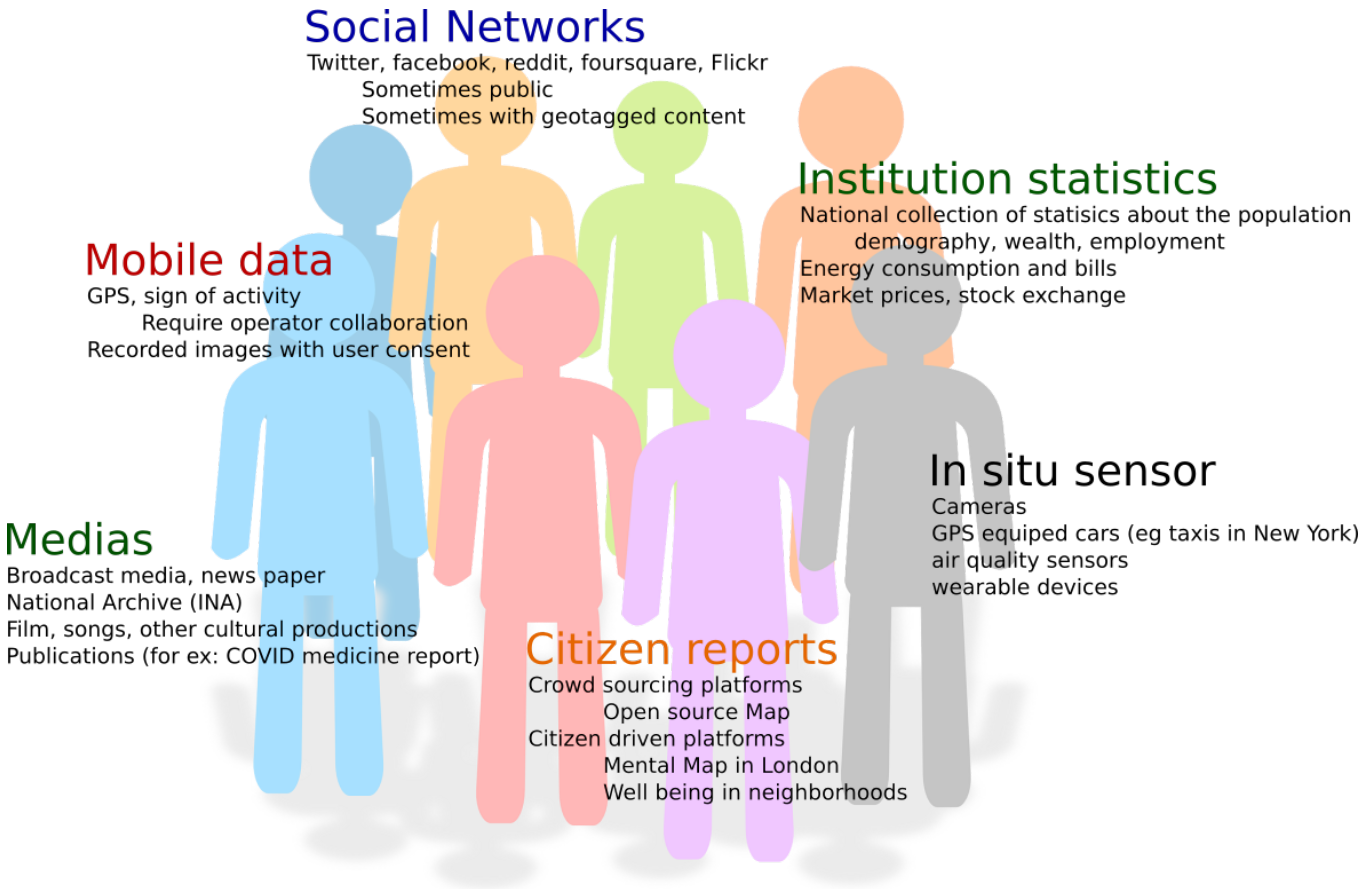
How to build NLP models for social sciences ?

Paul Gay

GreenAI U.P.P.A.

