Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-1 Issue-3, June 2021

Plagiarism Detection Technique using www and Wordnet

Kamlesh Sharma, Nidhi Garg, Arun Pandey, Daksh Yadav, Nikhil

Abstract*: *Plagiarism is an act of using another person's words, idea or information without giving credit to that person and presenting them as your own. With the development of the technologies in recent years, the act of Plagiarism increases significantly. But luckily the plagiarism detection techniques are available and they are improving day by day to detect the attempts of plagiarizing the content in education. The software like Turnitin, iThenticate or SafeAssign is available in the markets that are doing a great job in this context. But the problem is not fully solved yet. These software(s) still doesn't detect the rephrasing of statements of another writer in other words. This paper primarily focuses to detect the plagiarism in the suspicious document based on the meaning and linguistic variation of the content. The techniques used for this context is based on Natural language processing. In this Paper, we present how the semantic analysis and syntactic driven Parsing can be used to detect the plagiarism.

Keywords*: *Natural Language Processing (NLP), Information Retrieval (IR), Cross-Language Information Retrieval (CLIR), Computational Linguistics, Wordnet, world wide web.

I. INTRODUCTION

With the development of the technologies in recent years, the problem of plagiarism increases consequently. So there is a need for a detection mechanism that identifies the plagiarized content in the digital form. This process started in the 1990s which was initiated by studying the copy detection techniques in the digital content. Later then it was detected with the help of the programs written in C and Pascal programming languages.

The algorithms for this method are mainly based on the plagiarism detection with the help of textual similarity and later on based upon the number of lines, variables, statements, and other parameters. But in recent years the methodologies have changed. Now, this detection is done with the help of Natural Language Processing (NLP), Information Retrieval (IR), Cross-Language Information Retrieval (CLIR), Computational Linguistics [2] and Artificial Intelligence. There are several methods for doing this job and some of these methods are very effective like substring matching using various algorithms like Rabin-Karp, Knuth-Morris or with the finite automata. In (Through) this method we find the maximum matches such that the substring works as a plagiarism Identifier. Next method is Fingerprint Analysis [7].

In this analysis we divide the whole document into a bunch of keywords known as chunks. Now this chunk will be compared with the Text document.

In this paper we present a method to check plagiarism using natural language processing techniques like Lexical Analysis and Semantic Analysis using the parse trees.

II. PROBLEM IN PLAGIARISM

A. Translation -

This problem of Plagiarism is mainly done with the help of Translation. In this method, the original content is translated from one language to another without giving credit to the original content. The retranslated content is reconstructed with the help of Google Translate and manual translation by the person who can speak both languages.

In the given Example the original content is first translated to German with the help of Google Translate and again reconstructed with it.

It is obvious that the retranslated sentence may have poor English, but now this reconstructed content shows very less plagiarism on detection.

- **Original Language**:-"A Computer is an electronic device which made human life easy. It is capable to complete more than one task in few times. The first computer was a mechanical device originally developed by Charles Babbage. The data is taken up by input devices and the result is shown up on Output devices."
- **Translated Language**:-"Computer ist ein elektronisches Gerät, das das menschliche Leben leicht gemacht hat. Es ist in der Lage, mehrere Aufgaben in wenigen Fällen auszuführen. Der erste Computer war ein mechanisches Gerät, das ursprünglich von Charles Babbage entwickelt wurde. Die Daten werden von Eingabegeräten übernommen und das Ergebnis wird auf Ausgabegeräten angezeigt"
- **Retranslated Language**:-"Computer is an electronic device that has made human life easy. It is able to perform several tasks in a few cases.

Manuscript received on 31 March 2021 | Revised Manuscript received on 17 April 2021 | Manuscript Accepted on 15 June 2021 | Manuscript published on 30 June 2021.

***** Correspondence Author

Dr. Kamlesh Sharma*, Associate Professor, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, India. Email: kamlesh.fet@mriu.edu.in

Nidhi Garg, Assistant Professor, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, India. Email: nidhigarg.fet@mriu.edu.in

Arun Pandey, Student, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, India. Email: pandeyelectrical73@gmail.com

Daksh Yadav, Student, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, India. Email: yadav.daksh2@gmail.com

Nikhil, Student, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad, India. Email: k.nikhil1103@gmail.com

The first computer was a mechanical device originally developed by Charles Babbage. The data is taken from input devices and the result is displayed on output devices".

