Chapter 21
Machine-Learning-Based External Plagiarism Detecting Methodology From Monolingual Documents:
A Comparative Study

Saugata Bose
University of Liberal Arts Bangladesh, Bangladesh

Ritambhra Korpal
Savitribai Phule Pune University, India

ABSTRACT

In this chapter, an initiative is proposed where natural language processing (NLP) techniques and supervised machine learning algorithms have been combined to detect external plagiarism. The major emphasis is on to construct a framework to detect plagiarism from monolingual texts by implementing n-gram frequency comparison approach. The framework is based on 120 characteristics which have been extracted during pre-processing steps using simple NLP approach. Afterward, filter metrics has been applied to select most relevant features and supervised classification learning algorithm has been used later to classify the documents in four levels of plagiarism. Then, confusion matrix was built to estimate the false positives and false negatives. Finally, the authors have shown C4.5 decision tree-based classifier's suitability on calculating accuracy over naive Bayes. The framework achieved 89% accuracy with low false positive and false negative rate and it shows higher precision and recall value comparing to passage similarities method, sentence similarity method, and search space reduction method.

INTRODUCTION

In this present Internet era, academics, as well as researchers are deeply concerned with plagiarism issue.

Plagiarism refers to copying from someone else's document without providing proper acknowledgements (Cosma & Joy, 2008). According to the Merriam-Webster online dictionary, plagiarism means

DOI: 10.4018/978-1-5225-8057-7.ch021

stealing and passing off (the ideas or words of another) as one's own, using (another's production) without crediting the source, committing literary theft or presenting as new and original an idea or product derived from an existing source.

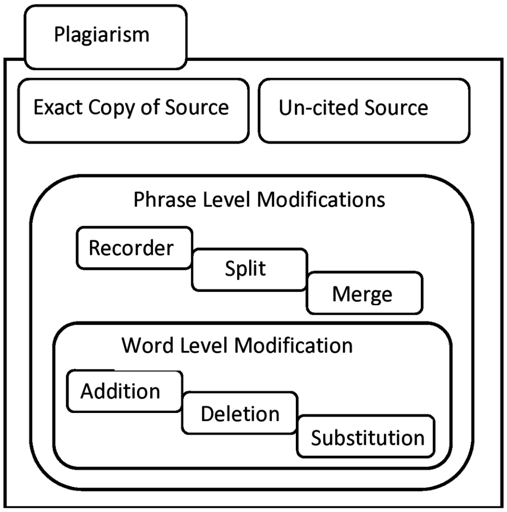
Plagiarism can be of many forms as shown in Figure 1. Either it can be an exact copy of the source document or some form of modified (addition, deletion, substitution in word level or in phrase level) version of source document, without properly acknowledging the source.

The severity of this copying can be understood by a finding (McCabe, 2002) where it is identified that 10% of American college students have been involved in partial copying their assignments whereas in high schools 52% of students have been involved in some form of plagiarism. To counter this problem, a study was conducted on why students are involved in plagiarism and it was found that 'means and opportunity' are their motivation (Bennett, 2005). Manually detecting plagiarized document is a humongous task, as well as a drain of academicians' precious time. As a result, academicians look for tools which can detect plagiarisms automatically. In recent years, many commercial detection tools have been developed such as Turnitin (iParadigms, 2010) and CopyCatch (CFL software, 2010) or MOSS (Aiken, 1994) for detecting plagiarism in computer programming source code (Chong, Specia, & Mitkov, 2010). In this paper, we concentrate on checking plagiarism in written text documents because there is a 'challenge of distinguishing true cases of plagiarism from mere coincidental similarity of wording' (Buruiana, Scoica, Rebedea, & Rughinis, 2013).

For developing a detection tool, one cannot simply rely on 'exact-word or phrase matching' (Reddy, 2013). Paraphrasing or rearranging words of a sentence makes the task even more complex. Furthermore, academicians categorize plagiarism in two sections: external plagiarism where suspicious documents are compared with original ones and intrinsicplagiarism where one tries to find plagiarized passages within a document without accessing potential original documents.

As shown in Figure 2, the plagiarism detection methods are classified in three categories: fingerprinting, term occurrences and style analysis (Eissen, Stein, & Kulig, 2006). Among these, "term occurrence" is the familiar style, developers follow. According to Reddy, 'Plagiarism detection is a process of find-

Figure 1. Forms of plagiarism
Reddy, 2013.


Plagiarism
Exact Copy of Source
Un-cited Source
Phrase Level Modifications
Recorder
Split
Merge
Word Level Modification
Addition
Deletion
Substitution