National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Anusha Ananth
Student ID: x21131465

School of Computing
National College of Ireland

Supervisor: Mr Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Anusha Ananth
Student ID:	x21131465
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Mr Vladimir Milosavljevic
Submission Due Date:	01/02/2023
Project Title:	Configuration Manual
Word Count:	511
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:	
Date:	1st February 2023

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

Anusha Ananth
x21131465

1 Introduction

This document contains comprehensive information on the hardware and software components used in the project. It also describes the packages utilized when running the application. Furthermore, it provides a detailed step-by-step approach for the project "Role of Social Media in Start-ups Success using Machine Learning Approaches: Twitter".

2 System Configuration

2.1 Hardware Configuration

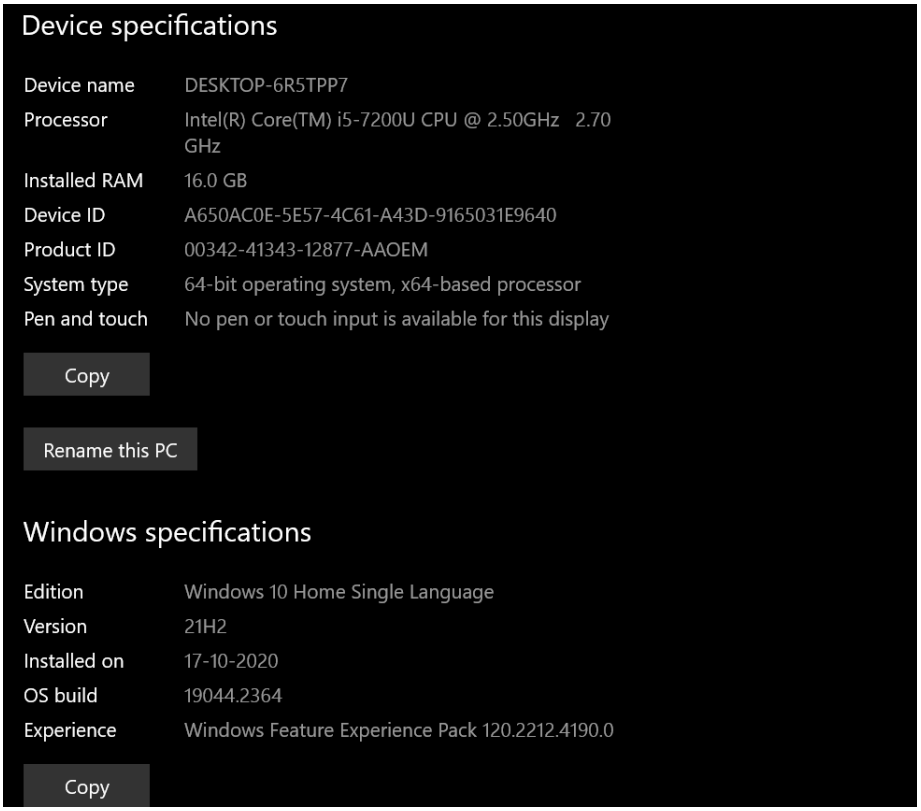

Figure 1: System Specification

3 Applications used in this Research

To execute the project Jupyter Notebook was used and all the machine learning algorithms were developed using the python programming language. The following were the main libraries used in this project.

- **Jupyter Notebook:** 6.4.11
- **python language:** 3.8.5

4 Python Libraries

- **Numpy:** 1.22.3
- **Pandas:** 1.4.2
- **Keras:** 2.11.0
- **matplotlib:** 3.5.1
- **sklearn:** 1.1.3
- **Tensorflow:** 2.11.0
- **Selenium Webdriver:** 4.1.3
- **Beautifulsoup:** 4.11.1

The libraries used in the project are explained in the section below:-

- **Numpy:** Numpy is a python library function that helps in doing complex array and matrix operations on a complex mathematical function.
- **Pandas:** It is a python library that helps complex analytical operations which are used in the fields of data analytics.
- **sklearn:** This helps in solving regression and classification algorithms.
- **Keras:** Keras is a python package which acts as an interface between the neural network.
- **Selenium Webdriver:** Scrapes the website as scripts and these scripts have to be extracted.
- **Beautifulsoup:** Is used to scrape the data from the website it acts between the API and the programming language.

5 Data sources

In this research, there are two datasets were utilized. One of the datasets is from Kaggle which is an open-source platform where every day thousands of data sets are uploaded by data scientists and researchers use the dataset for their research purposes. And another dataset was scraped from Twitter using Selenium Webdriver.

6 Code Snippets

```
months = {'Jan':1, 'Feb':2, 'Mar':3, 'Apr':4, 'May':5, 'Jun':6, 'Jul':7, 'Aug':8, 'Sep':9, 'Oct':10, 'Nov':11, 'Dec':12}
def extract_dates(date_str):
    if len(date_str) < 8:
        return None
    yr = date_str[-4:]
    mon_str = date_str[:3]
    mon_num = None
    for k in months.keys():
        if k == mon_str:
            mon_num = months[k]
    if not yr.isdigit() or mon_num is None:
        return None
    mon_num = (mon_num - 1)
    if mon_num == 0:
        mon_num = 12
    end_date = str(yr) + '-' + str(mon_num) + '-15'
    start_date = str(int(yr) - 10) + '-' + str(mon_num) + '-15'
    return (start_date, end_date)
```

Figure 2: Twitter scraping based on date

Figure 2 shows the code snippet scraping the data based on the date required for analysis.

```
Percent progress 3.33667000333667 running for 1.672506332397461e-05 mins
Percent progress 4.170837504170837 running for 3.926356633504232e-05 mins
Percent progress 5.005005005005005 running for 5.592902501424154e-05 mins
Percent progress 5.839172505839173 running for 7.505416870117188e-05 mins
Percent progress 6.67334000667334 running for 9.327332178751627e-05 mins
Percent progress 7.5075075075075075 running for 0.00010836919148763021 mins
Percent progress 8.341675008341674 running for 0.00012508630752563476 mins
Percent progress 9.175842509175842 running for 0.00014170010884602865 mins
Percent progress 10.01001001001001 running for 0.00015912453333336783 mins
Percent progress 10.844177510844178 running for 0.00017580191294352214 mins
Percent progress 11.678345011678346 running for 0.00019243955612182617 mins
Percent progress 12.512512512512513 running for 0.00021291971206665038 mins
Percent progress 13.34668001334668 running for 0.00022954543431599934 mins
Percent progress 14.180847514180847 running for 0.0002462506294250488 mins
Percent progress 15.015015015015015 running for 0.00026549895604451496 mins
Percent progress 15.849182515849183 running for 0.00028410752614339195 mins
```

Figure 3: The output of Twitter Scraping

Figure 3 shows the output of scraping the data.

Figure 4 shows the epochs of the RNN model. There were 50 epochs run for this model which gave an accuracy of 92%.

Figure 5 shows the code snippet of the RNN model which calculates the accuracy, loss and score of the RNN model.

```
Epoch 40/50
2719/2719 - 328s - loss: 0.1823 - accuracy: 0.9239 - 328s/epoch - 121ms/step
Epoch 41/50
2719/2719 - 340s - loss: 0.1810 - accuracy: 0.9256 - 340s/epoch - 125ms/step
Epoch 42/50
2719/2719 - 330s - loss: 0.1812 - accuracy: 0.9266 - 330s/epoch - 121ms/step
Epoch 43/50
2719/2719 - 329s - loss: 0.1766 - accuracy: 0.9277 - 329s/epoch - 121ms/step
Epoch 44/50
2719/2719 - 327s - loss: 0.1775 - accuracy: 0.9261 - 327s/epoch - 120ms/step
Epoch 45/50
2719/2719 - 332s - loss: 0.1766 - accuracy: 0.9265 - 332s/epoch - 122ms/step
Epoch 46/50
2719/2719 - 329s - loss: 0.1779 - accuracy: 0.9265 - 329s/epoch - 121ms/step
Epoch 47/50
2719/2719 - 332s - loss: 0.1770 - accuracy: 0.9279 - 332s/epoch - 122ms/step
Epoch 48/50
2719/2719 - 332s - loss: 0.1752 - accuracy: 0.9274 - 332s/epoch - 122ms/step
Epoch 49/50
2719/2719 - 327s - loss: 0.1741 - accuracy: 0.9289 - 327s/epoch - 120ms/step
Epoch 50/50
2719/2719 - 334s - loss: 0.1731 - accuracy: 0.9287 - 334s/epoch - 123ms/step
```

Figure 4: RNN Epochs

```
validation_size = 1500

X_validate = X_test[-validation_size:]
Y_validate = Y_test[-validation_size:]
X_test = X_test[:-validation_size]
Y_test = Y_test[:-validation_size]
score,acc = model.evaluate(X_test, Y_test, verbose = 2, batch_size = batch_size)
print("score: %.2f" % (score))
print("acc: %.2f" % (acc))
```

```
1292/1292 - 28s - loss: 0.6511 - accuracy: 0.8125 - 28s/epoch - 22ms/step
score: 0.65
acc: 0.81
```

Figure 5: RNN Accuracy

7 Steps of file execution

- **1.TwitterScraping.ipynb** must be executed which consists of the Twitter scraping
- **2.EDA.ipynb** must be executed which contains the exploratory data analysis of the dataset.
- **3.TweetFeatureExtraction.ipynb** this file contains code that the features are extracted from the scraped data.
- **4.RandomForestRegression.ipynb** contains the code for the Random Forest regressor model.
- **5.SentimentAnalysis.ipynb** contains the code for the sentiment analysis of Twitter tweets.
- **6.CorrelationAnalysis.ipynb** code for the correlation analysis between the dependent an the independent variables.
- **7.SVR.ipynb** The Support Vector Regressor model is developed.
- **8.RegressionAnalysis.ipynb** Regression Analysis code is written and the RMSE value is calculated.
- **9.ScaledRegressionAnalysis.ipynb** Here the scaled features are used for the regression analysis.
- **10.RNN.ipynb** Recursive Neural Network is discussed here.