Spelling Corrector

Objective:

The main objective of these project is to build an end to end nlp system to correct wrongly inputed words using the state of the art models

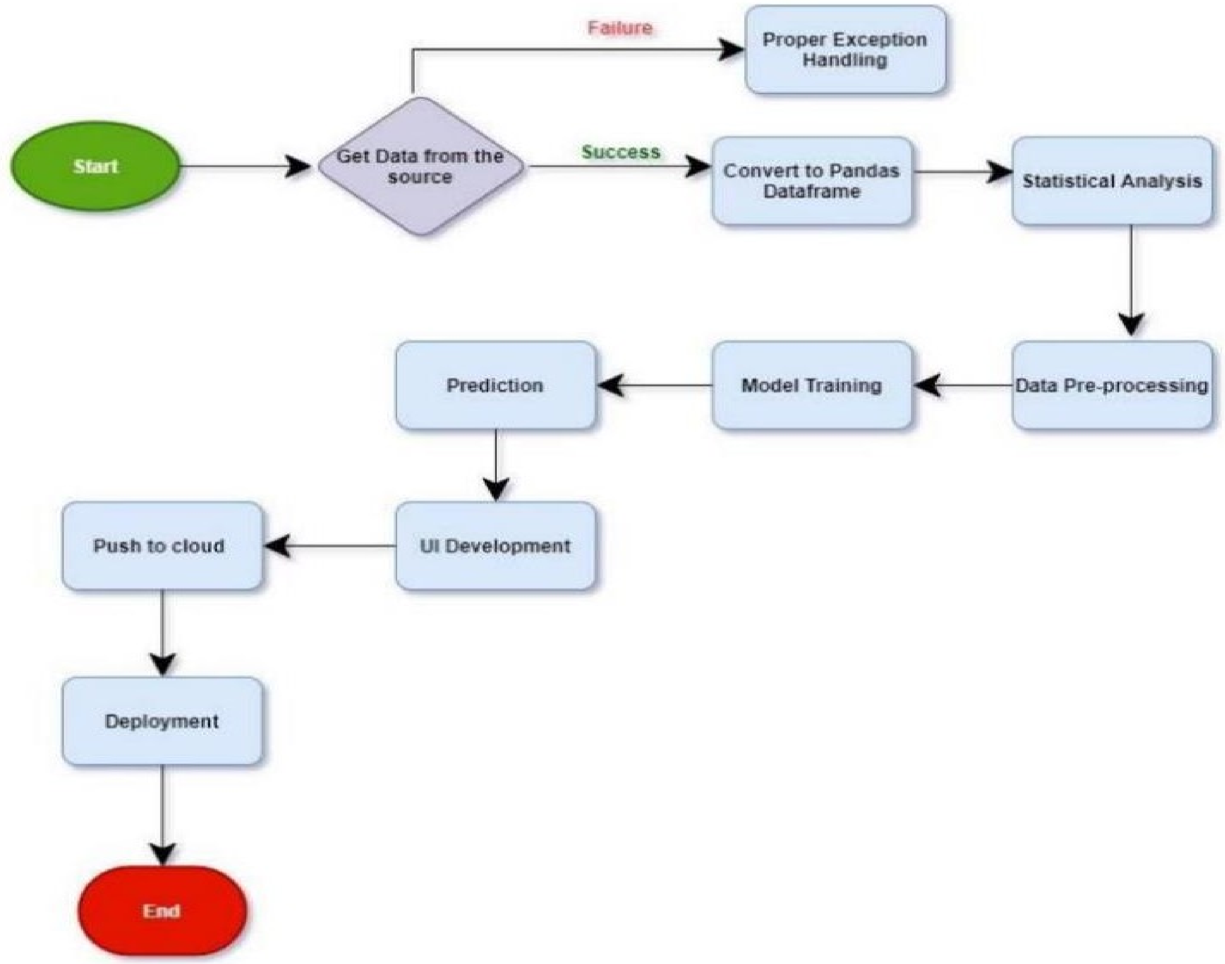
Benefits:

