

**TRIBHUWAN UNIVERSITY**

**INSTITUTE OF ENGINEERING**

**PURWANCHAL CAMPUS**

**A MAJOR PROJECT PROSPOSAL ON PLAGIARISM DETECTION USING SUPPORT VECTOR MACHINE (SVM) AND NLP**

**Submitted By:**

ANISH BANJARA (PUR074BCT004)

KEWAL SINGH MEYANGBO (PUR074BCT014)

PANKAJ KUMAR SONI (PUR074BCT019)

RAJ KUMAR SAH (PUR074BCT026)

**Submitted To:**

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

PURWANCHAL CAMPUS

DHARAN

1st December, 2021

# ACKNOWLEDGEMENT:

In preparing this proposal, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. First of all, we would like to take this opportunity to express our sincere gratitude to the **Department of Electronics and Computer Engineering, Purwanchal Campus**for providing us the opportunity to undertake this project by including it as a part of curriculum in B.E. Computer Engineering and giving us a good guideline for proposal.

We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in writing this proposal. We would also like to express our gratitude to our teachers and colleagues whose support and suggestions have been an invaluable contribution. Their continuous encouragement and valuable comments are truly appreciable. Many people, especially our classmates and team members itself, have made valuable comment suggestions on this proposal which gave us an inspiration to improve our project proposal. We thank all the people for their help directly and indirectly to complete our proposal.

# ABSTRACT:

Plagiarism is incredibly challenging problem in the academics. Students copy assignments, assessments, reports, thesis, etc. and getting credit for what they have not done. In this report we proposed a plagiarism detection system using machine learning approach. We used Support Vector Machine (SVM) as learning algorithms. Learning features used in the method are words similarity, fingerprints similarity, latent semantic analysis (LSA) similarity, and word pair. Those features are adapted from some state-of-the-art methods in detailed analysis of a plagiarism detection system.

# **TABLE OF CONTENTS:**

[ACKNOWLEDGEMENT: II](#_Toc89779089)

[ABSTRACT: III](#_Toc89779090)

[TABLE OF CONTENTS: IV](#_Toc89779091)

[INTRODUCTION: 1](#_Toc89779092)

[PROBLEM STATEMENT: 1](#_Toc89779093)

[OBJECTIVE: 2](#_Toc89779094)

[LITERATURE REVIEW: 2](#_Toc89779095)

[BACKGROUND: 2](#_Toc89779096)

[PLAGIARISM DETECTION PROCESS: 2](#_Toc89779097)

[PREPROCESS: 2](#_Toc89779098)

[FEATURE EXTRACTION: 3](#_Toc89779099)

[MODEL LEARNING: 3](#_Toc89779100)

[CLASSIFICATION: 4](#_Toc89779101)

[METHODOLOGY: 4](#_Toc89779102)

[BLOCK DIAGRAM: 4](#_Toc89779103)

[WORKING PRINCIPLE: 4](#_Toc89779104)

[FLOWCHART: 6](#_Toc89779105)

[SYSTEM REQUIREMENT: 6](#_Toc89779106)

[HARDWARE REQUIREMENT: 6](#_Toc89779107)

[SOFTWARE REQUIREMENT: 6](#_Toc89779108)

[PROJECT GANTT CHART: 7](#_Toc89779109)

[EXPECTED OUTPUT 7](#_Toc89779110)

[REFERENCES: 8](#_Toc89779111)

# INTRODUCTION:

With increasing volume of digital data in internet, plagiarism and copyright infringement is also increasing along with it. General availability of the internet and easy access to textual information enhances the need for automated plagiarism detection. Plagiarism means to present others work, ideas and words as their own or without giving acknowledgments of the source. Plagiarism is not illegal like copyright infringement, but it is immoral and against the norms of academia. It is of two types: 1) Textual plagiarism and 2) Source code plagiarism.

