

Configuration Manual

MSc Research Project
Msc Data Analytics

Sreenivas
Student ID:X23326603

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

**National College of Ireland
Project Submission Sheet
School of Computing**



Student Name:	Sreenivas
Student ID:	X23326603
Programme:	Msc Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	11/08/2025
Project Title:	Configuration Manual
Word Count:	1044
Page Count:	10

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:	sreenivas
Date:	11th August 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):	

Configuration Manual

Sreenivas
X23326603

1 Introduction and Research Purpose

This configuration manual documents the setup, configuration, and execution of the project **Comparative Sentiment Analysis using GPT-3.5 Labeling, SVM, and BERT with SHAP Explainability**.

The project aims to answer:

1. How do traditional ML models (SVM) compare with transformer-based models (BERT) in interpreting sentiment from student reviews?
2. How can explainable AI (SHAP) provide insight into model decision-making at both token and global feature levels?

The pipeline uses a dataset sourced from Kaggle's "Course Reviews on Coursera" dataset¹.

The manual is intended for researchers, developers, and system administrators who wish to set up, run, and maintain this pipeline.

2 System Requirements

2.1 Hardware

- CPU: Quad-core or better
- RAM: 16 GB minimum
- GPU: NVIDIA CUDA-enabled GPU (recommended for BERT)
- Storage: 10 GB free space

2.2 Software

- OS: Windows 10+, Ubuntu 20.04+, macOS 12+
- Python 3.9+
- CUDA 11.x (if using GPU)

¹https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera?select=Coursera_reviews.csv

3 Dataset Information

- **File:** Coursera_reviews.csv
- **Source:** Kaggle
- **Size:** 137,000 rows
- **Key Columns:**
 - reviews - Student feedback text
 - course name - Name of course
 - rating - Numeric rating
- **Usage in Pipeline:** The reviews column is labeled using GPT-3.5 sentiment analysis and used for model training and SHAP explanations.

4 Installation and Environment Setup

1. Ensure you have Python 3.9 or higher installed on your system.
2. Create a new project directory in your preferred location:

```
mkdir sentiment_project
cd sentiment_project
```

3. Create a virtual environment:

```
python -m venv venv
source venv/bin/activate    # Linux/Mac
venv\Scripts\activate      # Windows
```

4. Install the required Python packages:

```
pip install openai pandas tqdm scikit-learn joblib shap torch \
transformers nltk matplotlib seaborn
```

5. Download the dataset from Kaggle:

<https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera>

Save it in the project directory with the name:

Coursera_reviews.csv

6. Create the Python scripts described in the following sections:

- gpt.py
- svm_model.py
- shap_svm_tree.py
- bert_model.py
- bert_shap_final.py

using the code provided in this manual.

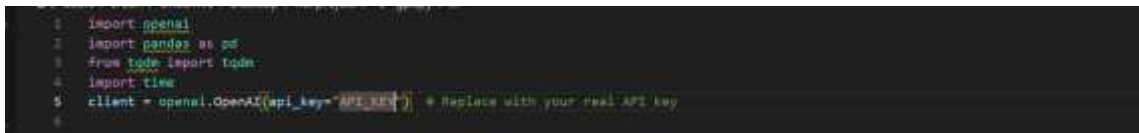
5 Project Structure Overview

```
project/
python_scripts/                                # Local code files (not in Overleaf)
  gpt.py                                       # GPT-3.5 labeling
  svm_model.py                               # SVM training
  shap_svm_tree.py                           # SVM SHAP explainability
  bert_model.py                              # BERT fine-tuning
  bert_shap_final.py                         # BERT SHAP explainability
data/                                         # Dataset and outputs
  Coursera_reviews.csv                       # Original dataset
  GPT3.5_All_Reviews_Labeled.csv
  shap_linear_svc_top1000.csv
  svm_global_importance_from_1000_reviews.csv
  bert_predictions.csv
  shap_bert_token_level_explanations.csv
  shap_bert_global_importance.csv
logos/                                       # Logos and flowchart images
  editor_mermaid_chart.png
  NCI_Logo_colour.png
text/                                       # LaTeX section files
  section1.tex
  section2.tex
  ...
  section10.tex
titlepage.tex                               # Title page layout
refs.bib                                    # References
configManual.tex                            # Main LaTeX file for manual
```

