National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Sakshi Kacheshwar KalungePatil
Student ID: x23122366

School of Computing
National College of Ireland

Supervisor: Prof. Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sakshi Kacheshwar KalungePatil
Student ID:	x23122366
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Prof. Rejwanul Haque
Submission Due Date:	11/08/2025
Project Title:	Customer Churn Prediction using RAG-Based Sentiment Analysis with LLMs and CatBoost
Word Count:	679
Page Count:	5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	*Sakshi KalungePatil*
Date:	12th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Churn Prediction using RAG-Based Sentiment Analysis with LLMs and CatBoost

Sakshi Kacheshwar KalungePatil
x23122366

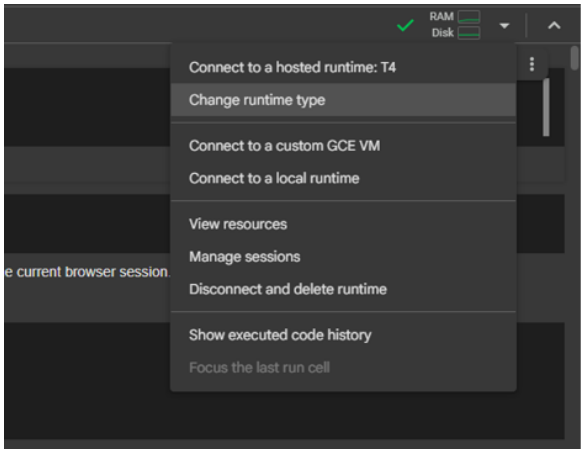
1 Introduction

This configuration manual demonstrates a step-by-step guide, providing the required environment setup, and necessary libraries used to execute the research project conducted for Customer churn prediction using RAG-based sentiment analysis with fine-tuned TinyLLaMA combined with CatBoost Classifier. The model integrates both sentiment-rich textual reviews and behavioral user metadata for churn prediction. This manual includes a detailed guide to replicate the research work by installing the required libraries used for data preprocessing, model implementation, model training and testing, dataset selection source, and environment setup for code execution using Google Colab framework.

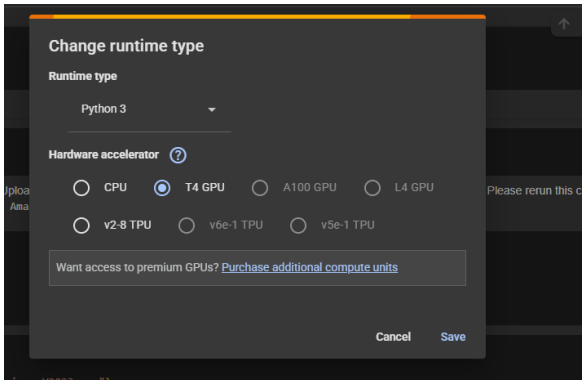
2 Google Colab environment Setup

This research project requires intensely high processing power for training and testing two distinct models: RAG-based TinyLLaMA and CatBoost model that uses approximately 15,000 sentiment-rich textual customer reviews for data cleaning process to final evaluation of churn prediction. By default, the **Hardware runtime type** is set to **CPU**, but to process such a large dataset, it is recommended to change the hardware runtime type to **T4 GPU**. The following steps can be executed to setup the Google Colab environment as follows:

- Ensure the system has stable internet access for login and installation purposes.
- Log in to Google Colab via the respective Google account.
- Once logged in, by default the hardware runtime will be set to CPU. In this case, when tried running the code cells with CPU, it resulted in crashing the environment. Hence, change the runtime to T4 GPU as shown in the following figures.



(a) Change runtime type



(b) Change Hardware accelerator

Figure 1: Google Colab Environment Setup

3 Dataset Selection

This section provides the dataset requirements used in the research project. The selected dataset is a publicly available dataset on **Kaggle** [1] featuring real-time customer reviews of millions of products from the dataset - Amazon Reviews 2023. To access the dataset, ensure that you have a Kaggle account. For this research project, the following Kaggle dataset was selected to predict customer churn by implementing sentiment analysis using the RAG-based LLaMA model and the CatBoost model.

