

**HINDUSTAN AERONAUTICS LIMITED**

**AIRCRAFT DIVISION**

**NASHIK**

TRAINING PROJECT REPORT

**“QUESTION ANSWERING MODEL BASED ON**

**AI AND NATURAL LANGUAGE PROCESSING”**

Mr. SIDDHIVINAYAK SAHOO

B.E (CSE) 3RD YEAR

SVKM NMIMS

MPSTME ,SHIRPUR

PROJECT GUIDE:

Mr. M. RAJENDER

DGM (DESIGN)

**AIRCRAFT UPGRADE RESEARCH AND DESIGN CENTRE**

**CERTIFICATE**

This is to certify that Mr. SIDDHIVINAYK SAHOO, SAP ID- 70021119051, of B.E. (CSE) 3rd Year from SVKM NMIMS MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMNT AND ENGINEERING SHIRPUR, has successfully completed training and project work during the period 4th MAY to 4th JUNE 2021 on ***“QUESTION ANSWERING MODEL BASED ON ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING”*** at Design Quality department at Aircraft Upgrade Research and Design Centre (AURDC), Hindustan Aeronautics Limited (HAL), Nashik, as a part of Industrial Internship.

PROJECT GUIDE HEAD OF DEPARTMENT

Mr. M. RAJENDER Mr. S.J.BHOLE

DY. GENERAL MANAGER ADDL. GENERAL MANAGER

(DESIGN) (DESIGN)

# **ACKNOWLEDGEMENT**

This is the report of the project I undertook during Internship for duration of 4 weeks at Hindustan Aeronautics Limited, Nashik.

I would like to express my gratitude to my project guide **Mr. M. RAJENDER,DY. GENERAL MANAGER (DESIGN)** who has guided me through each phase of my project. His guidance, suggestions and invaluably problem solving and friendly advice during our discussions, are much focused to give me a proper direction to analyze and understand the project objective.

I am grateful to **TRAINING DEPARTMENT** for giving me the opportunity to undertake the project work at “M/s Hindustan Aeronautics Limited, Aircraft Division, Nashik, India”.

I would also like to extend my sincere thanks to all the staff members and employees.

**ABSTRACT**

The Aircraft Manufacturing and Overhaul is a complex and technical intensive job. At various stages of aircraft manufacturing and overhaul faults do occur due to various reasons. To resolve these faults technician have to go through a lot of manuals and paper work .

Since this is monotonous and time consuming sometimes takes very long time to go through all the documentation and find out the desired solution. It is proposed to implement Machine Learning/Artificial Intelligence/Natural Language Processing model to automate this long procedure .Due to this project technician can upload the paperwork and manuals on their local machine and directly search for answers.

This project is an implementation of Natural Language Processing (NLP) field of Artificial Intelligence. This project aims to develop, train, and train a question answering model . This is a simple implementation of search engine

**TABLE OF CONTENTS**

1.ACKNOWLEDGEMENT

2.ABSTRACT

3.ABOUT THE COMAPANY

4.INTRODUCTION

5.SYSTEM CONFIGURATION

6.PROJECT

1.PROBLEM STATEMENT

2.MODEL 1

* UNDERSTANDING THE DATA
* EXPLORING THE ENTITIES
* PREPROCESSING
* NATURAL LANGUAGE PROCESSING
* OUTPUT

3. MODEL 2 :YES OR NO ANSWERING ON LIMITED VOCABULARY USING NEURAL NETWORKS

* UNDERSTANDING THE DATA
* TOKENINZING THE DATA
* VECTORIZING THE DATA
* CREATING THE MODEL
* MODEL SUMMARY
* MODEL ACCURACY PLOT
* TESTING THE MODEL

4. MODEL 3:USING PRETRAINED BERT MODEL ON SQUAD DATASET

* TRANSFER LEARNING AND BERT MODEL
* PREPROCESSING
* CREATING MODEL AND FINE TUNING
* EVALUATING ON TEST DATASET
* PREDICTIONS

7. CONCLUSIONS

8. REFERENCES

# **ABOUT COMPANY**

Hindustan Aeronautics Limited (HAL) is a premier Aerospace company in Asia which is engaged in Design, Development and Manufacture of military and civil aircraft for over 75 years. It was established as Hindustan Aircraft in Bangalore in 1940 by Seth Walchand Hirachand to produce military aircraft for Indian Air force.

