Low Level Design Document (LLD)

BBC News Articles Sorting

Revision Number: 1.3
Last date of revision: 24/07/2022

Rushikesh Darge

Contents

Introduction **4**

- Why this Low-Level Design Document? 4
- Scope 5
- Constraints 5
- Out of Scope 5

Technical specifications **6**

- 2.1 Dataset 6
- 2.1.1 Diabetes dataset overview 6
- 2.2 Predicting Category 6

Technology stack **7**

Proposed Solution **8**

Model training/validation workflow **9**

User I/O workflow **11**

Key performance indicators (KPI) **12**

Abstract

Text documents are one of the richest sources of data for businesses: whether in the shape of customer support tickets, emails, technical documents, user reviews or news articles, they all contain valuable information that can be used to automate slow manual processes, better understand users, or find valuable insights. However, traditional algorithms struggle at processing these unstructured documents, and this is where machine learning comes to the rescue!

1 Introduction

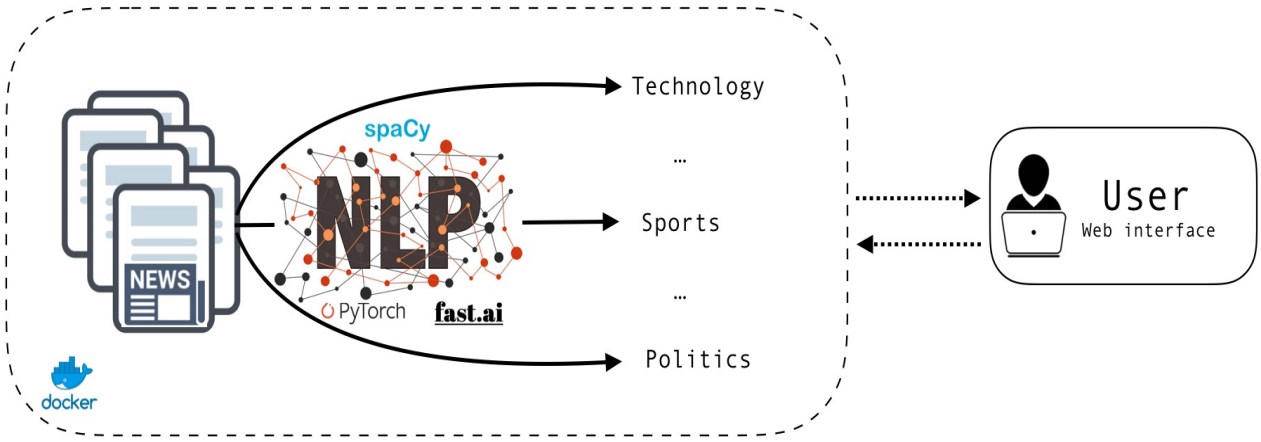
1.1 Why this Low-Level Design Document?

The purpose of this document is to present a detailed description of BBC News Articles Sorting System. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the higher management for its approval.

The main objective of the project is to predict the category of news based on the content of the news.

By automating sorting of news:

- We can use this system in different fields like scraping text and then sorting category wise.



This project shall be delivered in two phases:

Phase 1: All the models and its accuracy.

Phase2: Integration of UI with model..

