Configuration Manual

MSc Research Project
Data Analytics

Gulbahar Erol
Student ID: x23136235

School of Computing
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Gulbahar Erol		
Student ID:	23136235		
Programme:	Msc Data Analytics	**Year:**	2023/2024
Module:	Research Project		
Lecturer:	Naushad Alam		
Submission Due Date:	12 August		
Project Title:	The Impact of Deep Learning on Multilingual Toxic Data Analysis Review		
Word Count:	……………………………………… **Page Count:** ………………………………………………		

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: Gulbahar Erol

Date: 11.08.2024

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

Gulbahar Erol
Student ID: x23136235

1 Introduction

This file has been prepared to show the software and hardware requirements for the Deep Learning Multilingual Toxic comment analysis project, as well as the necessary steps for implementation.

2 System Configuration

2.1 Hardware Requirements

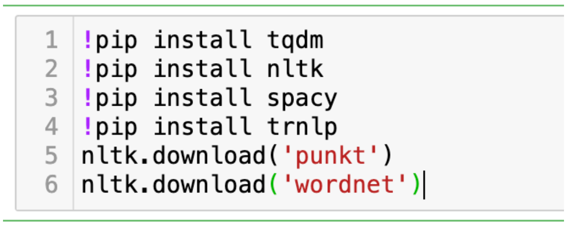
A multi-core processor (Intel i5 or higher) and 16GB RAM are recommended.

2.2 Software Requirements

Compatible with Windows, macOS, or Linux. For project implementation, Jupyter Notebook version 6.3.0 with the Python kernel version 3.8.8 and Google Colab Pro+ are used.

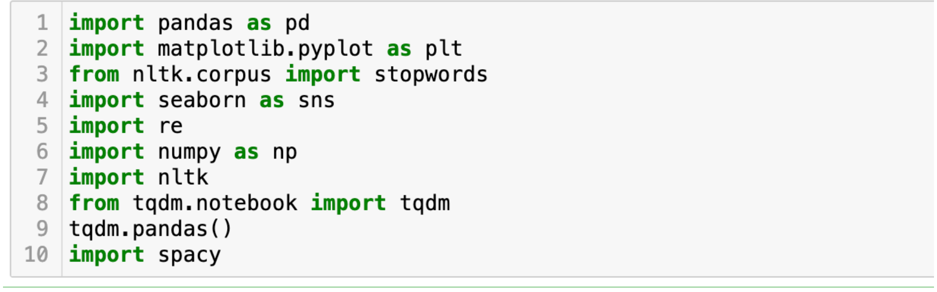
3 Project Implementation

Downloading required libraries. In figure 1,2 shown libraries used in EDA analysis in Jupyter notebook.



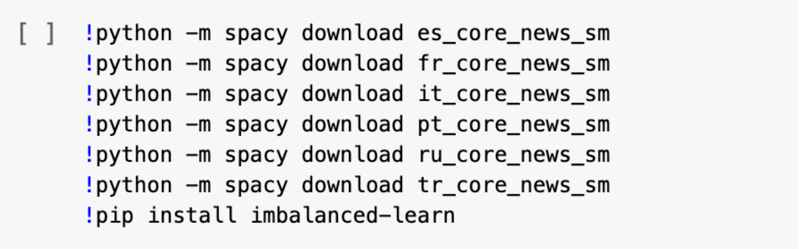
```
!pip install tqdm
!pip install nltk
!pip install spacy
!pip install trnlp
nltk.download('punkt')
nltk.download('wordnet')
```