1. Use as spell corrector for end users
2. Can be use in many applications
3. Can be used in an android application app

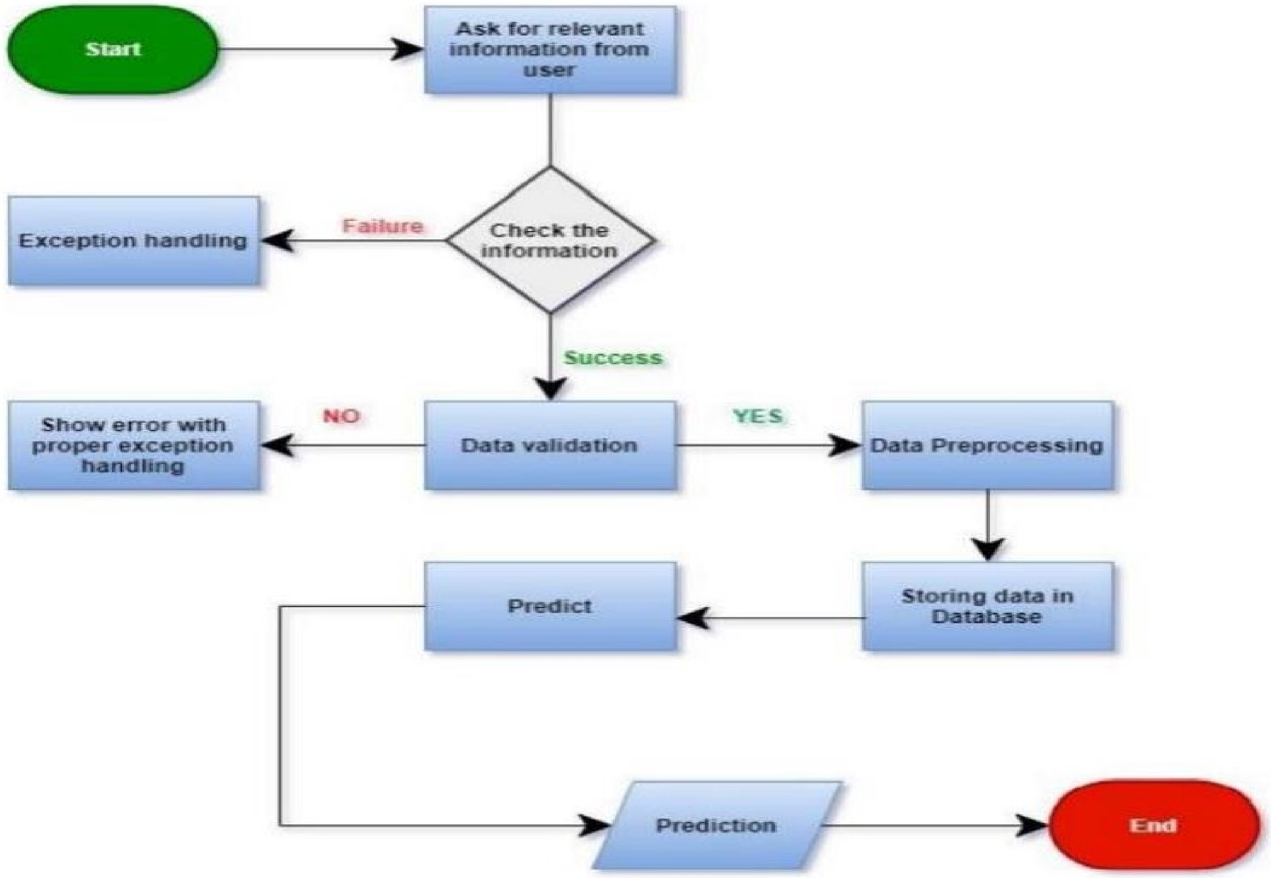
Data:

https://github.com/neuspell/neuspell#Datasets
10000 plus text records

Architecture
Machine Learning Model



6. User I/O Workflow



Model Training:

* Data exported from text file

Data Preprocessing:

