iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** II **Month of publication:** February 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.40363

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: 7.538*
*Volume 10 Issue II Feb 2022- Available at www.ijraset.com*

# Compliance Verification Process Automation

Siddique Irshad Ahmed[1], Kanse Vinse Ramdas[2], Asst. Prof. Ahlam Ansari[3]
[1, 2, 3]*Department of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India*

***Abstract: Compliance Documents are documents which include records, reports, observations, and verbal responses to establish or confirm compliance with a regulatory requirement by a program or facility. But these documents are very detailed and lengthy and difficult to comprehend in a short amount of time. Thus it creates many problems while adding new regulations or modifying the documents.***
***Also, verifying these documents is very exhausting. Our product will provide organizations a web platform to quickly manage compliance documents. It works as an assistant to deal with compliance documents. This assistant is an AI-powered chatbot that can take queries from users, employees and other stakeholders related to compliance of the organizations and can solve them. It also helps to cross verify the document based on a given standard and gives results as a percentage of its correctness and generates reports as pdf.***
***Keywords: OCR, Natural Language Processing (NLP), Artificial Neural Network (ANN), Knowledge Graph (KG), Cloud Computing***

## I. INTRODUCTION

In today's world, in order to make a reputation in the business world companies must provide customer satisfaction which leads to quick resolution of customer queries.. Hence automation is required to make most of the task in hand spontaneous. Over the last decade every employee had to manually perform a compliance process. But as the technology advanced there have been many successful implementations of compliance automation.
Compliance automation is the automation of compliance processes that companies had previously done manually. Compliance automation uses artificial intelligence features and technology to make compliance procedures easier. This gives organizations compliance-related workflow capabilities such as appropriate action planning, control testing and control analysis.
Initially, compliance automation tools receive a company's security policies. The regulations that pertain to an organization's industry, configurations, accounts, inventories, and security procedures are copied into the compliance automation software to identify violations.
An organization can change or update this database of compliance regulations and security standards at any time. The Question Answering Chatbot allows the user to upload relevant documents on which two main functions, namely answer extraction and question generation are performed, after converting the documents to a suitable knowledge base. The Question Answering module employs ranking functions to extract relevant answers from the knowledge base, out of which top K answers are further fed to a neural network which chooses the final answer.

This product have following major features:

1) *Verify Compliance:* The selected document can be verified by comparing with the Standard Compliance framework. Verification is a very crucial process in the compliance world as the legitimacy and genuinity of the documents are checked here.
2) *Generate Reports:* Based on the verification result a report is generated containing the analysis of the document. Reports are documented on the basis of parameters that affected the verification process and it also contains the status of the document that is currently being verified.
3) *QA Chatbot:* Chatbot will allow users to query specific parts of the document which will help users to find things quickly.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue II Feb 2022- Available at www.ijraset.com*

Table 1: Table of Existing Methodology

| Cited Work | Year | Author(s) | Overview |
|---|---|---|---|
| [1] | 2018 | Tien Fui Yong, Saiful Azad, M. Mostafizur M. Mostafizur, Kamal Zamli, Gollam Rabby | This paper provides fast processing, user-friendliness, and inferring solutions. They have used a Natural Language Processing based PDF-to-text(NLPDF) conversation system with all datasets of A4 size pdfs. Starting with defining sections like header, footer, paragraph, columns and other differently oriented text. After that it started parsing text line by line format generating a new line(\n) at the end of line except if there is '-' in the end. Their model gives an average precision of 94.1%. |
| [2] | 2021 | Nikolaos Livathinos, Cesar Berrospi, Maksym Lysak, Viktor Kuropiatnyk | This paper uses Recurrent Neural Network(RNN) to parse documents directly from PDF printing commands which helps to reduce misinterpretation and time to parse documents. First they classify command into 10-20 fine-grained lebal which helps to increase accuracy and gives document a more detailed structure. They also claimed 97% accuracy on pdf articles related COVID-19. |
| [3] | 2021 | Yuta Koreeda, Christopher D. Manning | The system is designed on Visually Structured Documents(VSDs) a NLP based model as they don't have text in sequential order but scattered on whole which as easily interpreted visually, but on parsing text are totally jumbled. They proposed to formulate the task as prediction of transition labels between text fragments that maps the fragments to a tree, and developed a feature-based machine learning system that fuses visual, textual and semantic cues. For documents like VSD their accuracy score is around 95% accuracy. |
| [4] | 2019 | Simon Butler, Jonas Gamaliel Sona, Björn Lundell, Christoffer Brax | Apache PDFBox is used to extract text from PDF files using Optical Character Recognition (OCR). From the generated text document, question-answer pairs are generated using the Overgenerating Transformations and Ranking algorithm. The user input is matched against the question-answer pairs using pattern matching to fetch the answer. |
| [5] | 2020 | Nazakat Ali | A feed-forward neural network is used to allocate scores to different passages according to the relevance of the passage to the question. InferSent representations are used to capture the general semantics of a question and the passage. In this paper, word features are used to generalize different retrieval tasks. The basic model uses TF-IDF vector space along with improvements such as low rank representations, correlated feature hashing, and sparsification. |

