**BERT-BILSTM model for hierarchical Arabic text classification**

**If you use software from this page, please cite us as follows:**

Benamar HAMZAOUI, Djelloul BOUCHIHA, Abdelghani BOUZIANE, and Noureddine DOUMI. BERT-BILSTM model for hierarchical Arabic text classification. Submitted (still under review).

**If you need help, contact** Djelloul BOUCHIHA**:**

[bouchiha@cuniv-naama.dz](mailto:bouchiha@cuniv-naama.dz); [bouchiha.dj@gmail.com](mailto:bouchiha.dj@gmail.com); [djelloul.bouchiha@univ-sba.dz](mailto:djelloul.bouchiha@univ-sba.dz);

**Before using any software from this page, please follow carefully the following notes:**

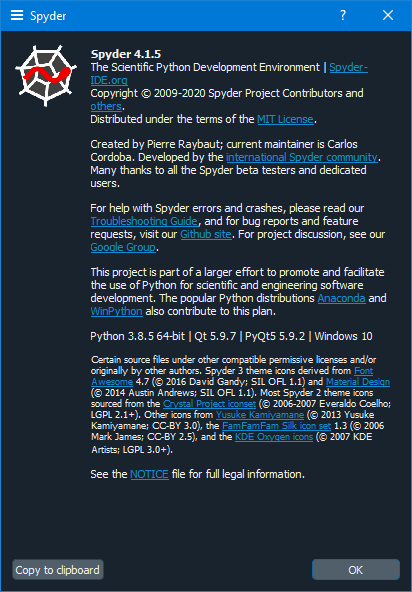
First you have to download the WiHArD dataset from <https://data.mendeley.com/datasets/kdkryh5rs2/2>. Download the whole dataset as one CSV file ([WiHArD.csv](https://data.mendeley.com/datasets/kdkryh5rs2/2#:~:text=CSV-,WiHArD.csv,-3%20MB))

You also need to download:

<arabic-stop-words.txt>

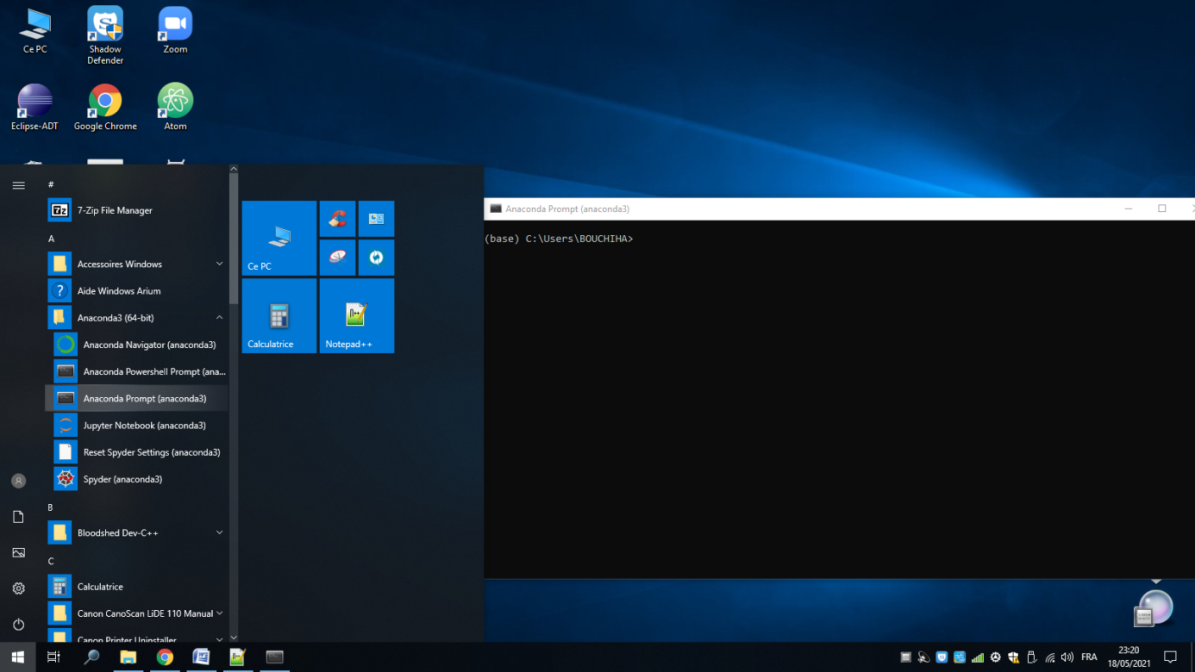
<WiHArD.hierarchy.xml>

**Note that all classifiers have been implemented using the Scientific Python Development Environment (Spyder IDE, version 4.1.5):**



**Before running any of the classifier, you have to install some additional Python packages:**

Since we are using Anaconda environment including Spyder (Python editor), then we add packages through Anaconda Prompt terminal:



For the first time, check the already installed packages with:

C:\...>pip list

For BERT embedding, you have to install:

For english language:

C:\...>pip install -U tensorflow

C:\...>pip install -U tensorflow\_hub

C:\...>pip install -U tensorflow\_text

For Arabic language:

C:\...>pip install -U tensorflow

In the case of Ali Safaya (<https://huggingface.co/asafaya/bert-base-arabic>)

C:\...>pip install -U transformers

Look at (<https://pytorch.org/get-started/locally/>) to install pytorch

C:\...>conda install pytorch torchvision torchaudio cudatoolkit=10.2 -c pytorch

**Next, you will find two classifiers:**

<ARBERT-BILSTM-Model.py>

<Doc2Vec-DecisionTreeModel.py>

**Next are some error and warning messages that you may meet when dealing with our classifier:**

When launching the Arabic BERT model, if you get the following error message:

RuntimeError: The size of tensor a (532) must match the size of tensor b (512) at non-singleton dimension 1

So, you must reduce the string length introduced to the BERT model. For example: Arabic\_Bert\_Model\_T(i[0:2000])

If you receive the following error message:

ValueError: The first argument to `Layer.call` must always be passed.