1.2 Scope

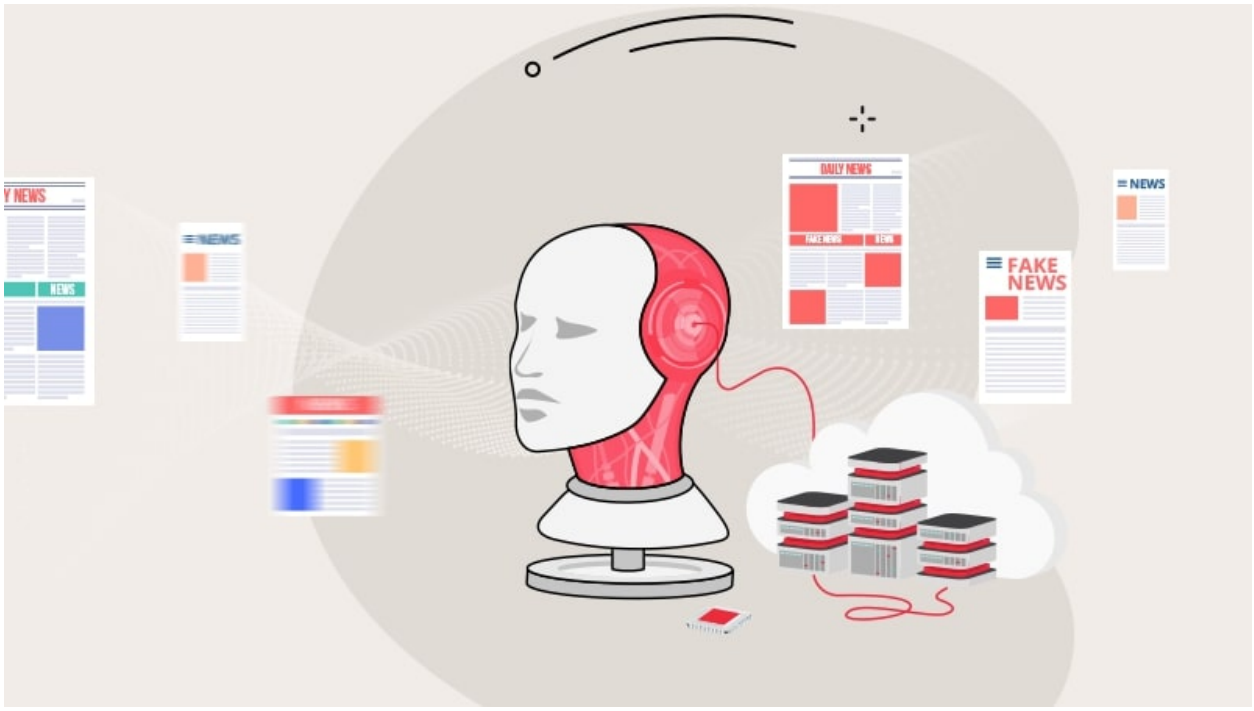
This software system will be a Web application This system will be designed to detect the category of news. This is done using the news content or it can be a headline or summary of specific news.

1.3 Constraints

We will only be classifying 5 categories of news.

1.4 Out of Scope

Creating new categories or identifying out of train categories is out of scope.



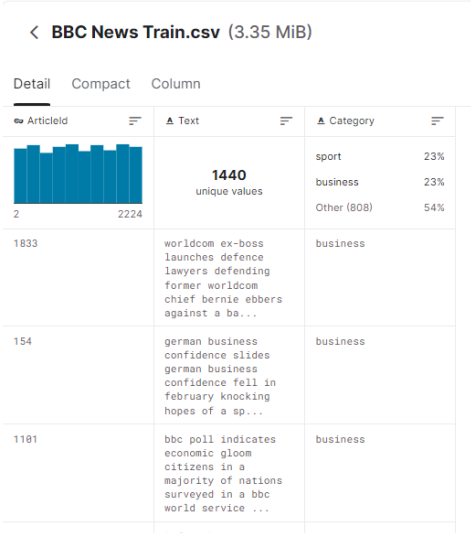
2 Technical specifications

2.1 Dataset

Dataset for this project is available on Kaggle. Which contain 1490 records with their 5 categories which are politics, business, entertainment, tech and sport.

2.1.1 Diabetes dataset overview

Dataset consists of 3 columns. Those are ArticleId, text and categories.



BBC News Train.csv (3.35 MiB)

Detail Compact Column

ArticleId	Text	Category
2 — 2224	1440 unique values	sport 23% business 23% Other (808) 54%
1833	worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a ba...	business
154	german business confidence slides german business confidence fell in february knocking hopes of a sp...	business
1101	bbc poll indicates economic gloom citizens in a majority of nations surveyed in a bbc world service ...	business

2.2 Predicting Category

We need to predict Categories for each news text.

2.3 Deployment

We are using Streamlit to build and host our model.

3 Technology stack

Front End	HTML/CSS/JS
Backend	Python Django
Database	MongoDB/MySql
Deployment	Streamlit

4 Proposed Solution

Ref :
https://www.researchgate.net/publication/326425709_Text_Mining_Use_of_TF-IDF_to_Examine_the_Relevance_of_Words_to_Documents

In this paper, the use of TF-IDF stands for (term frequency-inverse document frequency) is discussed in examining the relevance of key-words to documents in corpus. The study is focused on how the algorithm can be applied on a number of documents. First, the working principle and steps which should be followed for implementation of TF-IDF are elaborated. Secondly, in order to verify the findings from executing the algorithm, results are presented, then strengths and weaknesses of the TF-IDF algorithm are compared.