HAL is an organization where integrated airborne systems in the form of fighter aircraft and helicopters are conceived, developed, manufactured and serviced. It is one of the few corporate giants in Asia whose capabilities span the entire range of activities from product conception to after sale support. HAL is also involved in manufacture and assembly of structure required for India’s Space programs.

Hindustan Aeronautics Limited is the largest Public Sector Unit (PSU) under Ministry of Defense and is a Navratna Company. The company takes up maintenance and overhaul services to cover the life cycle requirement of all old and new products. Presently, 13 types of aircraft/helicopters and 17 types of engines are being overhauled. Additionally, facilities for repair/overhaul of various accessories and avionics integrated on aircraft of Russian, Western and Indigenous designs are also provided. HAL has 20 production Divisions, 10 Research and Development Centers and one Facility Management Division.

HAL Nashik has 3 divisions:

* Aircraft Manufacturing Division (AMD)
* Aircraft Overhaul Division (AOD)
* Aircraft Upgrade Research and Design Centre (AURDC)

Aircraft Division Nasik, established in the year 1964 for license manufacture of MiG-21FL aircraft & K-13 Missiles, is located at Ojhar, 24 kilometers from Nasik and approximately 200 kilometers from Mumbai in the state of Maharashtra. The division since then manufactured other MiG variants; viz MiG-21M, MiG-21 BIS, MiG-27 M and the state-of-the-art aircraft i.e. Su-30 MKI. Along with manufacturing, the division also carries out overhaul of the MiG series aircraft and started ROH of Su-30 MKI.

With the introduction of New a state-of-the art project i.e. Su-30 MKI and for smooth activities and operation of the division then  Aircraft  Division, Nasik is de-lineated into two Divisions viz Aircraft Manufacturing Division (AMD) for manufacturing activities  and Aircraft Overhaul Division (AOD) for Repair & Overhauling (ROH) activities.

|  |  |
| --- | --- |
|  |  |
| **SU-30MKI** | **SU-30MKI Touchdown** |

The Division is currently engaged in the following activities:

* Manufacturing of Su-30 MKI Aircraft.
* Supply of spares / units /consumables and aerospace fasteners.
* Support to HAL and Non-HAL made aircraft of Russian Origin in terms of Life extension, Modifications, Site Repairs and investigations.
* Diversification into Civil Aircraft manufacture by taking up work packages.
* Export of MiG spares to Egypt, Syria, Vietnam, Malaysia, Algeria, Poland & Russia.

|  |  |
| --- | --- |
| Russian official confirms talks to sell 21 MiG-29 fighters to IAF - The Week |  |
| **MiG 29** | **MiG 27 flying in Formation** |

**AURDC Division:**

* AURDC was established as Design Department in 1964 to provide design support to the manufacturing program, it has grown into a full-fledged R&D center named Aircraft Upgrade Research and Design Centre (AURDC). With vast experience on Su30MKI, MiG series aircraft in design and development for product improvements, role capability enhancement, indigenization, technology up-gradation, structural integrity studies for life extension, flight test analysis and mid-life upgrade are its strengths.
* AURDC over the years gained experience in design and development of new airborne system, Integration of systems and weapons in the aircraft. Evolved many modifications in terms of reliability, maintainability, operation, capability, enhancement and etc.

# **INTRODUCTION**

Aircraft manufacturing and overhauling is a diverse and vast process. In each step of the process, we need to make sure the quality and standards are maintained with utmost importance. Furthermore, each part or equipment of the aircraft is manufactured locally at the Aircraft Manufacturing Division (AMD) of HAL, Nashik, or has to be sourced from another HAL manufacturing facilities, and some specialized licensed parts are also sourced from. With this level of the abstruse, multistage process it fairly regular for minute things to go wrong , and resolving each of this error technician have to refer to a lot of documentation , manuals which consumes a lot of time and slows down the process of production . We will try to automate this process of searching for the answers by implementing question answering models using natural language processing .

[Question](https://en.wikipedia.org/wiki/Question) answering (QA) is a computer science discipline within the fields of [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval) and [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) (NLP), which is concerned with building systems that automatically answer questions posed by humans in a [natural language](https://en.wikipedia.org/wiki/Natural_language).

Question answering research attempts to deal with a wide range of question types including: fact, list, [definition](https://en.wikipedia.org/wiki/Definition), *How*, *Why*, hypothetical, semantically constrained, and cross-lingual questions.