## II. LITERATURE SURVEY

### A. *Survey Of Existing Systems*

Compliance documents are very critical and confidential documents to one's own company. It contains all of the vulnerabilities and private data which if fallen in wrong hands a company's existence is at risk.

Currently available solutions for compliance verification takes at least 3 days for the verification of documents, which is a rather time consuming and very resource intensive process.

While the documents are being verified, a company expects its document to be verified as soon as possible but unfortunately they have to wait at least 3 *working* days for the process to be completed. So there's a lack of transparency between the company and the concerned authorities.

There have been many situations where it takes a pretty long time for the documents to be verified , hence companies are bound to get frustrated as the document contains confidential information.

There is also a lack of communication between the company and the authorities. This creates a barrier between the company and the authorities and thereby the concerned company doesn't get the status of the verification process or any alert regarding the process.

### B. *Limitations or Research Gap*

*1)* They have restrictions on the types of the pdf or documents that can be processed.

*2)* The existing compliance platform requires many other documents and proof just to verify the compliance document is valid as per standard compliance framework.

*3)* As in the existing system most of the work is done manually, they require a lot of time just to get done with a single document.

*4)* Existing systems have a lot of manual work which is done in the coordination of multiple people, which can cause errors because of many reasons like miscommunication.

## III. PROPOSED SYSTEM

### A. *Hardware / Software and Technologies used details:*

*1) Hardware / Software*

*a)* One or more machines running one of:

- Ubuntu 16.04+
- Windows 8+
- Mac OS 8+

*b)* 2 GB or more of RAM per machine.

*c)* 20GB of free disk space.

*d)* Full network connectivity between all machines in the cluster.

*e)* Browser.(Google Chrome, Microsoft Edge, Mozilla Firefox)

*2) Technologies*

Table 2: Technology stack used

| | |
|---|---|
| Frontend Stack | ReactJS + Material UI |
| Backend Stack | Python + NLP + Django |
| Database | PostgreSQL |

*B. Design Details*

Use case diagram represents the user scenario and at its a user's interaction with the system in accordance with its association between the user and the different possibilities in which the user is participating.

