High Level Design (HLD)

News Summarization

Document Version Control

Date Issued	Version	Description	Author
05/12/2022	1	HLD – V1.0	Diana Laveena DSouza

Contents

Abstract

Making news is hard enough, even if you don’t think about tight deadlines and thorough fact-checking. A news piece must meet specific editorial criteria, such as accuracy, timeliness, and availability of sources. On its own - writing a news piece is not a big deal. But there is much stuff going on in the world. And the reality is that a news media platform must deliver news in time to remain competitive and engage with the target audience. Thus, Text Summarization is a cost-effective and time-saving option for the media and journalists. Think about it as a helping hand for the journalist.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions before coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the system's structure, including the database architecture, application architecture, application flow and technology architecture. The HLD uses non-technical to mildly-technical terms, which should be understandable to the system's administrators.

1.3 Definitions

Term	***Description***
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment

GCP	Google Cloud Platform

2 General Description

2.1 Product Perspective

News Summarization is a Deep Learning Technology to save time, enhances readability and is cost-effective for media and journalists.

2.2 Problem Statement

To create a model that should perform extractive and abstractive summarization of the news articles from different reading categories.
.

2.3 Proposed Solution

The solution proposed here is to create a deep-learning model that summarizes various news articles from different reading categories.

2.4 Data Requirements

CORNELL NEWSROOM is a large dataset for training and evaluating summarization systems. It contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 38 major publications. The summaries are obtained from search and social metadata between 1998 and 2017 and use various summarization strategies combining extraction and abstraction.

We use the Cassandra database by inserting records into it and then exporting them to a CSV file for further use.

2.5 Tools used

Python programming language and frameworks such as NumPy, Pandas, and transformers are used to build the whole model.