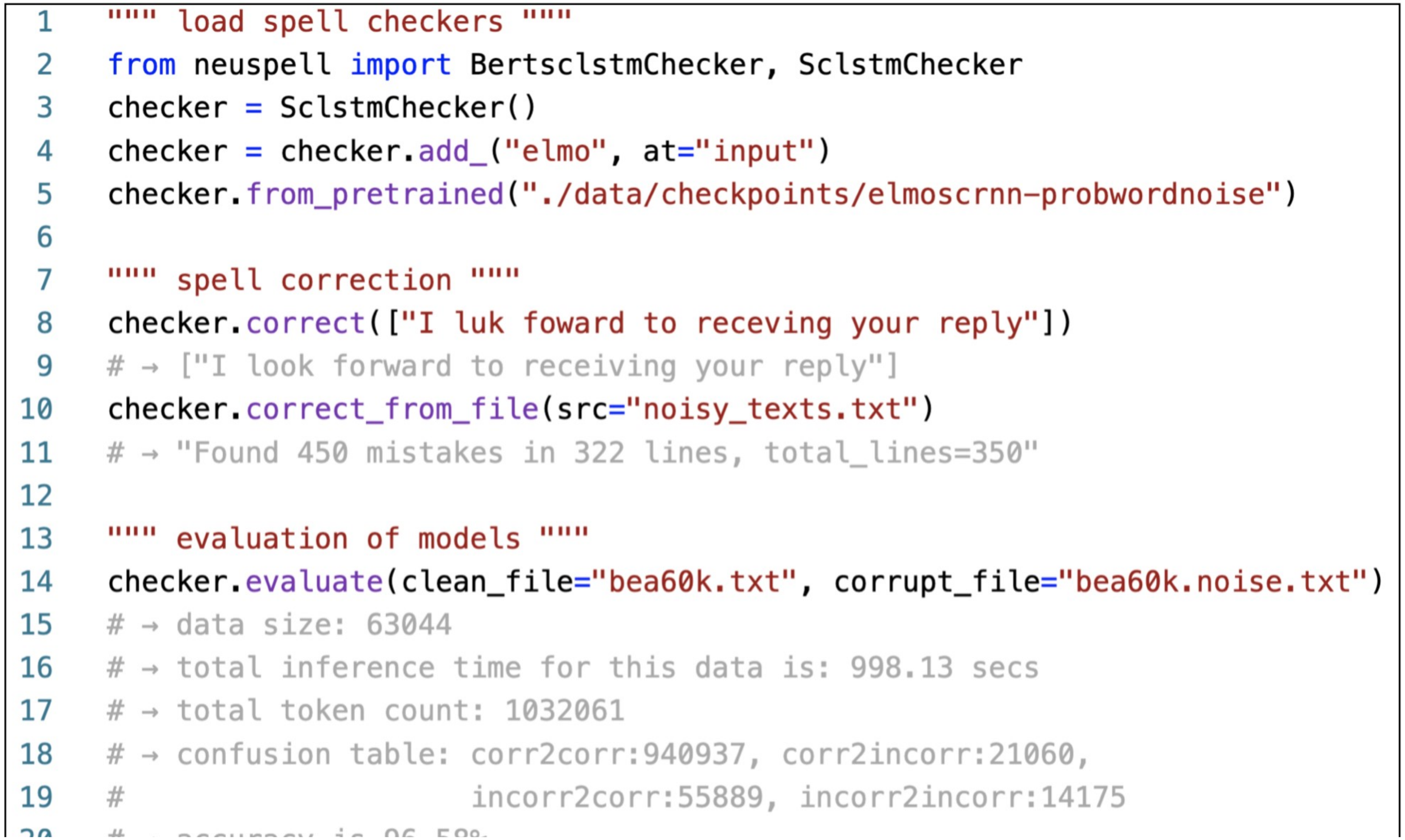
*Performing text preprocessing which includes removing punctuations marks, converting into tokens .

*Checking if any other things accept words are there in the dataset using all unique tokens code.

Train and Test Split:

* Train data 75% of whole data.
* Test data   25% of whole data.
* Data is randomly spitted.
* There is only train and test data.

Model Selection:



```
""" load spell checkers """
from neuspell import BertsclstmChecker, SclstmChecker
checker = SclstmChecker()
checker = checker.add_("elmo", at="input")
checker.from_pretrained("./data/checkpoints/elmoscrnn-probwordnoise")

""" spell correction """
checker.correct(["I luk foward to receving your reply"])
# → ["I look forward to receiving your reply"]
checker.correct_from_file(src="noisy_texts.txt")
# → "Found 450 mistakes in 322 lines, total_lines=350"

""" evaluation of models """
checker.evaluate(clean_file="bea60k.txt", corrupt_file="bea60k.noise.txt")
# → data size: 63044
# → total inference time for this data is: 998.13 secs
# → total token count: 1032061
# → confusion table: corr2corr:940937, corr2incorr:21060,
#                    incorr2corr:55889, incorr2incorr:14175
# → accuracy is 96.58%
```

Q&A

1. What is the source of the Data.

A: The data for training is provided by client (ineuron) in form of text file and source of file.

2. What are types of data.

A: The combined of text words and certain punctuation marks

3. What the complete flow you followed in this project.

A: Refer slide 4th and 5th slide for better understanding.

4.How are logs managed.
A: We are using different logs as per the steps that we follow in validation and modelling like file validation, Data insertion, Model Training, Prediction Log.

5. What are techniques were you using Data-Preprocessing.
A: * Checking if the data contains punctuations marks or any other things other than words and removing it
Converting text into tokens to be able to fed to the transformer model of hugging face
Fine tuning already train model on our data

6. How training was done and model was used.

A: * Before diving the data in training and test set we performed pre-processing order to get better.

* As per the model the training and test are valid.

7. What are different stages of deployment.

A: * When model is ready we deployed it local environment, where UAT is perfomed.

* Then project uploaded in GitHub account.

* Deployed in AWS Cloud Platform using CiCd github actions