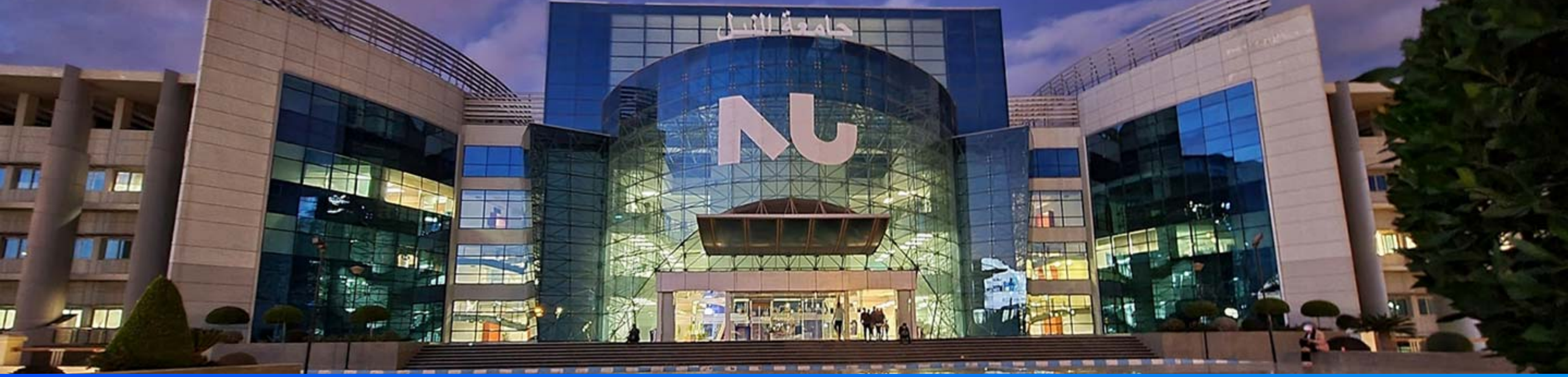


NILE UNIVERSITY

AI HACKATHON ORIENTATION

16-18 / 02 / 2023

FEDERATED TEAM

ADVANCED LEVEL

NUAIH Named Entity Recognition

Natural Language Processing of hadith (Arabic Language)

The Federated Team

Evaluation criteria

Time

RULES

Sumission format

Exchange ideas

Data usage

EVALUATION CRITERIA & INFERENCE TIME

01 How to deal with the problem / Problem Solution

02 State of the art models/methods

03 Why your solution outperforms the others?

04 If you try other solutions, explain why these solutions didn’t fit with the problem

05 Record your inference time

06 Explain the results

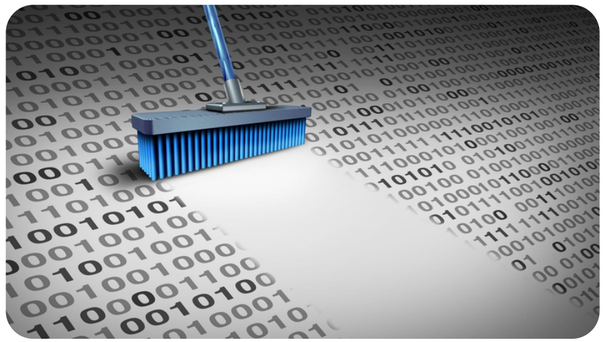
07 Error analysis

08 How long does it take your model/method to predict the outcome?

09 A trade-off between the inference time and the score

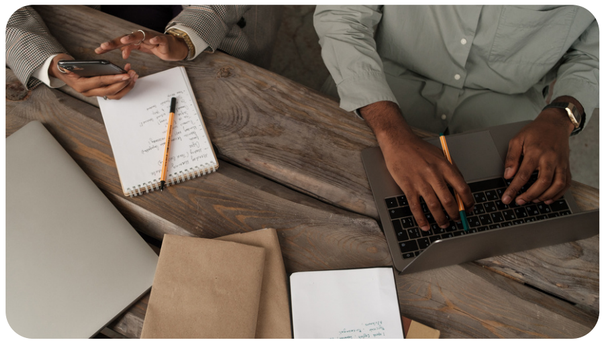
10 Like optimization problems

THE PROBLEM



DATA CLEANING

In most nlp tasks/projects the problem lays in the handling, and cleaning of the data. Some data like in class 4 mislabeled, it should be class 5



