

Time Series Anomaly Detection and RAG System for AI-Driven Governance Insights Projects by Aspire to Innovate (a2i)

by

Md. Aber Islam
20301102

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Aber Islam
20301102

Approval

The thesis titled “Time Series Anomaly Detection and RAG System for AI-Driven Governance Insights Projects by Aspire to Innovate (a2i)” submitted by

1. Md. Aber Islam(20301102)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 31, 2025.

Examining Committee:

Supervisor:
(Member)

Md. Anowarul Arif Khan
Results Management Analyst and Cluster Lead
Data & Research
Aspire to Innovate (a2i) Programme

Co-Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In the field of Machine learning, there are many applications where we can add our trained model to use. One of the most useful this is using LLM models but often these LLM models can not give us the latest answer or answer based on our needs. This is where RAG comes in this gives LLM models context. It works like a helping hand to the LLM models giving it proper context so that it performs based on our needs.

Many advancements in machine learning and artificial intelligence have occurred in recent years. To make data-driven decisions, we need a lot of data. Collecting data is a big part of Machine learning and data science. The large language model needs a lot of data to make meaningful answers. Web scraping is a big part of that. Finding data for any Bangla language can be hard as there are not many sources that provide it. For this, we are highly dependent on web scraping. Our project-specific data will be gathered by scrapping the popular Bangla news portal website. This project aims to scrap a news portal website from a given date range. The user can select the range of the data and which news portal they want to scrap from.

Time series analysis and prediction is also one of the crucial aspects of machine learning oftentimes time it does not follow traditional norms of prediction models like regression and classification. Throughout my internship, I have worked with time series prediction models which helps businesses to predict the future and be prepared.

Throughout my internship period, I was involved in various projects like time series analysis and prediction, sentiment analysis, creating a scrapping website, and making RAG with LLM models. This report contains all the necessary information and processes of those projects which shows my journey as a machine learning intern in a2i.

Keywords: Machine Learning; Anomaly Prediction; Web Scraping; RAG; LLM;

Dedication

I would like to dedicate my efforts and struggles in my educational life to my dear parents, who have helped and supported me throughout my life. I would also like to dedicate this internship report to my respected supervisor, Md. Anowarul Arif Khan, who guided and assisted me throughout my professional persona.

Acknowledgement

Before anything else, I want to express gratitude to the Almighty Allah, without whom my internship could not be concluded.

Furthermore, I would like to thank my supervisor, Md. Anowarul Arif Khan, Results Management Analyst and Cluster Lead, Data and Research, a2i, for his excellent guidance and support in helping me finish my internship duties and responsibilities. He supported me whenever I needed help.

Additionally, throughout the internship, I received guidance from Dr. Md. Golum Rabiul Alam Sir, who helped me understand the process and the internship guidelines. Finally, without our parent's unwavering support, it may not have been achievable. Thanks to their amazing encouragement and prayers.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Dedication	iv
Acknowledgment	v
Table of Contents	vi
List of Figures	viii
List of Tables	ix
Nomenclature	ix
1 INTRODUCTION	1
1.1 Preamble	1
1.2 Aims and Objectives	1
1.3 Information Sources	2
1.4 Document Overview	2
2 ORGANIZATION's PROFILE	3
2.1 Organization Overview	3
2.2 First day of joining a2i	4
3 WORKPLACE DETAILS	5
3.1 My Responsibility	5
3.2 Organization Structure	5
3.3 Communication and Meetings	5
3.4 Remote Office	5
3.5 Work-space Facilities	6
3.6 Conclusion	6
4 MY CONTRIBUTIONS	7
4.1 Overview	7
4.2 Getting Familiar with My Workflow	7
4.3 333 Data Analysis and Anomaly Prediction	8

4.3.1	Data Cleaning	9
4.3.2	Key Analysis	9
4.3.3	Feature Engineering	9
4.3.4	Daily Anomalies Over Time	11
4.3.5	Model Training	11
4.3.6	Model Evaluation	11
4.4	Bangal News Scraper	12
4.4.1	Architecture of Bangal News Scraper	12
4.4.2	Structure of the DataBase	14
4.4.3	User Interface	15
4.5	Bangla RAG System	16
4.5.1	Overview of User Query	16
4.5.2	Embedding of Database	18
4.5.3	Response of the system	18
4.6	Challenges	19
4.7	Learning	20
5	GROWTH	21
5.1	Professional Growth	21
5.1.1	Office environment	21
5.1.2	Teamwork	21
5.2	Technical Growth	21
5.2.1	NumPy and Pandas	21
5.2.2	Data Visualization Stack	21
5.2.3	YData Profiling	22
5.2.4	Scikit-learn Toolkit	22
5.2.5	Streamlit	22
5.2.6	Database Connectors	22
5.2.7	Web Scraping Stack	22
5.2.8	NLP and Embeddings	22
5.2.9	Vector Search	22
6	CONCLUSION	23
6.1	Conclusion	23
6.2	Recommendations	24
	Bibliography	25

List of Figures

4.1	Process of 333 project	8
4.2	Trends in the Number of Calls Over Time	10
4.3	Anomalies Over Time	10
4.4	Scraper Creation Flowchart	13
4.5	ER Diagram of the Website	14
4.6	User Interface of the Website	15
4.7	Process of User Query	17
4.8	Process of Data and Meta-Data	18

List of Tables

Chapter 1

INTRODUCTION

1.1 Preamble

An internship provides students a bridge between academic and professional life, where students can apply their theoretical knowledge to real-life problems. Current and fresh graduate students get the opportunity to work in their desired fields where they can gain valuable experience. This experience enhances their academic development and emphasizes skills like teamwork and communication. The experience gathered by the students can also be used to enhance their CV. Reconcilable companies also work with fresh graduates and students to know their requirements. Additionally, the companies work with young, talented individuals for future recruitment. During the internship, students get exposed to different aspects of the industry, like writing code, debugging software, and contributing to large-scale projects.

At BRAC University, it is mandatory to do CSE400, which can be either a thesis, internship, or project. Though the majority of the students prefer to do a thesis, some of the students are also interested in doing internships, as it has gained popularity in recent years. Even though a small percentage of students opt for internships, the university fully supports those students who take internships. According to the policy of BRAC University, a student must complete at least 72 credits to start the CSE400 internship and the internship must be in a reputable company with a track record to prove its reputation. The internship period must be 6 months long. Students can choose which organization they want to do an internship in as long as it is related to the major they are doing.

As a student of BRAC University's Computer Science and Engineering department, I chose to do an internship at a2i (Aspire to Innovate) as my internship organization. The Government of Bangladesh has developed a global catalyst for digital transformation, which is known as a2i. It's the Innovation Agency of the Government of Bangladesh. It is a major component of the government's Digital Bangladesh agenda, a special plan of the ICT Division and the Cabinet Division funded by UNDP. It is essential to the country's efforts to meet the 2030 Sustainable Development Goals and turn it into a developed nation by 2041 [11]

1.2 Aims and Objectives

Because we know that finishing a bachelor's degree requires an internship, I must present information in a report so that BRAC University may evaluate me. Through-

out my report, I have highlighted my employment, including the surroundings and office culture. I also explained what I learned from a2i and my own experience. It will provide a comprehensive summary of the activities I have performed and projects I have worked on throughout my internship. This report aims to narrate my work experience, the challenges I have faced, and the knowledge I have gained during my internship.

