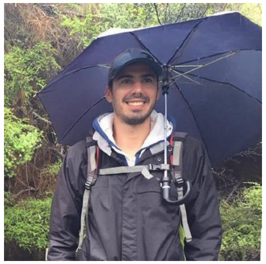
TETIS Text Mining

ITU GeoAI LMR

Rémy Decoupes
Nejat Arinik
Roberto Interdonato

Introduction



Rémy Decoupes

INRAE

Nejat Arinik

INRAE

Roberto Interdonato

cirad

Motivation

In the context of our research, we apply text mining:

- Epidemiology Event Based Surveillance
- Food Security

On tasks such as:

- Graph based analysis
- Text classification

… And we already use QCRI data (crisisNLP)

Methodology

Hypothesis

1. Fine-tune Bert like models available on HuggingFace

2. Use a Gazetteer (OSM) to improve results

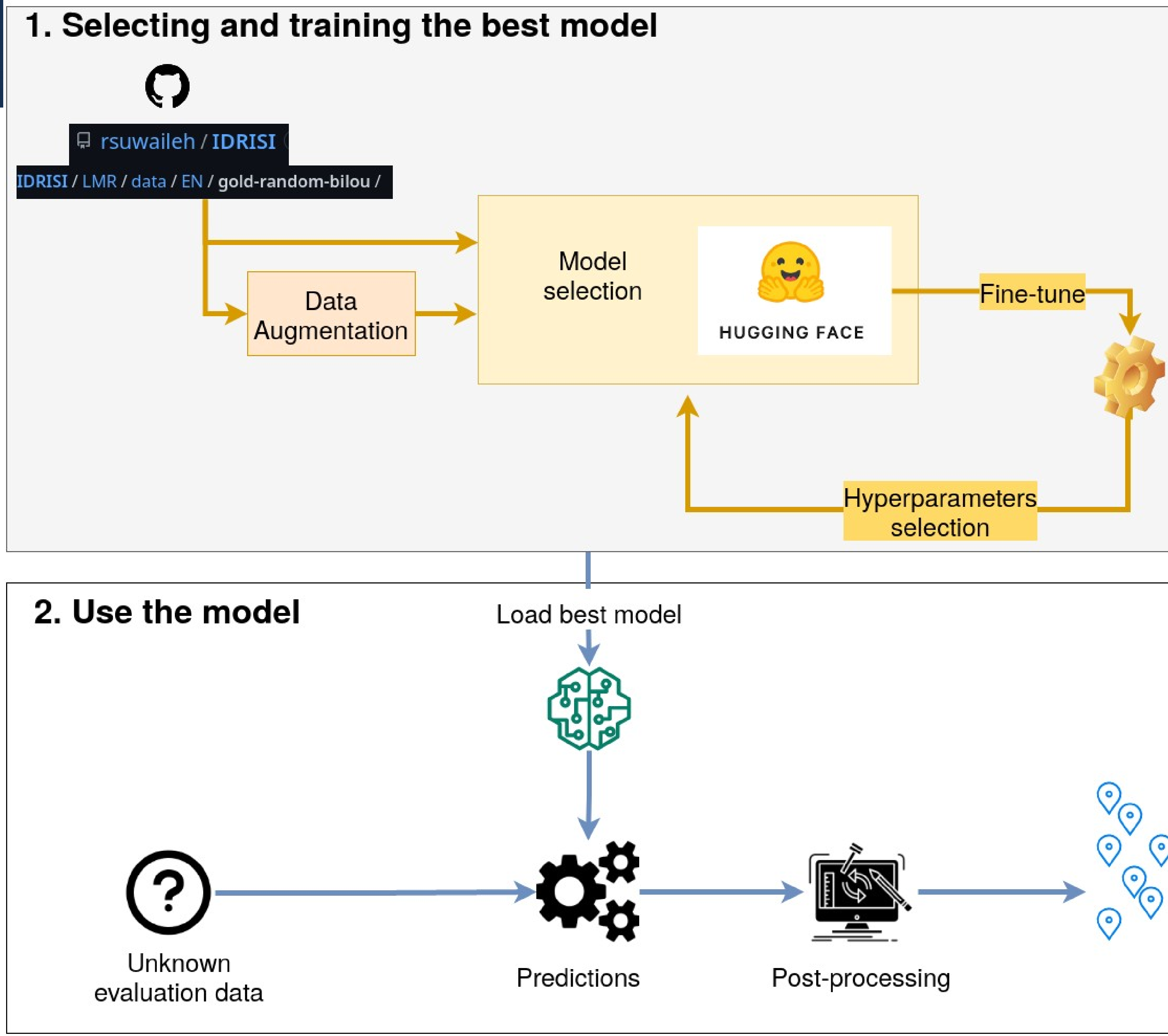
3. Apply Data augmentation to enlarge the training dataset