This means that the BERT model must be launched before building the Neural Network model.

If you get the following warning message:

UserWarning: The gensim.similarities.levenshtein submodule is disabled, because the optional Levenshtein package <https://pypi.org/project/python-Levenshtein/> is unavailable. Install Levenhstein (e.g. `pip install python-Levenshtein`) to suppress this warning. warnings.warn(msg)

Then, try to add python-Levenshtein package as follows:

C:\...>conda install -c conda-forge python-levenshtein

If you get the following warning message:

UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

This warning message disappears once the zero\_division parameter is set to 0 or 1. For example:

classification\_report(y\_true, y\_pred, target\_names=target\_names, zero\_division=0)

For LDA and QDA implementations, if you get the following warning message:

C:\...\anaconda3\lib\site-packages\sklearn\discriminant\_analysis.py:715: UserWarning: Variables are collinear

warnings.warn("Variables are collinear")

Open the file: C:\...\anaconda3\lib\site-packages\sklearn\discriminant\_analysis.py

Remove or set as comment the following statements:

rank = np.sum(S > self.tol)

if rank < n\_features:

warnings.warn("Variables are collinear")

When predicting a text’s class, if you get the following error message:

Cannot center sparse matrices: pass `with\_mean=False` instead. See docstring for motivation and alternatives.

This can be resolved by toarray conversion as follows:

clf.predict(x\_vec.toarray())

When executing RadiusNeighborsClassifier. If you get the following error message:

No neighbors found for test samples array([ …], dtype=int64), you can try using larger radius, giving a label for outliers, or considering removing them from your dataset.

This can be resolved by increasing the radius value as follows:

RadiusNeighborsClassifier(radius = 40)

When executing CategoricalNB (Naive Bayes), if you get the following error message:

index .. is out of bounds for axis .. with size ..

This can be resolved by increasing the value of the min\_categories parameter of CategoricalNB, till the error disappears. For example: CategoricalNB(min\_categories = 50)

Now, if you receive the following error message:

\_\_init\_\_() got an unexpected keyword argument 'min\_categories'

Then, you have to update scikit-learn package (https://scikit-learn.org/stable/install.html)

When executing CategoricalNB, MultinomialNB or ComplementNB (Naive Bayes), if you get the following error message:

Negative values in data passed to CategoricalNB (input X)

That means CategoricalNB does not admit negative vales, so you have to transform features by scaling each feature to a given range (0, 1 by default), by using the following code:

from sklearn import preprocessing

scaler1 = preprocessing.MinMaxScaler()

scaler1.fit(Xtrain)

Xtrain = scaler1.transform(Xtrain)

Scaler2 = preprocessing.MinMaxScaler()

Scaler2.fit(Xtest)

Xtest = scaler2.transform(Xtest)

When executing GaussianProcessClassifier, if you get the following error message:

C:\...\anaconda3\lib\site-packages\sklearn\gaussian\_process\\_gpc.py:472: ConvergenceWarning: lbfgs failed to converge (status=2):

ABNORMAL\_TERMINATION\_IN\_LNSRCH.

Increase the number of iterations (max\_iter) or scale the data as shown in:

https://scikit-learn.org/stable/modules/preprocessing.html

\_check\_optimize\_result("lbfgs", opt\_res)

This can be resolved by scaling data as follows:

from sklearn import preprocessing

scaler1 = preprocessing.StandardScaler().fit(Xtrain)

Xtrain = scaler1.transform(Xtrain)

scaler2 = preprocessing.StandardScaler().fit(Xtest)

Xtest = scaler2.transform(Xtest)

When executing MLPClassifier, if you get the following warning message:

C:\...\anaconda3\lib\site-packages\sklearn\neural\_network\\_multilayer\_perceptron.py:614: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (200) reached and the optimization hasn't converged yet.

warnings.warn(

This can be resolved by increasing the max\_iter parameter value of MLPClassifier. For example:

MLPClassifier(max\_iter=700)

When installing some packages, if you get the following error message:

ERROR: Could not install packages due to an EnvironmentError: [WinError 5] Accès refusé: 'C:\\...\\anaconda3\\Lib\\site-packages\\...'

Consider using the `--user` option or check the permissions.

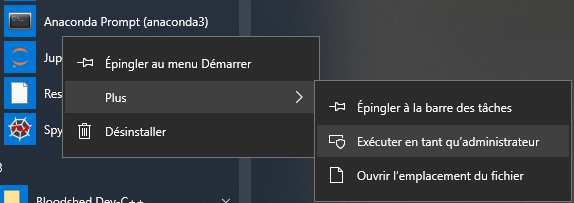
You can install the package for your user only, like this:

C:\...>pip install <package> --user

**Or**

You can install the package as Administrator, by following these steps:

1. Right click on the Command Prompt icon
2. Select the option Run This Program As An Administrator



1. Run the command pip install -U <package>

When launching the Arabic BERT model, if you get the following error message:

RuntimeError: The size of tensor a (532) must match the size of tensor b (512) at non-singleton dimension 1

So, you must reduce the string length introduced to the BERT model. For example: Arabic\_Bert\_Model\_T(i[0:2000])

If you receive the following error message:

ValueError: The first argument to `Layer.call` must always be passed.

This means that the BERT model must be launched before building the Neural Network model.