Textual plagiarism is copy pasting, paraphrasing, metaphor, idea, self-recycled, illegitimate and retweet plagiarism. Source code plagiarism includes reordering structure, manipulation and language switching plagiarism. To maintain academic integrity and intellectual property, we need automatic plagiarism detection tool since manual detection is no longer feasible. For this, computational methods have extremely helpful for text reuse, authorship, and direction identification. Different feature extraction tools, natural language processing techniques and machine learning approaches are helpful for successful classification and detection of plagiarism. For feature extraction, various selection approaches like threshold frequency selection method and multiple information ranking selection approaches can be implemented. LSA similarity value can be calculated on selected word pair features using cosine similarity function between sentences-context vectors of the two sentences in each data. We propose SVM (Support Vector Machine) learning algorithm for learning model because it works on smaller datasets, but on complex ones and it can be much stronger and powerful in building machine learning models. This algorithm analyzes the data, looks for a pattern, and then creates hyper plane separator to divide the data based on each class. Learning algorithm performs classification of input data whether it is plagiarized or not by using the learning model generated by modeling subsystem.

The biggest challenge in the plagiarism detection field is that most approaches are inadequate at detecting texts with substantial semantic and syntactic changes.

## PROBLEM STATEMENT:

Finding plagiarized parts of a thesis is terribly slow work for teachers and management. Even with a limited number of texts it relies on the teacher's ability to read and remember every submission. As the process of finding plagiarized parts in thesis are based on the teacher's ability to remember all that he or she has read, the results may be incomplete. Some clear cases of copy and paste may easily be overlooked. And since the workload cannot be shared between multiple assistants. Thus, we are introducing automated system for checking plagiarism in thesis.

## OBJECTIVE:

To detect plagiarism in the thesis paper submitted using natural language processing and SVM algorithm.

# LITERATURE REVIEW:

## BACKGROUND:

One of the earliest methods of plagiarism detection was introduced by Bird (1927), which investigated the application of statistical methods in detecting plagiarism of multiple-choice answers. Later methods developed through the 1960s were focused on detecting plagiarism in multiple-choice tests. Early plagiarism detection systems for written texts started to appear around the 1990s. These tools used statistical methods to calculate similarity between texts, and most tools focused on written-text plagiarism while some focused only on computer source code plagiarism.

## PLAGIARISM DETECTION PROCESS:

### PREPROCESS:

First thing conducted is the preprocessing of training data. The preprocessing takes the value of the sentences detected, the source-candidate sentence, and the label of each data. In this process, stop word removal and stemming are also performed.

Pre-processing is a major step in the process in which all the assignments are converted into an appropriate format. All these assignments collected must be in the same format. Other preprocessing task such as data cleaning, data integration, data reduction and data transformation and data discretization are performed if necessary to optimize data quality

### FEATURE EXTRACTION:

Feature selection is performed to get the number of features. Feature selection is performed to get the number of features that are not too big but still be able to represent the information required. Several feature selection methods are threshold frequency selection and mutual information ranked selection.

Frequency threshold is the minimum frequency limit possessed by a feature to be able to qualify for selection. Mutual information is a dependency value of a feature in deciding the value of the class. It will rank the features based on mutual information values. Top n-features are selected from the resulting ranking.

Then, the generation of word pairs features is performed. Word pairs are chosen to be the features in creating model and performing classification because word pairs that exist between two sentences can determine whether they are plagiarism or not. Word pairs features are expected to give information of different words that have similar context.

The extraction of each feature value is executed for each of the existing data in the training data. Value of the word pair feature is the number of occurrences of the word pair in each data. Then LSA similarity value can be calculated using cosine similarity function between sentences-context vectors of the two sentences in each data.  LSA similarity is chosen because it provides information in the form of conceptual similarity of context (semantics) between the two sentences. Conceptual similarity of context has a great influence in determining plagiarism between sentences. We calculate the cosine similarity between to-be-detected sentence semantic vector and source sentence semantic vector. The result is 1, so the value of this attribute is 1.

### MODEL LEARNING:

A learning model can be built from the features obtained by the previous process. Support Vector Machine (SVM) is learning algorithm it works on smaller datasets, but on complex ones, it can be much stronger and powerful in building machine learning models. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features) with the value of each feature being the value of a particular coordinate. SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. This algorithm analyzes the data, looks for a pattern, and then creates hyper plane separator to divide the data based on each class.

