system by using an artificial intelligence techniques. With the rapid growth of the internet, the availability of on-line text information has been considerably increased. The documents classification is usually done by selecting the feature extraction of the text document. As the classes are pre defined so it will be done by using supervised machine learning techniques. The supervised machine learning goal of it is to construct a concise model of the class labels in terms of predictor features. In supervised learning, each example is a pair consisting of an input object and a desired output value. Now a days the most commercial and official communication and documentation maintained in government organizations is in the form of textual documents. Automated Text Classification makes the classification process fast and more efficient since it automatically categorizes documents by mean of the count of features the class are defined.

Due to use of internet in forms of emails, blogs, search engines etc by this the information is

Overloaded for this efficient classification and retrieval of relevant content has become more important. For this system the machine learning model can be implemented by using the k-nearest neighbor algorithm, the k is the value of text or a word. Moreover, text posts on a blog do not strictly adhere to the blog topic. This introduces the need to develop a text classifier for a huge amount of text which has to classify. As a result, the text classification has become one of the key techniques for handling and organizing text data. Automatic text classification in the previous works is a supervised learning task, defined as assigning classes (pre-defined) to text documents based on the likelihood suggested by a training set of labelled documents. However, the previous learning algorithms have some problems. One of it is that they require a huge, often unaffordable, number of labelled training documents for the accurate learning. Therefore, the learning algorithm "k- nearest neighbour" is applying to classify and also to keep track on the count of the texted word which has been featured in a documents repeatedly.

The system performs a three-tier architecture which resembles that there will be a communication between client-server. The client provides web URL to the server. The server will fetch the html code from the web address. The Fetched HTML Code is processed and HTML Tags and code are removed and plain text remains. This plain text is processed and english keywords are removed using database stored keywords. Finally APA Scans the Remaining Text and assigns values to each remaining word and counts the iterations and compare phrases using the database classification data.

Automatic text classification has several useful applications such as classifying text documents in electronic format; spam filtering; improving search results of search engines; opinion detection and opinion mining from online reviews of products, movies or political situations; and text sentiment mining. This system can be implemented further for various application building, for example in page ranking, filtering text, and information retrieving.

II. RELATED WORKS

Mita K. Dalal and Mukesh A. Zaveri [1] has propose in their research paper that the process of text classification has been divided into four phases i.e., i) document pre-processing, ii) feature extraction / selection, iii) model selection, iv) training and testing the classifier.

i) Document pre-processing: Data pre-processing reduces the size of the input text documents. It involves activities like sentence boundary determination stop-word elimination and stemming.

ii) Feature extraction / selection: Feature extraction / selection help identify important words in a text document. This is done using methods like TF-IDF (term frequency-inverse document frequency) the text document is represented as a document vector.

iii) Model selection: An appropriate machine learning algorithm is used to train the text classifier. Naïve Bayes , Decision Tree, Neural Network, Support Vector Machine, Hybrid approach etc are used for feature/model selection.

iv) Training and testing the classifier: The trained classifier is tested using a test set of text documents. If the classification accuracy of the trained classifier is found to be acceptable for the test set, then this model is used to classify

ISSN: 2278 – 1323

new instances of text documents. The Automatic Text Classification system is a semi-supervised machine learning task that by design assigns a given document to a set of pre-defined categories based on its textual content and the extracted features [1].

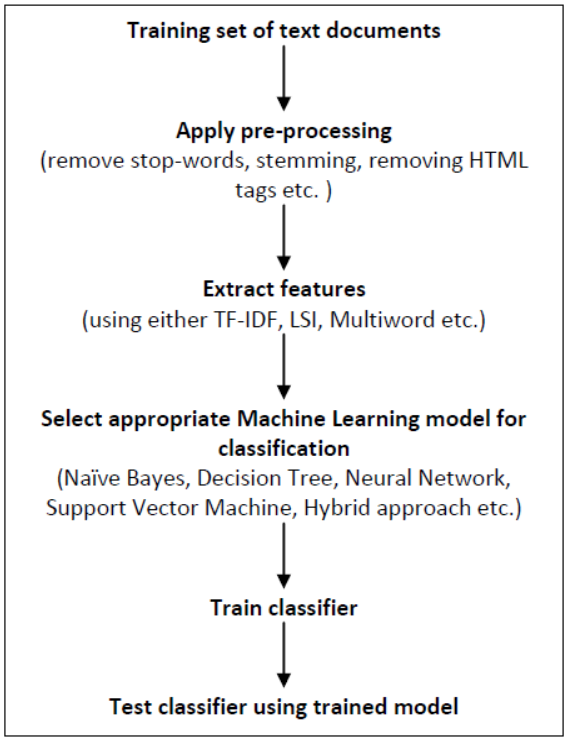

Fig. 1 Generic strategy for text classification [1]

In above figure, the process of classification is given as per the author had proposed.