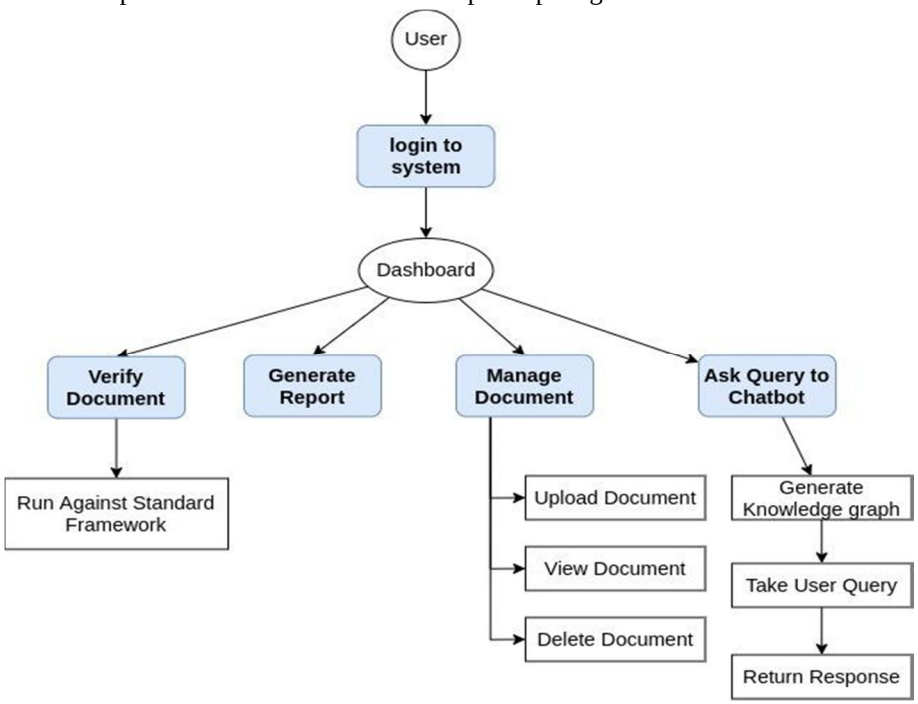



Figure 1: Data Flow Diagram

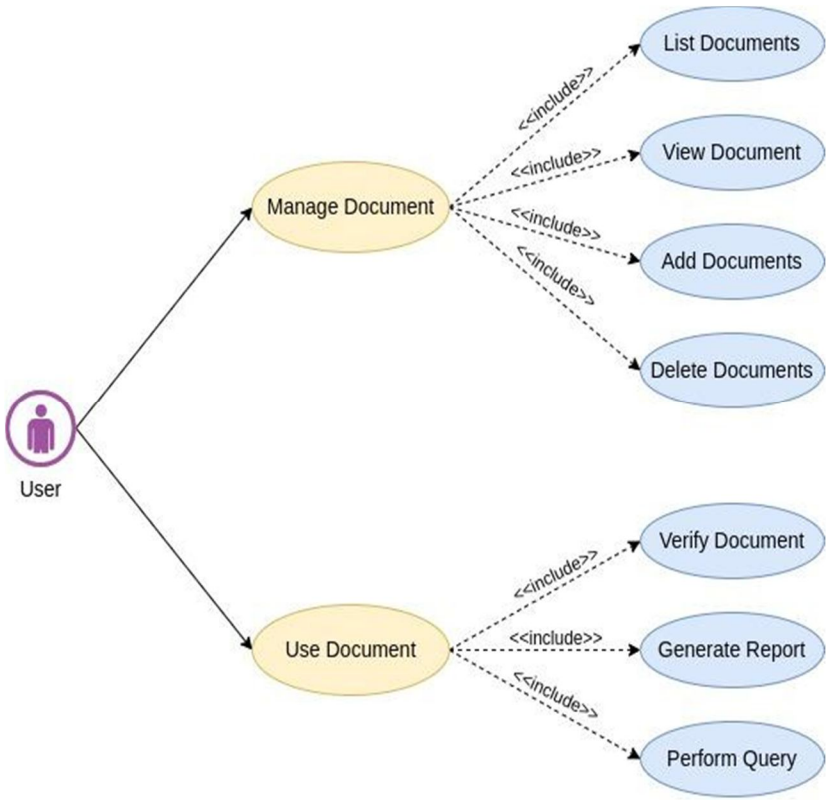



Figure 2: Use Case Diagram

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 10 Issue II Feb 2022- Available at www.ijraset.com*

*C. Proposed Algorithm*

The proposed system aims to make use of principles of the latest technology and resources to provide a more simple, error less, and automated approach. Thus, reducing various factors such as time consumption, human-based error, mistakes because of miscommunications and hassle of managing many documents. For understanding the system, we can divide it into various modules and understand them:

1) *Databases:* The database is the most fundamental part of the application as the entire record maintenance will be on the cloud. The database architecture for each operating company is separate with some exceptions The database for users will be centralized for reasons such as reducing latency while logging in. Now from the above description, one can easily conclude one or both two things:

- The generated database would be large.
- And how can everyone access it in real time?

Considering both the points, the database technology that will be used is MongoDB Atlas which stores data in the form of nested JSON objects, which are light in size and recommended way to save knowledge graphs. Along with giving us the benefits we need, it is easy to scale and relatively easy to work with real time access as it is already hosted in the cloud. With features such as complete atomic, consistent, isolated, durable transaction support; multi-version query support; and it is the go-to solution for full data integrity.

2) *Knowledge Graph:* A knowledge graph, also known as a semantic network, represents a network of real-world entities i.e. objects, events, situations, or concepts and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph." A knowledge graph is made up of three main components: nodes, edges, and labels. Any object, place, or person can be a node. An edge defines the relationship between the nodes. For example, a node could be a client, like IBM, and an agency like Ogilvy. An edge would be to categorize the relationship as a customer relationship between IBM and Ogilvy. Now Question is why are we using it?

