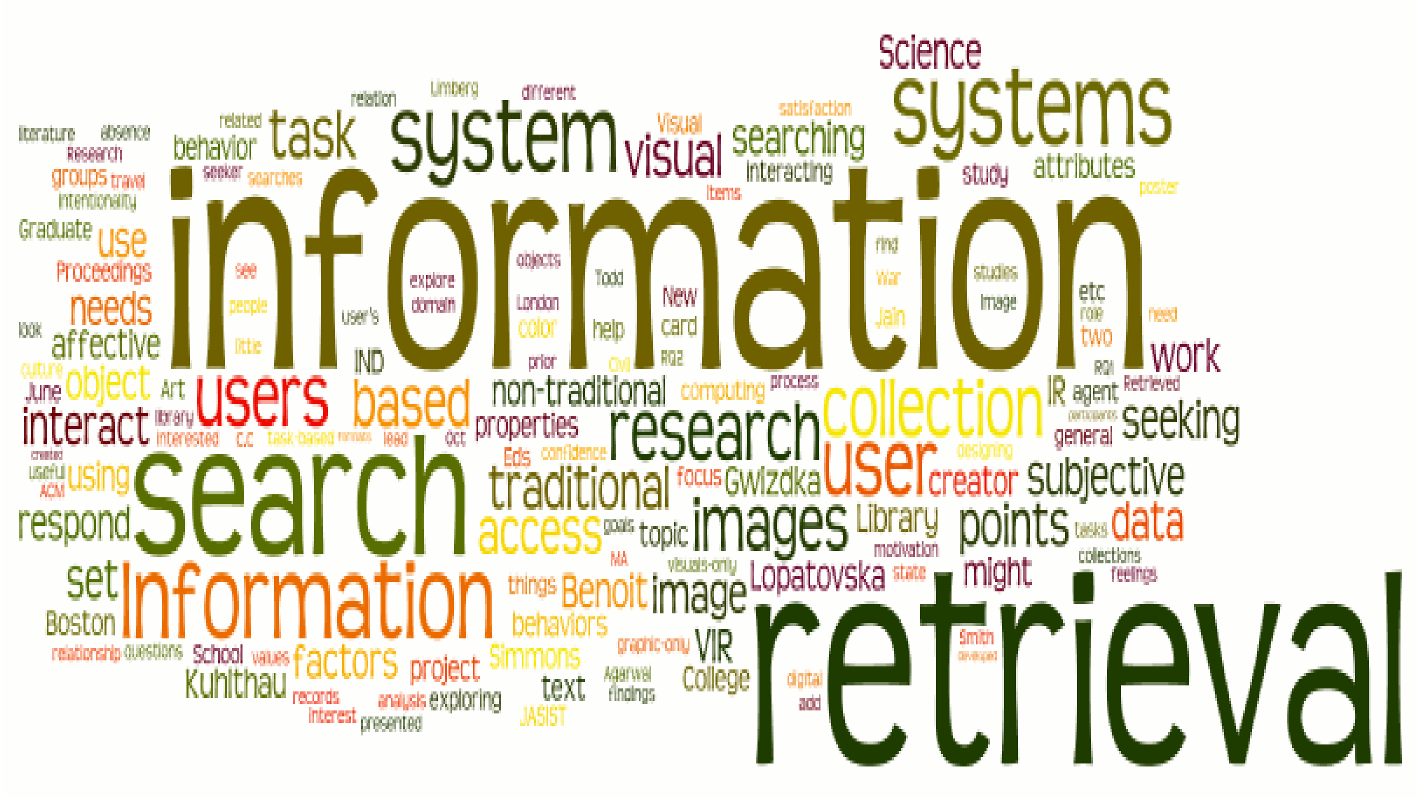
Maulana Azad National Institute of Technology



PROJECT SYNOPSIS

A FUZZY SIMILARITY BASED CONCEPT MINING MODEL FOR TEXT CLASSIFICATION

Text Document Categorization Based on Fuzzy Similarity Analyser and Support Vector Machine Classifier

Team Member: Final Year 2016

Anand Namdev	121112118
Amit Birla	121112030
Adish Jain	121112085
Suneel Kumar	121112021

Abstract

Text Classification is a challenging field and has great importance in text categorization applications. A lot of research work has been done in this field but here is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts. In this project, a new Fuzzy Similarity Based Concept Mining model (FSCMM) is proposed to classify a set of text documents into pre - defined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyser (FFCSA) is used to analyse each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model uses Support Vector Machine Classifier (SVMC) to classify correctly the training data patterns into two groups i.e. + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently great performance and high accuracy results.