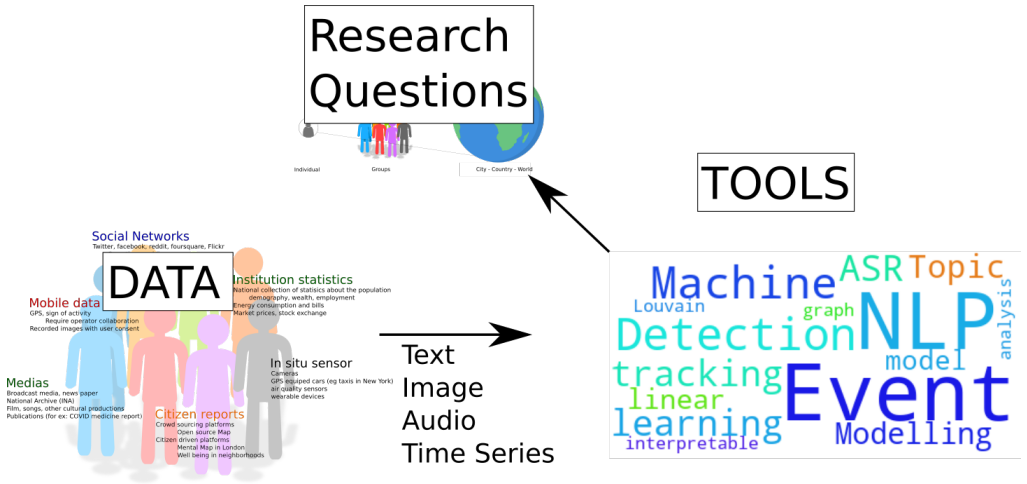
30-01-2022

We want to analyse data



Motivation

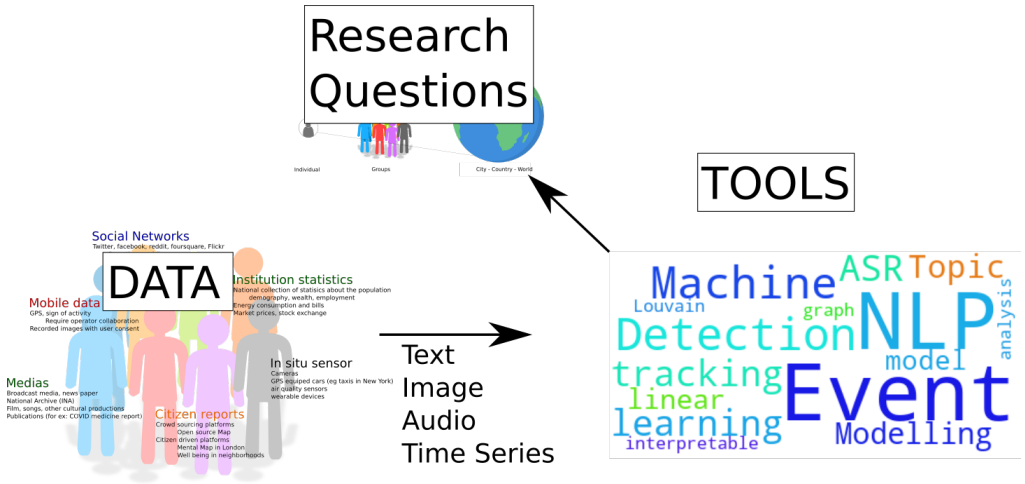
We want to analyse data to support social science



Goal: Estimate the weights of identified currents and their evolution

Motivation

We want to analyse data to support social science



Goal: Estimate the weights of identified currents and their evolution

- Classification of news, social networks data,...
- Zero shot and few shot learning

Objectives

On going work to digitalize social science researchers

- First feedbacks and workshop on hydrogen
- *Social Glue* Toward a benchmark suite to evaluate NLP models for social science

Unrolling of the methodology

- Starting to work with people which has some knowledge of programming, clustering and classification
- Iramuteq, but quick limitations

Focus on social network, because it is easy

- basic text manipulation
- Community detection
- Pretrained classifier : stance detection

Hardware is a limitation !

First collaborations with Tree

How hydrogen is discussed on Twitter ?

