iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6262

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

A Survey of Cross-Lingual Plagiarism Detection using Natural Language Processing

Dr. K. S. Aravind[1], Biradavolu Shanmukh[2], Guru Sri Charan[3], Chethan S Nellikoppad[4]
[1, 2, 3, 4]*Computer Science Department, Visvesvaraya Technological University*

Abstract: ***Plagiarism detection is gaining importance due to requirements for integrity in Research works especially when it comes to Cross-lingual plagiarism. In this paper, we have researched a new approach for Cross-Lingual sentence level plagiarism detection. The proposed approach works as initially the suspicious document is translated to a particular language and candidate document is retrieved for further processing. Pre-processing techniques will be selected using NB Classifier for better efficiency which will be used on document which includes POS tagging, Chunking, Stop-word removal, Lemmatization, Syntactic-semantic rule. Then Machine Learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes Classifier (NBC), Decision Tree (DT) are used to check plagiarism. The Evaluation and Analysis are performed using F-score, Accuracy, Precision and Recall. Our aim in this paper is to find the plagiarism and improve the accuracy of plagiarism detection between two languages by using concepts of natural language processing.***
Keywords: Plagiarism, Natural Language Processing, Cross-Lingual.

I. INTRODUCTION

Plagiarism is an unethical practice followed by person, where it is formally defined as copying another person's ideas, words or writing and pretending that there's own work. Plagiarism detection is a process of finding the plagiarized documents or plagiarized texts where a person doesn't give due credit to the original author of the document. These practices must be stopped by the authors and for this specific reason plagiarism detection is used. Plagiarism has two types, external plagiarism and intrinsic plagiarism on written texts. Here original documents are compared with the suspicious ones. Intrinsic plagiarism detection is used in finding documents which consists of plagiarized lines without acknowledging original documents. The main problem of the task is to compare with huge number of documents and dismiss the fraud. In this paper we are focused on the extrinsic plagiarism detection and how can it be reduced using our proposed algorithm. We see the plagiarism can be also be said when a person translates original author's document into another language and doesn't give its credit for the original author, this is known as cross lingual plagiarism. We mainly focus on Cross lingual plagiarism detection in this paper and how this problem can be solved suing NLP techniques. AI is used to deal with interaction between computers and humans using natural language which belongs to branch of NLP. These techniques help us to detect those documents or texts which are copied by the author. There are several researches made on cross-lingual plagiarism but not many research papers could give better accuracy and precision in terms of detecting the plagiarism.

In this paper, we report a new approach in detecting extrinsic cross-lingual plagiarism by using Natural language processing techniques such as POS Tagging, Chunking and Syntactic-semantic rule are used to distinguish the plagiarized texts from the original texts, Stop-word Removal and Lemmatization are the techniques used to filter the sentences and remove unwanted words and generalize them. Machine Learning algorithms are used in this paper to predict which documents belong to plagiarized category and which does not. Some of the ML algorithms used here are Support Vector Machine or SVM, Naïve Bayes classifier and Decision tree. The best fit algorithm will be considered for the further evaluation and analysis process, where we check the accuracy, precision, prediction score to say how good our model in terms of its result is. Since we are performing on cross-lingual we have to convert any one language to a particular language so that the following process can be performed. We also use another important technique called Candidate retrieval. Candidate retrieval is a process performed for each suspicious document before it computes pairwise document similarities to find potential sources of plagiarism. This is important as we enlarge the document database and thus use this database to perform pre-processing using NLP techniques.

Our objective in this paper is to find the plagiarism and improve the accuracy of plagiarism detection between two languages by using concepts of natural language processing. The structure of the article is organized as follows: Section 2 presents the Literature Survey of the research papers; Section 3 shows the Advantages and Disadvantages of the Existing algorithms and Section 4 is the future works.

II. LITERATURE SURVEY

In [1] aim of this research paper was to improve the accuracy of plagiarism detection by incorporating Natural Language Processing (NLP) techniques into existing approaches, they proposed a framework for external plagiarism detection which not only was used to analyze strings but also the structure of the text. As a baseline mechanism Ferret technique was used which performs Trigram comparisons between original and suspicious document pairs and computes similarity level for such document pair based on the number of matching words it contains. Besides Trigram, Language Model similarity scores, Longest Common Subsequence and Dependency Relations Matching measures were computed, then these similarity scores were given as indicators for a Machine Learning algorithm to learn models, to classify.

