IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Resolving Various Citation Issues in Publication

[1]Aakanksha Deshpande, [2]Dhananjay Epili, [3]Pratik Gadge, [4]Prof. Y.I Jinesh Melvin

[1]B.E Student, [2] B.E Student, [3] B.E Student, [4]Professor

[1]Information Technology, [2]Information Technology, [3]Information Technology, [4]Information Technology

[1]Pillai College of Engineering, Navi Mumbai, India, [2]Pillai College of Engineering, Navi Mumbai, India

[3]Pillai College of Engineering, Navi Mumbai, India, [4]Pillai College of Engineering, Navi Mumbai, India

Abstract: In the present age access, the world is growing with lots of easily accessible information available on almost every subject matter. The use of the freely available internet resource is causing a rise in easy copy and paste culture. Thus many people indudes plagiarism directly or indirectly in their paper due to improper citation. Citation simply means to give credit to the owner of the information. Research papers that are not properly cited are considered to be plagiarised. Plagiarism is considered as a serious offence and can have consequences. Thus the main aim of this project is to detect citation types of a research paper that includes formats such as IEEE, Elsiever , Springer , and many more. Also to convert the different citation formats that are detected into a single format that will be use for publication. Users can also detect the misplaced citations in the paper. This technique will reduce or can completely eliminate plagiarism from the research paper. The system makes used on Natural Language Processing (NLP) and machine learning algorithms and artificial intelligence techniques in order to make it more efficient.

Index Terms **- Citation, Plagiarism ,Natural Language Processing (NLP);**

I. INTRODUCTION

We are in the era of discoveries and innovations. Every day there is a constant growth in technology than the previous day. This is the result of the tremendous amount of research that is taking place from the last decades. In order for this knowledge to pass on to next generations we need to document the research properly. Hence we can say that the various research papers that are published play a huge role in our evolution and development. Research papers are published into various journals that are available worldwide to make sure the information is reached to every corner of the world. When working on a publication , we often do not rectify or neglect some things in a hustle of completion of the work. Citation is one of the things that are often overlooked . Having an improper citation can induce plagiarism in our publication. Simply defined plagiarism is to include someone else's content without giving him credit for his work.

Now-a-days detection of plagiarism is mostly done by various online available Plagiarism Detection tools. Unintentional plagiarism of even a sentence or two may have serious consequences. For students, plagiarism often means a failing grade, academic probation, or worse. While for a publication it can have very serious consequences. Though there are various tools to detect plagiarism they often have drawbacks that make them not very useful for Publications. Thus our project mainly aims to overcome the drawbacks of these existing systems and to reduce plagiarism.

II. LITERATURE SURVEY

Dr. Sudhir S. Patil and Dr.Hemant Yeole et.al [1] discuss about the plagiarism, Plagiarism Checker and Plagiarism detection tools from check research work through software like, Turnitin,Ithenticate, Plagiarism Checker, Viper, Duplichecker, Copyleaks, Paperrater, Plagium,Plagiarisma, Plagscan etc. The present study has used different plagiarism detection tools and checkers. The paper also describes the advantages and disadvantages of these tools.

In this paper by Vani K and Deepa Gupta et.al [2] a study on plagiarism is done with the focus on extrinsic text plagiarism detection, which is a fast emerging research area in this domain. The different extrinsic detection techniques and the methodologies involved are reviewed based on the current state of art. Further an overview of some of the available detection software tools, their features and detection efficiency is discussed with some of the output demos. The paper also throws light on the popular PAN competition, in the plagiarism domain and the major tasks involved in it. Further it attempts to identify the problems existing in available tools and the research gaps where immense explorations can be done.

There are a number of referencing styles that are used throughout the world, some of which are more common than the others. Author Mohsin Hassan Alvi et.al [3] includes the names of all the styles with a detailed description of a few more common and widely used styles. The paper elaborates the major differences that exist among various types of referencing styles.

In this paper, we quantify the contribution of this additional information to the reference extraction performance by an improved preprocessing using the information contained in PDF files and retraining sequence classifiers on an enhanced feature set. The authors Roman Kern and Stefan Klampfl et.al [4] found that the detection of columns, reading order, and decorations, as well as the inclusion of layout information improves the retrieval of reference strings, and the classification of reference token types can be improved using additional font information. These results emphasize the importance of layout and formatting information for the extraction of meta-data from scientific articles.

This paper by Sergey Parinov et.al [5] presents a method to process a content of research papers in binary PDF format at a server side that gives research information systems new features of citation content analysis. This method efficiently generates JSON versions of PDF documents that allows an easier recognition of papers' references, in-text citations, citation context, etc. As a result, one can parse an extended set of citation data, including a location of citations in a research paper's structure, frequency of mentioning for the same references, style of reference mentioning and so on.

Mausumi Sahu et.al [6] explains about types of plagiarism there are like text matching, copy paste, grammar based method etc. Proposes a new method implemented in a program ,where we utilise a text set to identify the copied part by comparing with some existing multiple files. Here we put the concept of a machine learning language i.e k-NN. It helps us to identify whether a paper is plagiarized or not. The k-Nearest Neighbor Algorithm is one of the simplest machine learning algorithms that is suitable for pattern recognition. The author has designed a process using a machine learning method i.e k-NN which improves the performance. Comparing all methods in this area, it can be concluded that the k-nearest neighbour method is very useful in pattern recognition as well as to find copied dataset to detect plagiarism. This method provides more accuracy and efficiency to detect plagiarism.

III. PROPOSED WORK

To detect different formats of citation in publication requires very fast, accurate and efficient machine learning and artificial intelligence algorithms . Preprocessing of documents include techniques such as sentence segmentation, tokenization, stop word removal, punctuation removal, lowercasing etc. Natural language processing (NLP) techniques mainly stemming and lemmatization are also employed in this stage. The various deep NLP techniques such as Part of Speech (POS tagging), Chunking, Semantic Role labelling (SRL), named Entity recognition (NER) and various other NLP and artificial intelligence (AI) techniques have been employed for improving the detection efficiency.

3.1 System Architecture:

The system architecture is given in Figure 1. Each block is described in this Section.

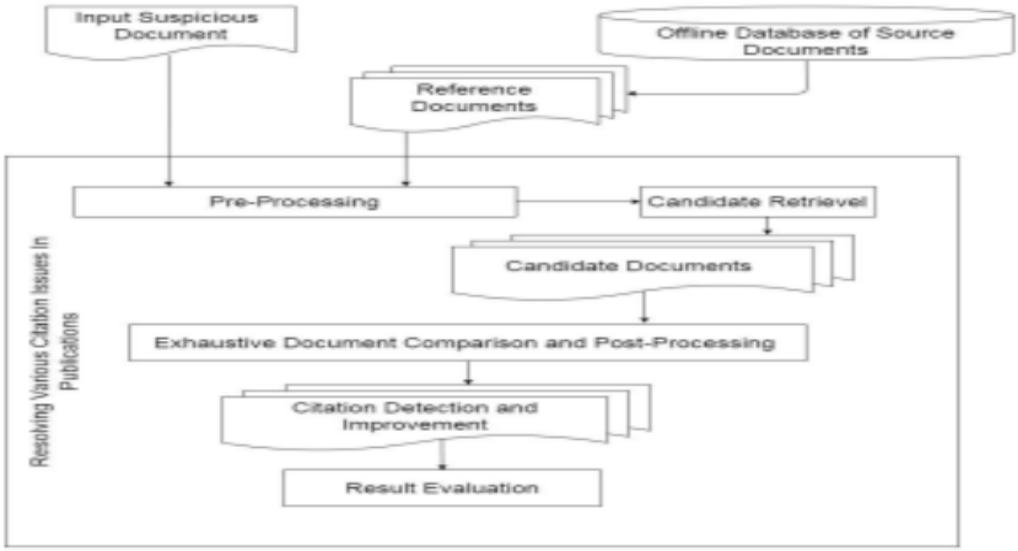

Fig 1. Proposed System Architecture

A. Input suspicious Document : This is the first part of the system. The system requires a user input of the Document or Publication that needs to be checked for citation errors. The user can insert the document as a PDF or as text. The Document are subjected to preprocessing before they are compared with the Reference Documents that are used to analyse the misplaced citations

B. Reference Documents :The second part of the system includes an offline database of the reference documents that contains documents used for making the research paper. These documents are also subjected to some preprocessing . The offline database is scanned to identify the documents that are near duplicates or have a certain amount of matching with the suspicious document inserted by the user.

C. Preprocessing :After the input document that is provided by the user and also the reference documents that are collected from the offline database , these documents at hand need to be preprocessed, where we remove any irrelevant information which makes the document easier to handle . These include techniques such as sentence segmentation, tokenization, stop word removal, punctuation removal, lowercasing, etc. Natural language processing (NLP) techniques mainly stemming and lemmatization are employed in this stage.Based on the models or techniques employed, pre- processing of the documents is done. Tokenization considers a document at word-level by dividing it into tokens. Further various NLP techniques are also employed for the effective document representation and handling. In pre-processing, the shallow procedures, viz., stemming or lemmatization are usually employed. Stemming is a heuristic process of removing the affixes from the words. Lemmatization produces the dictionary base forms of a word using vocabulary and morphology information. It is closely related to stemming but stemming operates only on a single word at a time while lemmatization operates on the full text. It can thus discriminate between words that have different meanings depending on part of speech

D. Candidate Retrieval :After pre-processing the next important stage is retrieving the near duplicate sources. To reduce this search space a document level comparison is done which retrieves the candidate sources for the given suspicious document at hand. In candidate retrieval tasks, the globally similar source documents with respect to a particular suspected document are retrieved. Thus each suspicious document is associated with a source set termed as candidate set. This process works similar to the information retrieval task in search engines, where the documents related to a particular query are retrieved.In case of offline database we have a hermetic system where each suspicious document is compared at a document level with the each of the sources in offline database to retrieve the source set associated with it. This document level comparison is done using the different methods of document retrieval and similarity analysis.Candidate retrieval stage plays an important role in deciding the overall efficiency of the system . If the candidate retrieval is not done, then each suspicious document has to be compared exhaustively with all the available sources which will be quite time consuming. Further there will be many sources which are completely unrelated to the suspicious document at hand. Thus a document level comparison is always appreciated before the actual in depth comparisons. Candidate retrieval task reduces the overall complexity , but at the same time implementing a well defined retrieval method is necessary. This is because any source document missed in this stage will not be accounted for in the further stages also. Thus retrieving all the related source candidates is essential while maintaining the accuracy.

E. Exhaustive Document Comparison and Post-Processing : Once the candidate documents are retrieved, each suspicious document is compared against its candidate set exhaustively. This is where the suspected plagiarized segments that have missing or improper citations and their corresponding source components are identified. In a detailed document comparison stage, each suspected document is compared against its source candidates using various methodologies and detection techniques.The comparisons can be on different levels including sentence level, N-gram level, word level and phrase levels. In this phase, deep NLP techniques such as Part of Speech (POS tagging), Chunking, Semantic Role labelling (SRL), named Entity recognition (NER) and various other NLP and artificial intelligence (AI) techniques has to employed for improving the detection efficiency. The source and suspicious components are compared using some similarity measures and fragments are selected. Once the fragments are obtained, post processing is done which mainly includes passage boundary detection phase. Here the deductions of source and suspicious passages are done based on certain boundary thresholds and some split-merge conditions. It is important that passages must be retrieved as a whole and not as pieces. Finally the system is evaluated on some standard data sets and performance is measured using standard metrics.

F. Citation Detection and Improvement :After the suspicious fragments are detected from the input document. We need to make the improvements in the original document. The output of the previous phase gives us details about the types of citations or the different formats like IEEE, Elsiever, etc that were detected and also about missing or misplaced citations from the paper by using various machine learning techniques. In this phase we will rectify this mistakes and convert into an improved document which the user can download . This will reduce or completely eliminate any plagiarism if present due to citation errors.