- Data collection
- Community detection
- Lexical analysis
- Stance detection

Hydrogen in Twitter data

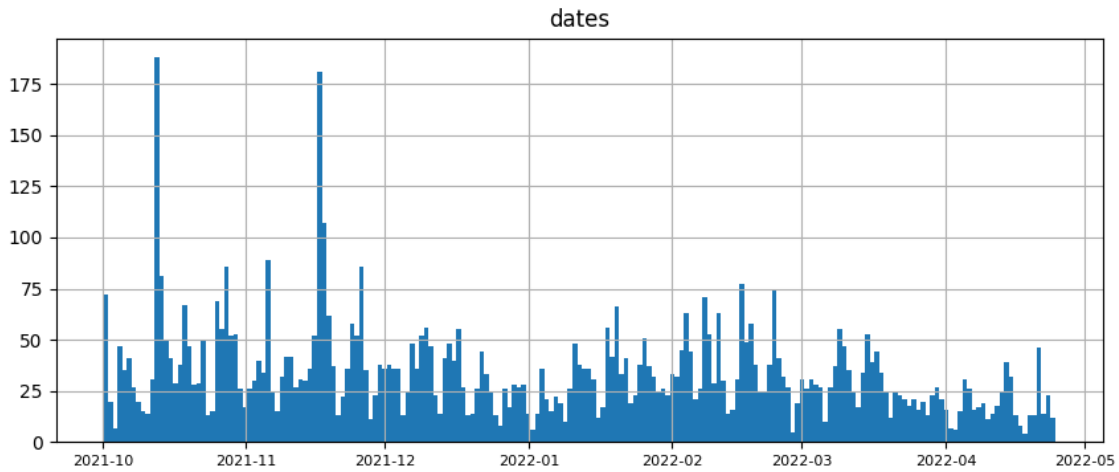


Figure 1: Occurences of word *hydrogène* in tweets

- October to April 2022 : 17M tweets, 6939 occurences
- September to December 2022 : 7M tweets, 3172 occurences
- Not a dedicated dataset but some events are visible
- Query extension difficulty due to multiple meanings

Stance detection with transformers

Negative

```
@RaphaelGirault "mobilité décarbonée
avec les autobus hydrogène de #Safra" ?
Un gâchis irresponsable.
```

Positive

```
Belle réussite de ce groupe français
Une vidéo à voir avec témoignage de
son président @GaussinChristo1 -
```

Neutral

```
Camions à hydrogène ou camions électriques:
qui aura le dernier mot?
```

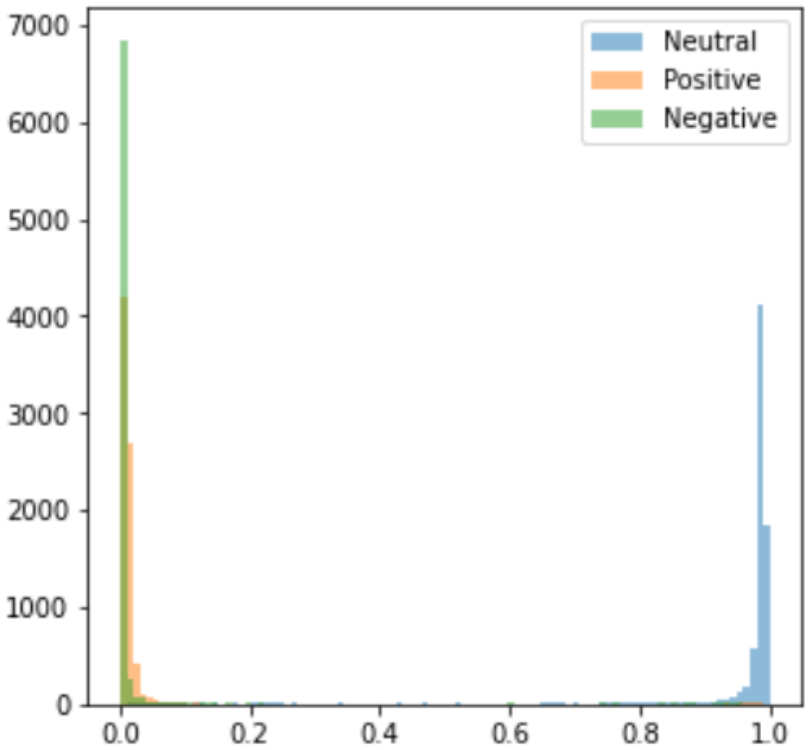


Figure 2: Histogram of stance scores

- Check with *relative* comparisons with different communities

Perspectives

- Relative comparisons with different communities
- Next steps with the Sobre Project
- Toward dedicated NLP models ...