B. Text Manipulation -

Plagiarism can be confused with the help of manipulating [2] the text and changing its presence. In the below example the bold texts are the words which are replaced by synonyms of the original text and short phrases are added in between the text to change its presence but not the main idea of the text. Paraphrasing the content while the keeping semantic of text requires citation around the plagiarized idea and cited the Authors.

Summarization of text, combination, reduction, reconstruction, paraphrasing, concept generalization, and concept specification is another form of text manipulation unless it is cited properly.

- **Original**: - Round Robin is a scheduling Algorithm used in the Operating systems and networks. In this algorithm equal time quanta or time slice is assigned to every process in a circular order, such that all process must be finished and no any process goes to starvation. Round Robin algorithm is quite easy to understand and implement. This algorithm can also be applied to other scheduling tasks, such as data packets scheduling in computer networks.
- **Plagiarized: -** Round Robin is a scheduling algorithm used in operating systems and networks. In this algorithm, **equal times**[synonym] **or time slots**[synonym] are entrust to each process in a **roundabout**[synonym] order, so that any process must be **completed**[synonym] and no process goes **unnoticed**[synonym]. The Round Robin algorithm is quite easy to understand and implement. This algorithm can also be applied to other **planning**[synonym] tasks, such as scheduling data packets in computer networks."

C. Character Manipulation -

In this method, the space between the sentences is replaced with any character to make a sentence as a whole word. Now the color of the replaced character is changed as the color of the background. Now, this content is human readable. But when this content will be inspected by the detector, it will show very less plagiarism or full unique content because the detector will be confused to take it as a whole word, not as a combination of words separated by a special symbol.

- **Original: -** Computer is an electronic device which made human life easy. It is capable to complete more than one task in few times. The first computer was a mechanical device originally developed by Charles Babbage. The data is taken up by input devices and the result is shown up on Output devices
- **Plagiarized**:-"Computer?is?an?electronic?device?which?made?human?life?easy.?It?is?capable?to?complete?more?than?one?task?in?few?times.?The?first?computer?was?a?mechanical?device?originally?developed?by?Charles?Babbage.?The?data?is?taken?up?by?input?devices?and?the?result?is?shown?up?on?Output?devices"
- **After changing the color:-** "Computer is an electronic device which made human life easy. It is capable to complete more than one task in few times. The first computer was a mechanical device originally developed by Charles Babbage. The data is taken up by input devices and the result is shown up on Output devices".

III. NLP TECHNIQUES USED

A. Pattern Matching:

Pattern matching is the process of interpreting the whole vocal expression [1] as parent rather than interpreting the single word one by one. That means the meaning is provided by the pattern matching of words of the input vocal sentence. A large number of patters are needed for deep level of analyzing the sentence.

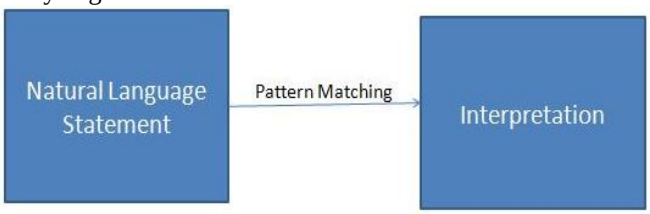


Figure 1: Pattern Matching

B. Syntactical driven Parsing:

A sentence is grouped according to some rules known as the syntax of that language. Syntactical driven parsing means conversion of words [2] present in the more numbers into the understandable language. Parsing is the process of analyzing the sentence, which is either in natural language or computer language according to the rule of grammar. This is done by converting the sentence into a different single unit of structure that gives a meaning for the sentence. It is done by creating a parse tree that cut the sentence into structured parts so that the computer can easily understand and process it.

C. Parse Tree:

Let us take a English statement: "The dog ate the bread.''