1.3 Information Sources

The majority of the information used in this report was gathered by me during my internship at a2i. Mostly, I learned about everything at the organization through conversations and teachings from coworkers, team managers, and members. Secondary sources gathered include the a2i website and other internet resources.

1.4 Document Overview

In this report, I detailed my professional experience with a2i. I also shared my accomplishments, expertise, and growth from working with the team. I briefly discussed the projects I was assigned. I also discuss the obstacles I had while working on these projects and how I overcame them. In summary, this report will present an in-depth analysis of a2i and particular insights into my personal development during my internship.

Chapter 2

ORGANIZATION'S PROFILE

2.1 Organization Overview

a2i (Aspire to Innovate) was established by the Bangladesh government in 2008 with the support of UNDP (United Nations Development Programme). The agenda of establishing a2i was to drive its digital transformation. It started operating at the Prime Minister's Office at first. Then, when a2i started to the group and more projects were taken, it needed more teams. That's why it shifted to the ICT Tower in Agargaon later. a2i plays a key role in the "Digital Bangladesh 2021" initiative.[11] By reimagining public services from a digital perspective, it seeks to replace the old administration with a more creative, citizen-centered approach. a2i mainly works to ensure that technology benefits all citizens equally, and for that, they are promoting digital equity, or "e-Quality," [12]. It generates a cooperative framework that promotes collaboration in planning, education, and action to transform Bangladesh's public service delivery. a2i offers a dynamic workplace where experts and developers create state-of-the-art digital solutions. By developing software that responds to the requirements of Bangladeshis, this dynamic team advances the government's goal and helps ensure a progressive and inclusive digital future

a2i was established by the Bangladesh government as an innovative organization. That is why they named it a2i which refers to Aspire to Innovate. The name a2i indicates their goal to foster innovation in processes, technologies, or services to better serve people and drive positive change. Therefore, an appropriate setup is required to provide the most effective outcomes and offer the best service. The blueprint that follows will describe how a2i operates thanks to the diligence and dedication of the team's skilled, knowledgeable, and specialized developers and engineers.

a2i is a government organization, which is why it is funded by the Bangladesh government. At present, a2i's growth initiatives are handled and supervised by the "Team Leader." Every team in the organization has a team leader who is in charge of managing the group's activities. They also make sure everything runs smoothly. This person should handle all formal communication and serve as the initial point of contact on a professional level.

Besides being a source of inspiration for digital governance, a2i also promotes an inclusive, citizen-centered culture of public service innovation while challenging traditional administrative mindsets. The program focuses on the Bangladeshi government's initiatives to advance "e-Quality," or digital equity, by creating an adaptable national framework for cooperative planning, action, and education. This strategy

promotes innovative modifications to the way public services are provided. To assist the government in executing innovative, goal-driven policies and utilizing whole-of-society ways to accomplish the Sustainable Development Goals (SDGs), a2i serves as an innovation intermediary through the use of a "whole-of-government" strategy. The "Digital Stewardship Community for Policymakers in Asia and the Pacific" is another initiative that a2i is leading along with the UNDP. Using the resources from UNDP Global, Regional, and a2i, this program aims to improve relationships between practitioners and policymakers, by offering a forum for exchanging best practices and possibilities that may unlock investments inside particular nations as well as throughout the Asia-Pacific region.

In addition, a2i is trying to encourage the digital growth of other developing countries as well as least-developed countries (LDCs) by providing funding, equipment, and direction. This is being carried out in collaboration with the South-South Network for Public Service Innovation and the UNDP's Accelerator Labs network. With strong relationships with countries in Africa, the Asia-Pacific region, and the Middle East, such as Bhutan, Fiji, Ghana, Jordan, The Gambia, the Maldives, the Philippines, São Tomé and Príncipe, Somalia, Turkey, Uganda, and Yemen, a2i has so far worked with more than 40 countries.

2.2 First day of joining a2i

Before starting my work, my supervisor told me which area I was interested in and what experience and projects I had done before. I was interested in Machine Learning, Natural language processing, and working with AI models. I have told my supervisor about the projects I did during my university and the projects I did on my own. I was told to revise machine learning algorithms and the techniques that are necessary for that. He also advised me to learn and research web scrapping, as we will be scraping data. My work will evolve and change during this internship, and depending on the project I do, I have to learn the specific technology related to it. I was also advised to be prepared to adapt and learn new technologies as my projects evolved and to be bold in implementing unfamiliar tools and methods. This approach would enable me to grow and develop my skills throughout the internship and to tackle new challenges with confidence.

Chapter 3

WORKPLACE DETAILS

3.1 My Responsibility

My responsibilities include collaborating with the Data team and completing the daily tasks assigned to me. I have been entrusted with both individual assignments and team projects, all of which I am required to submit by their respective deadlines. Additionally, my internship responsibilities encompass participating in team meetings and preparing structured reports regarding my projects.

3.2 Organization Structure

The organization currently has over 21 teams and is expanding to meet the needs of new government initiatives.

3.3 Communication and Meetings

Working at an office is an unusual experience for me. Our data team communicates online through a WhatsApp group. The group includes interns, data scientists, supervisors, and data experts. They used WhatsApp to communicate formal notices, meetings, and workshops, as well as allocate new work. Additionally, in the office, we communicate for many purposes. We hold weekly office meetings to review project progress and plan future efforts. After gathering comments, we plan the next steps. Our supervisor assigned us several tasks based on the discussion.

3.4 Remote Office

During the movement period from July to August 2024, it was challenging to conduct physical office operations due to safety concerns in the external environment, particularly for students. As a result, the organization allowed us to perform our office work remotely. However, in the event of an emergency requiring our presence in the office, a transportation system was arranged. From July 18 to August 25, we operated in a remote capacity. Tasks were assigned through our WhatsApp group, and we were expected to submit our work to our co-supervisor by the end of each day.

3.5 Work-space Facilities

Our department has a tea-coffee area with sugar, tea, powdered milk, coffee, and hot water. Anyone can enjoy coffee or tea whenever they want. There is a space to relax with coffee and enjoy the outdoors. The a2i office provides lunch facilities for its employees. We have a cafe on the ground floor, next to the parking lot, that serves a wide variety of foods at moderate prices. There we can eat breakfast, lunch, and snacks—all forms of food, from Bangla cuisine to fast food and desserts. Lunch normally begins at 12.30 p.m. every day.

3.6 Conclusion

In short, the working environment at a2i is incredibly employee-friendly, serene, and lively. Despite the heavy workload, this environment does not feel exhausted. The office's serene and relaxing setting ensures no boredom.