~~Hypothesis~~ Reality

1. Fine-tune Bert like models accessible on HuggingFace
→ The best trained models are the best models
(see IBM research: Choshen et al - 2022: ”Where to start? ...”)

2. Use a Gazetteer (OSM) to improve results
→ The models were good enough for City / State / Country.
For the others (Island, NPOI, …), the disambiguation could be too hard

3. Apply Data augmentation to enlarge the training dataset
→ It introduces too much noise


1. Selecting and training the best model
rsuwaileh / IDRISI
IDRISI / LMR / data / EN / gold-random-bilou /
Data Augmentation
Model selection
HUGGING FACE
Fine-tune
Hyperparameters selection
2. Use the model
Load best model
Unknown evaluation data
Predictions
Post-processing

Implementation details

python™

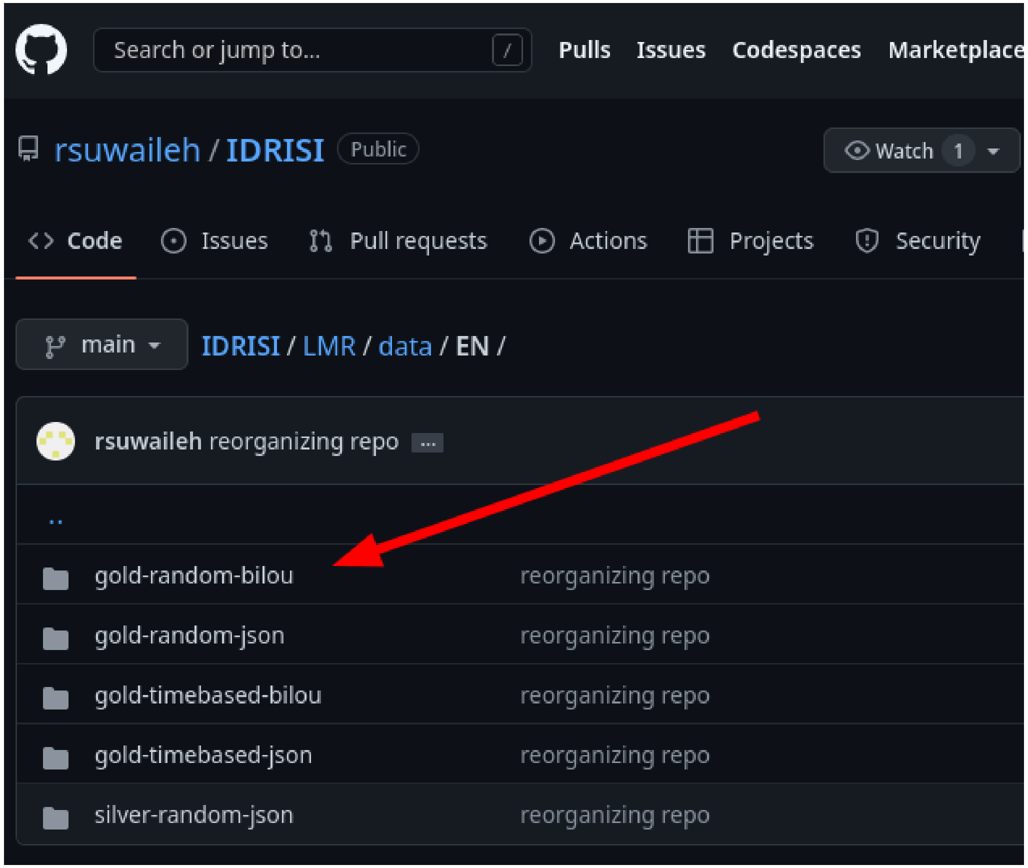
Hugging Face

roberta-base like 89

Fine-tuning with no pre-processing

Evaluation

Training data



Training data

```
label_encoding_dict =
{"B-CONT" : 1, "B-CTRY" : 2, "B-STAT" : 3, "B-CNTY" : 4, "B-CITY" : 5, "B-DIST" : 6,
"B-NBHD" : 7, "B-ISL" : 8, "B-NPOI" : 9, "B-HPOI" : 10, "B-ST" : 11, "B-OTHR" : 12, "I-
CONT" : 13, "I-CTRY" : 14, "I-STAT" : 15, "I-CNTY" : 16, "I-CITY" : 17, "I-DIST" : 18,
"I-NBHD" : 19, "I-ISL" : 20, "I-NPOI" : 21, "I-HPOI" : 22, "I-ST" : 23, "I-OTHR" : 24,
"L-CONT" : 25, "L-CTRY" : 26, "L-STAT" : 27, "L-CNTY" : 28, "L-CITY" : 29, "L-DIST" :
30, "L-NBHD" : 31, "L-ISL" : 32, "L-NPOI" : 33, "L-HPOI" : 34, "L-ST" : 35, "L-OTHR" :
36, "U-CONT" : 37, "U-CTRY" : 38, "U-STAT" : 39, "U-CNTY" : 40, "U-CITY" : 41, "U-DIST"
: 42, "U-NBHD" : 43, "U-ISL" : 44, "U-NPOI" : 45, "U-HPOI" : 46, "U-ST" : 47, "U-OTHR"
: 48,
"O":0}
```

```
B: Begining of a NER