Youngjoong Ko and Jungyun Seo Their proposed system consists of three modules; a module to pre-process collected documents, a module to create training sentence sets, and a module to extract features and to classify text documents.

First, the html tags and special characters in the collected documents are removed. And then, the contents of the documents are segmented into sentences. We extract content words for each sentence using only nouns. Because the proposed system does not have training documents, training sentence sets for each category corresponding to the training

Documents have to be created. We define keywords for each category by hand, which contain special features of each category sufficiently. Next, the sentences which contain pre-defined keywords of each category in their content words are chosen as the initial representative sentences. The unclassified sentences to their related category. This assignment has been done through measuring similarities of the unclassified sentences to the representative sentences. Word weights are computed using Term Frequency (TF) and Inverse Category Frequency (ICF). In Information Retrieval, Inverse Document Frequency (IDF) are used generally. But a sentence is a processing unit in the proposed method. Therefore, the document frequency cannot be counted [2].

Kim S., Han K., Rim H., and Myaeng S. H in their paper they uses porter steamer for stemming words and for term weighting they use term frequency and tf-idf for text classification they has used Naive Bayes algorithm[3]. They uses three steps for classifying text that is pre-processing(uses separating words, removing stop words, stemming and weighting) , feature extract/ selection(reduces size of feature set by selecting best features),learning algorithm(evaluates classification performance with some criteria).

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini , Chris Watkins they had proposed in their paper about classifying a text by string kernel. The kernel is an internal product in the feature space generated by all sub sequences of the length k. A substring is any ordered sequence of k characters in the text still not necessarily. The sub sequences are weighted by an exponentially decaying factor of their full length in the text, hence more noticeable occurrences that are close to neighbouring. A direct calculation of this feature vector would involve a prohibitive amount of computation even for reserved values of k, since the dimension of the feature space grows exponentially with k. Kernel methods (KMs) are an effective alternative to explicit feature extraction [4].

III. OBJECTIVES

The aim of Automatic Text Classification System involves assigning a text document to a set of pre-defined classes, using a machine learning technique. The classification is mainly done on the basis of significant words or features extracted from the text document. This features are distinguish by using a machine learning technique i.e, k- Nearest Neighbour.

IV. PROPOSED WORK

It has been observed from the literature survey that there are problems for using various machines learning algorithm and it becomes tedious in time consuming. To overcome the mentioned problems, a comparative analysis is done by using one algorithm “k-nearest neighbour” for a machine learning.

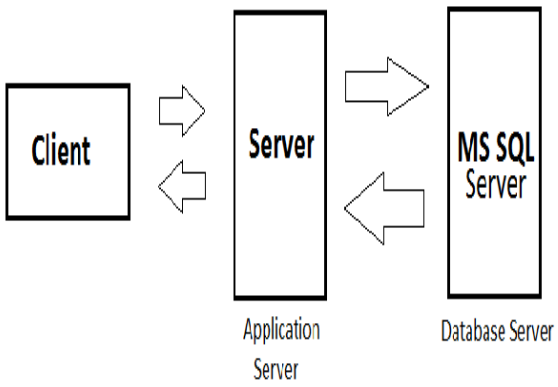

Fig. 2 Three-tier client-server architecture

The above figure is about Three-tier client-server architecture will has been used in the system. It enables an application from client machine to send commands to database through use of middleware service. The database processes command request, then sends a response to the middleware service which then forwards the reply to client.

ISSN: 2278 – 1323

Efficient because middle tier (application server) of the architecture handles the data processing operations between client and database server. It performs the following main activities:-

1. Translation between the different protocols.
2. Optimization of the load-balancing.
3. security control.
4. Management of the connections.

The middleware may contain several components. The components may reside on the server node, on the client node or on a new middleware node. The different types of middleware

- Database
- Network
- Application cooperation.

Hence the three tire architecture is a basic flow of the system in which there will be client-server communication will happen for data processing.

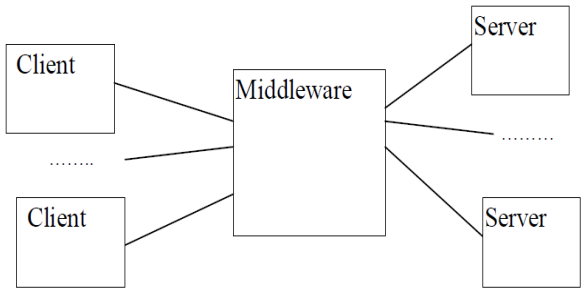

Fig. 3 Middleware process

V. SYSTEM ARCHITECTURE

Automatic Text Classification involves assigning a text document to a set of pre-defined classes automatically, using a machine learning technique. The classification is usually done on the basis of significant words or features extracted from the text document**.** The system involves four phases i.

i. Fetch html code from server.
ii. Remove the html code.
iii. Remove the English keywords.
iv. Classify the remaining text.