Chapter 4

MY CONTRIBUTIONS

4.1 Overview

Undertaking an internship is a significant milestone in a student's academic journey, as it helps us develop our careers. In today's highly competitive job market, gaining work experience is a valuable asset that can give individuals an edge. Additionally, internships help students transition smoothly into the workforce after completing their studies. However, applying theoretical knowledge in real-world settings can be challenging, and industry practices often vary, making it essential to learn and adapt. As part of my academic requirements at BRAC University, I joined a2i as an intern and was assigned to the Data and research team. This opportunity allowed me to acquire new skills and gain hands-on experience. Initially, I was unfamiliar with workplace dynamics, but my team leader and supervisor provided guidance and support, teaching me office etiquette and facilitating my integration into the team. Although a2i lacks a formal training program for interns, my supervisor was always available to address my questions and concerns. Due to confidentiality constraints and data being only available to the Bangladesh Government, in some of my projects, I will not be able to share some data related to the projects I did, but I will include the necessary pieces of information in this report to illustrate my work.

4.2 Getting Familiar with My Workflow

Before starting my work, my supervisor told me which area I was interested in and what experience and projects I had done before. I was interested in Machine Learning, Natural language processing, and working with AI models. I have told my supervisor about the projects I did during my university and the projects I did on my own. I was told to revise machine learning algorithms and the techniques that are necessary for that. He also advised me to learn and research web scrapping, as we will be scraping data. My work will evolve and change during this internship, and depending on the project I do, I have to learn the specific technology related to it. I was also advised to be prepared to adapt and learn new technologies as my projects evolved and to be bold in implementing unfamiliar tools and methods. This approach would enable me to grow and develop my skills throughout the internship and to tackle new challenges with confidence.

4.3 333 Data Analysis and Anomaly Prediction

333 is a free call service provided by the Bangladesh Government to connect citizens with government services and information. Citizens can get information, request any service, or get help through 333. The 333 helpline offers information on over 2000 government services. [1] My task was to analyze the data, understand the data, call patterns, and predict anomalies from the given data. My supervisor provided data with 1,833,670 entries from January 1st to October 7th, 2024. I will be cleaning the data and then analyzing the data and based on the data I will build a prediction model for forecasting anomaly. This involves time-series analysis with domain-specific classification. I will be using models like Isolation Forest and LGBM Classifier for anomaly detection. This project will follow processes like loading the data, cleaning the data, handling missing values, feature engineering, preprocessing features, training models, and evaluating models.

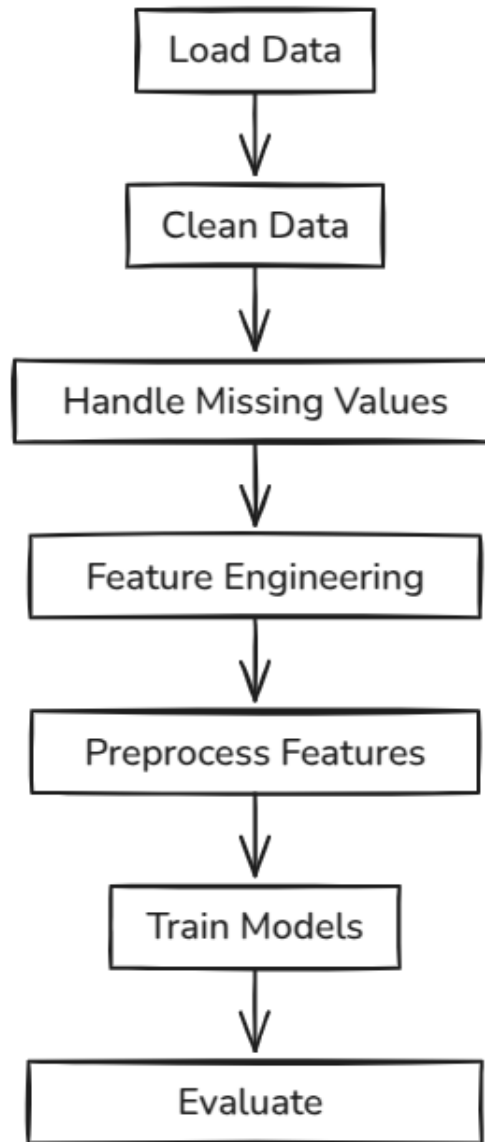


Figure 4.1: Process of 333 project

4.3.1 Data Cleaning

The dataset contains 1,833,671 rows and 19 columns. It includes features such as id, gender, physically challenged status, division, district, city, upazila, municipality, ward, mauza, village, location, comments, pcqc, scqc, call type, answered at, call duration min, and event date. However, some of these features are not essential for our analysis, and we will be removing the following six columns: municipality, ward, mauza, village, location, and comments.

To comprehensively analyze the dataset, we will employ YData Profiling, a widely used library for data quality profiling and exploratory data analysis. This tool provides critical insights into data quality, including identifying imbalanced classes, missing values, high correlations, and high cardinality variables. Additionally, it assesses the skewness of individual features. The generated report indicates a significant number of features with missing values, highlighting the presence of noise and incomplete data, which is typical in real-world datasets due to various factors such as failures during API calls. Consequently, we will remove rows containing missing values in the variables physically challenged, pcqc, scqc, and call type. We will also be imputing features like gender with the category "Unknown". On the other hand division, district, and upazila have more than 50% missing data and will be filling them up with 'Unknown'. Then we will turn so of the feature into category value which has been denoted as text value in the data.

4.3.2 Key Analysis

Something we want from our data is to know how many calls are not related to the services 333 provides. There are calls like prank calls or disconnected calls that are not related we want to know the percentage of 'Query Not Related To 333 Service'. And it is 68.70% of queries. Another thing we want to check is trends in the number of calls over time. For this, we want to visualize the number of calls.

We see an increased number of calls in April probably because of the heat wave that happened in April. Also we we see a decreased number of calls in early August.

4.3.3 Feature Engineering

In feature engineering, we also used feature extraction. I extracted information from event dates where we get year, month, and day. This will help us in training our model as this data contains time series. I also have to label anomalies by using our domain-specific knowledge. We have set limits for our anomalies where calls that are too short or too long are considered anomalies.

- Too Short (<0.2 min or 18 secs): Likely prank calls or disconnects.
- Too Long (>15 min): Potential system errors or anomalies.

We have considered Class 1 as the anomaly and Class 0 as the normal distribution. So the class distribution is:

- Anomalies (Class 1): 25.8% of calls.
- Normal (Class 0): 74.2% of calls.

It is also important to represent our anomalies visually for that we have visualized when we had anomalies over time.