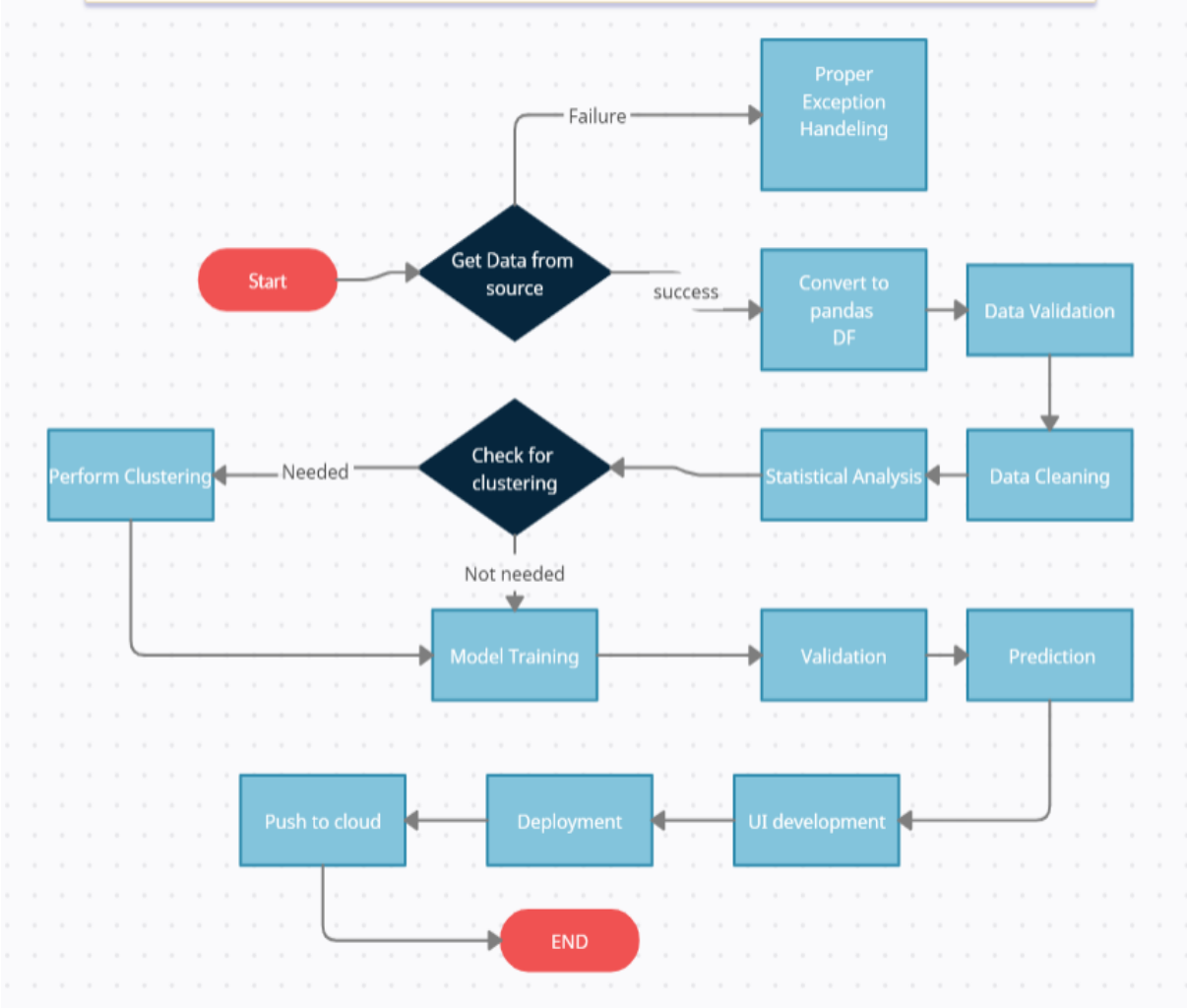
Ref:
https://www.researchgate.net/publication/351348412_Bengali_News_Classification_Using_Long_Short-Term_Memory

In this research paper researchers use LSTM to classify bengali news, and they got good F1 score for overall all the classes. They use the embedding layer, spatial dropout layer, LSTM layer, and dense layer to make their model. This model will help to classify five types of news. Using this model they try to help an online news portal.