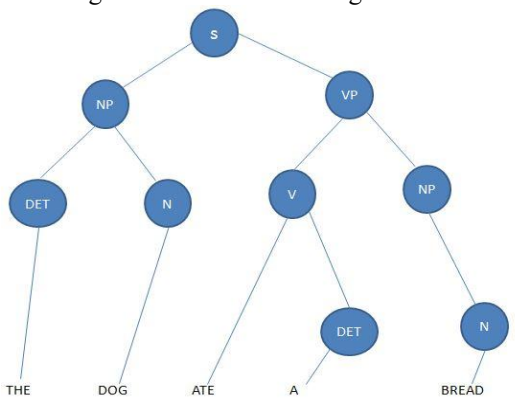


Figure 2: Parse Tree

Advantages of Parse Trees

1. Parse trees can quickly check the grammar of the sentence.
2. It is the central part of the semantic analysis of the sentence.
3. Parse trees are used for extraction of information from the chunk of sentences.

Retrieval Number:100.1/ijainn.B1015021221
DOI:10.54105/ijainn.B1015.061321
Journal Website: www.ijainn.latticescipub.com

Published By:
Lattice Science Publication (LSP)

4. This is beneficial for the sites who (that) use NLP for question-answering the exam, or for the online interviewing.
5. Parse trees can also use for translating the mechanical speech from the user

D. Semantic Grammars:

Semantic Grammar is sometime similar to the Syntactical driven parsing because it also involves the syntax of the language and the logic involved in the sentence. But it also features the semantic grammar of the sentence, which means it also check the logic of the grammar used in the sentence. [12]

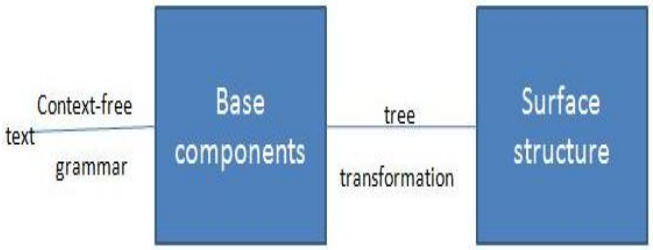


Figure 3: Text to Tree Transformation

E. Case Frame Instantiation:

Case frame expression is one of the very important parsing techniques and there are lots of research is going on. Case frame [4] instantiation has some very useful arithmetic properties such as its recursive properties and its ability to combine bottom up acknowledgement of key components with top down express of less structured components.

Natural language processing involves five stages which are as follow:-

1. **Lexical Analysis: -** The analysis of words in the sentence is known as Lexical analysis. The group of words, idioms or phrases in a language is called Lexicon. This is done by categorizing the whole content into Paragraph, Sentences or in words.
2. **Syntactic Analysis:-** The analysis of words in the sentence for grammar and arrangement of words in any predefined order according to the grammar. For example the sentence"Eat mango Ram" is not in the grammatical order so it is rejected.
3. **Semantic Analysis**: - At this stage we analyze the meaning of the words used in the sentence. The sentence is checked for its logic used in the sentence. The sentences "Do you have a tired blood" and, "The hot ice-cream" sentences without meaning. So they are rejected.
4. **Pragmatic Analysis**: - It analyze the sentences that are said earlier to reinterpret what it actually means? It involves those parts of language which require the real world knowledge.
5. **Discourse Integration**: - meaning of any new sentence of in any paragraph depends on the previous sentence. And this new sentence will provide meaning for the new sentence proceeding further.