I: Inside the current NER
L: Last: the final token of a multi-token
U: Unit: a single-token entity
O: Out: a non-entity token
```

Tuning the model

```
args = TrainingArguments(
    f"test-{self.task}",
    evaluation_strategy = "epoch",
    learning_rate=self.learning_rate,
    per_device_train_batch_size=self.batch_size,
    per_device_eval_batch_size=self.batch_size,
    num_train_epochs=self.num_train_epochs,
    load_best_model_at_end=True,
    seed=42,
    save_strategy = "epoch",
    # weight_decay=1e-5,

)
```

Compare the results

1. On the test dataset

On type-less:
- TETIS: 0.87
- CRF: 0.87

On type-full:
- TETIS: 0.83
- CRF: 0.81
- Bert: 0.79

2. On the evaluation dataset

	California Wildfires 2018	Canada Wildfires 2016	Cyclone Idai 2019	Ecuador Earthquake 2016	Greece Wildfires 2018	Hurricane Dorian 2019	Hurricane Florence 2018	Hurricane Harvey 2017	Hurricane Irma 2017	Hurricane Maria 2017	Hurricane Matthew 2016	Italy Earthquake Aug 2016	Kaikoura Earthquake 2016	Kerala Floods 2018	Maryland Floods 2018	Midwestern US Floods 2019	Pakistan Earthquake 2019	Puebla Mexico Earthquake 2017	Srilanka Floods 2017	**Average**
TETIS	0.92	0.75	0.89	0.94	0.93	0.88	0.77	0.93	0.87	0.91	0.95	0.9	0.92	0.89	0.91	0.93	0.88	0.93	0.93	0.90
baseline CRF	0.98	0.94	0.89	0.92	0.93	0.96	0.74	0.97	0.97	0.97	0.97	0.82	0.98	0.96	0.94	0.92	0.94	0.92	0.94	0.93
baseline BERT	0.98	0.93	0.92	0.95	0.92	0.97	0.78	0.98	0.96	0.97	0.97	0.87	0.98	0.96	0.93	0.94	0.95	0.93	0.93	0.94

Compare the results

1. On the test dataset

On type-less:
- TETIS: 0.87
- CRF: 0.87

On type-full:
- TETIS: 0.83
- CRF: 0.81
- Bert: 0.79

Question to the organizers:
Do the baselines were trained on each event separately ?

2. On the evaluation dataset

	California Wildfires 2018	Canada Wildfires 2016	Cyclone Idai 2019	Ecuador Earthquake 2016	Greece Wildfires 2018	Hurricane Dorian 2019	Hurricane Florence 2018	Hurricane Harvey 2017	Hurricane Irma 2017	Hurricane Maria 2017	Hurricane Matthew 2016	Italy Earthquake Aug 2016	Kaikoura Earthquake 2016	Kerala Floods 2018	Maryland Floods 2018	Midwestern US Floods 2019	Pakistan Earthquake 2019	Puebla Mexico Earthquake 2017	Srilanka Floods 2017	**Average**
TETIS	0.92	0.75	0.89	0.94	0.93	0.88	0.77	0.93	0.87	0.91	0.95	0.9	0.92	0.89	0.91	0.93	0.88	0.93	0.93	0.90
baseline CRF	0.98	0.94	0.89	0.92	0.93	0.96	0.74	0.97	0.97	0.97	0.97	0.82	0.98	0.96	0.94	0.92	0.94	0.92	0.94	0.93
baseline BERT	0.98	0.93	0.92	0.95	0.92	0.97	0.78	0.98	0.96	0.97	0.97	0.87	0.98	0.96	0.93	0.94	0.95	0.93	0.93	0.94

Issues to address

1. Type of location

City F1	Cnty F1	Cont F1	Country F1	District F1	Hpoi F1	Island F1	Nbhd F1	Npoi F1	Other F1	State	**Global F1**
0.76	0.76	0.7	0.92	0.24	0.18	0.77	0	0.54	0	0.89	0.82

2. Understand why some events have such bad results

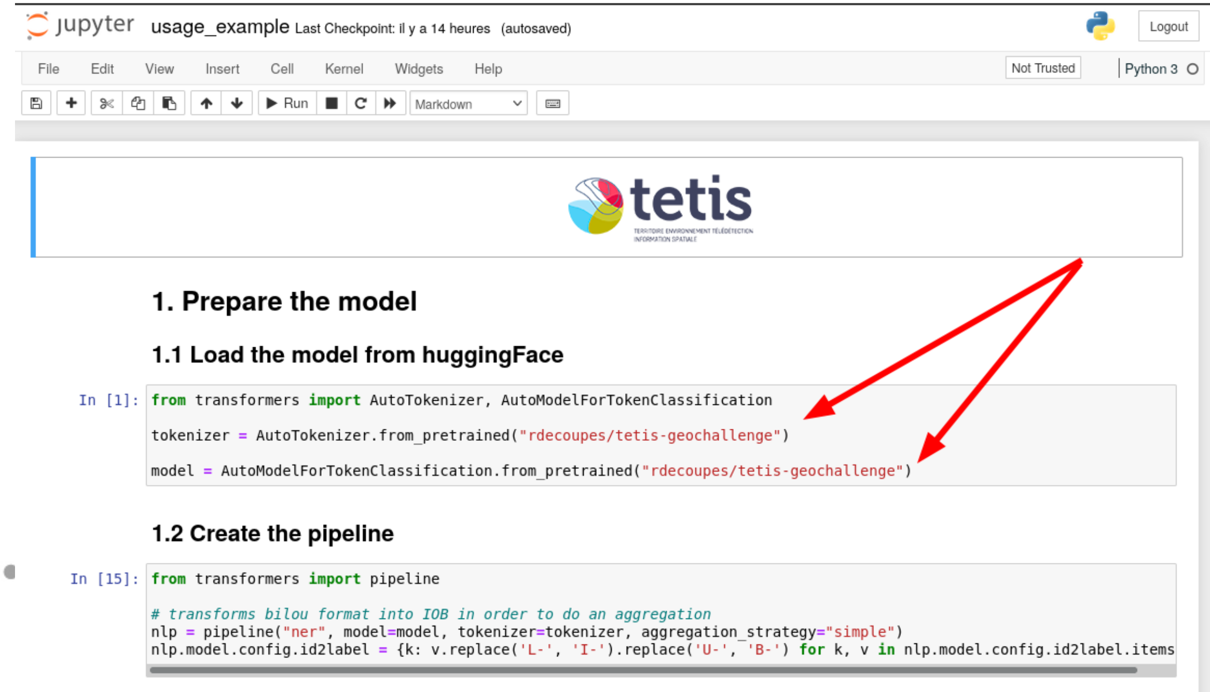
- Canada Wildfires (75%)
- Hurricane Florence (77%)

Takeaway messages

Good to know

1. Use a tool to track experiments (such as mlflow)
2. As always analyse manually the data
3. Our model could be reused ! (Check the repo)

Hugging Face



TERRITOIRE ENVIRONNEMENT TÉLÉDÉTECTION INFORMATION SPATIALE

Thank you for your attenion

Rémy Decoupes
Nejat Arinik
Roberto Interdonato

ITU GeoAI LMR