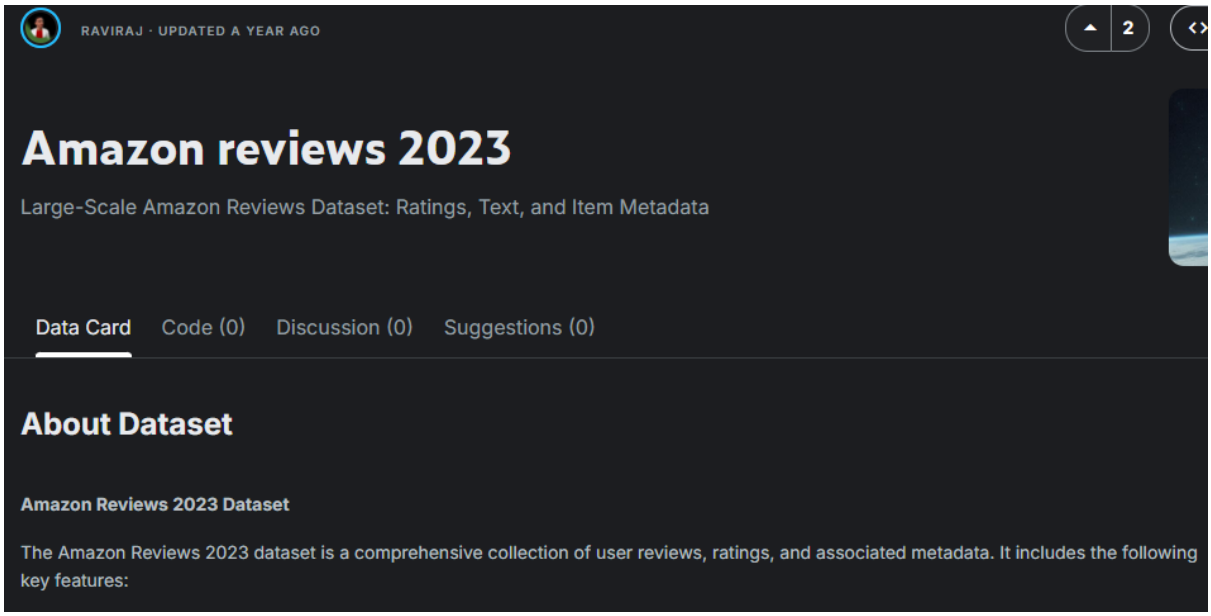


Figure 2: Amazon reviews 2023 - Kaggle dataset

[1] https://www.kaggle.com/datasets

4 Installation of Python libraries

Core Python libraries are installed for the implementation of the fine-tuned TinyLLaMA integrated with the retrieval augmented generation (RAG) module to perform sentiment analysis through textual customer reviews combined with CatBoost model to predict final churn retention using structural metadata. Few of the libraries were used from Hugging Face including **Hugging Face datasets**[2], **Hugging Face transformer**[3] and **Hugging Face peft**[4]

```
# Installing Hugging Face's transformers and libraries required for parameter-efficient fine-tuning using LoRA
!pip install transformers datasets peft accelerate bitsandbytes -q

# Installing necessary python libraries for combining semantic search and RAG module
!pip install faiss-cpu sentence-transformers transformers tqdm

# Installing catboost classifier for churn classification
!pip install catboost
```

Figure 3: Core Python libraries installation

```
# Installing necesssary pyhton libraries
import pandas as pd
import numpy as np
import re
import torch
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score,
    confusion_matrix, classification_report
)

# Python libraries required to fine-tune TinyLLaMA model
from datasets import Dataset, DatasetDict
from transformers import AutoTokenizer, AutoModelForCausalLM, TrainingArguments, Trainer, DataCollatorForLanguageModeling
from peft import get_peft_model, LoraConfig, TaskType
import torch
from sklearn.model_selection import train_test_split

# Python libraries used for implementing RAG module
from sentence_transformers import SentenceTransformer
import faiss
from catboost import CatBoostClassifier, Pool            # python library used for catboost classifier implementation
```

Figure 4: Required python libraries

[2]https://huggingface.co/docs/datasets/en/index
[3]https://huggingface.co/docs/transformers/en/index
[4]https://huggingface.co/docs/peft/en/index

```
pandas: 2.2.2
numpy: 2.0.2
re: built-in python library
torch: 2.6.0+cu124
seaborn: 0.13.2
scikit-learn: 1.6.1
datasets: 4.0.0
transformers: 4.54.0
peft: 0.16.0
sentence-transformers: 4.1.0
faiss: 1.11.0
catboost: 1.2.8
```

Figure 5: Python library versions

5 Mount Google Drive, File export and Download

This section describes the additional steps implemented for accessing the stored files, exporting the Python file in .csv format, and downloading the file. The Python file is mounted and saved on Google Drive for future accessibility, and the preprocessed data is exported and downloaded in .csv format to perform Tableau visualizations.

```
[ ]  # Saving the file on Google drive
     from google.colab import drive
     drive.mount('/content/drive')
```

(a) Mount on Google Drive

```
# Exporting the cleaned file
amazon_df.to_csv('cleaned_amazon_review.csv', index=False)

# Downloading the file
from google.colab import files
files.download('cleaned_amazon_review.csv')
```

(b) File export and download

Figure 6: Mount Google Drive, File export and Download

6 Tableau Setup

This section describes the Tableau setup done to create various visualizations such as bar charts, scatter plots, and pie charts to display the sentiment distribution, churn classification, and sentiment labels categorization. The following steps were implemented to create visualizations in Tableau.