A. Fetching the HTML code

The Human Admin/client provides the web URL. The web URL is the address box were human admin/client will give the address which human wants to know the information as per the knowledge of web, the web designing is done in HTML. After entering the address the APA firstly fetches the HTML Code from the Web-Address. As the web addresses are in form of html code.

B. Remove the html code

There are many tags in the code.The Fetched HTML Code is processed and HTML Tags and Code are removed and Plain Text Remains. The unused tags or repeatedly coming code are been removed i.e., the head tag part. Only the body of HTML is remained.

C. Remove the English keywords

All the stop words and the code which is repeatedly coming are also removed. After this process there will be only plain text, the plain text has many English words, verbs, adjectives which are not of use. This Plain Text is processed and English Keywords are Removed using Database stored Keywords.

D. Classify the remaining text

Finally APA Scans the Remaining Text and assigns values to each remaining word and counts the iterations and compare phrases using the database Classification data. The classifier will classify the featured word and will make a class in database with its information. In the last two stages there will be a "k nearest neighbour" algorithm which will be used to classify the featured word according to the iterations and will compare with the remaining classes if the class is present in the database the information will update and if not present then it will make a new class. The use of k nn algorithm is used. There is no use of any other algorithm for classification; due to this the classification is easier and maintainable.

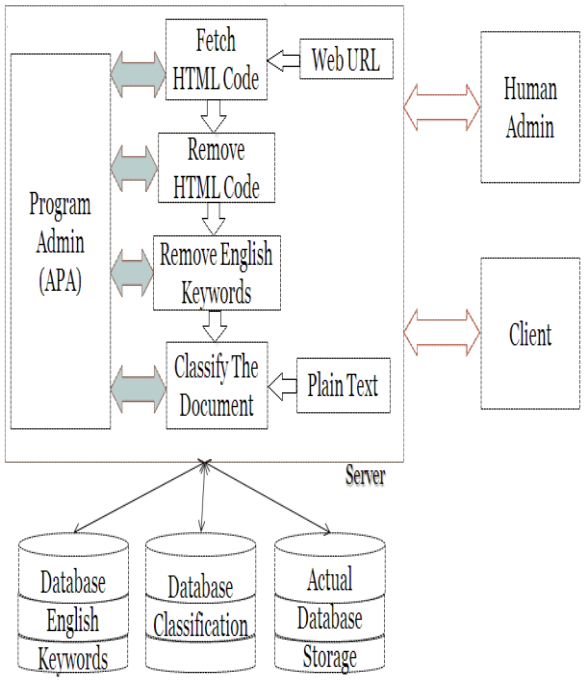

Fig. 4 System Architecture

VI. SOFTWARE REQUIREMENTS

i. IIS Server
ii. .Net Framework 4.5
iii. SQL Server 2008

A. Internet Information Services (IIS)

It was formerly called Internet Information Server – it is a web server application and set of feature addition modules created by Microsoft for using Microsoft Windows. It is the most used server in industries. The protocols support in IIS include: FTP, SMTP, FTPS,HTTP/HTTPS and NNTP. IIS was initially released as a set of web-based services for Windows IIS has features as:

ISSN: 2278 – 1323

a. HTTP modules: Used to perform tasks specific to HTTP in the request-processing pipeline, such as respond to information and investigate the sent in client headers, returning HTTP error, and redirecting requests.

b. Security modules: Security is used to perform tasks related to security in the request-processing pipeline, such as specifying authentication schemes, performing URL authorization, and filtering requests.

c. Content modules: Content module is used to perform tasks related to content in the request-processing pipeline, such as processing requests for static files, returning a default page when a client does not specify a resource in a request, and listing the contents of a directory.

d. Compression modules: Compression modules is used to perform tasks related to compression in the request-processing , such as compressing responses, applying Gzip compression transfer coding to responses, and performing pre-compression of static content.

e. Logging and Diagnostics modules: Logging and Diagnostics modules is used to perform tasks related to logging and diagnostics in the request-processing pipeline, such as passing information and processing status to http.sys for logging, reporting events, and tracking requests currently executing in worker processes.