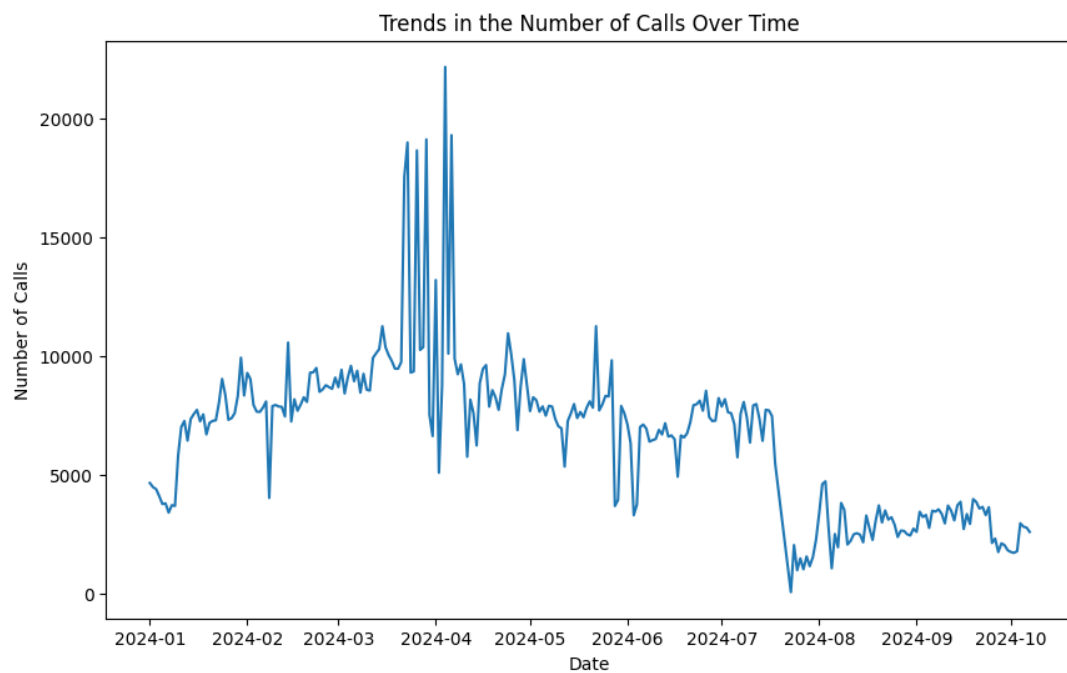


Figure 4.2: Trends in the Number of Calls Over Time

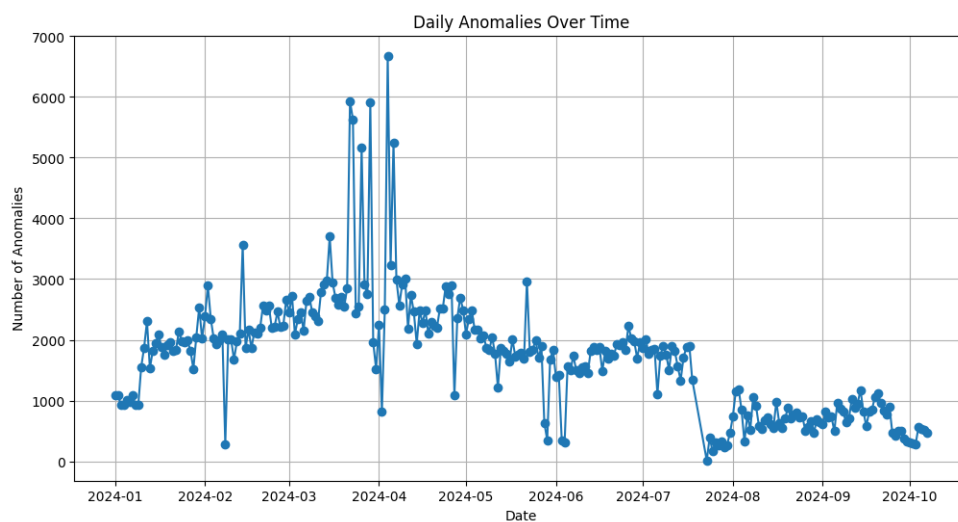


Figure 4.3: Anomalies Over Time

4.3.4 Daily Anomalies Over Time

Before pre-processing, it is essential to separate the test and training data to prevent data leakage. If data leakage occurs, the model may memorize the training data instead of learning from it, leading to overfitting and potentially resulting in diminished performance when applied to real-world data. We have used frequency encoding for high cardinal features like pcqc and scqc. Then for numeric data, we used a standard scaler, and for categorical data, we used one hot encoding. Then we used advanced techniques like PCA(Principal Component Analysis) to reduce defatation.[4] High Dimension is a curse for machine learning models. We have reduced the dimension to 20 dimensions as the model will train better with fewer dimensions or features. In PCA the features lose some of the data In our case we have retained about 99% of data for 20 dimensions.

4.3.5 Model Training

For model training, I have used two models Isolation forecast and LightGBM. Isolation forecast builds random trees to isolate anomalies faster. We have not adjusted any permeates or hyperparameter tuning for this model just to check how our data performs. The accuracy of the model is 56%. In classification accuracy is not how we measure a model's performance because some classification is more important than others. For classification problems, a good measure is their precision, recall, and F1 score. Then for LightGBM, I have used weight based on the number of anomalies as it contains imbalanced data. Also have used class validation with 3 folds to the best parameters for LightGBM.

4.3.6 Model Evaluation

Isolation Forest's results for class 0 (normal calls), precision is 0.74, which means when it predicts a normal call, it's correct 74% of the time. Recall is 0.64, so it correctly identifies 64% of all actual normal calls. F1-score is 0.69. For class 1 (anomalies), precision is 0.20 – only 20% of the predicted anomalies are real anomalies. Recall is 0.28, meaning it catches 28% of all actual anomalies. The F1-score here is 0.23. The accuracy is 56%, but that's misleading because of class imbalance. The macro average F1 is 0.46, weighted average F1 is 0.58.

For LightGBM Class 0 has precision 0.97, recall 0.64, F1 0.77. Class 1 has precision 0.45, recall 0.93, F1 0.61. Accuracy is 71%, macro avg F1 0.69, weighted avg F1 0.73, which performs significantly better than isolation forest because of weight adjustment and hyperparameter tuning.

Class 0 (Normal Calls):

- Precision (0.97): Nearly all predicted normal calls are correct (3% false positives).
- Recall (0.64): Missed 36% of normal calls (flagged incorrectly as anomalies).

Class 1 (Anomalies):

- Precision (0.45): 45% of predicted anomalies are true anomalies (55% false positives).

- Recall (0.93): Captures 93% of true anomalies (critical for minimizing missed cases).
- F1-Score (0.61): Reasonable balance for imbalanced data.

Overall:

- Accuracy (71%): Better than Isolation Forest but still skewed by class imbalance.
- Weighted Avg F1 (0.73): Stronger due to emphasis on class 0.

4.4 Bangal News Scraper

The Bangla news scraping website will be made with Streamlit, a popular Python web library for doing data science and machine learning-related projects. We can easily integrate many charts and deploy our website in a few lines of Python code.[10] In the middle, there will be a scrapping engine that will execute all the scripts necessary to scrap the news portal, format the data, and store it in a database. For this version of the website, we are storing the data in the MySQL database. The necessary data we need from scrapping are the Date of the publication, the Headline of the article, the body of the article, the URL of the article, and the Name of the news portal of that article. Involvement with this project has enhanced my expertise in data collection, web scraping techniques, and data pipelining processes.

