ABSTRACT

Plagiarism refers to the un-authorized use of the original work done by someone else and presenting as one’s own work, without giving any acknowledgements to the original author. The evil of plagiarism has been around since the human beings started documenting their work. Plagiarism in the literary and research articles has always been a discouraging factor for the genuine authors and researchers. The special tools for plagiarism detection on the other end came as a big relief to the researchers. These tools make use of numerous Natural Language Processing (NLP) techniques to detect the plagiarism in research publications, books, magazines or even software programs. These specialized software tools are available in the market as proprietary, commercial solutions or even in the open source form. Many such tools work efficiently with English language documents, but fail to give satisfactory results for the documents written in linguistically complex languages like Hindi, Punjabi and other regional languages of India.

In this thesis, an attempt is made to develop a software tool that uses the Natural Language Processing techniques (NLP) for detecting plagiarism in Punjabi text documents. The software is built using the prevalent open source technologies – PHP and MySQL and is by choice designed to be web based for the universal access. Besides string matching for plagiarism detection, the software also detects word switching, synonym replacement and re-phrasing of the Punjabi text for calculating the final plagiarism score. The software generates the reports which are interactive, user friendly and easy to understand.

The first chapter in this thesis provides an introduction to Plagiarism, describing its most common forms and the reasons behind the plagiarism being practiced by the researchers. The chapter also highlights the case studies related to the plagiarism instances and the mechanisms practiced by the universities and research journals for dealing with it. Towards the end of this chapter, some existing plagiarism detection tools available for English text are discussed. As the existing tools do not work with the Punjabi text documents, there is a need for a special software for plagiarism detection in Punjabi Text Documents. The
objectives, challenges and scope of the Plagiarism Detection Software for Punjabi Text Documents have also been discussed in this chapter.

The Second Chapter of the thesis gives an overview of the work already done in the area of Similarity Calculations and Plagiarism Detection using natural language processing techniques. This chapter also provides an in-depth review of the existing Plagiarism Detection Software available at the time of writing this thesis.

The third chapter discusses the system design and the proposed architecture of the “Uprala – The Plagiarism Detection Software for Punjabi Language”. The chapter discusses the methodology followed along with the details on each and every component of the system. The relevant comparison of the various similarity calculation techniques is also discussed in this chapter.

The fourth chapter enlists the outcomes of the various modules of the software developed. Various metrics used for the evaluation of the results at each step along with their resultant values are also discussed in this chapter. In the end of this chapter, a comparison with other existing Plagiarism Detection tools is presented.

The fifth chapter concludes the work presented in the thesis along with the final words mentioning the future scope of work that can be used for extending the current system as well as improving the performance of the system.