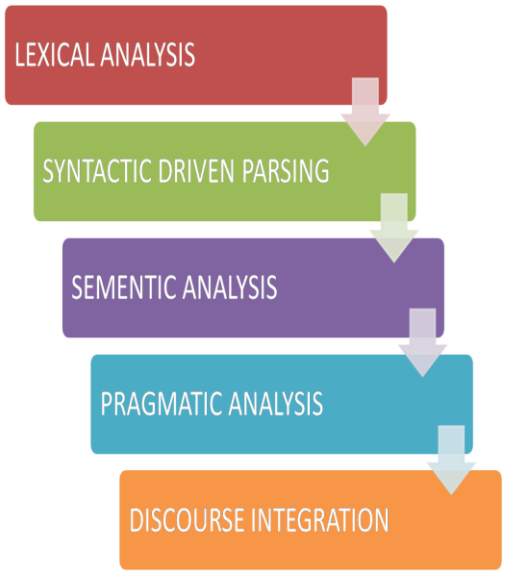


Figure 4: Stages in NLP

IV. PROPOSED SYSTEM

The proposed system for the detection of plagiarism is shown in the fig. n. This system uses Wordnet as the semantic translator for any word. The query String is fetched up by the system using the query string and using the natural language processing techniques, this string is converted into small chunks of words using the Tokenization process. Now these tokenized words are converted into Parse trees to provide the semantic checker all details for the string. Semantic checker uses WWW and Wordnet to check all the semantic words related to the tokenized word [11]. And the results are shown in the resulting Document with the percentage of plagiarism found.

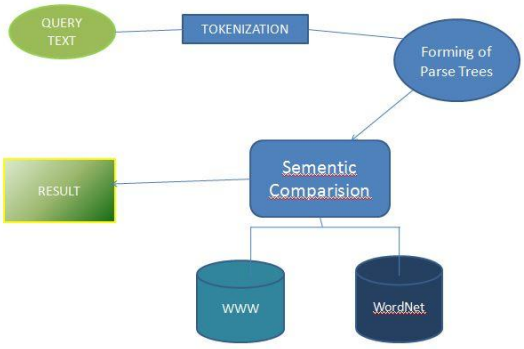


Fig. 5 Block Diagram of System

V. COMPONENTS OF SYSTEM

A. Query String:

A user uploads a document over the system to check how much percentage of plagiarism is in the document. This document contains the String to be checked against the Plagiarism over the internet. The file format of this document is usually like .txt, .doc, .pdf and another document format.

B. Tokenization:

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* [9], perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization:

```
from nltk.tokenize import word_tokenize

text = "Hello everyone. Welcome to our proposed
system."
word_tokenize(text)
```

['Hello', 'everyone', '.', 'Welcome', 'to', 'our', 'proposed','system' '.']

C. Forming of Parse Tree:

```
sent = ['Our', 'proposed', 'system', 'is', 'best"]
parser = nltk.ChartParser(groucho_grammar)
for tree in parser.parse(sent):
        print(tree)
```

D. Semantic Comparison:

This analyzes the meaning of the words used in the sentence. The sentence is checked for its logic used in the sentence. The sentences "Do you have a tired blood" and "The hot ice-cream" are the sentences without meaning, hence, they are rejected.

E. Wordnet:

Wordnet[10] is a lexical database available in the NLTK package of python programming language which is used to provide the synonyms and antonyms of the given word.

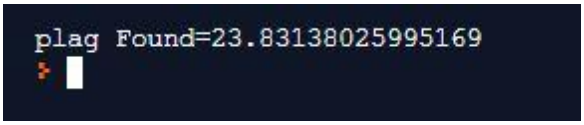
```
from nltk.corpus import wordnet
s="Our  Proposed System is best"
p="

for syn in wordnet.synsets("Comparision"):
   for l in syn.lemmas():
      synonyms1.append(l.name())
 for syn in wordnet.synsets("Similarity"):
   for l in syn.lemmas():
      synonyms2.append(l.name())
print(synonyms1.wup_similarity(synonyms2))
```

F. www

The World Wide Web is used to fetch the data from several web pages all over the internet and it acts a source for detecting the plagiarized content from the source document.

```
import nltk
from nltk.corpus import wordnet

a="Started his hearted any civilly. So me by marianne
admitted speaking. Men bred fine call ask. Cease one
miles truth day above seven. Suspicion sportsmen
provision suffering mrs saw engrossed something. Snug
soon he on plan in be dine some. "

p="Started his heart any civilly. So me by Marianne
admitted to speak. Men have very well called to ask.
Stop a day of truth miles above seven. Mistrust towards
athletes who suffered mrs saw something excited. Snug
soon he on the plane at dinner."

text = nltk.sent_tokenize(a)
plag= nltk.sent_tokenize(p)
txt1 = []
plg2 =[]
f=0
n=1
p=0.0

for sentence in text:
    for word,pos in
nltk.pos_tag(nltk.word_tokenize(str(sentence))):
        if (pos == 'VP' or pos == 'VBD' or pos == 'VB'):
            txt1.append(word)

for i in plag:
    for word,pos in
nltk.pos_tag(nltk.word_tokenize(str(i))):
        if (pos == 'VP' or pos == 'VBD' or pos == 'VB' ):
            plg2.append(word)

for j in range(len(txt1)):

  if(wordnet.synsets(txt1[j])!=None):
    syn1=wordnet.synsets(txt1[j])[1]
  for k in range(len(plg2)):
    if(wordnet.synsets(plg2[k])!=None):
      syn2=wordnet.synsets(plg2[k])[1]
      if(syn1.wup_similarity(syn2)!=None):
        f=f+float(syn1.wup_similarity(syn2))
        n=n+1

f=f/n
p=f*100
print("plag Found="+str(p))
```