Conclusions:

We decided at baseline models to use TF-IDF to convert text to numbers. And use machine learning algorithms to sort. Then we also experiment with LSTM layers with embedding layers.

5 Model training/validation workflow



TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Number of times term t appears in a doc, d

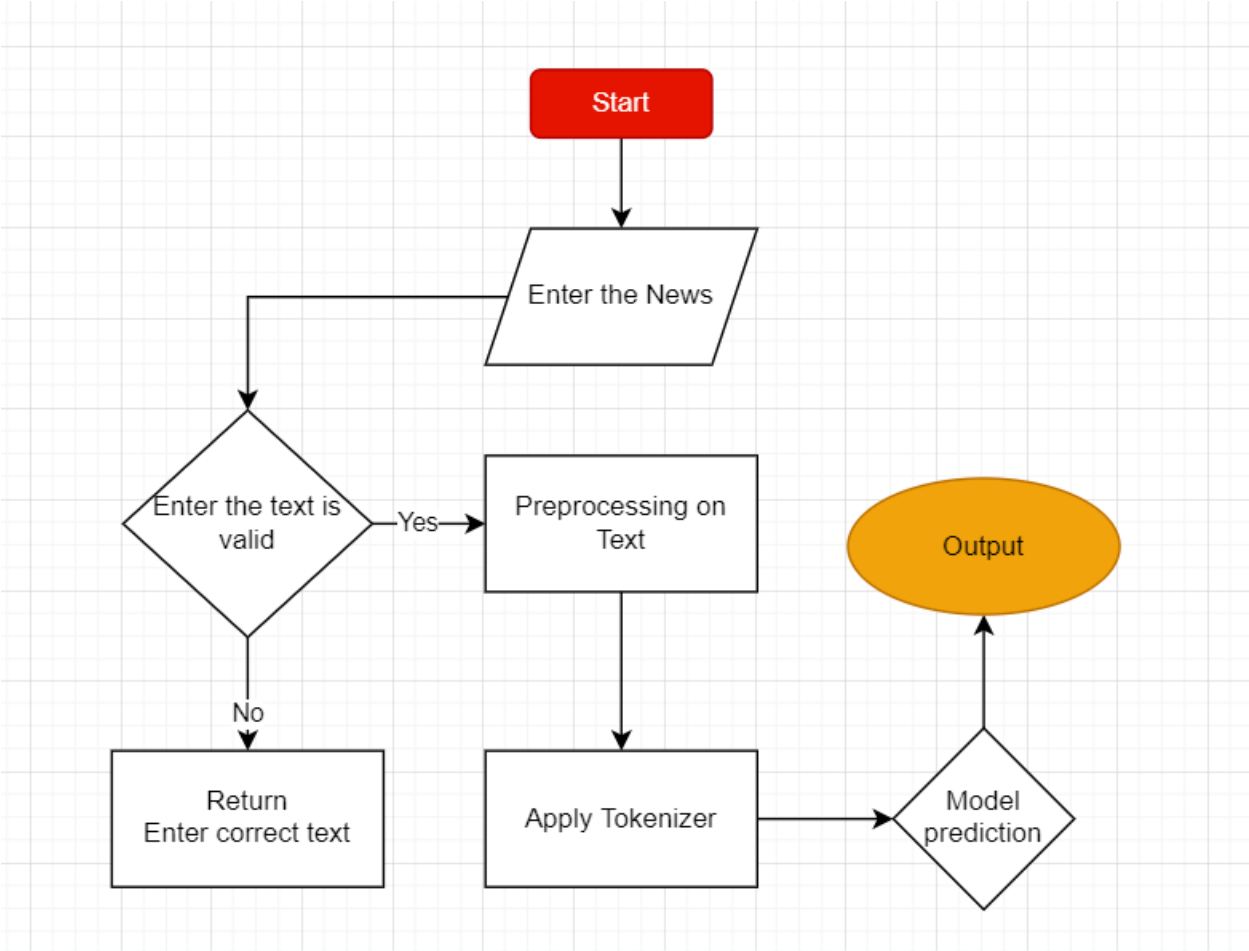
Inverse document frequency

$$\log \frac{1 + n}{1 + \text{df}(d, t)} + 1$$

n ← # of documents

df(d, t) ↖ Document frequency of the term t

6 User I/O workflow



7 Key performance indicators (KPI)

- Latency of model should be less because we need to sort lots of news articles
- Accuracy is also important.
- Interpretability is partially important.