6 Configuration Settings

6.1 API Keys

Open gpt.py and replace the placeholder API key with your own. Figure 1 shows the exact location in the script.



```

1 import openai
2 import pandas as pd
3 from tqdm import tqdm
4 import time
5 client = openai.OpenAI(api_key="API_KEY") # Replace with your real API key
6

```

Figure 1: Location in gpt.py to insert OpenAI API key (key masked for security)

6.2 Model Paths

- **SVM Model:** svm_model.pkl
- **TF-IDF Vectorizer:** tfidf_vectorizer.pkl
- **BERT Model Folder:** ./bert sentiment_model

6.3 Dataset Files

- Ensure Coursera_reviews.csv is present in the data/ directory.
- This dataset must be downloaded from Kaggle: <https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera>

7 Module Execution Guide

This section explains how to execute each component of the pipeline in the correct order, showing the key code portions and expected outputs.

7.1 Step 1: GPT-3.5 Sentiment Labeling

Script: gpt.py

Purpose: Labels each review in Coursera reviews.csv using GPT-3.5.

Command:

```
python gpt.py
```

Key Code Snippet:

```

client = openai.OpenAI(api_key="your_api_key_here")

def query_gpt_sentiment(review_text):
    prompt = (
        "Classify the sentiment of the following student review as "
        "Positive, Negative, or Neutral:\n\n"
        f"Review: \"{review_text}\"\n\nSentiment:"
    )
    response = client.chat.completions.create(
        model="gpt-3.5-turbo",
        messages=[{"role": "user", "content": prompt}],
        temperature=0,
        max_tokens=16
    )
    return response.choices[0].message.content.strip()

```

Expected Output:

All reviews labeled and saved.

7.2 Step 2: SVM Model Training

Script: svm_model.py

Purpose: Preprocesses reviews, applies TF-IDF, tunes hyperparameters, trains SVM, and saves the model.

Command:

```
python svm_model.py
```

Key Code Snippet:

```
tfidf_vectorizer = TfidfVectorizer(
    max_features=10000,
    ngram_range=(1, 2),
    min_df=2,
    max_df=0.95,
    sublinear_tf=True
)

param_grid = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf'],
    'gamma': ['scale', 'auto']
}

grid_search = GridSearchCV(
    SVC(random_state=42, probability=True),
    param_grid,
    scoring='f1_macro',
    cv=5
)
```

Expected Output:

```
Best parameters: {'C': 1, 'kernel': 'linear', 'gamma': 'scale'}
Final Test Accuracy: 0.85
Final Test F1-Score: 0.84
```

7.3 Step 3: SVM SHAP Explainability

Script: shap_svm_tree.py

Purpose: Generates local and global SHAP explanations for SVM predictions.

Command:

```
python shap_svm_tree.py
```

Key Code Snippet:

```
explainer = shap.LinearExplainer(model, X_top,
                                feature_perturbation="interventional")
shap_values = explainer.shap_values(X_top)
```

```
top_indices = row.argsort()[::-1][:5]
explanation = "; ".join([f"{feature_names[j]}: {row[j]:.4f}"
                        for j in top_indices])
```

Expected Output:

```
shap_linear_svc_top1000.csv
svm_global_importance_from_1000_reviews.csv
```

—

7.4 Step 4: BERT Model Training

Script: bert_model.py

Purpose: Fine-tunes BERT for sentiment classification.

Command:

```
python bert_model.py
```

Key Code Snippet:

```
model = BertForSequenceClassification.from_pretrained(
    'bert-base-uncased',
    num_labels=len(label_encoder.classes_)
)

optimizer = AdamW(model.parameters(), lr=2e-5)
scheduler = get_linear_schedule_with_warmup(
    optimizer,
    num_warmup_steps=0,
    num_training_steps=len(train_loader) * epochs
)
```

Expected Output:

```
Epoch 1/2 - Average training loss: 0.45
Test Accuracy: 0.88
Test F1-Score: 0.87
```

—

7.5 Step 5: BERT SHAP Explainability

Script: bert_shap_final.py

Purpose: Produces token-level and global SHAP explanations for BERT predictions.

Command:

```
python bert_shap_final.py
```

Key Code Snippet:

```
explainer = shap.Explainer(pipe, tokenizer)
shap_values = explainer(texts)

token_contribs.append((token_lower, float(val)))
top_tokens = sorted(token_contribs, key=lambda x: abs(x[1]), reverse=True)[:5]
```

Expected Output:


```
shap_bert_token_level_explanations.csv
shap_bert_global_importance.csv
```

8 Flowchart

The following flowchart illustrates the complete pipeline, from data collection to model explainability outputs.



Figure 2: Pipeline Flowchart for GPT-3.5 Labeling, SVM and BERT Training, and SHAP Explainability

9 Expected Outputs and Verification

After running the pipeline, the following outputs should be generated. These files can be found in the `data/` directory unless otherwise specified.

9.1 GPT Labeling Output

- **File:** GPT3.5_All_Reviews_Labeled.csv
- **Description:** Contains the original dataset with an additional column gpt_sentiment containing sentiment labels (Positive, Negative, Neutral).
- **Verification:** Open in Excel and check that the gpt sentiment column is populated for all rows.

reviews	reviewers	date_reviews	rating	course_id	gpt_sentiment
Pretty dry, but I was able to pass with just two complete w	By Robert S	Feb 12, 2020		4 google-cbrs-cpi-training	Positive
would be a better experience if the video and screen shots w	By Gabriel E R	Sep 28, 2020		4 google-cbrs-cpi-training	Negative
Information was perfect! The program itself was a little anno	By Jacob D	Apr 08, 2020		4 google-cbrs-cpi-training	Positive
A few gram	By Dale B	Feb 24, 2020		4 google-cbrs-cpi-training	Neutral
Excellent c	By Sean G	Jun 18, 2020		4 google-cbrs-cpi-training	Positive
Some of th	By Daniel F	Dec 23, 2019		4 google-cbrs-cpi-training	Negative
Solid pres	By Logan D	Sep 03, 2020		5 google-cbrs-cpi-training	Positive
Probably th	By Luis M C	Nov 21, 2019		5 google-cbrs-cpi-training	Positive
The Proct	By scott w	Sep 28, 2020		5 google-cbrs-cpi-training	Negative
Covered al	By Ryan H	Aug 26, 2019		5 google-cbrs-cpi-training	Positive
Great cour	By Samuel D	Jan 24, 2020		5 google-cbrs-cpi-training	Positive
The cours	By Scot A W	May 01, 2020		5 google-cbrs-cpi-training	Positive
Pretty well	By Jonathan L	Jan 17, 2020		5 google-cbrs-cpi-training	Positive
Great cour	By Geoff T	Jul 21, 2019		5 google-cbrs-cpi-training	Positive
The cours	By William J	Jun 07, 2019		5 google-cbrs-cpi-training	Neutral
Well prese	By Ariel R	Sep 27, 2019		5 google-cbrs-cpi-training	Positive
Course is e	By Austin N	Dec 27, 2019		5 google-cbrs-cpi-training	Positive
The instr	By Jayson P J	Nov 22, 2019		5 google-cbrs-cpi-training	Positive
Easy to fol	By Robert S	Nov 07, 2019		5 google-cbrs-cpi-training	Positive
Excellent T	By Scott C	Sep 02, 2020		5 google-cbrs-cpi-training	Positive
very helpf	By Wei W	Mar 23, 2020		5 google-cbrs-cpi-training	Positive
Material c	By Hugh H	Sep 01, 2019		5 google-cbrs-cpi-training	Positive

Figure 3: Sample of GPT-labeled dataset showing the added gpt_sentiment column

9.2 SVM SHAP Output

- **Files:**
 - shap_linear_svc_top1000.csv – Local feature contributions for the top 1000 long reviews.
 - svm_global_importance_from_1000_reviews.csv – Global feature importance across these reviews.
- **Verification:** Open in Excel and check for valid feature names and contribution values.

Review	Predicted_Sentiment	Actual_Sentiment	Top_Feature_Contributions
Pretty dry, but I was able to pass with just two complete w	Positive	Positive	module rest: 0.3514; effect extremity: 0.1400; shiller cheerful: 0.0570; absurde n
would be a better experience if the video and screen shots w	Negative	Negative	shiller changing: 0.1660; effect extremity: 0.1242; developer course: 0.0692; develop
Information was perfect! The program itself was a little anno	Positive	Positive	shiller changing: 0.1660; shiller brilliant: 0.0584; completely information: 0.0514; deal
A few grammatical mistakes on test made me do a double ts	Neutral	Neutral	curiosity: 0.4741; shiller changing: 0.1660; shiller brilliant: 0.0584; deal basic: 0.0276
Excellent course and the training provided was very detailed	Positive	Positive	shiller cheerful: 0.0738; deal basic: 0.0276; understand happening: 0.0170; dordresle
Some of the quizzes contained material not explicitly cover	Negative	Negative	quora: 0.6140; que trouze: 0.6035; understand general: 0.2090; shiller cheerful: 0.123
Solid presentation all the way through. I really appreciated th	Positive	Positive	can't jump: 0.5885; shiller changing: 0.1660; shiller brilliant: 0.0584; keep track: 0.628
Probably the best certification course I've taken in this resp	Positive	Positive	dot: 0.2703; finance humane: 0.1876; course took one: 0.1874; finance however: 0.16
The ProctorU.com system took 2 times the amount of time s	Negative	Negative	complex concept: 0.2275; module rest: 0.1410; cognitive: 0.1161; complex financial
Covered all of the required information in an easy to underst	Positive	Positive	shiller changing: 0.1660; understand general: 0.0910; shiller brilliant: 0.0584; everyth
Great course! Lectures were straight forward and easy to fol	Positive	Positive	can't jump: 0.1400; interaction want: 0.1091; shiller cheerful: 0.0670; understand

Figure 4: Sample of local SHAP explanations for the SVM model

Token	Mean_Contribution_Over_1000
download	0.00235247
never received	0.00235247
contact	0.00235247
password	0.00235247
email	0.00235247
please contact	0.00235247
password please	0.00235247
received	0.002246813
please	0.001854651
fine	0.000593623
grammatical	0.00017076
double take	0.00017076
do	0.00017076
double	0.00017076
made	0.00017076
test made	0.00017076
never	0.000111223
view	7.90E-05
better experience	7.90E-05
sho	7.90E-05
go	7.90E-05

Figure 5: Sample of global SHAP feature importance for the SVM model

9.3 BERT Predictions and SHAP Output

- **Files:**
 - bert_predictions.csv – True and predicted labels for the test dataset.
 - shap_bert_token_level_explanations.csv – Token-level SHAP importance.
 - shap_bert_global_importance.csv – Global SHAP feature importance.
- **Verification:** Open bert_predictions.csv in Excel and verify predictions match the label distribution. Check SHAP CSVs for meaningful token contributions.

Review	Top_Token_Contributions
Pretty dry, but I was able to pass with just two comple dry:	0.0082; pretty: -0.0030
would be a better experience if the video and screen s would:	0.0062
Information was perfect! The program itself was a littl information:	-0.0101
A few grammatical mistakes on test made me do a double take but all in all not bad.	
Excellent course and the training provided was very d course:	0.0202; excellent: -0.0055
Some of the quizzes contained material not explicitly covered in the course.	
Solid presentation all the way through. I really appreci presentation:	0.0132; solid: -0.0031
Probably the best certification course I've taken in thi probably:	-0.0102
The ProctorU.com system took 2 times the amount of proctor:	-0.0039
Covered all of the required information in an easy to u covered:	-0.0036
Great course, lectures were straight forward and eas course:	0.0418; great: -0.0185
The course was straight forward and prepared me for course:	0.0026
Pretty well designed course. Few things on the test w well:	0.2622; pretty: -0.0770
Great course, i have feedback about one question an course:	0.1043; great: -0.0326
The course was fine but I never received an email to d course:	0.0003
Well presented course material that takes you step b presented:	0.1068; well: -0.0321
Course is easy enough if you have a basic background course:	0.0123
The instructions were clear and concise. I enjoyed the instructions:	-0.0017
Easy to follow. Did a good job at teaching me the mat easy:	-0.0186
Excellent Training. More than enough to pass the test testing:	0.0055; excellent: 0.0221

Figure 6: Sample of token-level SHAP explanations for the BERT model

Token	Mean_Contribution_Over_1000
helpful	0.351221804
well	0.115071986
presented	0.106807666
training	0.095475554
covered	0.024481626
course	0.017953695
presentation	0.013231101
dry	0.008227465
would	0.006167066
interesting	0
tough	0
structured	0
poorly	0
find	-0.000806908
prefer	-0.001457863
instructions	-0.001653921
honestly	-0.003059377
solid	-0.003149404
proctor	-0.003914626
pace	-0.004003436
terrible	-0.006162287
lecturer	-0.006892094
co	-0.00831015
waste	-0.008591943
material	-0.008959499
information	-0.010105336
probably	-0.01023052
easy	-0.018607481

Figure 7: Sample of global SHAP feature importance for the BERT model