4.4.1 Atrcitracture of Bangal News Scraper

The Bangla news Scraper is designed in a way so that we can scrap multiple news articles from multiple news portals from a given date range. It stores the data in a database and uses a user-friendly interface for viewing and downloading the data. The scraping process is handled in parallel meaning multiple scrapers work at the same time to collect articles for a given date range then it makes sure that there is no error or duplicate data. It stores the data in a data frame and then it stores that data frame in the database. Then it uses Streamlit's User interface to visualize the article, download data, and filter by data or site.

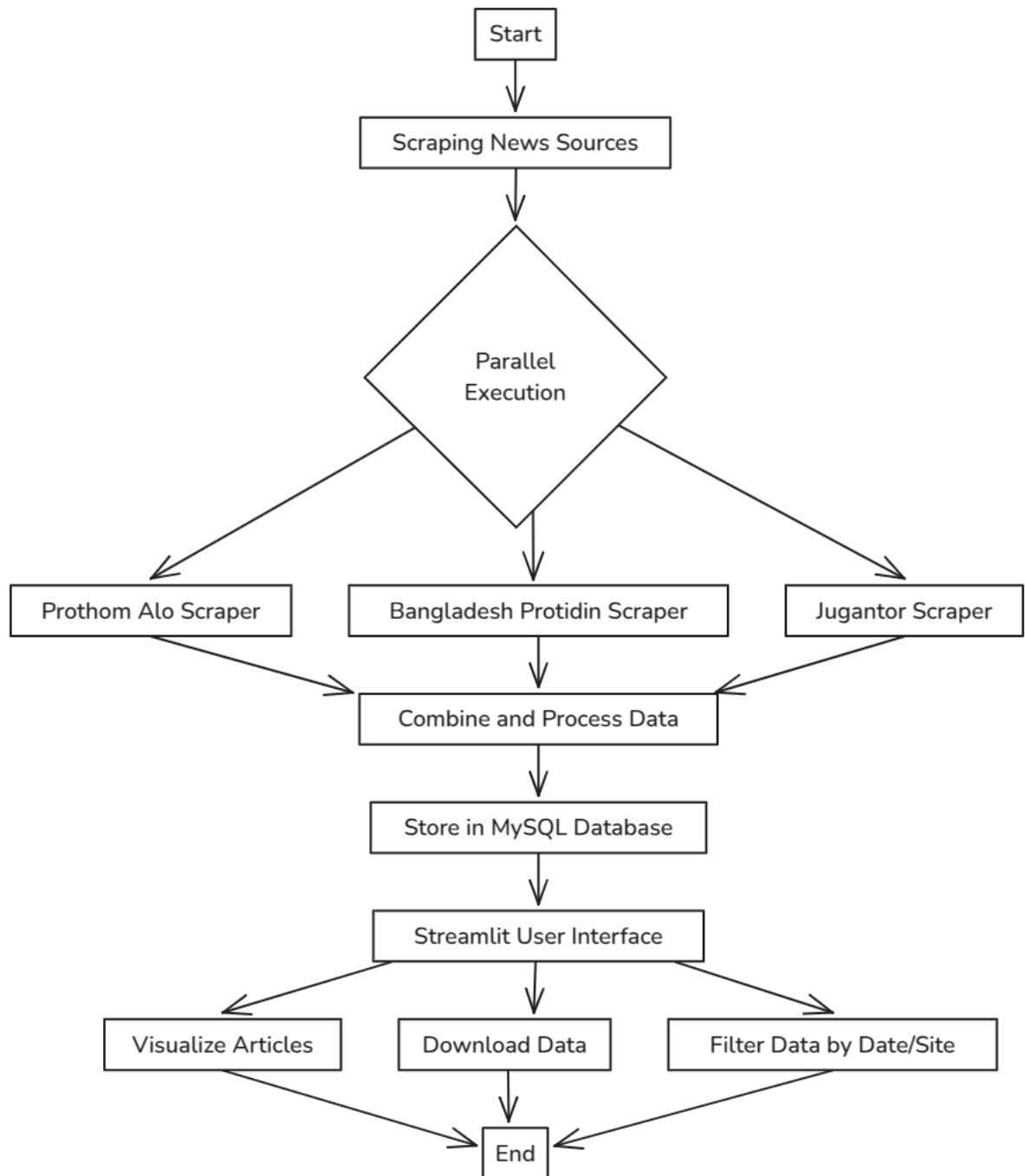


Figure 4.4: Scrapper Creation Flowchart

4.4.2 Structure of the DataBase

The ER Diagram shows the structure of the database used in this project to store data. It has two main entities: ARTICLES and SITES. The ARTICLES table stores details about each news article, such as its ID, publication date, headline, content, link, and the site where it was published. Each article is uniquely identified by its id, and the article link ensures no duplicate articles are stored. The SITES table holds information about the news sites, unique IDs, site names, and site URLs. There is a one-to-many relationship between SITES and ARTICLES. A single news site can have multiple articles associated with it.

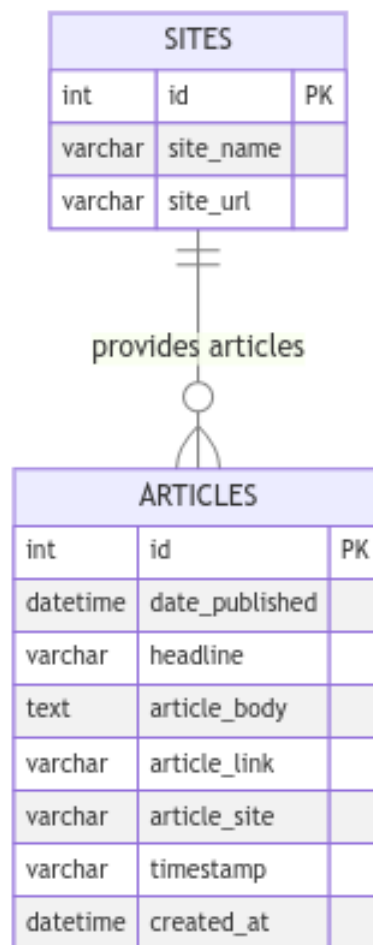


Figure 4.5: ER Diagram of the Website

4.4.3 User Interface

The user can the start date and date as input. There is also an option to just choose one news site or multiple news sites depending on needs. The website will show the number of total articles fetched from the database and show a provide in a data frame and a visualization of number of articles by each site. The user can also download the file as CSV. The user interface is modern and very simple to use.