In [2] the basic idea behind this work was to detect plagiarism by using Natural Language Processing (NLP) techniques into already present approaches, they have proposed a hybrid approach that combines the fundamentals of natural language processing and text mining to detect plagiarism in a document. This approach was found to be very effective in detecting synonyms and changes in the arrangement of words used in a sentence. The use of Jaccard's coefficient proved to be a very effective tool in computing the similarity coefficient between two sentences.

This approach was able to guard plagiarism detection algorithms against rewritten or spun content; as currently used tools are gullible to this small fix made in the material. The documents were classified using some special techniques, firstly the keywords were extracted from the text and were then sent to different mechanisms for understanding and processing.

In [3] the objective of this research paper was to enhance the latest designs for detecting paraphrasing with the capacity of recognizing derived versions of the same word, while computing plagiarism likelihood.

In this paper, the approach used in detecting external plagiarism was based on lexical analysis tools and n-gram technique. The advantage of this approach was that the effort for similarity computing remains the same, but the text processing was done only once per document, in a totally isolated preprocessing stage.

As a result, this plug-in property of the design allows further integration with different similarity algorithms like bag-of-words, SCAM, YAP etc. This paper was mainly focused on plagiarism detection performance with less execution speed and more on precision and recall.

In [4] This research paper aims to explore and compare the potency of syntactic-semantic based linguistic structures in plagiarism detection using natural language processing techniques. Proposed methodology was NLP based detailed passage level analysis where initially the documents were subjected to sentence segmentation using NLTK toolkit, this was followed by NLP based processing using the different NLP techniques like Part of speech tagging (POS),Chunking, Semantic role labelling along with two adjuncts which are chunking with tagging and semantic role labelling with tagging. In the proposed approach, sentence level comparisons were used and each sentence was processed using the above NLP techniques. Then a combined similarity metric that utilizes syntactic-semantic information was computed. Passage boundary detection was implemented based on sentence splitting and merging conditions.

In [5] This research paper gave results of a new proposed algorithm to detect plagiarism accurately in a text sample that is capable of dealing with extrinsic or intrinsic plagiarism. This approach showed that with several changes in detecting plagiarism algorithm could give a more accurate result:

1) Paraphrased plagiarism is mainly used with help of semantic parsing which will increase accuracy of plagiarism. 2) Syntactic Parsing is used in POS-tagger to detecting plagiarism. 3) Structuring the words in a sentence before doing the matching will give the efficient time. The documents were classified using some special techniques, majorly divided into three different phases. The first phase was Suspicious Journal Parsing Side where sentences are extracted from Suspicious journal. The second phase was Database Journals Parsing Side where sentences are extracted from journals present in database. Lastly, the third phase was Detection Processing Side where the Similarity score will be matched by the words on each para in duplicate document with each para in database. The results achieved using this method were below par and it required a lot of correction in the approaches used.

In [6] This paper investigates the effect of adjusting the default weighting method of LSA method, TF-IDF, to enhance the weight of class features.

The MCW are considered to be the best features for stylometric discrimination in classification tasks. The literature revealed that the usage patterns of the most common words differ from one author to another. However, they are always omitted or were given very low weight in most of pre-processing steps in many author detection applications. The corpus of English novels was used as an experimental dataset the results revealed that LSA with adjusted weighting for stylometric features performed better than traditional LSA. The paper presents a background of stylometry and LSA.