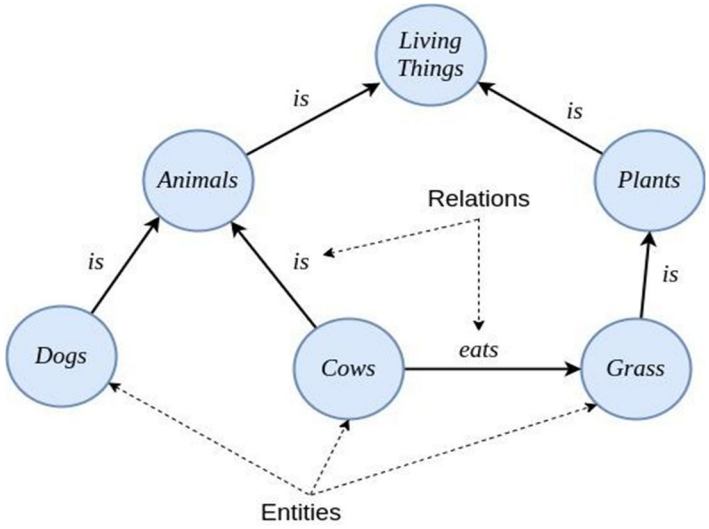



Figure 3: Example of knowledge graph

A knowledge graph brings together machine learning and graph technologies to give AI the context it needs. As we are parsing the large compliance document, we need something which can relate the extracted data from it so that we can return meaningful responses when a user query from it. This relation needs to be stored somewhere somehow so there won't be a need of parsing the document again and again every time we need it.

3) *Chatbot:* A chatbot is the best way for an application to communicate with users along giving them full freedom of query and operation they can do. We are using chatbot to give users the ability to perform a query on the document which has already been parsed and converted into the knowledge graph.

4) *Workflow of Application:* The workflow will be completely automated, and the platform remains as simple and as easy as possible, so the user will have less to no effort getting accustomed to the application. The application will work in the following ways:

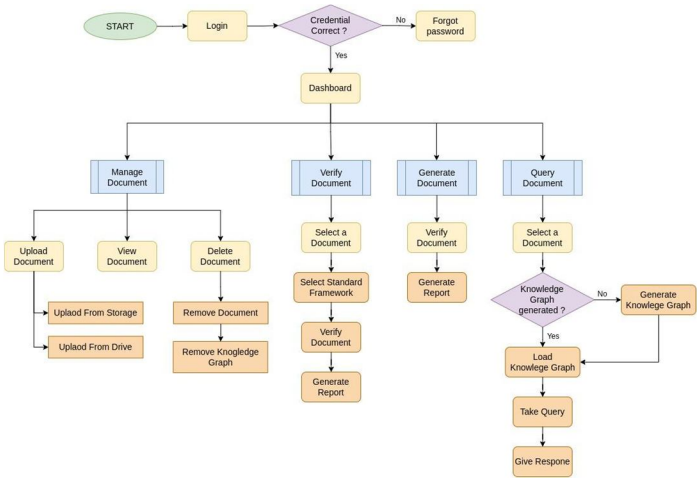



Figure 4: Flowchart

- Every user can see the list of uploaded documents, or view a specific document, or add new documents or delete the existing ones.
- They can select a specific document to match against their standard framework and get their -verification score and also generate reports for the same.
- Users can query on the document through chatbot, where the document is first converted into the knowledge graph(only once). Then the query passed through processing to get related keywords. This keyword is then matched against a knowledge graph, where more related entities are extracted then converted into meaningful sentences then returned as response

5) *Security:* Coming from all the above-mentioned points it is an obvious question how secure the application is. And to answer that question few things are required to be implemented (or will be implemented if absent), those are a good firewall, SSL integration, and a secure hosting platform. We also provide Authentication and Different Level of authorization and encryption to save confidential data.