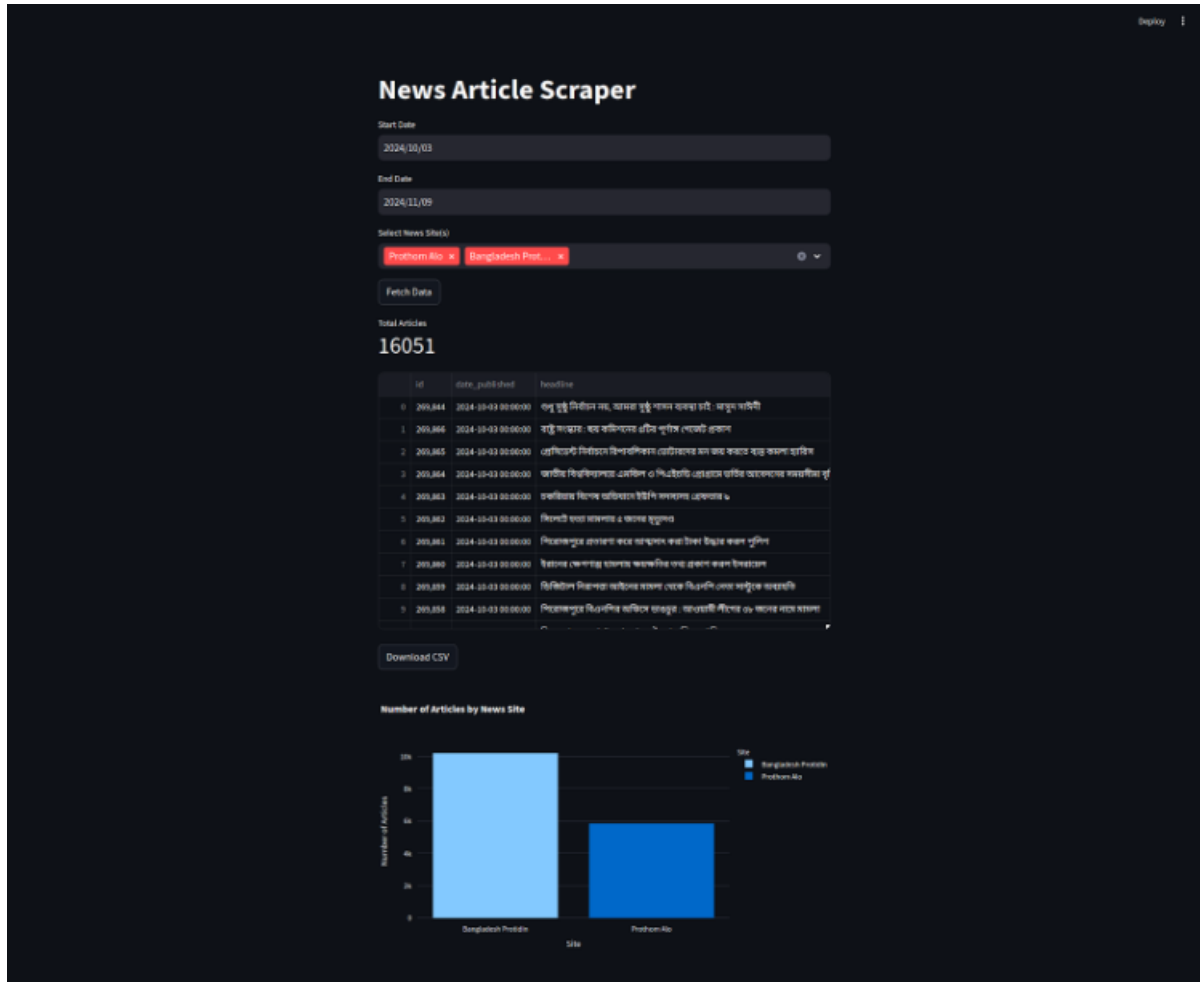


Figure 4.6: User Interface of the Website

4.5 Bangla RAG System

The Bangla RAG system utilizes the scrapped data from the newspaper scrapping website as the context for the RAG model. RAG is Retrieval Augmented Generation.[6] The LLMs can not give us information about recent events or sometimes fail to answer questions because it was trained in old data and do not have enough context to write. RAG's sole purpose is to make the LLM model context so we get an answer tailored to our needs. The project connects to a MySQL database, loads news articles, generates embeddings using a pre-trained model, creates an FAISS index, and then allows searching through the articles using the index. There's also a BanglaRAG class that handles the search functionality.

4.5.1 Overview of User Query

When a user submits a query, it undergoes a query embedding process. This process converts the text query into a numerical vector representation that encapsulates its semantic meaning. Essentially, it translates the words into a mathematical format that computers can effectively understand and compare.

Following the embedding, the system utilizes a FAISS (Facebook AI Similarity Search) index to search for similar vectors within large datasets. FAISS is a specialized library designed to efficiently identify comparable vectors.[7] Upon locating similar vectors, the system retrieves the corresponding metadata—supplementary information about each result that enhances context and relevance. Ultimately, these results are ranked according to their similarity to the query, ensuring that the most pertinent information is presented to the user first.

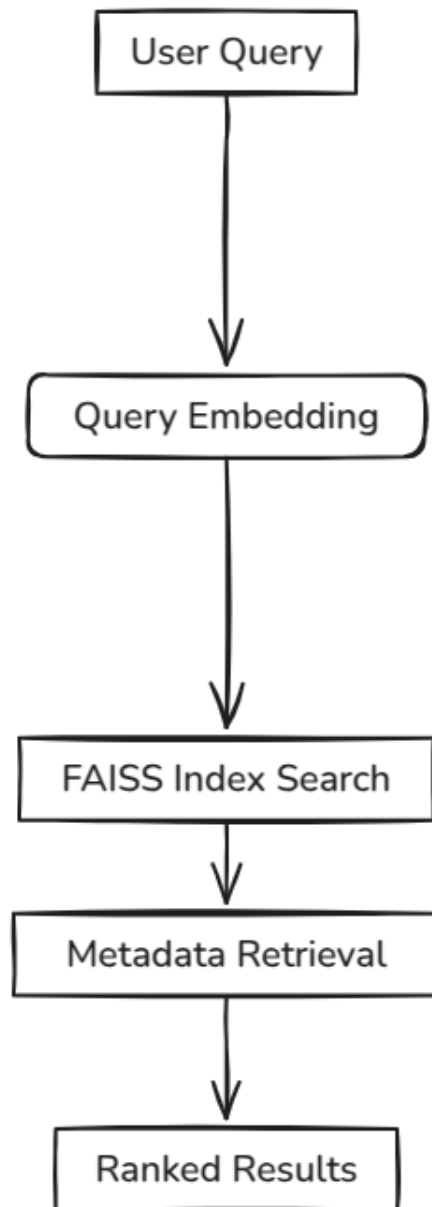


Figure 4.7: Process of User Query

4.5.2 Embedding of Database

Beginning with a database of documents, the system conducts batch processing to efficiently handle multiple documents simultaneously. During this process, two parallel operations are executed. First, the system generates embeddings for all the documents, similar to the method used for embedding the query. These embeddings are utilized to create the FAISS index, which is stored for future reference. This can be likened to constructing a detailed map of all documents within that mathematical space. Second, the system archives metadata associated with each document. This metadata may encompass elements such as titles, authors, timestamps, or any other pertinent information that aids users in understanding and assessing the search results.