Perspectives

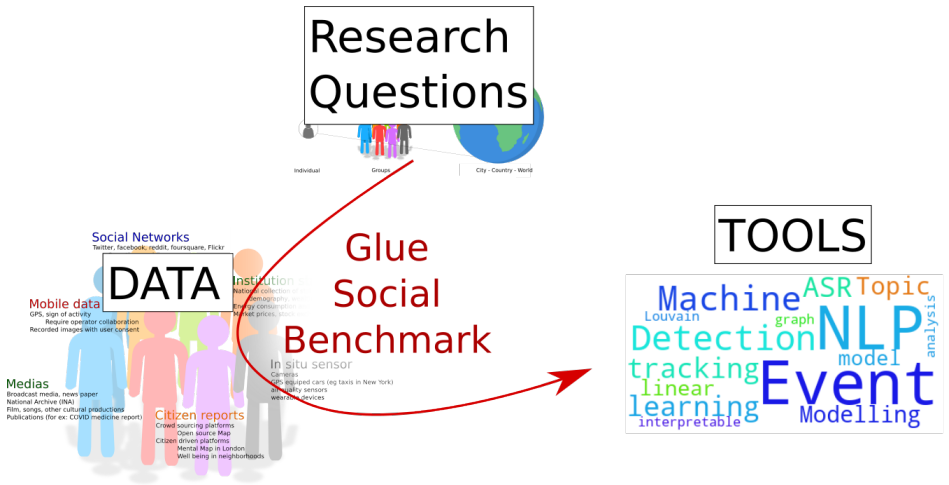
- Relative comparisons with different communities
- Next steps with the Sobre Project
- Toward dedicated NLP models ...

And to get these NLP models ?

Building a benchmark

Which models to use ?

- **Use case 1:** I have a new pretrained model I obtained somewhere, how good it is to understand data ?
- **Use case 2:** I want to build a new model, how can I train it and control this training



Desiderata

- Few shot or zero shot learning tasks
  - Proxy to control validation accuracy when training large models
  - Use cases don't tolerate large training data
- In French
- On topics linked with environmental transition
- Ideally, a corpus to pretrain the model ?

State of the art

Glue Benchmark dataset

- Collections of pre-existing datasets
- 9 sentence and sentence pair (of mostly) classification tasks

SentEval library

- Evaluate sentence encoding
- 17 downstream tasks and 10 probing tasks (sentence length, Verb tense prediction,...)

EleutherAI Language Model Evaluation Harness

- A framework for few-shot language model evaluation
- Currently 200 zero shot tasks