Introduction

Artificial Intelligence (AI) is an area of study that embeds the computational techniques and methodologies of intelligence, learning and knowledge to perform complex tasks with great performance and high accuracy.

Natural Language Processing (NLP) is the heart of AI and has text classification as an important problem area to process different textual data and documents by finding out their grammatical syntax and semantics and representing them in the fully structured form.

Text Mining (TM) consists of extracting regularities, patterns, and categorizing text in large volume of texts written in a natural language And NLP is used to process such text by segmenting it into its specific and constituent parts for further processing.

Text categorization (also known as text classification, or topic spotting) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles, filing patents into patent directories, spam filtering, identification of document genre, and even automated essay grading. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems rivals that of trained human professionals, thanks to a combination of information retrieval (IR) technology and machine learning (ML) technology.

Therefore, TM categorization is used to analyse and comprise of large volume of non - structured textual documents. Its purpose is to identify the main concepts in a text document and classifying it to one or more pre-defined categories. NLP plays an important and vital role to convert unstructured text into the structured one by performing a number of text pre-processing steps. This processing results into the extraction of specific and exclusive concepts or features as words. These features help in categorizing text documents into classified groups.

Text classification is performed on the textual document sets written in English language where words can be simply separated out using many delimiters like comma, space, full stop, etc. Most of the developed techniques work efficiently with English language where the words can be clearly determined by simple tokenization techniques. Such text is referred as segmented text. Hence, the proposed effort is to work only on the English text documents.

Problem Definition

Automated text classification frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. Also a lot of research work has been done in the field of text classification but there is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts or features using supervised learning paradigm and different classification algorithms. Thus the main objective of the project is to propose a concept mining model which can classify a set of text documents into predefined category groups (CG).

Problem Domain

Text categorization is one of the challenging field of machine learning which can be defined as the Field of study that gives computers the ability to learn without being explicitly programmed". Or it can be mathematically defined as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

The project has supervised machine learning as its base where the machine or the computer is presented with examples and their desired outputs is given by the "teacher" and the goal is to learn a general rule that maps input to output. The project also uses concept of Natural Language Processing which can be defined as the field of computer science concerned with the interaction between computers and human language. Many challenges in NLP involve natural language understanding that is enabling computers to derive meaning from human or natural language.

Prerequisite

Concept Mining

Concept Mining is used to search or extract the concepts embedded in the text document. These concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new text document is introduced to the system, the concept mining can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts. In this way, the similarity measure is used for concept analysis on the sentence, document, and corpus levels. These concepts are originally extracted by the semantic role labeller and analysed with respect to the sentence, document, and corpus levels. Thus, the matching among these concepts is less likely to be found in nonrelated documents. If these concepts show matching in unrelated documents, then they produce errors in terms of noise. Therefore, when text document similarity is calculated, the concepts become insensitive to noise.

Feature Extraction

In text classification, the dimension or the size of the feature vector is usually huge. The problem becomes worse when there is the problem of Curse of Dimensionality, in which the large collection of features takes very much dimension in terms of execution time and storage requirements. This is considered as one of the problems of Vector Space Model (VSM) where all the features are represented as a vector of n dimensional data. Here, *n* represents the total number of features of the document. This features set is huge and high dimensional. There are two popular methods for feature reduction: Feature Selection and Feature Extraction.

In feature selection methods, a subset of the original feature set is obtained to make the new feature set, which is further used for the text classification tasks with the use of Information Gain.

In feature extraction methods, the original feature set is converted into a different reduced feature set by a projecting process. So, the number of features is reduced and overall system performance is improved. Feature extraction approaches are more effective than feature selection techniques but are more computationally expensive. Therefore, development of scalable and efficient feature extraction algorithms is highly demanded to deal with high-dimensional document feature sets. Both feature reduction approaches are applied before document classification tasks are performed.

