User Guide  
Authorship Verification using Impostor Projections and Siamese Networks

**Purpose**  
This document is the User’s Guide / Operating Instructions (תיק למשתמש) for the capstone project. It focuses only on the nominal (successful) workflow of the system and is intended for end‑users who wish to run experiments and inspect past results.

# 1. Overview

The project implements an Impostor‑style Siamese Network built on fine‑tuned, domain‑specific BERT models to verify the authorship of short text chunks. The codebase is split into small, clean Python modules that are orchestrated by a single main.py controller. This lets you drive a complex end‑to‑end pipeline from either a command‑line menu or a headless Colab notebook.

# 2. System Capabilities – Nominal Flow

## 2.1 Fine‑Tuned BERT Artifacts

Every time the pipeline fine‑tunes a BERT model for an impostor‑author, it:  
1. saves the checkpoint under data.fine\_tuned\_bert\_model\_path (default saved\_trained\_models/berts/);  
2. pushes the same directory to the HuggingFace Hub repository configured in config.yaml (currently ElyMK1/bert-shakespeare-english-mlm).

3. A BERT Tokenizer is also uploaded along with the MLM model. We use this fine-tuned tokenizer for preprocessing an impostor-author.

This process guarantees a local backup plus a single cloud source‑of‑truth, so you can later load the model with transformers.from\_pretrained() from any machine.

## 2.2 Interactive Menu (main.py)

python3 main.py

1 Create Dataset  
2 Run Model Procedure (train Siamese networks)  
3 Classification Procedure (signal generation)  
4 DTW Procedure (distance matrices between signals for each iteration– trained model)  
5 Isolation Forest Procedure (anomaly scores from each distance matrix)  
6 Clustering Procedure (final clusters of all anomaly scores vectors)  
999 Fine‑Tune All BERTs  
0 Exit  
  
Nominal path: 1 → 999 (Once only! Our HF repository already has all 31 fine-tuned BERTs on the listed hyperparameters down below) → 2 → 3 → 4 → 5 → 6 → 0. Each step requests confirmation—answer yes to proceed.

# 3. Prerequisites

\* Python 3.10 and the exact package versions in Pipfile.lock  
\* pipenv ≥ 2024.4.1 (pip install pipenv)  
\* CUDA 11.8 / cuDNN 8.1.1 on local GPUs (for code testing) or CUDA 12.5 on Colab (for real use).  
\* HuggingFace access token (write scope)  
\* Weights & Biases API key (if wandb.enabled: true – this key should generally not be disabled as our entire result-based procedures rely on its storage, unless locally testing).

# 4. Configuration File – config/config.yaml

Below is the exact configuration file shipped with the repository. Modify values here (not the code) to redirect artefact locations, tune hyper‑parameters, or toggle services.

data:  
 organised\_data\_folder\_path: "data/data\_sources"  
 shakespeare\_data\_source: "dataset\_shakespeare\_collection.json"  
 impostors\_data\_source: "dataset\_impostors.json"  
 shakespeare\_path: "data/raw/shakespeare"  
 impostors\_path: "data/raw/impostors"  
 all\_impostors\_data\_source: "dataset\_all\_impostors.json"  
 classify\_text\_data\_source: "text\_to\_classify.json"  
 fine\_tuned\_bert\_model\_path: "saved\_trained\_models/berts/"  
 trained\_siamese\_path: "saved\_trained\_models/models/"  
 signals\_folder\_name: "signals\_per\_model"  
 isolation\_forest:  
 isolation\_forest\_folder\_name: "isolation\_forest\_per\_model"  
 all\_models\_scores\_file\_name: "all\_models\_isolation\_forest\_score.json"  
 clustering\_output\_file: "clustering\_results.json"  
 dtw:  
 output\_distance\_folder: "distance\_per\_model"  
 dtw\_file\_name: "distance\_matrix.json"  
 included\_text\_names\_file\_name: "included\_text\_names.json"  
 signals\_file\_name: "processed\_signals.json"  
 pairs: "pairs.json"  
 all\_signals: "all\_signals.json"  
  
training:  
 load\_pretrained\_model: False  
 optimizer:  
 initial\_learning\_rate: 1e-5  
 learning\_rate\_decay\_factor: 0.1  
 gradient\_clipping\_threshold: 1.0 # or 5.0  
 training\_batch\_size: 8  
 epochs: 10  
 early\_stopping:  
 monitor: "val\_accuracy"  
 patience: 4  
 baseline: 0.97  
 impostor\_chunk\_ratio: 4  
 test\_split: 0.25  
  
model:  
 chunk\_to\_batch\_ratio: 8  
 chunk\_size: 50  
 cnn:  
 filters: 500  
 padding: "valid"  
 kernel\_size: [3, 6, 12]  
 pool\_size: 1  
 bilstm:  
 dropout: 0.25 # 0.2 - 0.5  
 units: 500  
 fc:  
 in\_features: 512  
 out\_features: 512  
  
isolation\_forest:  
 number\_of\_trees: 100  
 percentile\_threshold: 5  
 anomaly\_score\_threshold: 0.000025  
  
clustering:  
 algorithm: "k-medoids"  
 k-medoids:  
 n\_clusters: 2  
 random\_state: 0  
  
bert:  
 model: "bert-base-uncased"  
 repository: "ElyMK1/bert-shakespeare-english-mlm"  
 max\_sequence\_length: 512  
 mlm\_probability: 0.15  
 train\_batch\_size: 16  
 num\_epochs: 5  
 save\_steps: 1000  
 save\_total\_limit: 2  
 logging\_steps: 500  
 logging\_dir: "logs/shakespeare"  
 token: <<your api key - feel free to use ours: hf\_PcvaVcEXwbReburAnISPgkDSCFhBdzYSlI>>  
  
wandb:  
 enabled: True  
 run\_name: "full-procedure-run"  
 project: "siamese-authorship-verification"  
 artifact\_name: "siamese-authorship-verification-branches"  
 api\_key: <<your api key - feel free to use ours: 8dc3df6ba5ef5bae0ad60917973e82163f35bdf8>>