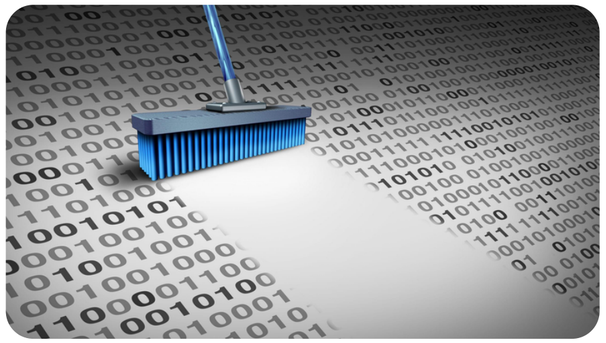
PICKING THE RIGHT ARCHITECTURE

There is multiple deep learning models that we used before, and performed amazingly specially with Arabic text, but a lot of these models require a lot of resources to train.

RESOURCES MANAGEMENT

We divided the team so each member has a task. With this approach we can correct mistakes faster, and iterate with different models faster too.

PROBLEM SOLUTION



DATA HANDLING & CLEANING

Because the state of the dataset, even if we had all the resources, and the best model, we would be training on data that is filled with noise, and data entries mistakes. No model will reach its optimal potential with this kind of data.

- **Example**: there is some hadith text that has the subtext as narrator, but actually it should have the subtext as content.

- **Example**: In the subtext the name of the prophet Mohamed is entered with different variation, and sometimes the different name of Mohamed is used. And there is unique challenges in every subtext group that we handled.

COMPARING OUR SOLUTION

OUR ADVANTAGE

We can say that our approach is better than others because of two reason:

- **First**: our model in our first try we used the dataset without any type of cleaning, or handling, and it was able to give us 58 f1 score on Kaggle, and again that without any kind of data cleaning. Because we got a great base model with decent training time, we are able to experiment with the data cleaning more, and get to know what affects the results more. while most teams will take time to build, and optimize their model.
- We are confident enough in it, that instead of spending time on waiting for the model to train. we spent on data cleaning.

OTHER SOLUTIONS

- Since we're confident in our model , we kept trying improving the results by understanding the data more. We kept trying to understand what affects the results. Most of our failures actually was an improvement. With every failure we understood the data more.
- **Example**: stemming every word in the column should gets us better results, but with stemming it actually works only on verbs, abut when comes to vocab, it removes very important information. like removing ال from the start of every word, it also removes ال from the end of the word, which changes its meaning. الله also starts with ال, but it also cant remove since it will become meanness.
- So, most of our experimentation was in the changing how we approach cleaning the data, what is noise? and what is useful information?

OTHER SOLUTIONS CON.

- In the beginning we used a LSTM model, because obviously it is the most used neural network to solve NLP problems.
- It gave us decent results, 53 f1 score. Our lstm architecture was light, around 500 thousand parameters. Which is the right size for colab, or our local machine to train on.
- It doesn’t take long to get good results from it, but it had less parameter than the needed to get the optimal score that we wanted.
- We tried the model with multiple variation of the dataset, cleaned, uncleaned, oversampling, and without oversampling.
- Next we moved to Arabic bert model minilm, and Arabic xlmroberta.
- In minilm it’s a pretrained model that is used for NLA, and text classification. One the most important things in this model that was trained on 16 languages, Arabic is one of these 16.

DATA PREPROCESSING

Raw data try

- Check nan values one record that has nan value in subtext, because of the size of the dataset we removed it.
- Duplicate values: 237 we decided to just remove the duplicate and keep the original
- Fixing the subtext data entry: there is some hadith text that has the subtext narrator, but actually it should have the

Compare Results

Locally / Leaderboard	MMini	AraBert	XLM Roerta AR
Results	75 % / 59 %	92 %/ 79 %	92 %/ 76.53 %

BACK TO AGENDA PAGE

Inference Time

Cell Time / Pred Time	MMini	AraBert	XLM Roerta AR	LSTM
Time for the Testing data	null	6.43 Sec / 5.16 Sec Best Model	null	null

```
CPU times: user 5.15 s, sys: 17.4 ms, total: 5.16 s
Wall time: 6.43 s
```

Note : This is the Colab configuration Free tier Colab Resources with GPU

BACK TO AGENDA PAGE

Server Resources

CPU	GPU	RAM
Intel Xeon 128 CPUs	2X Nvidia A40	256 GB

BACK TO AGENDA PAGE