Pseudo Thesaurus

The Pseudo Thesaurus is a predefined English Vocabulary Set which is used to check the invalid words or to remove extra words from a sentence while processing the sentence in the TDTP. It is also used for word stemming so that the exact word can be obtained. For example, consider three different words for the word research - researching, researcher and researches. When the word stemming is performed, research is the final resulting word with the feature frequency counted as 3.

Class Feeder

Text Classification is the process of assigning the name of the class to a particular input, to which it belongs. The classes, from which the classification procedure can choose, can be described in many ways. So classification is considered as the essential part of many problems solving tasks or recognition tasks. Before classification can be done, the desired number of classes must be defined.

Support Vector Machine Classifier (SVMC)

Support vector machines are supervised learning models with associated learning algorithms that analyse data and recognize patterns and are used for classification and regression analysis.
SVMC is a popular and better method than other methods for text categorization. It is a kernel method which finds the maximum margin hyper plane in the feature space paradigm separating the data of training patterns into two groups like Boolean Logic 1 and 0. If any training pattern is not correctly classified by the hyper plane, then the concept of slack measure is used to get rid out of it.

Approach

We have proposed a new Fuzzy Similarity Based Concept Mining Model (FSCMM) to classify a set of text documents into predefined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyser (FFCSA) is used to analyze each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model uses Support Vector Machine Classifier (SVMC) to classify correctly the training data patterns into two groups; i. e., + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently and effectively with great performance and high - accuracy results.

Project Work Flow

Here the proposed Fuzzy Similarity Based Concept mining model (FSCMM) is discussed. This model automatically classifies a set of known text documents into a set of category groups. The model shows that how these documents are trained step by step and classified by the Support Vector Machine Classifier (SVMC). SVMC is further used to classify various new and unknown text documents categorically.
The proposed model is divided into the two phases: Text Learning Phase(TLP) and Text Classification Phase (TCP).

Text Learning Phase

TLP performs the learning function on a given set of text documents. It performs the steps of first stage i.e. Text Document Training Processor (TDTP) and then the steps of second stage i.e. Fuzzy Feature Category Similarity Analyser (FFCSA). The TDTP is used to process the text document by converting it into its small and constituent parts or chunks by using NLP concepts at the Sentence, Document and Integrated Corpora. Then, it searches and stores the desired, important and non-redundant concepts by removing stop words, invalid words and extra words. In the next step, it performs word stemming and feature reduction. The result of sentence level preparation is low dimensional Reduced Feature Vector (RFV). Each RFV of a document is sent for document level preparation, so that Integrated Reduced Feature Vector (IRFV) is obtained. To obtain IRFV, all the RFVs are integrated into one. Now, Reduced Feature Frequency Calculator (RFFC) is used to calculate the total frequency of each different word occurred in the document. Finally, all redundant entries of each exclusive word are removed and all the words with their associated frequencies are stored in decreasing order of their frequencies. At the integrated corpora level, the low dimension Integrated Corpora Feature Vector (ICFV) is resulted.

In such a way, feature vectors at each level are made low dimensional by processing and updating step by step. Such functionality helps a lot to search the appropriate concepts with reduced vector length to improve system performance and accuracy.