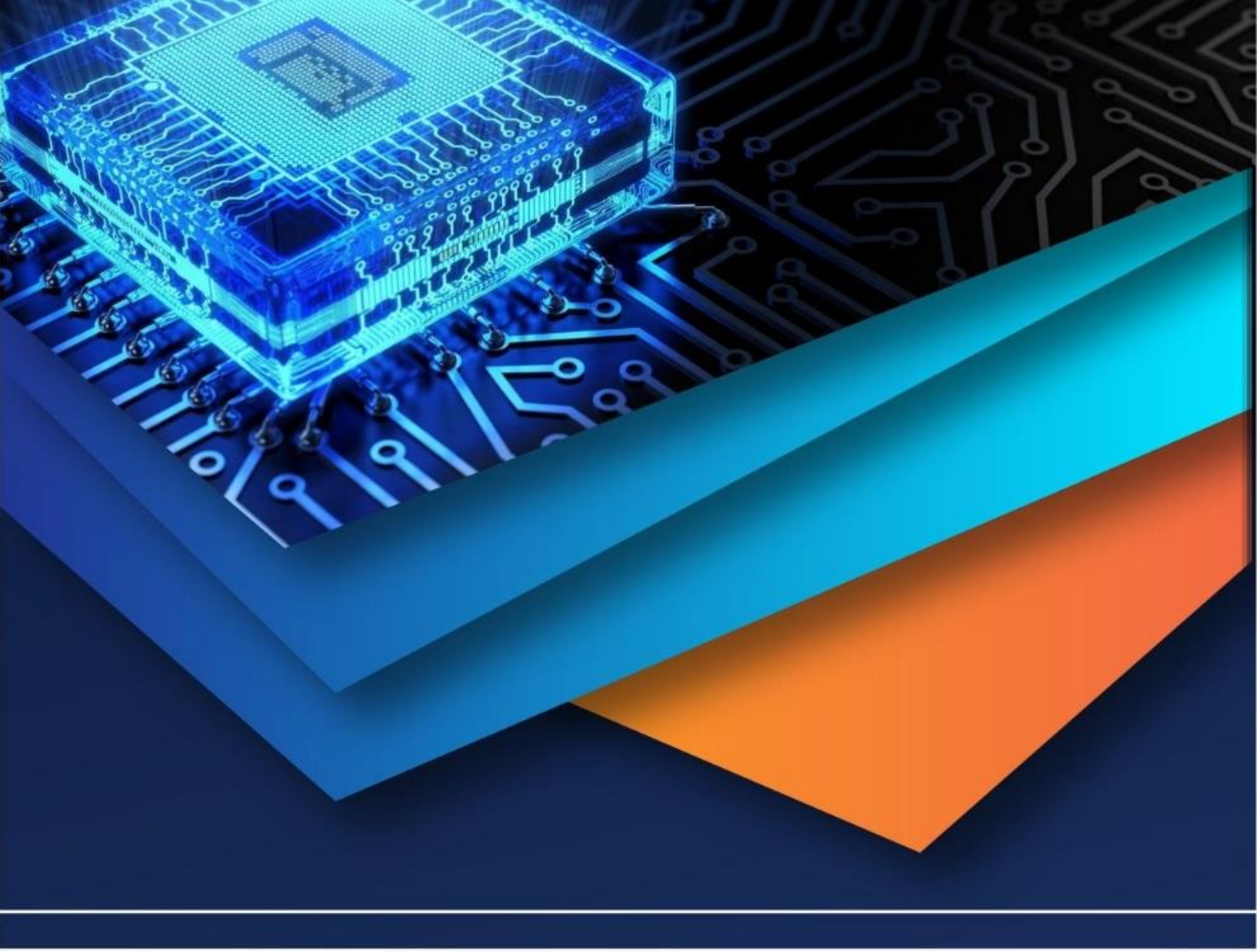
6) Other than these the application provides secure logins, load balancing for simultaneous logins, access only to the necessary ports, communication within the application to avoid any data leak/breach that can happen, saves logs to trace the origin of the issue if any happens, etc. depending upon the requirements.

## IV. CONCLUSION

The final developed product will increase the efficiency of the organization to manage the compliance documents. It will help users to easily work with documents with a simple interface and fast working chatbot which will help tackle user queries in a very efficient way. The manual intervention for managing the compliance process of an organization is completely removed, thereby inculcating a more efficient approach. The application saves a knowledge graph of parsed documents for future use, thus removing the need of going through the whole document or parsing it again and again whenever required.

## REFERENCES

[1] Madisch, I. (2008, November 08) A Highly Accurate PDF-To-Text Conversion System.
[2] Livathinos, N. M. (2021). Robust PDF Document Conversion Using Recurrent Neural Networks. PDF conversion, 1(2).
[3] Koreeda, Y. (2021). Capturing Logical Structure of Visually Structured Documents with Multimodal Transition Parser. Document Parser, 2(5).
[4] Koetter, F. (2019). Conversational Agents for Insurance Companies. Conversational Assistant, 4(2).
[5] Ali, N. (2020). Chatbot: A Conversational Agent employed with Named Entity Recognition Model using Artificial Neural Network. Conversational Chatbot, 3(2).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY