

Configuration Manual

MSc Research Project
Artificial Intelligence

Phanisai kotha
Student ID: 23415754

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Phansai Kotha

Student ID: 23415754

Programme: : Artificial Intelligence

Year: 2025

Module: MScResearchProject

Lecturer: Rejwanul Haque

Submission Due

Date: 28/01/2026

Project Title: Network based social media sentimental analysis on twitter

Word Count: 540

Page Count: 6

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: Phanisai kotha

Date: 28/01/2026

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

Phanisai Kotha
23415754

Network-Based Social Media Sentiment Analysis on Twitter

1 System Requirements

1.1 Hardware Requirements

- Minimum 8 GB RAM; recommended 16 GB
- Dual-core CPU or above
- GPU optional (training supported on CPU)
- Minimum 10 GB free storage

1.2 Software Requirements

- Windows 10/11, macOS, or Linux
- Python 3.10 or above
- pip for package installation
- Optional IDE: Jupyter Notebook, VS Code, PyCharm
- Browser: Chrome or Firefox for viewing Streamlit UI

2 Project Folder Structure

Place files in the following structure:

```
PROJECT_ROOT/
├── archive/
│   └── Sentiment140.csv
├── app/
│   ├── streamlit_app.py
│   ├── logreg_model.pkl
│   ├── logreg_vectorizer.pkl
│   ├── svm_model.pkl
│   ├── svm_vectorizer.pkl
│   └── bert_model/
│       ├── config.json
│       ├── tokenizer.json
│       ├── vocab.txt
│       └── model.safetensors
```

```
|  
|— Sentimental_analysis.ipynb  
|— requirements.txt
```

Maintaining this structure ensures all paths resolve correctly.

3 Installation and Environment Setup

3.1 Creating a Virtual Environment

```
python -m venv sentiment_env
```

Activate:

Windows

```
sentiment_env\Scripts\activate
```

macOS/Linux

```
source sentiment_env/bin/activate
```

3.2 Installing Dependencies

Ensure the following packages exist in `requirements.txt`:

```
pandas  
numpy  
matplotlib  
seaborn  
scikit-learn  
nltk  
torch  
transformers  
datasets  
streamlit  
pyvis  
networkx
```

Install via:

```
pip install -r requirements.txt
```

Download NLTK resources:

```
python -c "import nltk; nltk.download('stopwords')"
```

4 Dataset Configuration

4.1 Placing the Dataset

Move the Sentiment140 dataset to:

archive/Sentiment140.csv

The notebook loads it using latin-1 encoding.

4.2 Sampling Process

The system uses:

- Approximately 1.51 million tweets after duplicate removal
- A **20% stratified sample** (around 303,487 tweets)
- An **80/20 train-validation split**

These steps execute automatically through notebook code.

5 Preprocessing Configuration

The cleaning pipeline applies:

```
def clean_text(s):  
    s = str(s).lower()           # converting into lowercase  
    s = re.sub(r'http\S+|www\.\S+', '', s) # this removes URLs  
    s = re.sub(r'@\w+', '', s)    # this removes mentions  
    s = re.sub(r'#\w+', '', s)    # this removes hashtags  
    s = re.sub(r'^a-z\s]', '', s) # keep letters & spaces  
    s = re.sub(r'\s+', ' ', s).strip() # collapse spaces  
    return s
```

Stopword handling is performed internally by the TF-IDF vectoriser rather than manually in code.

BERT uses its own tokenizer; NLTK is used only for EDA visualisation and token frequency inspection.

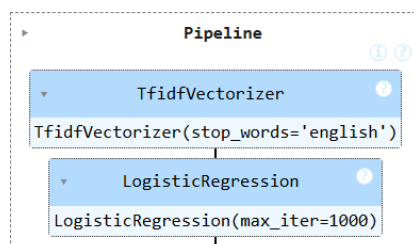
No manual adjustments are required here.

6 Model Training Configuration

Launch Sentimental_analysis.ipynb to train all models.

6.1 Logistic Regression Training

Configuration:



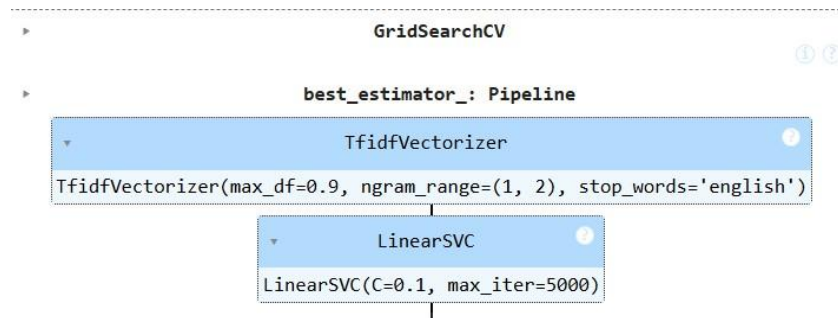
- TF-IDF with English stopwords
- n-gram range of (1,1)
- Logistic Regression with `max_iter=1000`

Outputs:

- `logreg_model.pkl`
- `logreg_vectorizer.pkl`
- Confusion matrix
- ROC curve
- Classification report

6.2 SVM (LinearSVC) Training

GridSearchCV evaluates:



- C values: 0.01, 0.1, 1
- n-gram ranges: (1,1) and (1,2)
- 5-fold cross-validation
- scoring using macro F1

Outputs:

- `svm_model.pkl`
- `svm_vectorizer.pkl`
- confusion matrix
- ROC curve
- evaluation metrics

6.3 BERT Fine-Tuning

Configuration:

- model: `bert-base-uncased`
- epochs: 1
- batch size: 32

- max sequence length: 64
- optimizer: AdamW at lr=2e-5
- scheduler: linear decay

BERT produces:

- batch-level loss values
- epoch-level average loss
- bert_model/ folder with all files needed for deployment

7 Evaluation Configuration

The notebook automatically generates:

- accuracy
- precision
- recall
- F1-score
- ROC-AUC
- confusion matrices
- ROC curves
- training and prediction times
- consolidated model comparison table

No manual configuration is required.

8 Streamlit Application Configuration

8.1 Preparing the Application

Copy into /app/:

```
logreg_model.pkl
logreg_vectorizer.pkl
svm_model.pkl
svm_vectorizer.pkl
bert_model/
```

8.2 Running the Streamlit App

```
cd app
streamlit run streamlit_app.py
```

Access through:

`http://localhost:8501`

8.3 Features Provided

- real-time sentiment prediction

- batch CSV prediction should be less than 200 mb and evaluation
- probability bar chart
- confidence gauge
- prediction logging
- interactive knowledge graph explorer
- filtering tweets by keyword or user

9 Troubleshooting Guide

9.1 Missing Transformers Module

```
pip install transformers
```

9.2 BERT Model Not Loading

Verify presence of:

```
app/bert_model/config.json
```

If missing, copy entire model directory again.

9.3 Streamlit CSV Error

Ensure CSV contains a column named either **text** or **tweet**

9.4 Memory Errors

Reduce:

- Number of TF-IDF features
- BERT batch size