B. Microsoft .NET Framework

The Microsoft Windows operating system uses The Microsoft .NET Framework as a component. It provides a huge of pre-coded solution to common program requests, and manages the execution of programs written particularly for the Framework. The .NET Framework is used by many new applications created for the Windows platform. The pre-coded solutions make the class library and cover a huge range of programming requests in areas including: data access, user interface, database connectivity, web application development, cryptography, numeric algorithms, and network communications. The function of the class library is used by user's (programmer) who wants to build software they combine them with their own code to make necessary applications. Programs written for the .NET Framework execute in a software background that manages the program's runtime requirements. .NET provides a program interface, Web services, to communicate with remote components. The library provides classes that can be used in the court to accomplish a range of common programming tasks, such as string management, data collection, database connectivity, and file access. One of the most important features of the .NET Framework class library is that it can be used in a consistent manner across multiple languages.

C. Microsoft SQL Server 2008

SQL Server 2008 was released in October 2005. It included subject support for controlling XML data. For this reason, it will define an XML data type that can be used as a data type in literals or database columns in queries. SQL Server 2008 has also been better with syntax, indexing algorithms and improved error recovery systems. Data pages are verify summed for better error resiliency, and optimal concurrency support which has been added for better performance. SQL Server 2008 introduced Database Mirroring. Database mirroring is a high ease of access option that offers redundancy and failover capability at the database. Failover can be performed automatically or can be configured for regular failover; failover needs a witness partner and an operating mode of synchronous (also known as high-safety or full safety).

VII. CONCLUSION

Due to an increase in the number of blogs, websites and electronic storage of textual data, the commercial importance of automatic text classification applications has increased and much research is currently focused in this area. Text classification can be automated successfully using machine learning techniques, there are various techniques in machine learning and but for proposing this system only one algorithm is use that is "K Nearest Neighbour". The algorithm has a great domain by which various operations can be done. Much facility that in the past have been done with plenty of time and financial costs can be able by this techniques and algorithms with low costs and effectiveness performance.

This system has Use of Virtual Private networks which has more security. In our network we can use this system for storing the information and maintaining the database and tracking the Documents without human Interaction. The system can avoid the use of the unnecessary overheads of the previous system which slows the data transmission, i.e. logging to the website and downloading the code and formatting later. Security Aspects make the system to keep threats (Network) away**.**

REFERENCES

[1] Mukesh A. Zaveri and Mita K. Dalal, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications (0975 – 8887), Volume 28– No.2, August 2011.

[2] Youngjoong Ko and Jungyun Seo "Automatic Text Categorization by Unsupervised Learning" KOSEF under Grant No. 97-01-02 03-01-03.

[3] Kim S., Han K., Rim H., and Myaeng S. H., "Some effective techniques for naïve bayes text classification", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466., 2006.

[4] Huma Lodhi, Craig Saunders , John Shawe-Taylor, Nello Cristianini, Chris Watkins **,** "Text Classification using String Kernels" Journal of Machine Learning Research 2 (2002) 419-444, Published 2/02.

[5] Zhang W., Yoshida T., and Tang X."Text classification using multi-word features". In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524. 2007.

[6] HaoLili., and HaoLizhu.. "Automatic identification of stopwords in Chinese text classification". In proceedings of the IEEE international conference on Computer Science and Software Engineering, pp. 718 – 722. 2008.

[7] Porter M. F."An algorithm for suffix stripping". Program, 14 (3), pp. 130-137.1980.

[8] Liu T., Chen Z., Zhang B., Ma W., and Wu G. Improving text classification using local latent semanticindexing. In proceedings of the 4th IEEE international conference on Data Mining, pp. 162-169. . 2004.

[9] M. M. Saad Missen, and M. Boughanem. "Using WordNet" s semantic relations for opinion detection in blogs". ECIR 2009, LNCS 5478, pp. 729-733, Springer-Verlag Berlin Heidelberg.

[10] Balahur A., and Montoyo A. "A feature dependent method for opinion mining and classification".In proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering, pp. 1-7. 2009.

ISSN: 2278 – 1323

[11] Cho K. and Kim J. "Automatic Text Categorization on Hierarchical Category Structure by using ICF(Inverted Category Frequency)".Weighting. In Proceedings of KISS conferencepp.507-510. 1997.

Authors Profile:

Vrusha U.Suryawanshi pursuing the Bachelor degree (B.E.) in Computer Science and Engg in 2015 from RGCER, Nagpur.

Pallavi Bogawar pursuing the Bachelor degree (B.E.) in Computer Science and Engg in 2015 from RGCER, Nagpur.

Pallavi Patil pursuing the Bachelor degree (B.E.) in Computer Science and Engg in 2015 from RGCER, Nagpur.

Priya Meshram pursuing the Bachelor degree (B.E.) in Computer Science and Engg in 2015 from RGCER, Nagpur.

Komal Yadav pursuing the Bachelor degree (B.E.) in Computer Science and Engg in 2015 from RGCER, Nagpur.

Under the guidance of:

Prof. Nikhil S. Sakhare
Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Research, Nagpur, India.

ISSN: 2278 – 1323