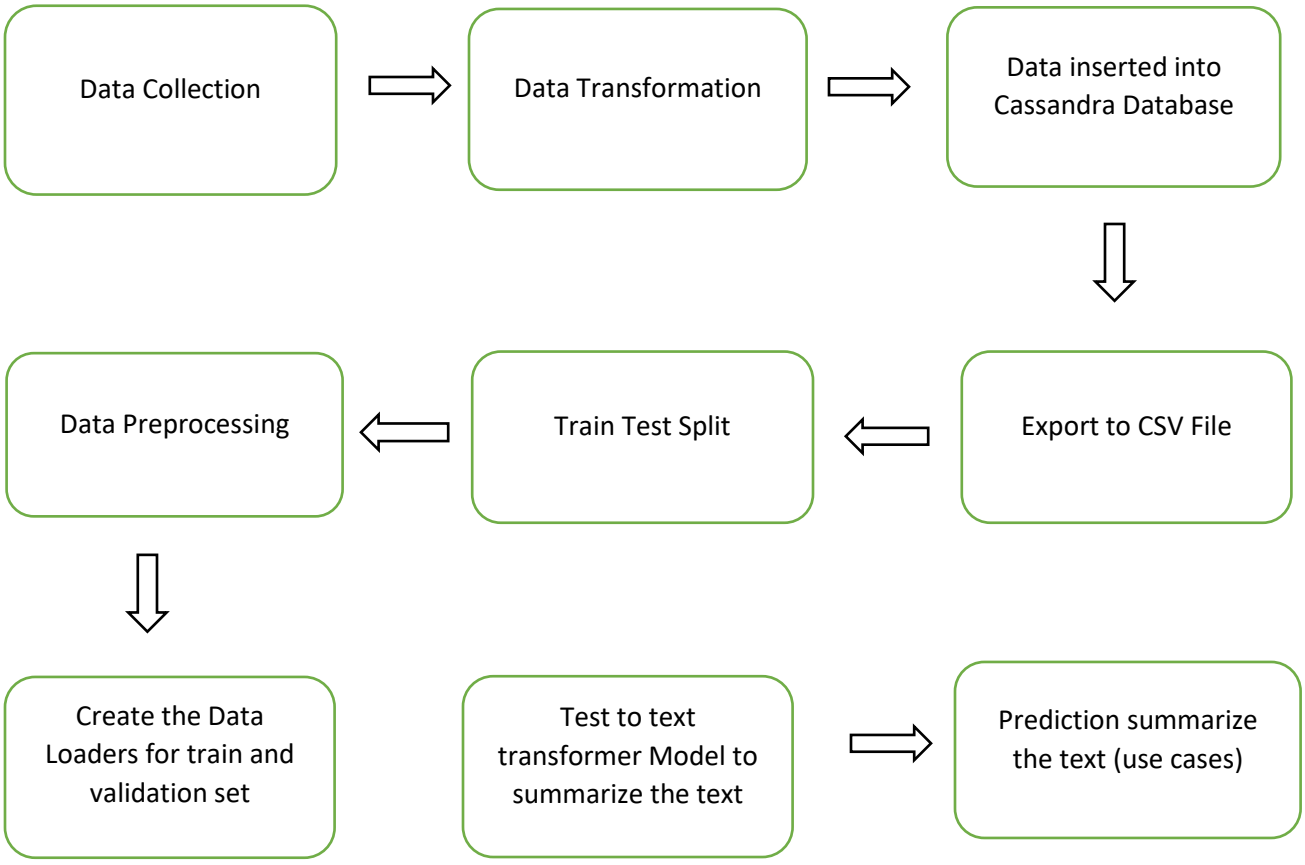

- PyCharm is used as IDE.
- GCP is used for the deployment of the model.
- Cassandra is used to retrieve, insert, delete and update the database.
- Front-end development is done using HTML.
- Python Flask is used for backend development.
- GitHub is used as a version control system.

3 Design Details

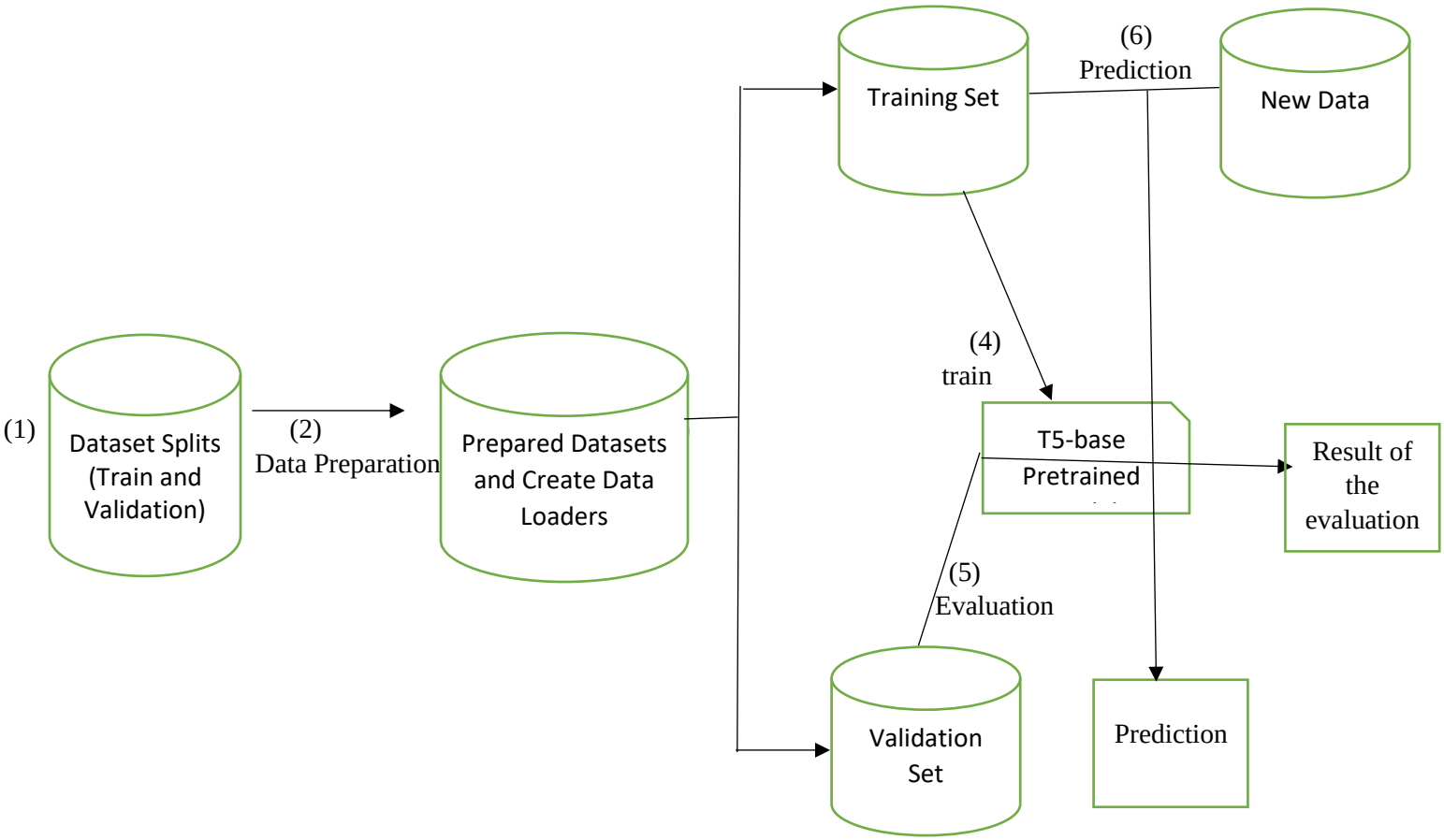
3.1 Process Flow

To summarize the article, we will use the deep learning-based transformer model. Below is the process flow diagram is shown below.