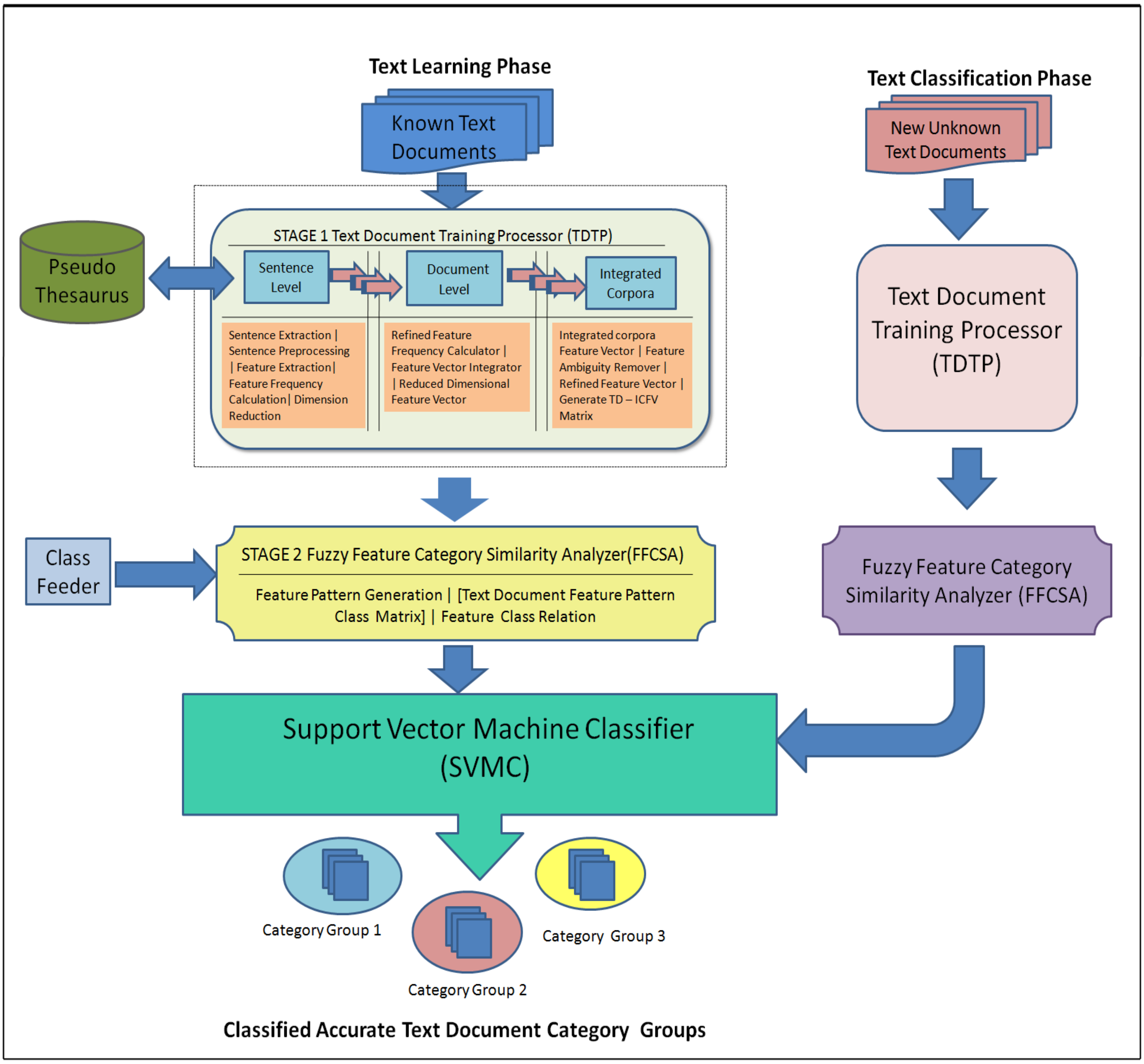
FFCSA performs similarity measure based analysis for feature pattern (TD – ICFV) using the enriched fuzzy logic base. The membership degree of each feature is associated with it. Therefore, an analysis is performed between every feature of a text document and class.

SVMC is used for the categorization of the text documents. It uses the concept of hyper planes to identify the suitable category. Furthermore, SVMC accuracy is checked by providing some new and unknown text documents to be classified into the respective Category Group (CG). This task is performed in TCP.

Text Classification Phase

To check the predictive accuracy of the SVMC, new and unknown text document is used, which is independent of the training text documents and is not used to construct the SVMC. The accuracy of this document is compared with the learned SVMC’s class. If the accuracy of the SVMC is acceptable and good, then it can be used further to classify the future unseen text documents for which the class label is not known. Therefore, they can be categorized into the appropriate and a suitable category group.

Work Flow Diagram


Text Learning Phase
Known Text Documents
STAGE 1 Text Document Training Processor (TDTP)
Sentence Level
Document Level
Integrated Corpora
Sentence Extraction | Sentence Preprocessing | Feature Extraction| Feature Frequency Calculation| Dimension Reduction
Refined Feature Frequency Calculator | Feature Vector Integrator | Reduced Dimensional Feature Vector
Integrated corpora Feature Vector | Feature Ambiguity Remover | Refined Feature Vector | Generate TD – ICFV Matrix
Pseudo Thesaurus
Class Feeder
STAGE 2 Fuzzy Feature Category Similarity Analyzer(FFCSA)
Feature Pattern Generation | [Text Document Feature Pattern Class Matrix] | Feature Class Relation
Text Classification Phase
New Unknown Text Documents
Text Document Training Processor (TDTP)
Fuzzy Feature Category Similarity Analyzer (FFCSA)
Support Vector Machine Classifier (SVMC)
Category Group 1
Category Group 2
Category Group 3
Classified Accurate Text Document Category Groups

Application

This task has several applications, including automated indexing of scientific articles according to predefined technical terms, filing patents into patent directories, selective exposure of information to information consumers, spam filtering, identification of document genre, and even automated essay grading. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved.