- Log in to Tableau Public, a free platform that is used to create, explore, and publicly share visualizations.
- After Log in, click on **"New Data Source"** from the toolbar and **"Text file"** to connect the csv file format or click on **"Connect to data"** for easy access.
- Once the data is connected, explore the number of visualizations provided by combining various rows and columns from the data.

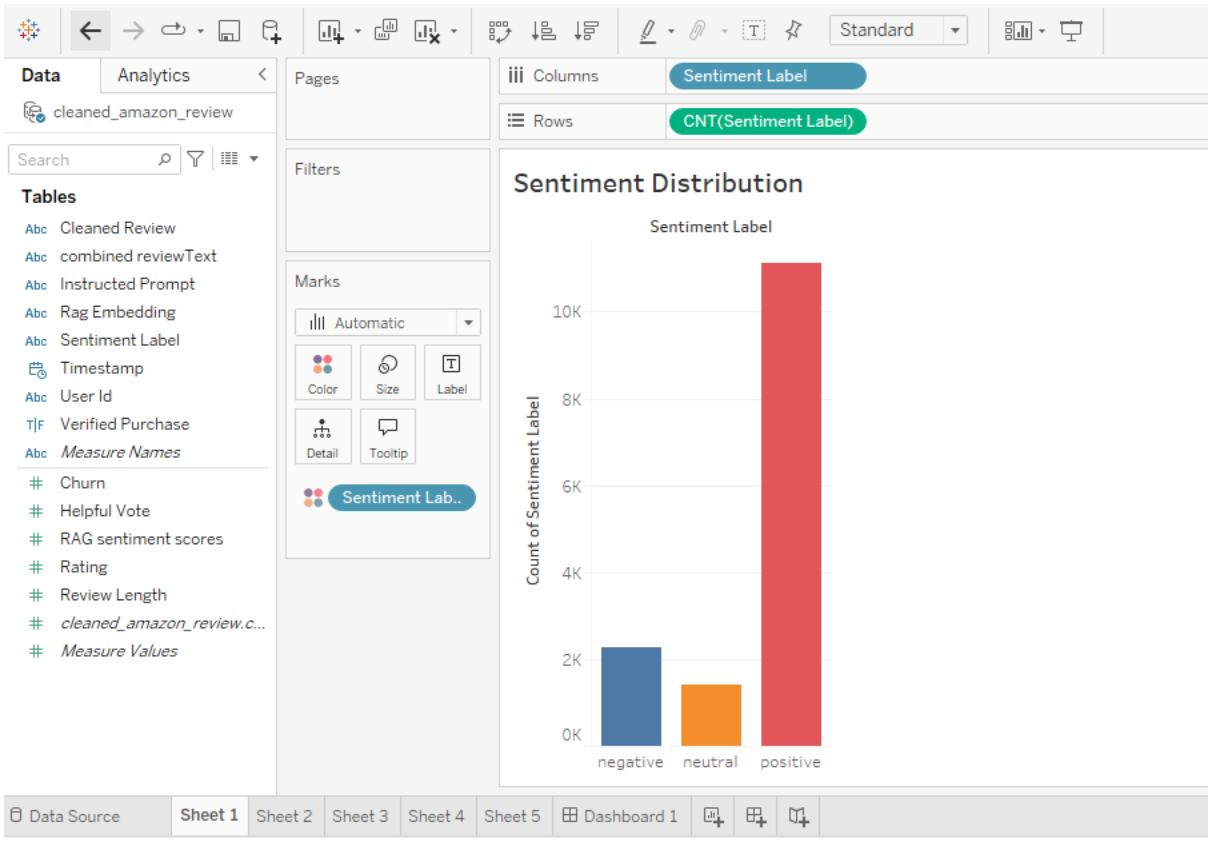


Figure 7: Tableau visualizations

References

Find Open Datasets and Machine Learning Projects | Kaggle (n.d.).
URL: *https://www.kaggle.com/datasets*

Google Colab (n.d.).
URL: *https://colab.research.google.com/*

Raviraj (n.d.). Amazon reviews 2023.
URL: *https://www.kaggle.com/datasets/ravirajbabasomane/amazon-reviews-2023*

Tableau Public | Find inspiration and improve your data skills (n.d.).
URL: *https://www.tableau.com/products/public*