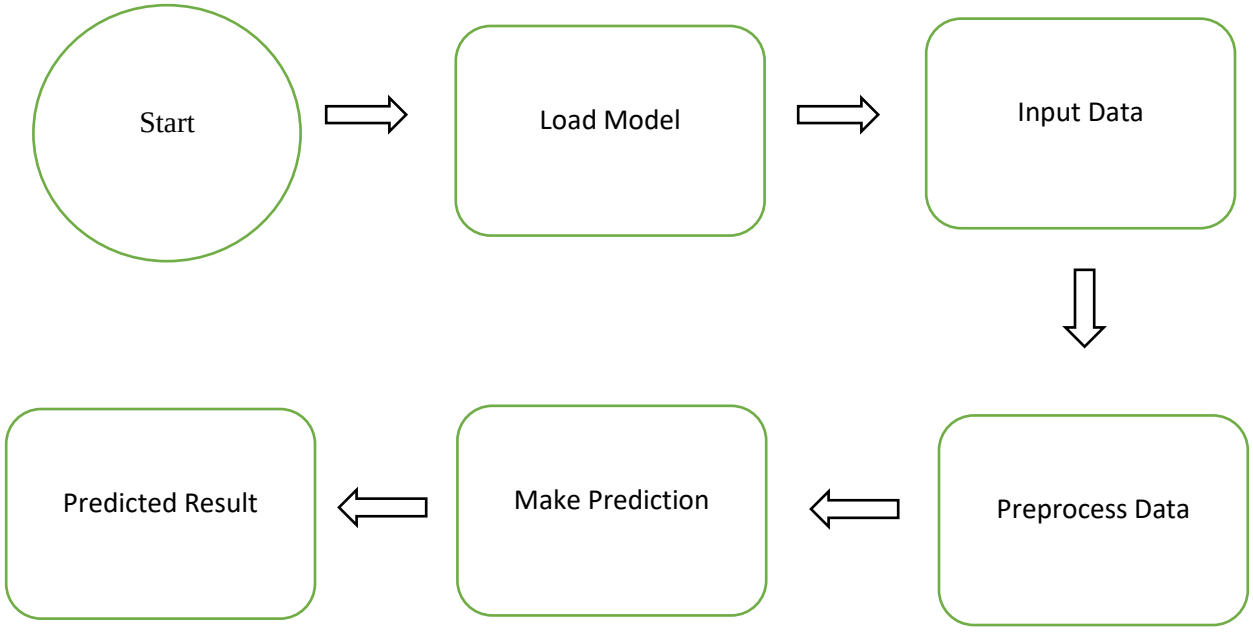
Proposed Methodology



3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event Log

The system should log every event, so the user will know what process is running internally.

Initial Step-By-Step Description:

1. The System identifies at what step logging is required.
2. The System should be able to log each system flow.
3. Developer can choose the logging method. You can choose database logging/File logging as well.
4. System should not hang even after using so many loggings. Logging just because we can easily debug issues, so logging is mandatory too.

3.3 Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong. An error will be defined as anything outside the normal and intended usage.

4 Performance

The t5-base transformer model is used for text-to-text generation. The model will summarizes the news articles belonging to different categories. It will save time and be cost-effective for the media and journalists.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will use Python as an interface between them. Each component will have its task to perform, and it is the job of Python to ensure the proper transfer of information.

4.3 Resource Utilization

When any task is performed, it will likely use all the preprocessing power available until that function is finished.

4.4 Deployment

5 Conclusion

The deep-learning model summarizes various news articles from different reading categories.

6 References

1. https://lil.nlp.cornell.edu/newsroom/download/index.html
2. https://www.kaggle.com/code/raryan/t5-abstractive-text-summarization
3. https://www.youtube.com/watch?v=SFJnm_ZLMrA
4. https://www.kaggle.com/code/abhinavkrjha/extractive-and-abstractive-text-summarization
5. https://www.kaggle.com/code/staefff/extractive-text-summarization-with-pagerank
6. https://huggingface.co/course/chapter7/5?fw=pt