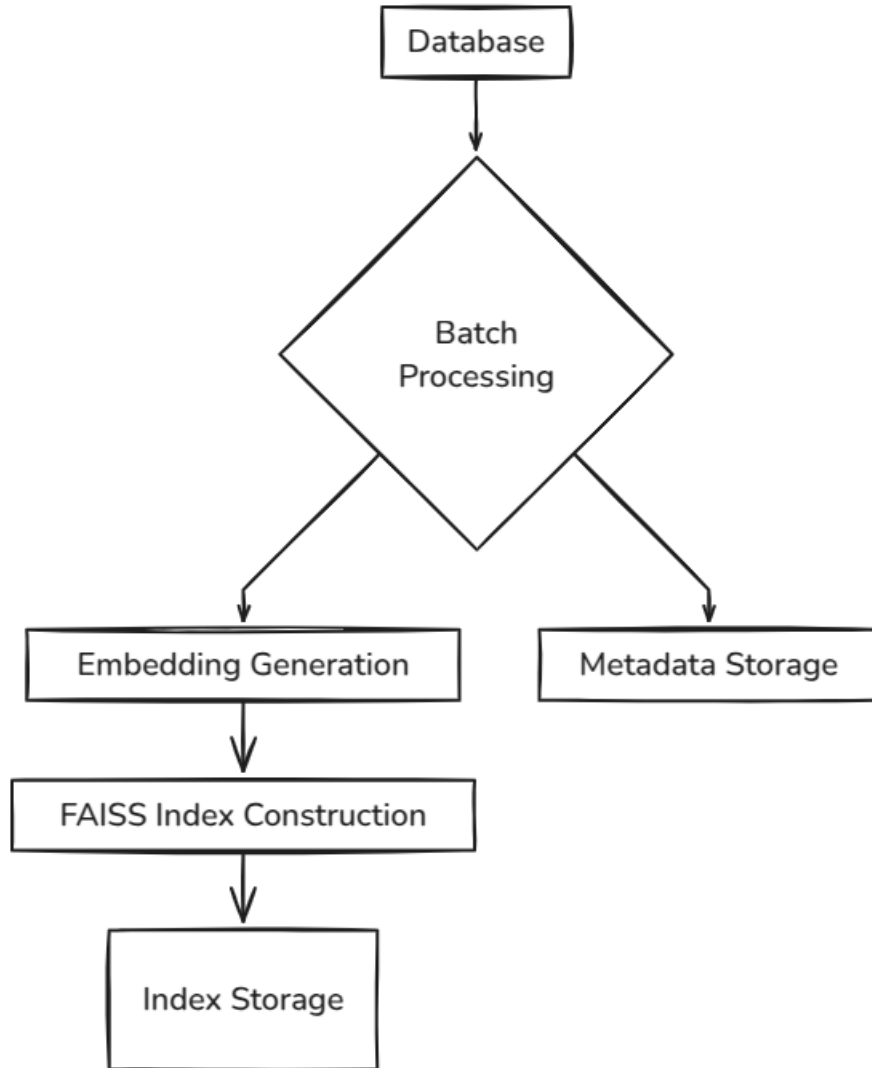


Figure 4.8: Process of Data and Meta-Data

4.5.3 Response of the system

The response of the RAG system should be statistical meaning it will prioritize numbers. For example, if the query is “How many road accidents happened in November

2024” the result should prioritize statistical answers like “There were 70 road accidents that happened in November”. FAISS is particularly clever in organizing these vectors, which makes searches extremely fast, even with millions of documents. The system initially transforms the input into an embedding, similar to the process applied to the documents. This query embedding is subsequently compared to the indexed document embeddings utilizing FAISS. The comparison employs mathematical metrics, such as cosine similarity, to identify the closest matches. Once FAISS identifies the most similar vectors, the system retrieves the corresponding metadata to deliver relevant results to the user. The ranking process may take into account various factors beyond mere vector similarity, including document recency, user feedback, and other relevance indicators. This system is capable of understanding semantic meaning rather than relying solely on keyword matching.

4.6 Challenges

There are a lot of things that were unknown to me at first throughout this project I have faced many problems here are some of the problems I faced:

- Building accurate time series anomaly detection models requires balancing domain expertise with statistical methods, especially challenging when defining what constitutes a true anomaly versus normal variation in your specific domain.
- Addressing imbalanced datasets in anomaly detection posed challenges for model evaluation, as conventional metrics often fell short of effectively reflecting true model performance.
- Bengali text processing consistently presented challenges across all projects, particularly in ensuring proper Unicode encoding and maintaining text integrity across various system components.
- Scraping content from Bengali newspapers required robust error handling to accommodate different website structures, implementation of rate limiting, and meticulous management of encoding during the data collection process.
- Database design required careful attention to efficiently store and retrieve Bengali text while ensuring proper character encoding and indexing for rapid searches.
- Implementing FAISS for Bengali text necessitated careful optimization to uphold search accuracy while considering the nuances of Bengali language semantics.
- Conducting vector similarity searches for Bengali text involved extensive tuning to accurately capture semantic relationships, a task made particularly complex due to the morphological richness of the Bengali language.
- Feature engineering for time series data involved developing meaningful representations capable of effectively capturing both temporal patterns and domain-specific anomalies.

- Constructing a high-performance end-to-end system entailed optimizing each component to ensure seamless collaboration, spanning from data collection to storage and retrieval.

4.7 Learning

Throughout my internship, I have gained knowledge in the different aspects of machine learning, data collection, and optimization. A few of the learnings are given below:

- Acquired extensive expertise in processing non-English text, specifically Bengali, across various technical applications, including web scraping and natural language processing.
- Developed a strong understanding of feature engineering for time series data, with a focus on creating features that effectively integrate domain knowledge.
- Gained proficiency in all aspects of data pipeline development, encompassing web scraping, database design, and frontend development using Streamlit.
- Obtained practical experience in the implementation and optimization of vector similarity search systems utilizing FAISS for non-English language applications.
- Learned effective methodologies for addressing imbalanced datasets and integrating multiple modeling approaches to enhance anomaly detection.
- Developed specialized knowledge in database design for multilingual applications, particularly in optimizing the storage and retrieval of non-English text.
- Mastered advanced web scraping techniques to accommodate diverse website structures while ensuring performance efficiency and compliance with site policies.
- Built expertise in end-to-end system development, gaining insights into the interaction of various components and strategies for optimizing their integration.

Chapter 5

GROWTH

5.1 Professional Growth

5.1.1 Office environment

This is my first professional experience gained through working with a2i. Before this opportunity, I had limited knowledge of corporate life, the application of our skills in real-world settings, and office culture. This experience has significantly contributed to my personal and professional growth.

