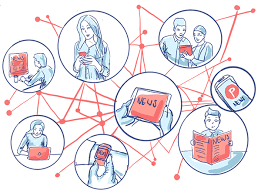
**Architecture Design**

News Articles Sorting



**Revision Number - 1.0**

**Last Date of Revision – 07-07-2022**

**SACHIN**

# Document Control

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** |  | **Version** | **Description** | **Author** |
| 07-07-2022 | 1.0 |  | Introduction,  Architecture,  Deployment | Sriphani |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**2**

**Contents**

Document Version Control…………………………………………………………………………2

1. Introduction……………………………………………………………………………………….4
   1. What is Architecture Design Document?..............................................................4
   2. Scope……………………………………………………………………………………4
2. Architecture……….……………………………………………………………………….……...5
   1. NLP Architecture………………………………………………………………….5
   2. Components of NLP Architecture……………………………………………….5

3. Deployment……….……………………………………………………………………….……...6

* 1. NLP Deployment………………………………………………………………….6
  2. Publish datasets and reports from Power BI Desktop………………………………7

# 1. Introduction

## 1.1 What is Architecture Design Document?

Any software needs the architectural design to represent the design of the software. IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.” The software that is built for computer-based systems can exhibit one of these many architectures.

Each style will describe a system category that consists of:

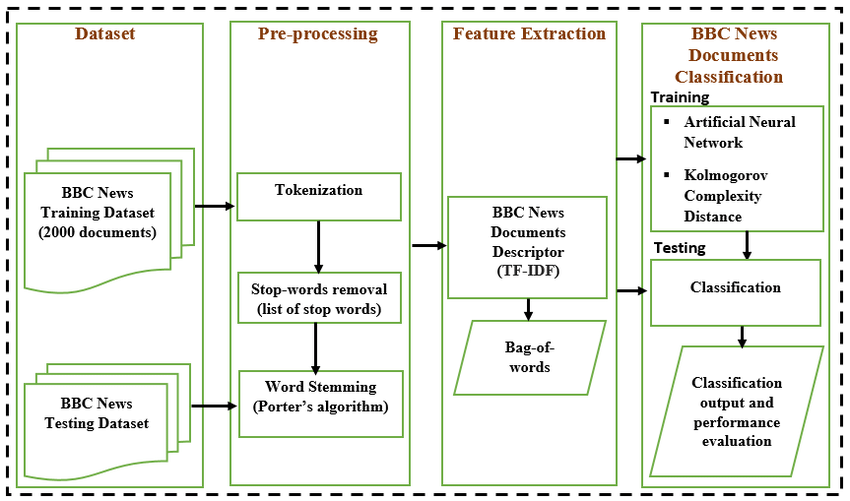
* A set of components (eg: a database, computational modules) that will perform a function required by the system.
* The set of connectors will help in coordination, communication, and cooperation between the components.
* Conditions that how components can be integrated to form the system.
* Semantic models help the designer to understand the overall properties of the system.

**1.2 What is Scope?**

Architecture Design Document (ADD) is an architectural design process that follows a step-by-step refinement process. The process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the design principles may be defined during requirement analysis and then refined during architectural design work.

# 2. Architecture

## 2.1 NLP Architecture



News Collection

The first step of news classification is accumulating news from various sources. This data may be available from various sources like newspapers, press, magazines, radio, television and World Wide Web and many more. But with the widespread network and information technology growth internet has emerged as the major source for obtaining news. Data may be in available in any format i.e. it may in .pdf, .doc, or in .html format.

News Pre-processing

After the collection of news text pre-processing is done. As this data comes from variety of data gathering sources and its cleaning is required so that it could be free from allcorrupt and futile data. Data now needs to be discriminated from unrelated words like semicolon, commas, double quotes, full stop, and brackets, special characters etc. Data is made free from those words which appear customarily in text and are known as stop words.

News Tokenisation

News tokenization involves fragmenting the huge text into small tokens. Each word in the news is treated as a string. The output of this step is treated as input for the next steps involved in text mining.

Stop Word Removal

The stop words language specific and does not carry any information. It generally includes conjunctions, pronoun and prepositions. They are contemplated of low worth and are removed eventually. These words need to be percolate before the processing of data.

Stop words can be removed from data in many ways. There removal can be on the basis of concepts i.e. the removal will be of the words which provide very fewer information about classification. Another way of removal of stop words is the removal of the words that are present in the list of English stop words. The list is made up of approx 545 stop words and is provided by Journal of Machine Learning Research. Stop words can also be abolished depending upon the frequency of their occurrence. In this method frequency of occurrence of words is computed and then weights are assigned to words. Then depending on these

weights the stop words are dropped.

Word Stemming

After the removal of stop words the next activity that is performed is stemming. This step reduces a word to its root. The motive behind using stemming is to remove the suffixes so that the number of words would be brought down.

be brought down. For example the words like user, users, used, using all can be reduced to the word “USE”. This will reduce the required time and space.

For stemming there exists many stemmers like S-Stemmers, Lovins Stemmer , Porter Stemmer , Porter Stemmer, Paice/Husk Stemmer. Among these stemmers M.F. Porter is mostly used.

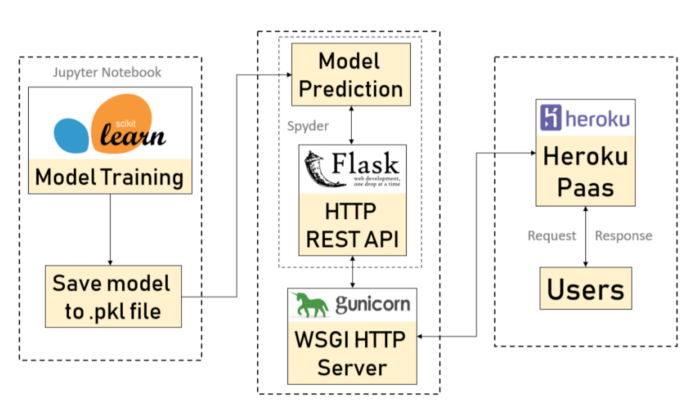
Porter Stemmer: This is the most widely used stemmer because of its high precision and simple algorithm. It includes five steps which are easy to implement and are applied to each and every word.

.

# 3. Deployment

## 3.1 NLP Deployment

The deployment process lets you clone content from one stage in the pipeline to another, typically from development to test, and from test to production.



.

## Publish datasets and reports from Heroku

