Low Level Design

News Article Sorting

Written By	Rajat Chaudhari
Document Version	1.0
Last Revision Date	14/04/2022

Table of Contents

News Article Sorting 1

1. Introduction 3
2. Scope 3
3. Architecture 3
4. Architecture Description 4
   - A. Data Description 4
   - B. Data Insertion into Database 4
   - C. Data Pre-processing 4
   - D. Model Building 4
   - E. Data Validation 4
   - F. Deployment 4
5. Test Case 5

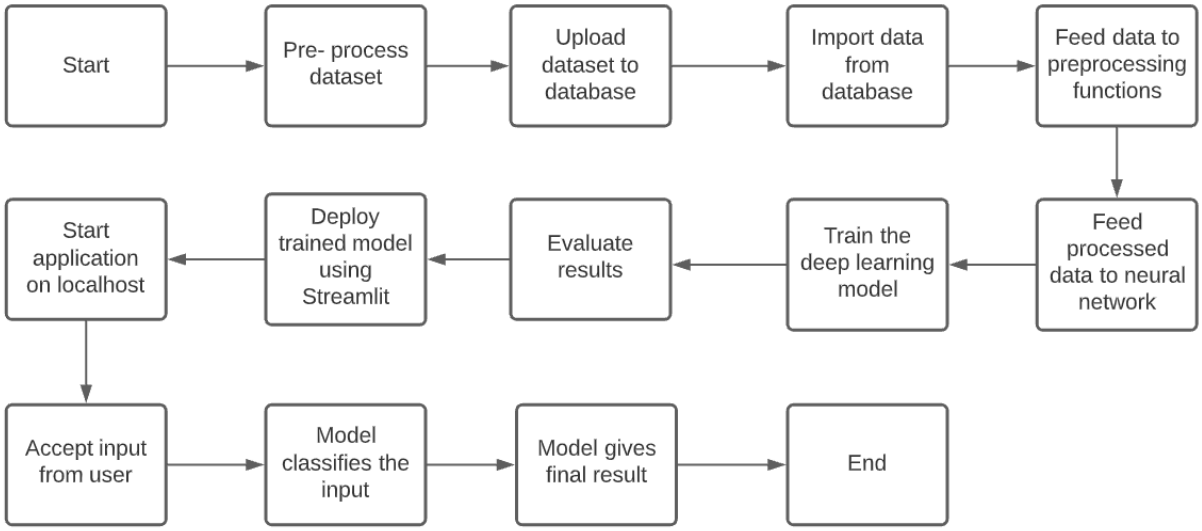
1. Introduction

The purpose of this Low Level Design Document (LLD) is to give the internal logical design of the program code for News Article Sorting System. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

2. Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

3. Architecture



4. Architecture Description

A. Data Description

The training set consists 1490 records, and the test set consists of 736 records. There are three data fields in the dataset:

I. ArticleId
II. Article
III. Category

B. Data Insertion into Database

DataStax Astra is a Cassandra database, which is used for this project to store the data. The data is uploaded to the database in the form of a csv file. A script is written to connect to the database, and retrieve the data from the database.

C. Data Pre-processing

The text data is pre-processed before being fed to the model. The following operations are performed on the text:

i. Lower case
ii. Removing special characters
iii. Removing stop words
iv. Lemmatization
v. Tokenization

D. Model Building

K-Fold cross validation is used to validate the model on the dataset. The model used for the news article sorting system is a neural network. The model accepts text data as input and one of the five categories as an output.

E. Data Validation

The users will have to do data validation of the input themselves before supplying it to the model.

F. Deployment

The model is deployed on the cloud using Streamlit package.

5. Test Case

Test ID	Description	Expected Result	Result
T-001	The text box of the website has placeholder text for the category – ‘sports’. Simply click on the ‘Predict’ button and see what the prediction is. This tests the main functionality of the system, prediction of a news article.	Bar chart for percentage of confidence model has for each category. The ‘sports’ category should have the highest percentage. Below the bar chart, prediction ‘sport’ should be displayed in a green rectangular element.