* *Closed-domain* question answering deals with questions under a specific domain (for example, medicine or automotive maintenance), and can exploit domain-specific knowledge frequently formalized in [ontologies](https://en.wikipedia.org/wiki/Ontology_(information_science)). Alternatively, *closed-domain* might refer to a situation where only a limited type of questions are accepted, such as questions asking for [descriptive](https://en.wikipedia.org/wiki/Descriptive_knowledge) rather than [procedural](https://en.wikipedia.org/wiki/Procedural_knowledge) information. Question answering systems in the context of machine reading applications have also been constructed in the medical domain, for instance related to Alzheimer's disease.
* [*Open-domain*](https://en.wikipedia.org/wiki/Open_domain#References) question answering deals with questions about nearly anything, and can only rely on general ontologies and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer.

In this project we will try to implement the closed domain question answering model where the model is suppose to answer the question based on the given context .

# **SYSTEM CONFIGURATION, SOFTWARE AND EQUIPMENT**

PC Model: HP ZBook 17

Processor: Intel(R) Core(TM) i7-4600M 2 Core(s) 4 Logical Processor(s) @2.9GHz

RAM: 8 GB

OS: Windows 7 Professional 64-bit (6.1, Build 7601)

Display: 1920\*1080 (32 bit) (60Hz)

Software used:

* Anaconda Navigator/Prompt
* Jupyter Notebook

Programming Language:

* Python 3.8.5

Datasets:

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

# 

# **PROJECT**

## Problem Statement

***To Design and Develop a Machine Learning (Artificial Intelligence) Model for implementing Comprehension Question Answering Model using Natural Language Processing and neural network. Squad dataset is used for both training and validation of the model***

***MODEL 1 : Basic Model using NLP Libraries Spacy ,NLTK***

This model works a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities.

For this model an article from Wikipedia is given as a comprehension and based on the comprehension model tries to answer the asked questions in the form of a chatbot.

**UNDERSTANDING THE DATA:**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**EXPLORING THE ENTITIES:**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**PREPROCESSING:**

While preprocessing the following changes in the sentences of the comprehension is been done:

1. Senetence.lower : To convert the sentence to lower case
2. “ . ” are replaced by space
3. Stop words , punctuation marks, numbers are removed from the article :Stop words are the commonly used words in a language i.e. a, the, is , are ,there , here..etc

Graphical user interface

Description automatically generated

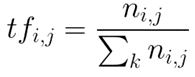
**NATURAL LANGUAGE ROCESSING**

TFIDFVectorizer:

TFIDF works by proportionally increasing the number of times a word appears in the document but is counterbalanced by the number of documents in which it is present. Hence, words like ‘this’, ’are’ etc., that are commonly present in all the documents are not given a very high rank. However, a word that is present too many times in a few of the documents will be given a higher rank as it might be indicative of the context of the document.

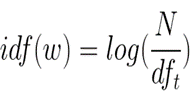
*Term Frequency:*

Term frequency is defined as the number of times a word (i) appears in a document (j) divided by the total number of words in the document.

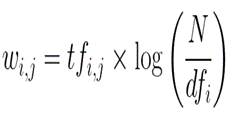


*Inverse Document Frequency:*

Inverse document frequency refers to the log of the total number of documents divided by the number of documents that contain the word. The logarithm is added to dampen the importance of a very high value of IDF.



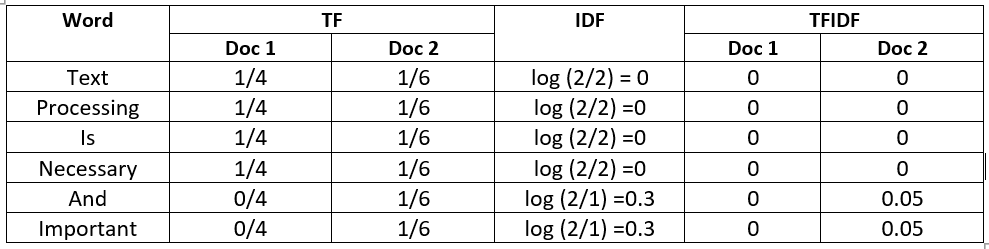
TFIDF is computed by multiplying the term frequency with the inverse document frequency.



Let us now see an illustration of TFIDF in the following sentences, that we refer to as documents.