III. IMPLEMENTATION

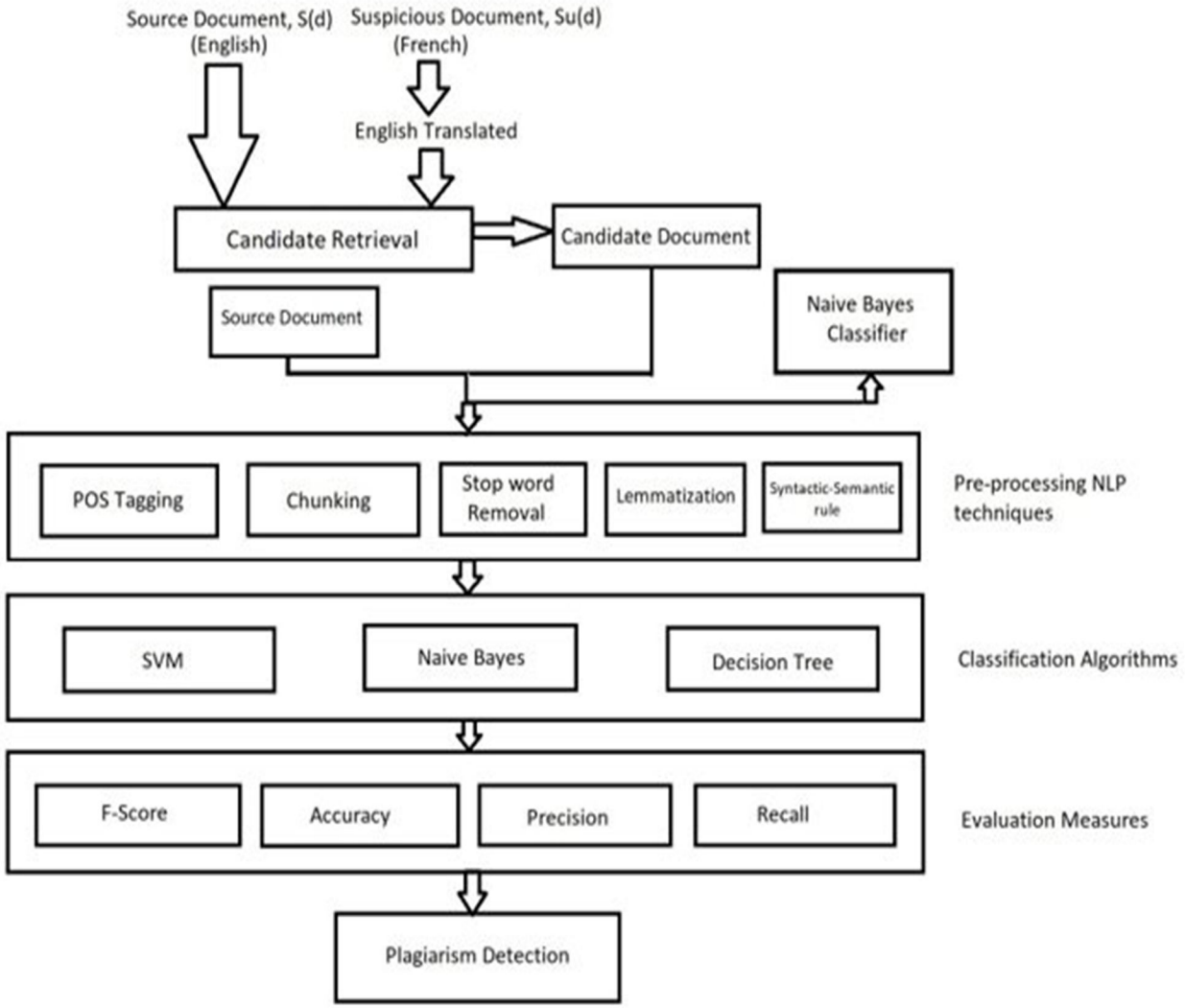


Fig. Cross Lingual Plagirism Detection

In our model, we are focused mainly on extrinsic plagiarism detection. We extract the data from well-known corpus called PAN(2012).We extract source documents and suspicious documents from PAN corpus. Since our model is based on cross lingual, we convert our data to one particular language. Then the process of candidate retrieval is performed. Candidate retrieval is a process of finding only suspicious sentences or phrases from a document. It stores all the suspicious document into one single document called candidate document. Candidate document along with Source document is used in comparing in next stage called pre-processing stage. This stage is base for our model since we use NLP pre-processing techniques for plagiarism detection. There are several pre-processing techniques but with the help of Naive-Bayes classifier we use specific techniques which are required in our program. Pre-processing helps in identifying those documents which are plagiarized and non-plagiarized. After receiving this data, we process into our next stage called Classification Algorithms. We use Machine learning classification algorithms such as Naive-Bayes classifier, decision tree and SVM. This helps us to predict the output as plagiarized and non-plagiarized. With help of this output, we go to our last stage, Evaluation measures. This stage, we evaluate the model and determine its efficiency. Accuracy score, recall score and f-score are used to determine the efficiency of the model.

IV. ADVANTAGES AND DISADVANTAGES

SL No	Papers	Advantages	Disadvantages
1	Using Natural Language Processing for Automatic Detection of Plagiarism by Miranda chong, Lucia Specia, Ruslan Mitkov.	Successfully incorporated NLP techniques for existing algorithms. Produced significant improvement compared to other algorithms.	Fail to process multilingual plagiarism. Sentence structure generalization were not addressed in this paper.
2	A Hybrid Approach for Detection of Plagiarism using NLP by Takshak Dessai, Udit Deshmukh, Mihir Gandhi, Lakshimi Krup.	It has high recall rate of 90% and precession rate nearly 94. Cosine similarity index was used for computing the lexical similarity for each pair of sentences.	Maximizing detection performance in terms of precision and recall leads to less oriented on the execution.
3	NLP Applications in External Plagiarism Detection by Sorin Avram, Dan Caragea, Theoder Borangiu.	The use of semantic parsing increases the accuracy of detecting plagiarism especially in paragraphed plagiarism. Time Efficient.	Not efficient for external plagiarism checking.
4	Unmasking text-plagiarism using syntactic-semantic based on NLP techniques by Vani K, Deepa Gupta.	Effective in synonyms detection. This approach was able to guard plagiarism against rewritten content.	Replacing words with their synonyms. Challenges such as time complexity, database inclusion, and word sense disambiguation were not dealt with properly in this project.
5	Plagiarism detection algorithm using NLP based on grammar analyzing by Ronald Adam, Suharjitho.	Based on evaluation results, it is found NLP techniques that extract syntactic-semantic outperformed the other systems with highest plg_score.	NLP techniques perform well with simulated sets while present a decreased performance with artificial plagiarism.

V. CONCLUSION

Our current research represents a technological endeavor in plagiarism detection, beyond its primitive form. In many cases, plagiarism continues to exists which are so hard to identify just by using the traditional tools. The amount of time increases with size of documents.

Our research shows that the accuracy of our model is much better than previous existed models. Use of NLP techniques helps to give better precision and accuracy and machine learning algorithms helps in detection.

At present our studies are formally based on two languages which are English and French or other mainstream languages. Our future studies lie in exploring the regional languages where we see good scope in improving our model to detect plagiarism.

VI. ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance.

So, with gratitude I acknowledge all those whose guidance and encouragement served as beacon of light and crowned our effort with success.

I am thankful to our Management for being a constant inspiration and providing all the facilities that needed throughout the Internship.

I would like to thank our Principal Dr.P. Mahabaleshwarappa, for his constant guidance, advice and encouragement to complete the Internship.

I consider it a privilege and honor to express my sincere gratitude to our beloved HOD, Mrs. Prameela Devi for her constant encouragement and all the support provided during this Internship.

I convey my sincere thanks to my guide Dr. K S Arvind, Assistant Professor, Dept. of CSE for his valuable guidance throughout the tenure of this Internship, and those support and encouragement made this work possible.

It's also a great pleasure to express my deepest gratitude to all my faculty members of my department for their cooperation and constructive criticism offered, which helped me a lot during my project work.

Finally, I would like to thank all my family members and friends whose encouragement and support was invaluable.

REFERENCES

[1] Miranda chong, Lucia Specia, Ruslan Mitkov," Using Natural Language Processing for Automatic Detection of Plagiarism" (2010), Research group in computational linguistics, Vol. 62 pp. 1-11.

[2] Takshak Dessai, Udit Deshmukh, Mihir Gandhi, Lakshimi Krup , "A Hybrid Approach for Detection of Plagiarism using NLP" (2016), Information retrieval on Natural Language Processing, pp. 1-6.

[3] Sorin Avram, Dan Caragea, Theoder Borangiu, "NLP Applications In External Plagiarism Detection" (2014), U.P.B.Sci.Bull, Series, Vol. 76, pp. 29-36.

[4] Vani K, Deepa Gupta, "Unmasking text-plagiarism using syntactic-semantic based on NLP technique" (2017), Information Processing Management, Vol. 67 pp. 408-432.

[5] Ronald Adam, Suharjitho, "Plagiarism detection algorithm using NLP based on grammar analyzing" (2014), Journal of theoretical and applied information technology, Vol. 63, pp. 168-180.

[6] Muna Alsallal, Rahat Iqbal, Saad Amin, Anne James, Vasile Palade, "An Integrated machine Learning Approach for Extrinsic Plagiarism Detection" (2016), 9th International Conference on Development in E-Systems Engineering, Vol. 63, pp. 203-208.

[7] Meysam Roostaee, Mohammad Hadi Sadreddini and Sayed Mostafa Fakhrahmad, "An effective approach to candidate retrieval for cross-language plagiarism detecting" (2016), Information Processing and Management, Vol 52, Issue 2, pp. 1-19.

[8] Vani K, Deepa Gupta, "Unmasking Text Plagiarism using Syntactic-Semantic based natural language processing techniques: comparisons, analysis and challenges" (2018), Information Processing and Management, Vol 54, Issue 3, pp. 408-432.

[9] Stein B, Rosso P, "An Evaluation Frame Work for Plagiarism Detection" (2010), Proceedings of 23rd International Conference on Computational Linguistics, pp. 997-1005.