Figure 1 Installation Necessary Libraries



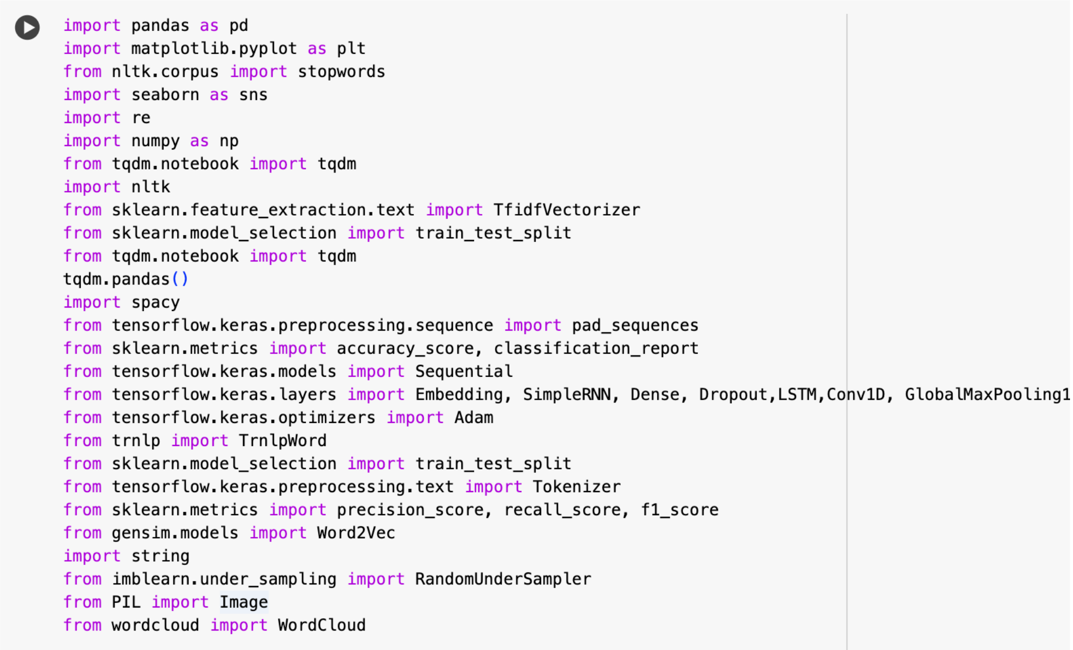
```
import pandas as pd
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
import seaborn as sns
import re
import numpy as np
import nltk
from tqdm.notebook import tqdm
tqdm.pandas()
import spacy
```

Figure 2 Importing Necessary Libraries



```
[ ]  !python -m spacy download es_core_news_sm
     !python -m spacy download fr_core_news_sm
     !python -m spacy download it_core_news_sm
     !python -m spacy download pt_core_news_sm
     !python -m spacy download ru_core_news_sm
     !python -m spacy download tr_core_news_sm
     !pip install imbalanced-learn
```

Figure 3 Installation libraries



```
import pandas as pd
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
import seaborn as sns
import re
import numpy as np
from tqdm.notebook import tqdm
import nltk
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from tqdm.notebook import tqdm
tqdm.pandas()
import spacy
from tensorflow.keras.preprocessing.sequence import pad_sequences
from sklearn.metrics import accuracy_score, classification_report
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, SimpleRNN, Dense, Dropout,LSTM,Conv1D, GlobalMaxPooling1
from tensorflow.keras.optimizers import Adam
from trnlp import TrnlpWord
from sklearn.model_selection import train_test_split
from tensorflow.keras.preprocessing.text import Tokenizer
from sklearn.metrics import precision_score, recall_score, f1_score
from gensim.models import Word2Vec
import string
from imblearn.under_sampling import RandomUnderSampler
from PIL import Image
from wordcloud import WordCloud
```

Figure 4 Importing Libraries

3.1 Data Collection

Data was obtained from Kaggle. 6 different csv files were imported. Language was defined under the 'language' heading in a new column and 6 versions were collected in a single data frame.

3.2 EDA

Jupyter notebook was used for data visualization and general analysis. Data types, distributions by languages, toxic comment distributions were visualized. Special characters, stop words, capital letter numbers, unique word numbers were visualized.

Six new columns are created which are number of words in each class, number of characters in each class, number of unique words in each class, number of special characters in each class, number of stopwords in each class, number of capital letters in each class. After visualization, columns were removed as they were not required for analysis.

3.3 Data Preprocessing

Since the data size was large, Google Colab Pro+ notebook was used for the next steps. Special characters, stop words, punctuation, URLs in the texts were removed. Capital letters were replaced with lower case letters. NAs were removed. Random under sampling was applied because the data set was not balanced.

The lemmatization step was applied to convert the text data into a format that the models can understand. Languages were grouped and separate lemmatization was applied for each group. Since there was no Turkish support in the Spacy library, a separate lemmatization step was applied for Turkish.

3.4 Modelling

Vectorization was applied with word2vec for each model. After the tokenization process, vectorization was added to the embedding layer in the model application step. LSTM, RNN, CNN models were applied sequentially.

3.5 Evaluation

Due to the imbalance of the data set, precision, recall and f1-score metrics were also included in the evaluation.