**Document 1:** Text processing is necessary.

**Document 2:** Text processing is necessary and important.



The above table shows how the TFIDF of some words are zero and some words are non-zero depending on their frequency in the document and across all documents.

The limitation of TFIDF is again that this vectorization doesn’t help in bringing in the contextual meaning of the words as it is just based on the frequency.

COSINE SIMILARITY:

**Cosine similarity** is a metric, helpful in determining, how similar the data objects are irrespective of their size. We can measure the [similarity between two sentences in Python](https://www.geeksforgeeks.org/python-measure-similarity-between-two-sentences-using-cosine-similarity/) using Cosine Similarity. In cosine similarity, data objects in a dataset are treated as a vector. The formula to find the cosine similarity between two vectors is –

Cos(x, y) = x . y / ||x|| \* ||y||

where,

* **x . y** = product (dot) of the vectors ‘x’ and ‘y’.
* **||x||**and**||y||** = length of the two vectors ‘x’ and ‘y’.
* **||x|| \* ||y||** = cross product of the two vectors ‘x’ and ‘y’.

WORKING:

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

OUTPUT:

Table

Description automatically generated

***MODEL 2 : Yes or No Answering on a limited vocabulary using deep learning and neural network***

Loading the dataset:Graphical user interface, text, application

Description automatically generatedUNDERSTANDING THE DATA:Graphical user interface, text, application

Description automatically generated

Creating the vocab :

Graphical user interface, text, application, email

Description automatically generated

TOKENIZING THE DATA:

Graphical user interface, text, application

Description automatically generated

Preparing the training dataset:

Graphical user interface, text, application

Description automatically generated

VECTORIZING THE DATA:

Graphical user interface, text, application

Description automatically generated

Creating the Model

1. Creating the encoders

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

1. Creating the model:

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with low confidence

Model.summary:

Table

Description automatically generated

MODEL ACCURACY PLOT:

Chart

Description automatically generated

Testing the model:

Graphical user interface, text, application, email

Description automatically generated

Making prediction using he vocabulary :

Graphical user interface, text, application, email

Description automatically generated

***MODEL 3 :Using pretrained bert model on the SQuAD dataset***

TRANSFER LEARNING AND BERT MODEL:

Transfer learning:

Transfer learning is a machine learning method where we reuse a pre-trained model as the starting point for a model on a new task. To put it simply—a model trained on one task is repurposed on a second, related task as an optimization that allows rapid progress when modeling the second task.

BERT Model:

BERT is an open source machine learning framework for [natural language processing](https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP) (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection. (In NLP, this process is called attention.

PREPROCESSING:

Text, timeline

Description automatically generated

CREATING MODEL AND FINE TUNING :

Graphical user interface, text, application, email

Description automatically generated

EVALUATING ON TEST DATASET :

Text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

PREDICTIONS: Graphical user interface, text, application

Description automatically generated

Scatter chart

Description automatically generated

Graphical user interface, application

Description automatically generated

# **CONCLUSION**

The project titled “**QUESTION ANSWERING MODEL BASED ON ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING**” has been completed within the stipulated time of the project. The objective of the project was to Design and Develop an Artificial Intelligent model which should be able to answers questions from the given context , SQUAD dataset has been used for training and testing model. The model was successfully developed and is ready to be put in real-world systems.

This internship has given me a new perspective on various topics regarding military aircrafts. It has given a lot of exposure to the industrial side of a corporation. Through this, I got acquainted with the workflows, procedures, and safety measures involved in the industry. The Training Department gave me a brief knowledge about HAL, Nashik. The AURDC division gave me an in-depth overview of aircraft designing and various avionic system integration in aircrafts.

# **REFERENCES**

To gather information about the company and fighter aircrafts-

<https://hal-india.co.in/>

<https://en.wikipedia.org/wiki/Hindustan_Aeronautics_Limited>

<https://en.wikipedia.org/wiki/Sukhoi_Su-30MKI>

Theoretical knowledge and understandings-

<https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

<https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>

<https://scikit-learn.org/stable/modules/naive_bayes.html>

https://huggingface.co/blog/bert-101

Documentation of various python 3 libraries used in the process-

<https://docs.python.org/3/>

<https://pandas.pydata.org/docs/>

<https://numpy.org/doc/>

<https://scikit-learn.org/stable/>

<https://matplotlib.org/>

<https://seaborn.pydata.org/>