IV. REQUIREMENT ANALYSIS

The implementation detail is given in this section.

4.1 Software

Natural Language Processing (NLP) :

Natural Language Processing (NLP) is a process of manipulating or understanding the text or speech by any software or machine. An analogy is that humans interact and understand each other's views and respond with the appropriate answer. In NLP, this interaction, understanding, and response are made by a computer instead of a human.

Natural language toolkit (NLTK):

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

IV. RESULTS AND DISCUSSION

After performing various experiments the following observations are taken:

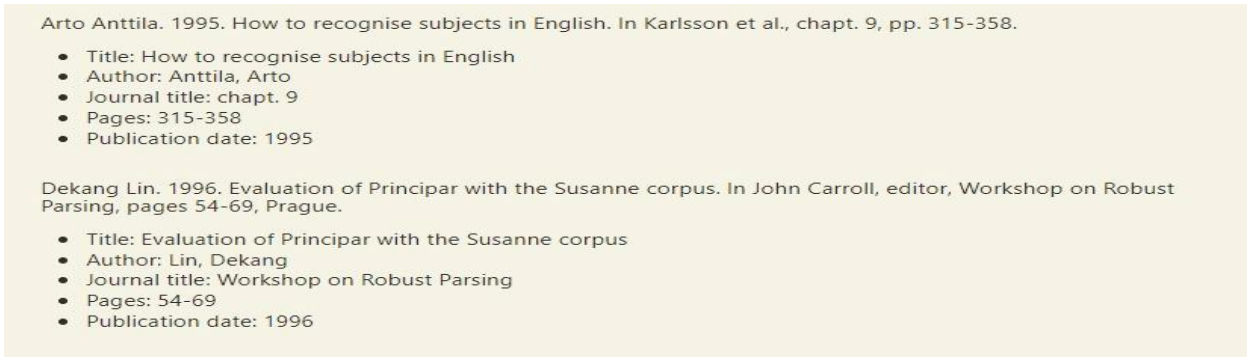

Fig 2. Citation Analysis for Publication

The figure above shows the breakdown of the external citation used by the user into different components like author name, journal name, paper title, publication date, etc. This data is also used to find out about the citation style and changing citation style from one to another (Ex. IEEE to APA, etc)

V. CONCLUSION AND FUTURE SCOPE

Citation and Plagiarism are important concepts for any publications. In this work, we have done analysis if the citations in a research paper and identify incorrect and misplaced citations using NLP. Future Scope of the work is very wide and robust and has a vast research potential , it consists of enabling deeper citation analysis, more accurate citation prediction, and increased knowledge discovery.

VI. ACKNOWLEDGMENT

We authored this paper on resolving citation issues, with the support of my external guide, Prof. Y.I Jinesh Melvin, Professor, Pillai College of Engineering. We are thankful to our Head Of Department Dr. Satishkumar Verma for providing us with the necessary resources and information. We would like to express our heartfelt gratitude to our Principal Dr. Sandeep Joshi for providing us students guidance and support.

REFERENCES

[1] Dr. Sudhir S. Patil and Dr.Hemant Yeole, Overview of Plagiarism Checkers and Plagiarism Detection Tools: A Study, 2019
[2] Vani K and Deepa Gupta ,Study on Extrinsic Text Plagiarism Detection Techniques and Tools, Journal of Engineering Science and Technology Review ,10 September 2016
[3] Mohsin Hassan Alvi , A Manual for Referencing Styles in Research , 2016
[4] Roman Kern and Stefan Klampfl, Extraction of References Using Layout and Formatting Information from Scientific Articles, D-Lib Magazine Volume 19, Number 9/10 , September/October 2013
[5] Sergey Parinov, Extraction and visualization of citation relationships and its attributes for papers in PDF, International Journal of Metadata Semantics and Ontologies , January 2017
[6] Mausumi Sahu ,Plagiarism Detection Using Artificial Intelligence Technique In Multiple Files, International Journal Of Scientific & Technology Research Volume 5, April 2016