Architecture Design Document

News Article Sorting

Written By	Rajat Chaudhari
Document Version	1.0
Last Revision Date	15/04/2022

Contents

1. Introduction

1.1 Architecture Design

The architecture design document is a technical document describing the components and specifications required to support the solution and ensure that the specific business and technical requirements of the design are satisfied.

1.2 Scope

This document serves as the implementation plan project. This document identifies points of contact for the project, lists implementation requirements, provides a brief description of each of the document deliverables, deliverables, and provides an overview of the implementation process for the data-center virtualization project.

1.3 Constraints

This News Article Sorting project relies on only text information to classify the articles in one of the five categories (business, sport, technology, entertainment, and politics).

2. Technical Specifications

2.1 Dataset

The data consists 1490 records for training set, 736 records for test set. The data fields in the dataset are:

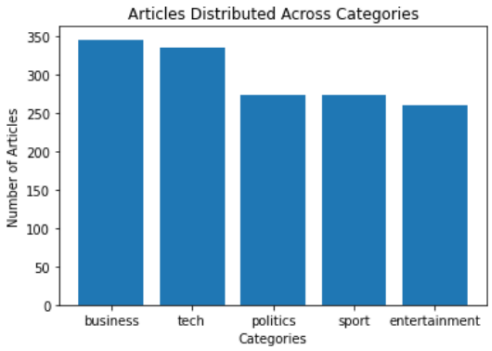
(i) ArticleID

(ii) Article

(iii) Category

	article_id	news	category
0	1584	alicia keys to open us super bowl r&b star ali...	entertainment
1	1863	johnson uncertain about euro bid jade johnson ...	sport
2	1765	nintendo ds aims to touch gamers the mobile ga...	tech
3	2062	turkey deal to help world peace a deal bring...	politics
4	1199	sales fail to boost high street the january ...	business

The articles are distributed across the 5 categories in the following way.



The articles were preprocessed to remove special characters before getting uploaded on the database.

2.2 Logging

Python library logging is used to log different kinds of messages to a log file.
There are 5 levels of logging messages in the logging python library, namely:

1. DEBUG
2. INFO
3. WARNING
4. ERROR
5. CRITICAL

For this project, the level is set to INFO. All the logging messages are written to logs.txt file.

2.3 Database

A Cassandra database is used to store the data for the News Article Sorting system. The data is uploaded on the database in the form of a csv file. Astra Datastax database is used for this purpose.

2.4 Deployment

Streamlit python package is used to deploy the News Article Sorting model on the cloud.

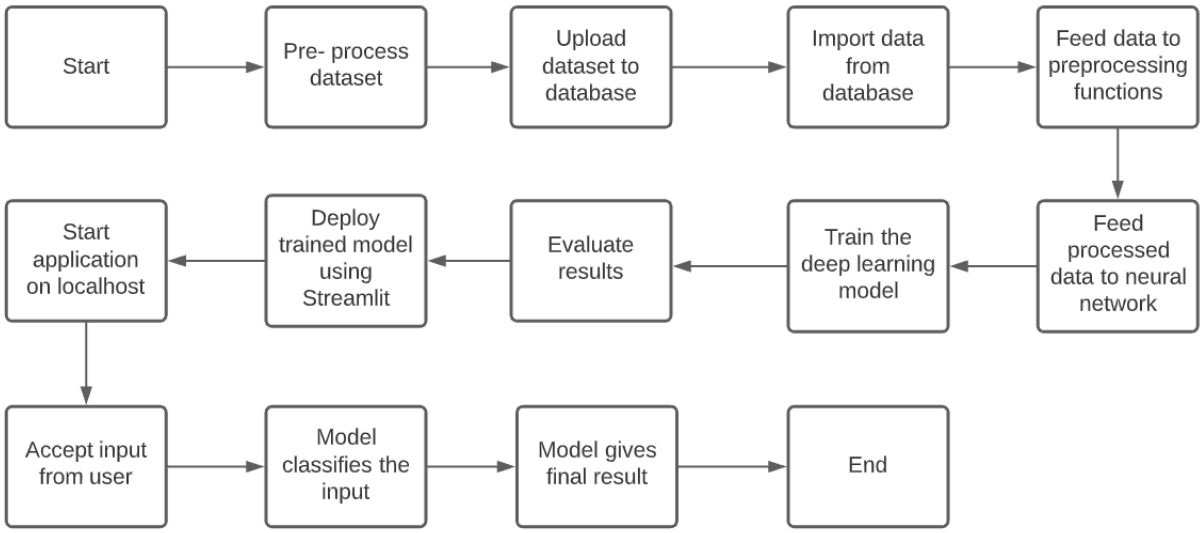
3. Technology Stack

Frontend + Backend	Streamlit
Deep Learning Framework	Tensorflow
Database	Astra DataStax

4. Proposed Solution

The data will be pulled from the database, preprocessed, and EDA will be performed on it. Then a classifier will be trained on the data, and tuned to get the highest accuracy. The classifier model will then be hosted on the cloud using Streamlit, a service which will provide a UI (user interface) to enable ease of use for the model. Streamlit will then accept an input, send it to the classifier model to process, and then will provide the prediction as the output.

5. Architecture



6. User Input/ Output Workflow