G. Result:



VI. ADVANTAGES OF PROPOSED SYSTEM

Figures The detection based on the Natural Processing techniques has advantages over the other proposed system which are as follows

1) The older detection methods use Wordnet and Wikipedia for their knowledge basis which can solve the problem of synonyms, semantic similarity, and paraphrasing problems. But the proposed system fails to overcome the syntax-based problems. But as the proposed system uses Ngram comparisons and tokenization, it overcomes this problem also.

2) Plagiarism can be confused with the help of manipulating the text and changing its presence. But as the whole text is first converted into chunks, and after the removal of stop words, the detection is now possible.
The tokenized words are then used as the source of semantic similarity calculation. Thus the problem of semantic similarity can be overcome.
3) The translated sentence from one language to another language can be the source of plagiarism. This problem can also be solved with the help of natural language processing techniques.
4) The proposed system is quite reliable, easy to understand, and most importantly easy to implement. The inbuilt packages are available in many languages to implement the system. The POS tagger which is the main tool used for this system is available in the NLTK package of the Python Language.
5) In 2010 Chong et al. applied several NLP techniques on short paragraphs to analyze the structure of the text to automatically detect the plagiarized text. They proved that NLP techniques can increase the efficiency of the detection, although there were several problems present like the problem of synonym disambiguation and sentence structure disambiguation. The proposed system can overcome all these problems.

VII. DISADVANTAGES OF PROPOSED SYSTEM

1) 1. There may be the drawbacks of this approach mainly due to the corpus or the database from which we use the comparison for the detection. If the corpus contains too much data words, we want to compare it with our chunks. The searching and comparisons will be increases rapidly also. This leads to the inconsistency in the proposed system.
2) 2. The system uses inbuilt packages for the tokenization and Tagging with the text document. The stop word removal also uses the functions of these packages. This may be overhead for the system to corporate with the real world requirements.
3) 3. This system cannot detect the plagiarized texts that are the Citing sources that were not actually referenced or used.

VIII. CONCLUSION AND FUTURE SCOPE

Avoiding plagiarism is important. The author must use anybody's idea to give the full credit to that author. It presents the viewers how much respect you give to that content. Most importantly, we give credit when the credit is due. You do not deceive the person who reads it by falsely believing that the job belongs to you. This is a growing attraction among students and an invariable complication for teachers in dealing with the issue. And therefore, the pain for someone who is caught for plagiarism can be severe. So we have to understand the consequences of plagiarism. The given system can detect the plagiarism over the source document weather it contains the semantics of that word. But we have to synchronize that technique with the techniques used in the plagiarism tools. The comparison of the software and tools has shown that still now their no software and tools that can detect or to prove that the document has been plagiarized 100%. The future work involves adding more capability and features to the current software and tools to detect the plagiarized document very efficiently.