### CLASSIFICATION:

Text classification should be performed to extract and separate the parts of a sentence into alternative words. With the help of this key words from a sentence can be found. The classification of the inputs is performed. Learning algorithm performs classification using the learning model generated by modeling subsystem. The result of the classification is the decision whether the input data is plagiarism or not.

# METHODOLOGY:

## BLOCK DIAGRAM:

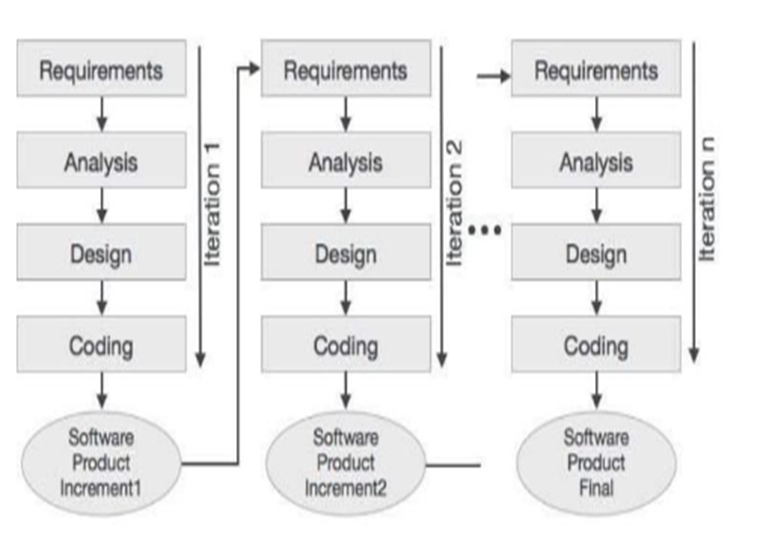
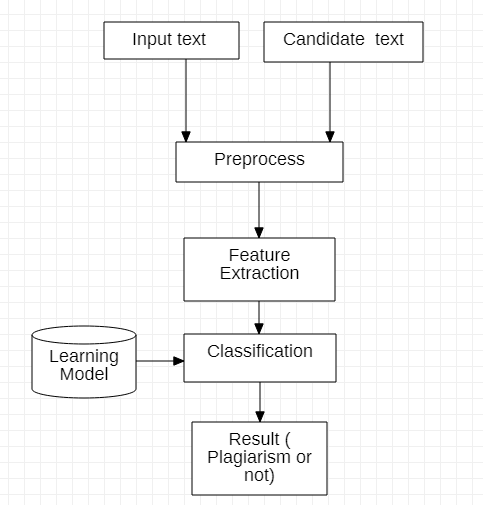


Fig: Iterative and Incremental Development Model

## WORKING PRINCIPLE:

A software development methodology refers to the framework that is used to structure, plan and control the process of developing a new system. Nowadays there are several software development methodologies each one having its own strengths and weaknesses. Developing a new software requires selecting the most appropriate methodology based on the nature of the new system. Even though there is no ideal software process, the nature of the project is such that Iterative and Incremental Development might be considered as the most suitable methodology over others. Iterative and Incremental Development is an approach in which the overall life cycle of the development process is divided into a number of iterations into sequence. Each iteration is considered as a stand-alone project composed of activities. Thus, the life cycle phases such as requirement analysis, design and implementation are carried out at each iteration in turn. As soon as an iteration is 13 fully completed, the results are integrated into the overall system which grows incrementally. Despite the fact that the waterfall model is widely used for software development, for the purpose of this project it cannot be considered as the ideal approach. It is relatively easy to define and well documented at each stage. In order to move to the next stage, it requires the stage to be fully implemented. However, this is not a safe approach to a project, since it is almost certain that at a point during the software cycle, it will be required to go back to make adjustments and refinements. Conversely, in the Iterative and Incremental Development approach smaller incremental chunks are used to build up the new system. Each chunk is assigned some activities and a limited amount of time to be completed. Once those activities are fully completed this chunk is added to the system which is compared to the earlier specifications. This cycle continues until the system is fully developed. In this case, errors during the iterations, development process will not continue until there is a fully functional outcome. In the worst-case scenario by following the Iterative and Incremental Development methodology the minimum requirements will be successfully delivered to the user. Based on the advantages described above we have chosen the Iterative and Incremental Development approach. That is, the android application to be developed will be broken-down into features, with each one developed in turn. Once a feature is fully functional, a prototype of the application has to be distributed to the students in order to evaluate it. Finally, if there are no errors with the specific feature, the new part of the application will be integrated in the existing code. This process will continue until the development of the entire application.