Existing datasets so far

Social glue benchmark

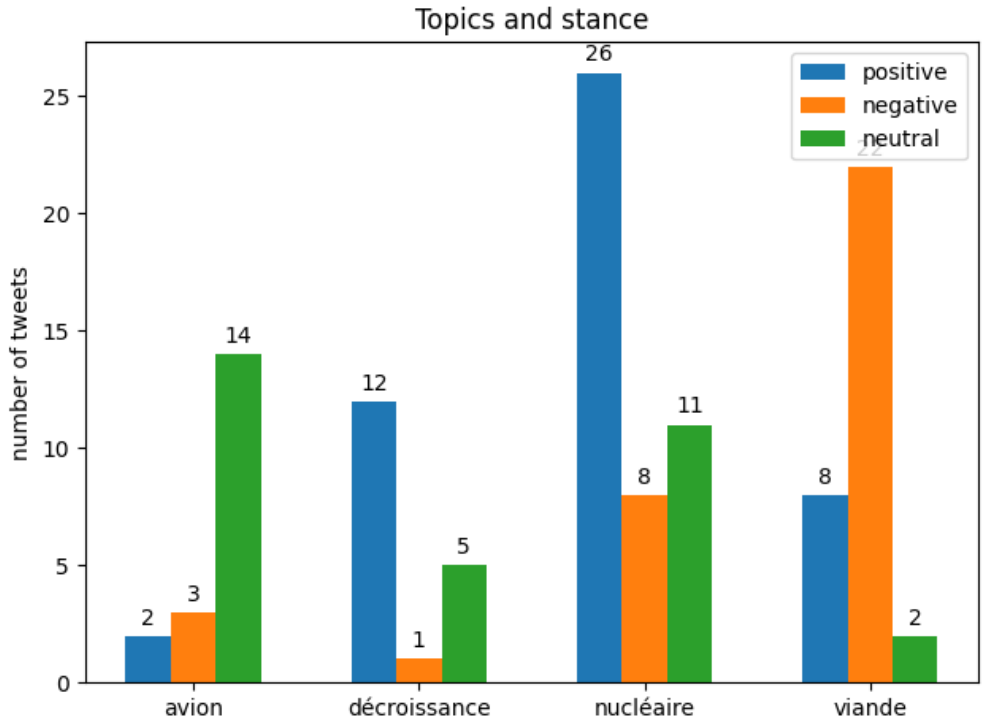
- *La dernière bibliothèque:* Retrieval
- *Overton window:* Topic and stance detection
- CrisisMD Tweet classification
- New to come with the *sobre* project.

- $queries = \{(q_1, a_1), \ldots, (q_n, a_n)\}$ where a_i is the correct content for query q_i
- $n = 40$ queries available
- Possibility to evaluate according to different tags
- TopK metric

$$Top_K = \frac{\sum_i \mathbb{1}_{r_i \in preds_i^k}}{\#queries},$$

where $preds_i^k$ are the top-k contents retrieved by the system for query q_i.

Stance detection



- Is the content on topic ?
- Stance annotation : *Positive*, *Neutral*, *Negative*

Some insights on stance detection

- it helps to know your topic very well, because it is all about cultural references
- Or depends on who said this ß External information is required (Transformer culture ?)

```
@DidierMaisto @alancelin @HelenaMorna
@cparisidf @_vd_ @BernardCadeau
@MurielHermine @c_laverdet @Sachaboyer
@Dany_Mauro_ @Guy_Marty_ @pda_tv
Merci ! Nous sommes fans !
```

To be refine with social science researchers...

How good are available pre-trained models for retrieval ?

	Top1	Top2	Top5	Top10
Fasttext (Wiki)	0.04	0.125	0.21	0.29
Camembert large	0.04	0.04	0.04	0.04
Camembert large last 8	0.04	0.04	0.04	0.12
bloom-560m	0	0.04	0.04	0.04

Table 1: Retrieval results for *la dernière bibliothèque*

Something is wrong with our transformers...

How good are you after fine-tuning ?

	Top1	Top2	Top5	Top10
Fasttext (wiki)	0.04	0.125	0.21	0.29
Fasttext Thomas	0.12	0.21	0.37	0.50
Fasttext Twitter	0.08	0.08	0.16	0.33

Table 2: Retrieval results for *la dernière bibliothèque*

- Probably data selection is the key
  - Fasttext Thomas : Gray litterature + Twitter
  - Fasttext Twitter : Twitter + more twitter data

Transformers : Is it a matter of information selection :

	Top1	Top2	Top5	Top10
Camembert large	0.04	0.04	0.04	0.04
Camembert large last 8	0.04	0.04	0.04	0.12
Camembert Oracle EE	0.08	0.08	0.12	0.37

Table 3: Retrieval results for *la dernière bibliothèque*

Searching which layer is usefull :

- Oracle selecting the layer with the best results
- venue for early exit ?
- Compromise between word matching and topic matching ?

Perspectives

- Frame better tasks with the coming project
- Obtain a methodology to train a transformer, and perform these tasks efficiently

What about code ?

- Github repo
- Interprétation des résultats
  - Contrastive examples
  - Connection with our active learning project

Thanks!

References I