## 4.1 Field‑by‑Field Explanation

# Field‑by‑Field Explanation

| **Section** | **Key** | **Description & Effect** |
| --- | --- | --- |
| **data** | organised\_data\_folder\_path | Root folder grouping raw & processed data. |
|  | shakespeare\_data\_source / impostors\_data\_source | Names of JSON files with canonical texts for each author pool. |
|  | shakespeare\_path / impostors\_path | Where raw plaintext files are stored before dataset creation. |
|  | all\_impostors\_data\_source | Aggregated impostor collection used during to fine-tune BERTs. |
|  | classify\_text\_data\_source | JSON file of unknown text used during inference. |
|  | fine\_tuned\_bert\_model\_path | Local directory saving per‑author fine‑tuned BERT checkpoints before they are uploaded to HuggingFace. |
|  | trained\_siamese\_path | Folder for Keras .h5 weights produced by Run Model Procedure. |
|  | signals\_folder\_name | Sub‑dir created under each model folder containing JSON arrays of numerical signals. |
|  | isolation\_forest.\* | Location & filenames for anomaly scores produced by Isolation Forest. |
|  | clustering\_output\_file | Final JSON with cluster labels after Clustering Procedure. |
|  | dtw.\* | Folder + file naming pattern for DTW distance matrices and related helper files. |
|  | pairs | JSON file containing all pairs by impostor-author names, used to train models and use them during the inference procedures.  It also defines the last iterations, allowing us to pick up where we left off during the last runtime. |
| **training** | load\_pretrained\_model | If true, skip training models for impostor pairs, if they exist. Otherwise, train models and upload the version as the latest tag in the artifacts registry. |
|  | optimizer.\* | Learning‑rate schedule and gradient‑clipping settings. |
|  | training\_batch\_size / epochs | Core loop batch size & number of epochs. |
|  | early\_stopping.\* | Keras early‑stopping callback parameters. |
|  | impostor\_chunk\_ratio | Multiplier for data augmentation of the number of chunks from each impostor during pre-processing. |
|  | test\_split | Validation dataset split ratio. |
| **model** | chunk\_to\_batch\_ratio / chunk\_size | How many chunks are in a batch, and how many tokens are in a chunk. |
|  | cnn.\* | 1‑D convolution filter bank capturing local n‑gram features. |
|  | bilstm.\* | Bi‑directional LSTM width and dropout for contextual fusion. |
|  | fc.\* | Width of final fully‑connected layer before computing pairwise distance. |
| **isolation\_forest** | number\_of\_trees / percentile\_threshold / anomaly\_score\_threshold | Hyper‑parameters governing anomaly detection. |
| **clustering** | algorithm | Currently supports k‑medoids (robust to outliers). |
|  | k-medoids.\* | Cluster count (n\_clusters) and seed (random\_state) for deterministic medoid selection. |
| **bert** | All keys | Control HuggingFace Trainer for domain fine‑tuning (sequence length, MLM probability, logging, checkpoint cadence). |
| **wandb** | All keys | Enable/disable W&B, set run names and artefact naming conventions. |

# 5. Monitoring & Visualising Experiments with Weights & Biases

Key URLs:

Dashboard (Runs): <https://wandb.ai/authorship-verification-siamese-network/siamese-authorship-verification>  
Artifacts: <https://wandb.ai/authorship-verification-siamese-network/siamese-authorship-verification/artifacts/>  
Table view: <https://wandb.ai/authorship-verification-siamese-network/siamese-authorship-verification/table?nw=nwuserwezkc21>  
Fine-tuned BERT hub: <https://huggingface.co/ElyMK1/bert-shakespeare-english-mlm>

**Workflow:**  
1. Login once: wandb login $API\_KEY  
2. Run any pipeline stage; metrics & artefacts sync live.  
3. View Runs, Artifacts (trained Siamese .h5, fine‑tuned BERT checkpoints, DTW matrices, Signal plots, T-SNE Graph of Clusters, etc.), and Table for comparisons.  
4. Download: wandb artifacts get siamese-authorship-verification-branches-{impostor 1 name}\_{impostor 2 name}:latest.  
Or a specific tag. Though, generally the latest is the intended one for use.

# 6. Google Colab Notebook – Authorship\_Verification\_using\_Impostor\_Projections\_and\_Siamese\_Networks\_Model.ipynb

The notebook is a **bootstrapper**, not a research playground. It contains just five executable cells:

1. Reset & Python 3.10 setup – Cleans workspace, installs Python 3.10 via apt.  
2. CUDA 12.2 & cuDNN install – Adds NVIDIA repository, installs driver libs, verifies with nvcc. These installations take ~30 minutes, but are quite necessary to be able to have a compatible host system with the latest Tensorflow version (2.18.0). Maintenance guide elaborates on this matter.

3. Clone repo & pip install – git clone and pip install of project requirements.  
4. Launch main.py in headless mode – python main.py --pipeline nominal (or interactively).