## FLOWCHART:



# SYSTEM REQUIREMENT:

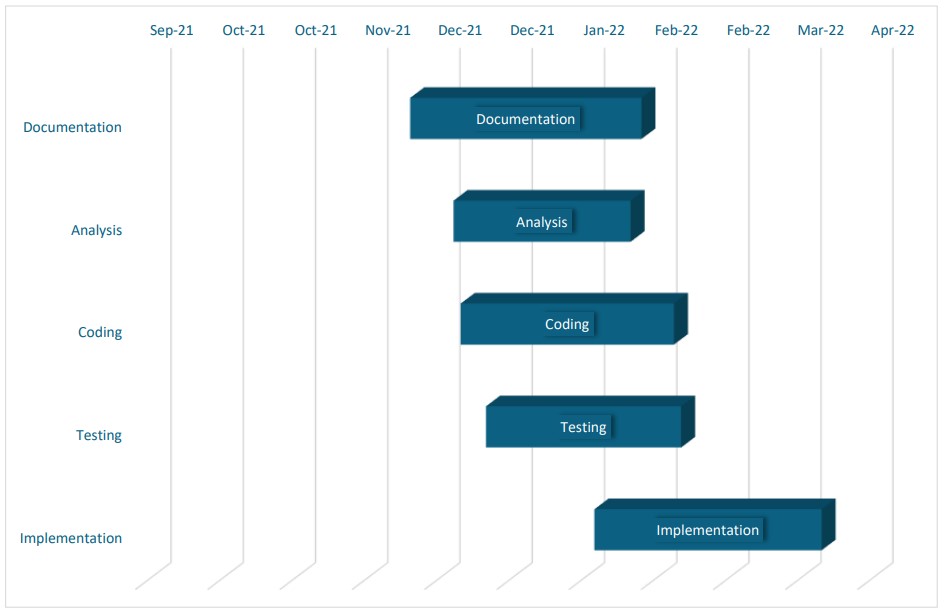
## HARDWARE REQUIREMENT:

* Processor: Intel core i5
* Hard Disk Drive: 100 GB and above
* RAM: 8 GB recommended
* Monitor: LED, LCD or CRT
* Mouse: USB optical
* Keyboard: USB

## SOFTWARE REQUIREMENT:

* Language: Python
* Operating System: Windows10

# PROJECT GANTT CHART:



# EXPECTED OUTPUT

1. Literal plagiarism sentence pairs should be detected
2. Partial literal plagiarism sentence pairs should be detected
3. Paraphrased plagiarism sentence pairs should be detected
4. Changed structure plagiarism sentence pairs should be detected
5. Translated plagiarism sentence pairs should be detected

# REFERENCES:

[1] Zakiy Firdaus Alfikri, Ayu Purwarianti, “Detailed Analysis of Extrinsic Plagiarism Detection System Using Machine Learning Approach (Naive Bayes and SVM)”, *TELKOMNIKA Indonesian Journal of Electrical Engineering*, Vol. 12, No. 11, pp. 7794 ~ 7804, Nov 2014.

[2] Hussain A Chowdhury and Dhruba K Bhattacharyya, “Plagiarism: Taxonomy, Tools and Detection Techniques”, Dept. of CSE, Tezpur University.

[3] Taresh Bokade, Tejas Chede, Dhanashri Kuwar, Prof. Rasika Shintre, “Online Assignment Plagiarism Checking Using Data Mining and NLP”, *International Research Journal of Engineering and Technology (IRJET),* Vol 08, No. 01, pp. 0056-0072, Jan 2021.

[4] Man Yan Miranda Chong, “A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques”, 2013.

[5] Wikipedia