National College of Ireland

# Configuration Manual

MSc Research Project

Data Analytics


Namrata Suryawanshi

Student ID:21197741

School of Computing

National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

# National College of Ireland

## MSc Project Submission Sheet

## School of Computing

| | |
|---|---|
| **Student Name:** | **Namrata Ashok Suryawanshi** |
| **Student ID:** | **X21197741** |
| **Programme:** | **Msc Data Analytics** |
| **Year:** | **2022-2033** |
| **Module:** | |
| **Supervisor:** | **Teerath Kumar Menghwar** |
| **Submission Due Date:** | |
| **Project Title:** | **Configuration Manual** |
| **Word Count:** | **755** |
| **Page Count:** | **8** |

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:** | **Namrata Suryawanshi** |
| **Date:** | |

**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):** | |

# Configuration Manual

Namrata Suryawanshi

x21197741

## 1. Introduction

In The Internet has recently accelerated the pace of change in people's lifestyles. Quora stands out among the numerous internet platforms as a vibrant area that encourages shared learning. It gives users the chance to post queries and interact with a community that shares insightful opinions and thoughtful responses. Platforms for asking and answering questions online have become crucial to the global exchange of knowledge. These platforms are useful tools for fostering community interaction and knowledge sharing. Nevertheless, despite their advantages, these platforms frequently encounter problems with content quality. Numerous websites put the production of material first without taking its quality into account, which contributes to problems like spam, racism, and online animosity. The spread of such harmful behaviors interferes with user experience, violates platform policies, and has a detrimental effect on the norms and health of online communities. Due to the spread of false information, this tendency may even lead to the emergence of polarized groups. Quora, a popular question-and-answer website with a sizable user base, draws about 300 million active visitors each month. This popularity highlights its function as a focal point for information sharing and group learning. Researchers have looked into ways to reduce biased, harmful, and dishonest content on these platforms as a reaction to this dilemma. The detection and classification of phony inquiries has been one area of focus. This endeavor is best shown by the Quora Insincere Questions Classification research challenge. A well-known question-and-answer website called Quora has struggled with fake or harmful queries meant to trick or incite other users. Maintaining the platform's integrity, user confidence, and level of interaction quality depends on identifying and handling such fake information. The problem of phony inquiries has been addressed in a number of research. Transformer-based models, like BERT, provide attention-based techniques that are excellent at handling noisy datasets, enabling improved generalization. The goal of this study is to improve the classification of insincere questions by fine-tuning pre-trained models using the Quora Insincere Questions Classification dataset.

## 2. Environmental Setup

## 2.1 Hardware Requirements

- 16GB RAM.
- 500 GB HardDrive.
- Intel. Core i5

## 2.2 Software Requirements

- Windows 11
- Python

## 2.3 Programming Prerequisites

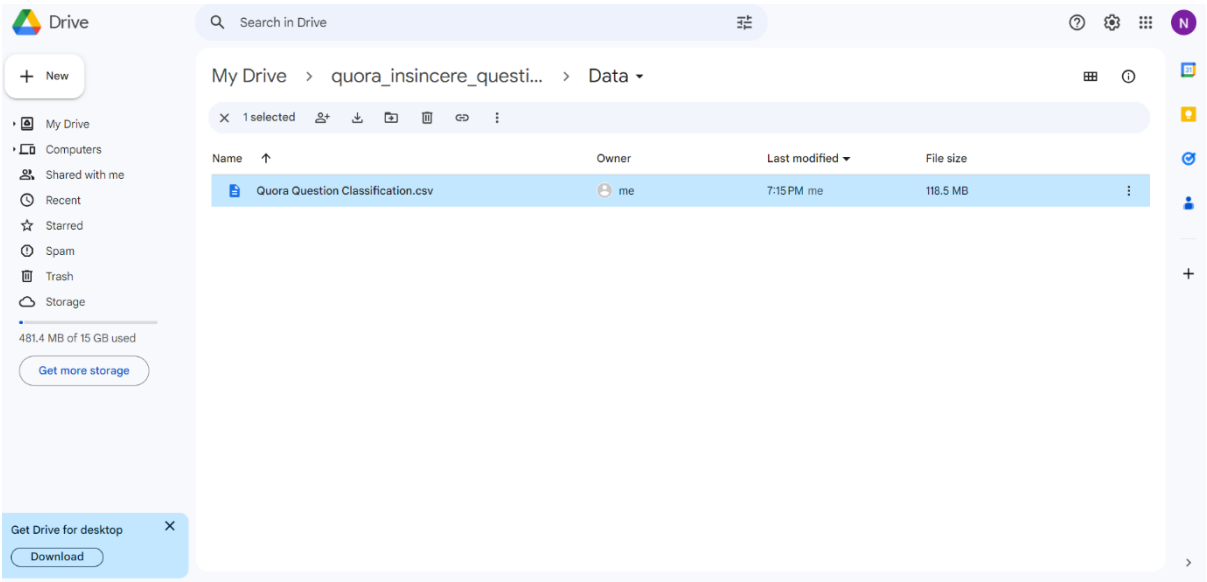
- Python
- Google Colab

Since Google Colab is hosted in the cloud and the code can be viewed from anywhere at any time, they were chosen to implement the code because they have all the necessary packages on hand.

The study was completed using a Tesla T4 GPU-equipped Colab Notebook. According to Colab Notebook, a RAM or GPU shortage won't cause the model training session to be halted.

| | |
|---|---|
| **IDE** | **Google Colab Notebook on Cloud** |
| **Program Language** | **Python** |
| **Computation** | **Tesla T4 GPU** |
| **Visualization Library** | **Matplotlib, WordCloud, Seaborn** |
| **Modelling Library** | **Keras, tensorflow, sklearn** |
| | |

## 3. Dataset Loading:




## 4. Coding Implemetation:

- Mounting Google drive:

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

    Mounted at /content/drive
```

- Importing the Libraries:

```
#importing all libraries
import os
import re
import string
import pickle
import itertools
import math,nltk
import numpy as np
import pandas as pd
import seaborn as sns
from tqdm import tqdm
import tensorflow as tf
import keras.backend as K
from keras.models import Model
import matplotlib.pyplot as plt
from keras.models import Model
from keras.models import load_model
from nltk.corpus import stopwords
from keras.models import Sequential
from nltk.stem import SnowballStemmer
from keras.utils import to_categorical
from sklearn import feature_extraction
from keras.layers import Conv1D,GRU
```

```
from keras.models import Sequential
from imblearn.over_sampling import SMOTE
import plotly.graph_objects as go
import plotly.express as px
import plotly.figure_factory as ff
from keras.layers import Layer
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from keras.preprocessing.text import Tokenizer
from sklearn.preprocessing import LabelBinarizer
from sklearn.feature_extraction.text import TfidfVectorizer
from tensorflow.keras.models import load_model
from tensorflow.keras.preprocessing.sequence import pad_sequences
from keras.layers import Dense, concatenate,Flatten,LSTM, Embedding, Input
from keras.layers import Dropout,Embedding,GlobalMaxPooling1D, MaxPooling1D, Add, Flatten
from sklearn.metrics import confusion_matrix,accuracy_score,classification_report
```

- Load Training Data:

```
#loading dataset
data=pd.read_csv("/content/drive/My Drive/quora_insincere_question_classification/Data/train.csv")
data.head(10)
```

| | qid | question_text | target |
|---|---|---|---|
| 0 | 00002165364db923c7e6 | How did Quebec nationalists see their province... | 0 |
| 1 | 000032939017120e6e44 | Do you have an adopted dog, how would you enco... | 0 |
| 2 | 0000412ca6e4628ce2cf | Why does velocity affect time? Does velocity a... | 0 |
| 3 | 000042bf85aa498cd78e | How did Otto von Guericke used the Magdeburg h... | 0 |
| 4 | 0000455dfa3e01eae3af | Can I convert montra helicon D to a mountain b... | 0 |
| 5 | 00004f9a462a357c33be | Is Gaza slowly becoming Auschwitz, Dachau or T... | 0 |
| 6 | 00005059a06ee19e11ad | Why does Quora automatically ban conservative ... | 0 |
| 7 | 0000559f875832745e2e | Is it crazy if I wash or wipe my groceries off... | 0 |
| 8 | 00005bd3426b2d0c8305 | Is there such a thing as dressing moderately, ... | 0 |
| 9 | 00006e6928c5df60eacb | Is it just me or have you ever been in this ph... | 0 |

## 5. Results:

### 1. CNN+LSTM+GRU without Pre-Trained Word Embedding

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.65 | 0.66 | 16107 |
| 1 | 0.67 | 0.69 | 0.68 | 16217 |
| accuracy | | | 0.67 | 32324 |
| macro avg | 0.67 | 0.67 | 0.67 | 32324 |
| weighted avg | 0.67 | 0.67 | 0.67 | 32324 |

### 2. CNN+LSTM+GRU with Pre-Trained Word Embedding (GloVe)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.87 | 0.85 | 0.86 | 16107 |
| 1 | 0.85 | 0.87 | 0.86 | 16217 |
| accuracy | | | 0.86 | 32324 |
| macro avg | 0.86 | 0.86 | 0.86 | 32324 |
| weighted avg | 0.86 | 0.86 | 0.86 | 32324 |

### 3. BILSTM without added Attention Mechanism and without Pre-Trained Embedding

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.69 | 0.81 | 0.74 | 16107 |
| 1 | 0.77 | 0.64 | 0.70 | 16217 |
| accuracy | | | 0.72 | 32324 |
| macro avg | 0.73 | 0.72 | 0.72 | 32324 |
| weighted avg | 0.73 | 0.72 | 0.72 | 32324 |

## 4. BILSTM with added Attention Mechanism and Pre-Trained Embedding

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.87 | 0.89 | 16107 |
| 1 | 0.88 | 0.90 | 0.89 | 16217 |
| accuracy | | | 0.89 | 32324 |
| macro avg | 0.89 | 0.89 | 0.89 | 32324 |
| weighted avg | 0.89 | 0.89 | 0.89 | 32324 |