5.1.2 Teamwork

During my university experience, I engaged in group work and participated in teamwork on several occasions. However, the collaboration I experienced in a professional office setting was notably different. The office environment features individuals of varying ages and levels of work experience, all working together as a cohesive team. As an intern, this was a new and valuable experience for me. I felt both excited and nervous about meeting the expectations of the role. Fortunately, my supervisor, co-supervisor, and fellow team members provided tremendous support and guidance, which facilitated my adaptation to the workplace.

5.2 Technical Growth

5.2.1 NumPy and Pandas

NumPy facilitated efficient numerical computations and array operations for managing embedding vectors. Pandas organized structured data through DataFrames, supporting database interactions and metadata processing.

5.2.2 Data Visualization Stack

Matplotlib/Seaborn produced static visualizations for data analysis, while Plotly created interactive charts for the Streamlit dashboard.[9]

5.2.3 YData Profiling

The "ProfileReport" features automated exploratory data analysis (EDA) by generating thorough data quality reports that include statistics and visualizations.[5]

5.2.4 Scikit-learn Toolkit

- IncrementalPCA: Enabled dimensionality reduction for large datasets.
- IsolationForest: Identified anomalous news articles.
- classification report: Assessed model performance metrics.[3]

5.2.5 Streamlit

Streamlit served as the foundation for the web interface, enabling real-time news search and system monitoring through interactive widgets and visualizations.[10]

5.2.6 Database Connectors

- MySQL Connector ("mysql.connector"): Established a direct bridge between Python and MySQL.
- SQLAlchemy("create engine"): Implemented an Object-Relational Mapper (ORM) for efficient database operations.

5.2.7 Web Scraping Stack

- Requests: Facilitated HTTP calls with built-in retry logic ("HTTPAdapter", "Retry").
- BeautifulSoup: Provided HTML parsing capabilities for news extraction.[8]
- NewsPlease: Specialized in extracting content from news articles.[2]

5.2.8 NLP and Embeddings

Sentence Transformers generated multilingual text embeddings using the "paraphrase-multilingual-MiniLM-L12-v2" model.[6]

5.2.9 Vector Search

FAISS enabled rapid similarity searches through optimized index structures IndexFlatIP for efficient news retrieval.[7]

Chapter 6

CONCLUSION

6.1 Conclusion

The process of translating theoretical knowledge into practical solutions for real-world challenges has proven to be both intellectually demanding and highly meaningful. This comprehensive project, which includes anomaly detection in governmental call center 333, automated news scraping, and the development of a Bangla Retrieval-Augmented Generation (RAG) model, illustrates the potential for data-driven systems to enhance public service delivery and improve access to information in Bangladesh. By tackling significant issues such as bureaucratic inefficiencies, data transparency, and linguistic inclusivity, this initiative aligns with global trends in digital governance, while also addressing specific local needs.

The anomaly detection framework in the 333 project enables institutions to proactively identify irregularities in administrative processes, thereby minimizing delays and reducing the potential for corruption. In conjunction with the automated scraping of newspaper websites, this system facilitates real-time monitoring of public sentiment and emerging issues, equipping policymakers with actionable insights during critical situations. Additionally, the Bangla RAG model addresses the language barrier by providing citizens with intuitive access to complex governmental data through natural language queries—a pioneering advancement for Bengali speakers. Collectively, these elements create an ecosystem where paperless workflows, predictive analytics, and AI-driven accessibility converge to promote transparency and equity.

My collaborative experience on this project with a2i, a leader in Bangladesh’s digital transformation, has been profoundly impactful. It has taught me the importance of harmonizing technical precision with societal impact—whether it involves refining anomaly detection algorithms to reflect the nuances of Bangladesh’s bureaucracy, optimizing web scraping techniques to accommodate various news formats, or iterating the RAG model to effectively understand colloquial Bangla. I have learned that addressing real-world challenges requires not only analytical rigor but also cultural understanding: datasets need to be interpreted within the context of local circumstances, and the success of an AI system relies on its adaptability to users’ linguistic and technological environments.

This experience has reinforced my commitment to democratizing technology. By ensuring that non-technical citizens can easily query policies or report issues using the Bangla RAG interface, we minimize reliance on intermediaries and empower

marginalized communities. The anomaly detection system serves as a mechanism for accountability, while the scraped news data provides valuable insights into public priorities—both of which are essential for responsive governance. As Bangladesh progresses toward its “Smart Bangladesh 2041” vision, initiatives like this will be critical in establishing trust in digital systems and ensuring that no citizen is overlooked.

Looking ahead, I am eager to further advance this work by enhancing the RAG model’s multilingual capabilities, integrating predictive analytics for disaster response, and advocating for ethical AI frameworks that prioritize user privacy while delivering societal benefits. This internship has not only refined my skills in machine learning and data engineering but has also instilled in me a lifelong dedication to utilizing technology as a catalyst for progress.

6.2 Recommendations

I strongly endorse a2i for individuals seeking to develop and enhance their skills. The projects undertaken by a2i are primarily government initiatives aimed at improving the welfare of the people of Bangladesh. Additionally, the work environment is positive, with a welcoming and collegial atmosphere. Unlike traditional government offices, a2i fosters a culture of productivity and collaboration.

Bibliography

- [1] a2i, “333 innovation brief 2024,” a2i, Tech. Rep., 2024. [Online]. Available: <https://a2i.gov.bd/wp-content/uploads/2024/04/333-Innovation-Brief-2024.pdf>.
- [2] F. Hamborg, *News-please*, Accessed January 31, 2025, 2024. [Online]. Available: <https://github.com/fhamborg/news-please>.
- [3] scikit-learn developers, *Scikit-learn: Machine learning in python*, Accessed January 31, 2025, 2024. [Online]. Available: <https://scikit-learn.org/stable/>.
- [4] scikit-learn developers, *Sklearn.decomposition.IncrementalPCA*, version stable, Part of scikit-learn documentation. Accessed January 31, 2025, scikit-learn, 2024. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html>.
- [5] *Ydata profiling documentation*, Accessed January 31, 2025, 2024. [Online]. Available: <https://docs.profiling.ydata.ai/latest/>.
- [6] Hugging Face, *Paraphrase multilingual minilm-l12-v2*, Accessed January 31, 2025. [Online]. Available: <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>.
- [7] Meta AI, *Faiss: A library for efficient similarity search*, Accessed January 31, 2025, Meta. [Online]. Available: <https://ai.meta.com/tools/faiss/>.
- [8] L. Richardson, *Beautiful soup documentation*, version 4, Accessed January 31, 2025. [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.
- [9] *Seaborn: Statistical data visualization*, Accessed January 31, 2025. [Online]. Available: <https://seaborn.pydata.org/tutorial/introduction.html>.
- [10] *Streamlit*, No URL provided in original reference.
- [11] a2i, *About us*, Accessed January 31, 2025, n.d. [Online]. Available: <https://a2i.gov.bd/about-us/>.
- [12] a2i, *E-quality centre for inclusive innovation*, Accessed January 31, 2025, n.d. [Online]. Available: <https://a2i.gov.bd/a2i-trending-insights/e-quailty-centre-for-inclusive-innovation/>.