[10] Bela Gipp, Norman Meuschke, "Citation Pattern Matching Algorithms for Citation Based Plagiarism Detection" (2011), Proceedings of the 11th ACM Symposium on Document Engineering, pp.1-6

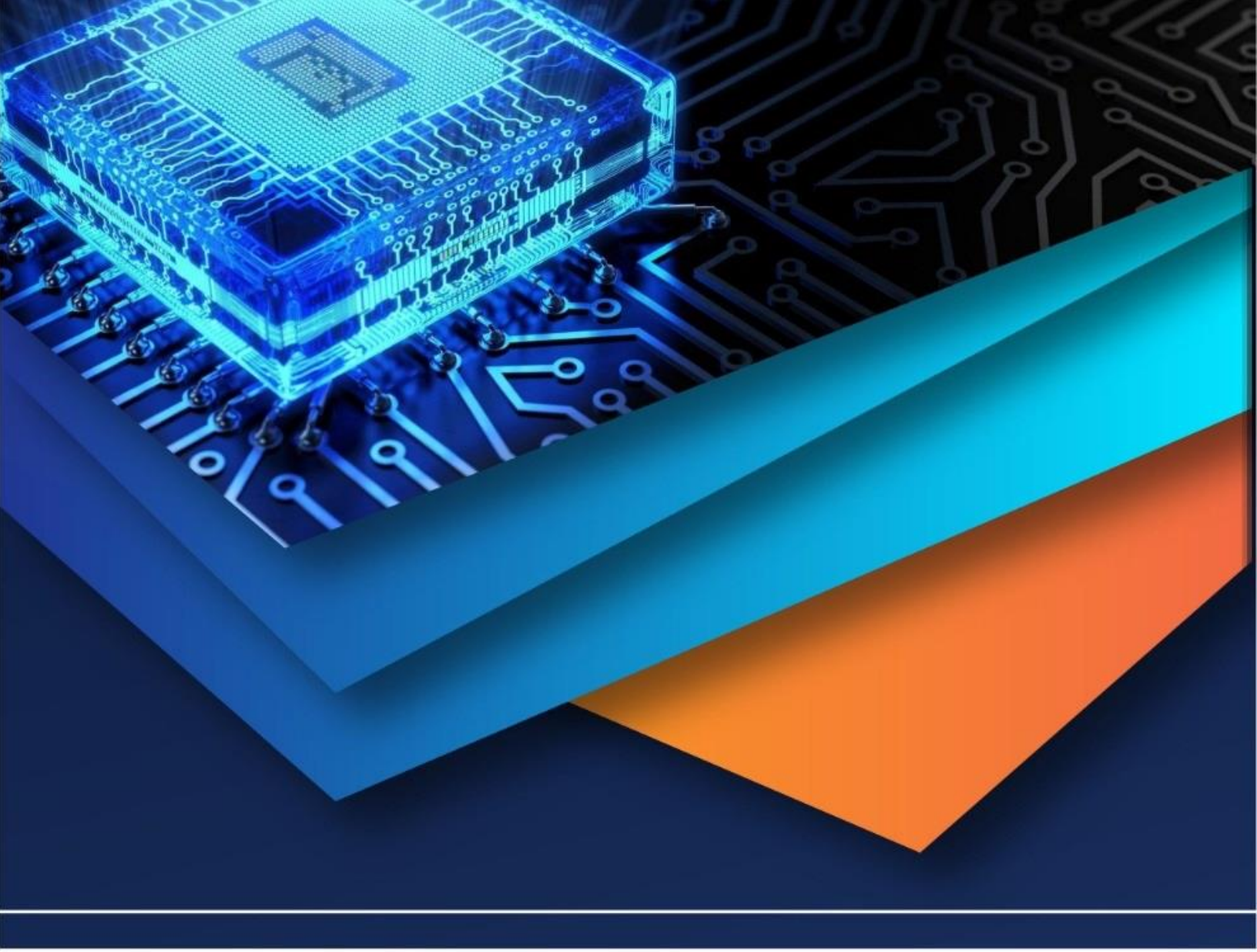
[11] Vani K, Deepa Gupta, "Exploration of fuzzy C means clustering algorithm in External Plagiarism Detection System" (2015), International Symposium on Intelligence System Technologies and applications, Vol 384, pp. 137-138.

[12] Vani K, Gupta, "Detection of Idea Plagiarism using syntax-semantic concept Extraction with genetic algorithm" (2017), Expert Systems with Applications, Vol-73, Issue-1, pp.11-26.

[13] https://academic.oup.com/jamia/article/18/5/544/829676, BASICS OF NATURAL LANGUAGE PROCESSING.

[14] www.writingsimplified.com/2009/11/plagiarism-introduction.html, PLAGIARISM INRODUCTION.

[15] https://smallseotools.com/plagiarism-checker, TOOL FOR CHECKING PLAGIARISM.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY