Similarity Analysis & Plagiarism Detection for Punjabi Documents using NLP

Akanksha Chuchra, Iknoor Singh, Vijay Paul Singh, Naveen Aggarwal
DIC, UIET, Panjab University, Chandigarh

Introduction

Plagiarism can be defined as the unauthorized use or reproduction of the work originally done by someone else and thereby presenting it as one's own original work. The copied work may be from some researcher's research work or it may be a copy or a partial copy of some literary articles.
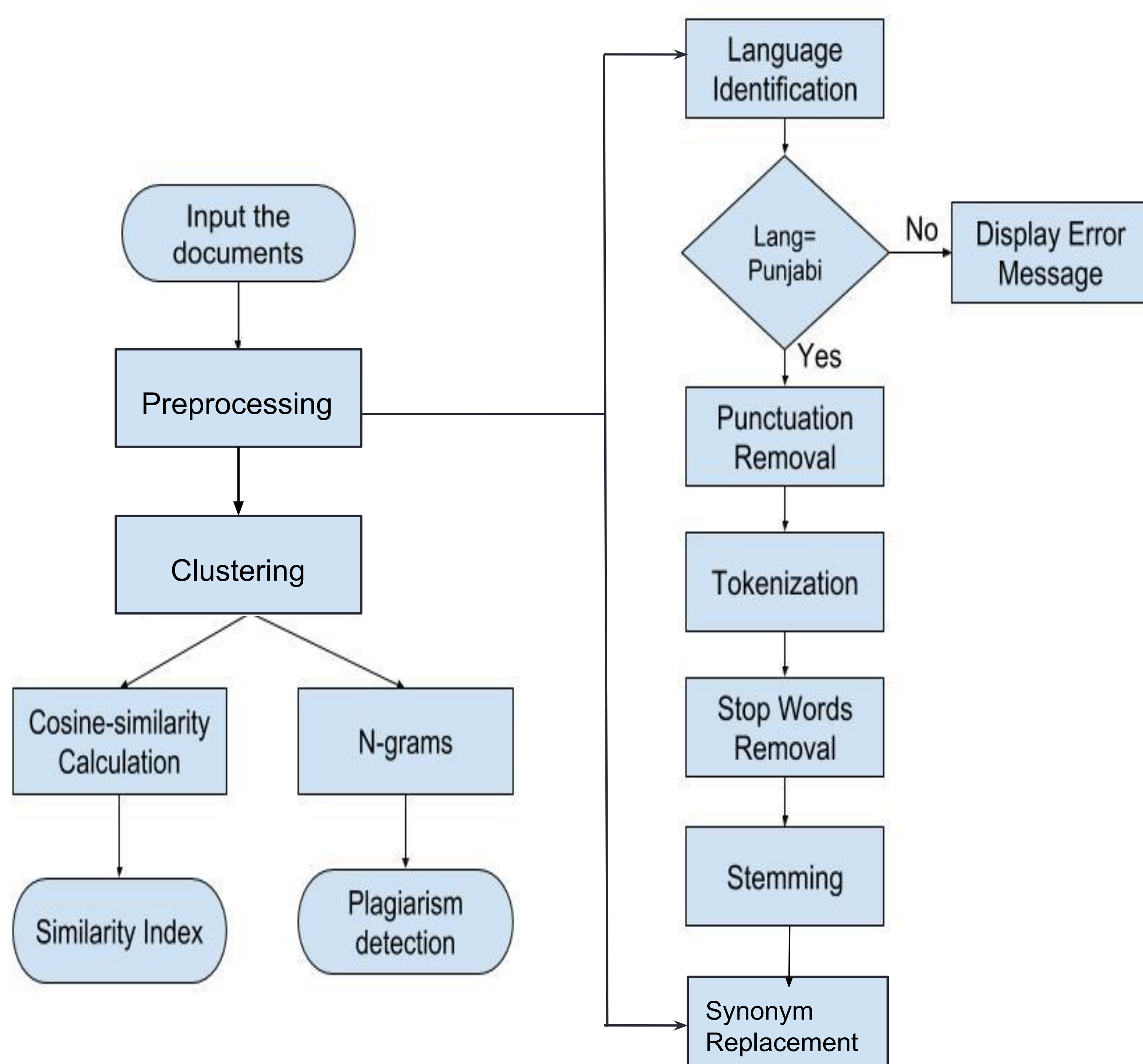
Objective

The main aim of this system is to develop a similarity and plagiarism detection system for Punjabi. It aims to calculate the similarity and plagiarism in a document containing plain Punjabi text in Gurmukhi script.
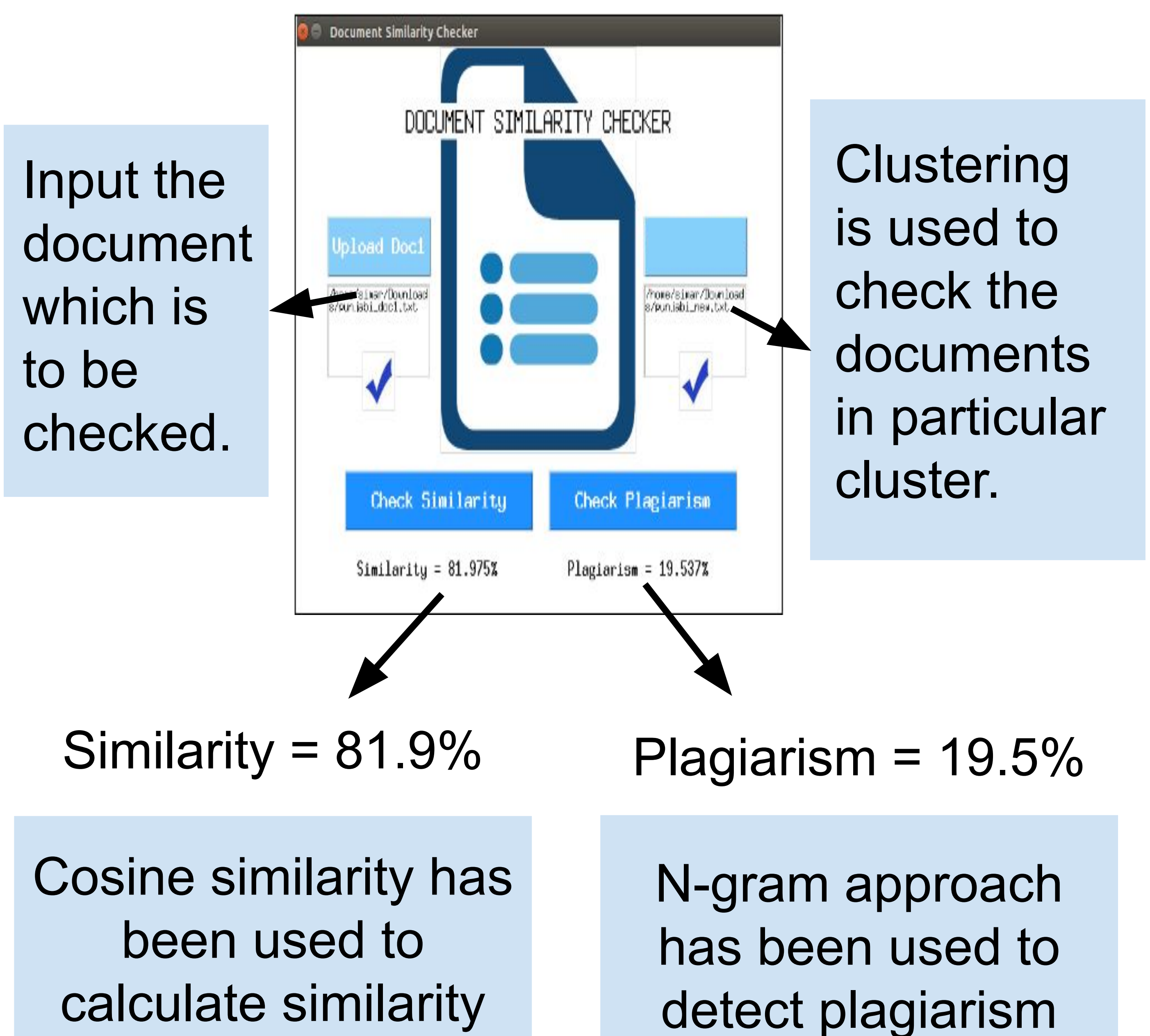
Significance

➤ Similarity detection is important to check whether two documents are related to the same topic or not. This finds applications in clustering of documents based on the topic to which they belong.

➤ In order to prevent copying of research works or projects etc. and maintaining information uniqueness, we need to detect plagiarism.

➤ Plagiarism detectors allow the users to analyze the text and find out the percentage of non-unique fragments in a given document.

➤ Many free online tools are available to detect plagiarism for English but there aren't any for Punjabi.

Flowchart

GUI Design

Easy to understand and user friendly GUI has been developed in order to demonstrate the working of the project.

Input the document which is to be checked.

Clustering is used to check the documents in particular cluster.

Similarity = 81.9%

Plagiarism = 19.5%

Cosine similarity has been used to calculate similarity

N-gram approach has been used to detect plagiarism

Results

We were able to find Similarity and plagiarism between Punjabi documents written in Gurumukhi script. The system displays the similarity percentage as well as the percentage of plagiarized content in the documents.

Future Scope

➤ It has various applications including auto-categorization of documents, keyword extraction etc.

➤ This project can also be made to work on the local documents as well as those on web.

➤ The plagiarized content can be highlighted and the source document can be found out.

Contact Details

For further information contact us at:
DIC, Uiet Block 1, Sector 25, Chandigarh, 160015
- akansha.chuchra96@gmail.com
- iknoorpro@gmail.com
- vpaul9678@gmail.com
- navagg@gmail.com
- dicuietpu@gmail.com

Links - http://www.uiet.puchd.ac.in/dic/