Some other applications are summary creation, clustering, language identification, term extraction and categorization, electronic mail management, document management, and market research with an investigation

Performance Measures

Performance measurement is the process of collecting, analysing and reporting information regarding the performance of an individual, group, organization, system or component.

Accuracy

Accuracy is used to describe the closeness of a measurement to the true value. Trueness is the closeness of the mean of a set of measurement results to the actual (true) value.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

Precision

Precision is a description of a level of measurement that yields *consistent* results when repeated. It is associated with the concept of "*random error*", a form of observational error that leads to measurable values being inconsistent when repeated. It is the closeness of agreement among a set of results.

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

F- Measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Project Scheduling Time graph

1) **Project Learning Phase**: It is the phase where concepts of Natural Language Processing are studied. It includes the detailed study of concept used in the modelling of the project. It includes learning python, learning NLTK which is the library of NLP in python, learning machine learning concept which includes SVMC (support vector machine classification). This phase will be covered in 1month duration.

2) **Algorithm Design**: This phase includes the detailed study of project and designing the appropriate algorithm for the project. The algorithm designed will be such that it completes the intended task in the best way possible and free from any logical errors. This phase will be covered in 1/2month (15days).

3) **Implementation Phase**: This phase includes the coding phase of the project where proper implementation of the project will be done according to the designed algorithm. The code written should be clear and easy to understand, which completes our intended task and also is free from any logical errors. This phase is divided in two part :

3. a) The first part of the implementation phase includes coding the learning part of the project which includes learning and training the machine. Here the machine is trained to give appropriate result for testing subjects. This part of the implementation phase will be covered in 25 days.

3. b) The second part of the implementation phase includes coding the SVMC which is support vector machine classifier. It is the final stage of the learning phase where the machine is finally trained to give correct result for new test documents. It has to be completed in 20days.

4) **Testing Phase**: It includes testing the machine with new documents. It includes 10-fold document testing and 5-fold document testing , which includes 9 document training and 1 document testing phase. This method is repeated for easy document and the results are noted for calculating the efficiency measure of the machine. This part will be covered in 25days.

5) **Documentation Phase**: It includes documenting the entire project. It includes proper detailed description about every concept used in the project. It includes description of the learning phase, about the machine training part, all the steps in correct order to beautifully explain the steps of the learning phase. It also explains the machine coding phase and then the testing phase in the end. Finally the documentation will be concluded with results and efficiency parameters. It will be covered in 10days.

Roles of Team Members

1. **Learning Phase**

   Learning phase of the project includes learning the concept and material required in building the project. It includes learning python, learning NLTK which is the library of NLP in python, learning machine learning concept which includes SVMC (support vector machine classification). This phase will be covered by all the members of the team.

2. **Algorithm Design**

   This phase includes the detailed study of project and designing the appropriate algorithm for the project. This phase will be completed by Anand Namdev & Amit Birla.

3. **Implementation Phase**

   This phase includes the coding phase of the project where proper implementation of the project will be done according to the designed algorithm.

   The first part of the phase which is training and learning of the machine will be completed by Anand Namdev,Amit Birla & Adish Jain.

   The second part of the phase which includes coding the SVMC will be completed by Anand Namdev & Amit Birla.

4. **Testing Phase**

   It includes testing the machine with new documents. It includes 10-fold document testing and 5-fold document testing , which includes 9 document training and 1 document testing phase. This phase will be completed by Adish Jain & Suneel Kumar.

5. **Documentation Phase**

   It includes documenting the entire project. It includes proper detailed description about every concept used in the project. This phase will be covered by Adhish Jain & Suneel Kumar.

Conclusion

Text classification is expected to play an important role in future search services or in the text categorization. It is an essential matter to focus on the main subjects and significant content. It is becoming important to have the computational methods that automatically classifies available text documents to obtain the categorized information or groups with greater speed and fidelity for the content matter of the texts.