Enhancing automatic plagiarism detection: Using Doc2vec model

Imene Setha
Research center on scientific and technical information
CERIST
Algiers, Algeria
sethaimene1@.com

Hassina Aliane
Research center on scientific and technical information
CERIST
Algiers, Algeria
ahassina4@gmail.com

***Abstract*—Academic institutions define plagiarism as an act of cheating and stealing other's ideas to pass as their own. Therefore, a huge interest is conducted into plagiarism detection field using multiple techniques. In this article, we propose a method to automatically detect different types of plagiarism from two languages. This method is based on sentence modelling to try to extract plagiarized parts from documents using Doc2Vec model which predicts semantic similarity between documents and phrases.We use the PAN corpus for English plagiarism detection and AraPlagDet for Arabic. Both PAN and AraPlagDet corporas provide a set of suspicious documents that are manually and artificially plagiarized along with their sources.**

***Keywords*—Paragraph vector, Doc2Vec, Word2Vec, LSA, PAN corpus, AraPlagDet corpus.**

I. Introduction

Along with the available large amounts of data on the web today, plagiarism detection becomes of a great importance especially for academic institutions. Indeed, cheating and stealing information without properly citing the right authors, implies directly the educational process for both students and supervisors. Developing tools and techniques for preventing such acts hence becomes quite a big concern for NLP researchers. Plagiarism may take multiple forms such as paraphrasing, direct copying, stealing thoughts, cloning codes in platforms etc. . . and researchers face several challenges. Nevertheless, many powerful tools and systems are available nowadays especially for the English language(Enago Plagiarism Checker,Duplichecker, Easybib. . . etc) using multiple methods and techniques such as linguistic, statistical and machine learning. In this work, we propose a new approach to automatic plagiarism detection based on recent deep learning models. Our main goal is to find new ways for plagiarism detection based on understanding sentence context, and document checking especially for Arabic language. Therefore we apply paragraph vector model on two corporas to create both sentence and document representation vectors to predict similarity between plagiarized parts. The first corpus is for English language (PAN-2009) and the second is for Arabic(Ara-Plag-Det). This paper is organized as follows: section 2 is dedicated to related works on recent semantic plagiarism detection research. In section 3, we describe our approach and introduce Doc2vec model. Then Experimentations and results are presented for both English and Arabic languages, along with our observations in a discussion section. The last section provides a conclusion and future work.

II. Related Work

Since plagiarism detection is a highly active academic research field, many approaches and methods have been used for this task. Plagiarism detection approaches fall into two major kinds: Intrinsic and extrinsic detection. For each kind of approach, several methods are used as illustrated in figure 1 which is inspired from [1]:

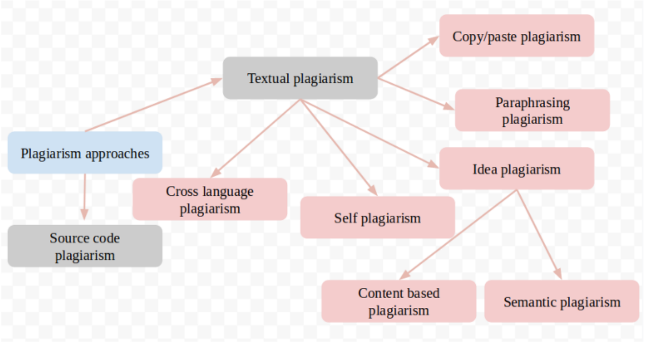

Fig. 1. Plagiarism detection types.

Semantic-based plagiarism detection generally presents terms as numerical vectors, to find semantic relatedness between them in the whole corpus using multiple techniques and measures. LSA is the most known method for plagiarism detection in many languages [2] [3] [4]. It transforms terms and documents into a huge matrix, decomposes it for searching in the corpus and then reduces it using SVD function.

In [5] LSA is used as term-document representation for Indonesian language, to handle intelligence plagiarized terms from a corpus. In comparison with SVM model, experiments show that LSA outperforms SVM learning model for the task. [6] integrated LSA with stylometry technique for intrinsic plagiarism detection. LSA was used to build the document-term matrix while stylometry was used for intrinsic approximation of human writing style. Machine learning models are also broadly used [7] [8].

Logistic regression is used in [9] to predict source code

2022 International Conference on Advanced Aspects of Software Engineering (ICAASE) | 978-1-6654-9202-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAASE56196.2022.9931542