REFERENCES

1. Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods", IEEE transactions on systems, man, and cybernetics part c: application and reviews,42(2), march 2012. [CrossRef]
2. Ahmed Hamza Osman, Naomie Salim and Albaraa Abuobieda, "Survey of text plagiarism detection", Journal of Computer Engineering and Applications ,1(1), June 2012 [CrossRef]
3. The Struggle with Academic Plagiarism: Approaches based on Semantic Similarity Tedo Vrbanec , Ana Mestrovi6.
4. S. Vukovi6 and B. Kopi6, "Plagiranje - Sveuciliste u Zarebu jos nije uvelo sustav provjere radova," Global 21, Zagreb, p. 24, Nov-20 16. 6. C. Grozea, C. Gehl, and M. Popescu, "ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection," in Proc. SEPLN, Donostia, Spain, 2012, pp. 10–18.
5. Binwahlan, M. S., Salim, N. & Suanmali, L. (2009b). Swarm based features selection for text summarization. IJCSNS International Journal of Computer Science and Network Security. 9(1), (pp. 175-179).
6. Manuel, Z., Marco, F., Massimo, M., & Alessandro, P. (2006). Plagiarism Detection through Multilevel Text Comparison. Paper presented at the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution.
7. Jiannan Wang, Guoliang Li, JianhuaFeng "Fast-Join: an efficient method for fuzzy token matching based string similarity join", Data Engineering (ICDE), 2011 IEEE 27th International Conference, March 2011.
8. Miranda Chong, Lucia Specia and Ruslan Mitkov, "Using natural language processing for automatic plagiarism detection", 4th International Plagiarism Conference, Northrumbia University, 2010.
9. Jiannan Wang, Guoliang Li, JianhuaFeng "Fast-Join: an efficient method for fuzzy token matching based string similarity join", Data Engineering (ICDE), 2011 IEEE 27th International Conference, March 2011.
10. YuriiPalkovskii, Alexei Belov, IrynaMuzyka,"Using WordNet based semantic similarity measurement in external plagiarism detection", Notebook Papers of CLEF , 2011.
11. Kamlesh Sharma, Dr. S. V. A. V. Prasad, Dr. T. V. Prasad, A Hindi Speech Actuated Computer Interface for Web Search, Int. J. of Advanced Computer Sc. App. (IJACSA), Vol. 3, Issue. 10, Oct, 2012. [CrossRef]
12. Kamlesh Sharma, Web Recognition of Spoken Hindi, Indian Journal of Science and Technology, Vol. 10, Issue. 35, Sep, 2017. [CrossRef]

AUTHORS PROFILE

Dr. Kamlesh Sharma is currently working as a Associate Professor, MRIIRS, Faridabad, India (more than 15 years teaching experience). MCA, M. Tech from MDU University and Ph. D. in Computer Science and Engineering from Lingaya`s Vidyapeeth, India. is currently Supervising five Ph. D. scholars. She has also supervised and guided research projects of M. Tech, B.Tech and application based projects for different competitions. She is also associated with four Govt. research projects in filed of health recommender system, IOT, Machine Learning, AI and NLP. She has published more than 55 research papers in field of NLP, IOT, Bigdata, Green Computing and Data Miningin reputed Journal (Web of Science, Scopus, UGC, Elsevier) and Conferences (ACM, IEEE). Her research area "Natural Language Processing" is based on innovative idea of reducing the mechanized efforts and adapting the software to Hindi dialect. She is associated with various professional bodies and renowned journals in varied capacities viz. CSI (Computer Society of India), Member, International Journal of Computer Networks and Applications (IJCNA) as Editoral Board Member, BJIT - BVICAM's International Journal of Information Technology, ISSN 0973 – 5658, Springer Index as Reviewer, International Journal of Computer Science and Information Security (IJCSIS), Google Scholar Index as Reviewer & Editorial board member, International Journal of Science & Engineering Development Research - IJSDR, UGC Approved Journal, Google Scholar Index as Member of referral/ review Management System.

Nidhi Garg, is currently an Assistant Professor in the Faculty of Engineering and Technology, Manav Rachna International Institute of Research & Studies, Faridabad. She received her Master's in Technology - Computer Science and Engineering from Maharishi Dayanand University, Rohtak in year 2012 and has 10+ years of teaching experience. Her current research interest includes Artificial Intelligence, Machine Learning and Image Processing. To add to her credits she has authored and co-authored many journals and conference papers in various computer science domains including Networking, Artificial Intelligence, and Machine Learning. She has also active member of IAENG and reviewer of conferences and journals like ICIMMI, ICCS, CIAIS'21 etc.

.

Retrieval Number:100.1/ijainn.B1015021221
DOI:10.54105/ijainn.B1015.061321
Journal Website: www.ijainn.latticescipub.com

Published By:
Lattice Science Publication (LSP)