The Augmented Social Scientist

Using Sequential Transfer Learning to Annotated Millions of Texts with Human-Level Accuracy (Do, Ollion & Shen, *SMR*, 2022).

I. General Presentation

More at www.css.cnrs.fr (click “Tutorial”)

Étienne Ollion (CNRS/ École polytechnique)
Rubing Shen (Sciences Po / École polytechnique)

Introduction

- Confessions of a (non-)Believer
- Figuring it Out
  - An Experiment
  - On a Social Scientific Research Question
  - In a Limited Amount of Time?
- Human Augmentation at the tip of our Fingers
- Bonus: Do we even need this in the era of chatGPT?

Research Questions

RQ1: Can a Model Achieve Human-Level Quality for Textual Annotation?

(+ which human? On which task?)

RQ2: What Role does the Expertise of Annotators play in the Training of the Model?

Research questions

On a real research topic: the rise of “Strategic News Coverage”

- Political Games over Political Measures or Ideas
- Revelation of Backstage Manoeuvers
- Lengthy Depiction of the Strategies of Politicians

Introduction

I. Data & Methods
II. The Experiment
III. Results
IV. Discussion
\+ Transition

Schedule

Augmented: General Presentation

Lab 1: Using the *Augmented* Package

BREAK

Lab 2: Conducting A Real Life Project

Augmented: Tips & Tricks

Data & Methods

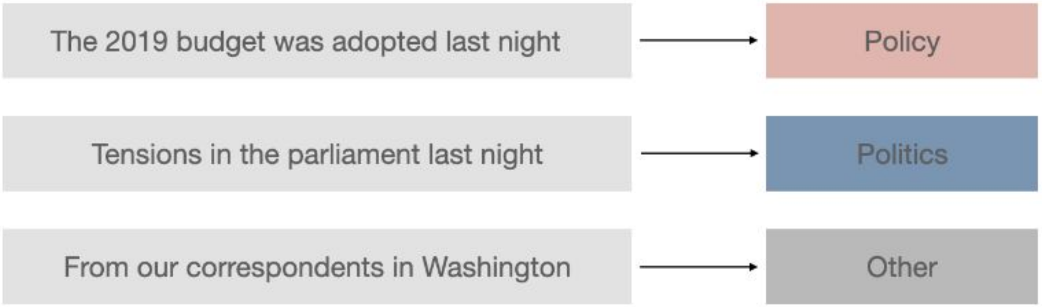
All articles in the Politics Section of the French daily *Le Monde*

1945-2015	61,511 articles	38,497,810 words

Data & Methods

Task 1: Policy vs. Politics

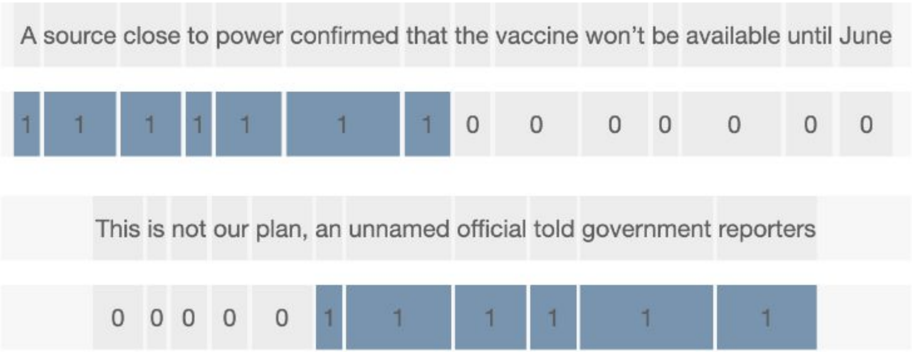
- Content of a Measure vs. Action of Politicians
- At the sentence level
- Complexity = high



Data & Methods

Task 2: Unattributed Quotes

- Prompts introducing unattributed quotes (“off the record”)
- Below the sentence level
- Complexity = medium high



Data & Methods

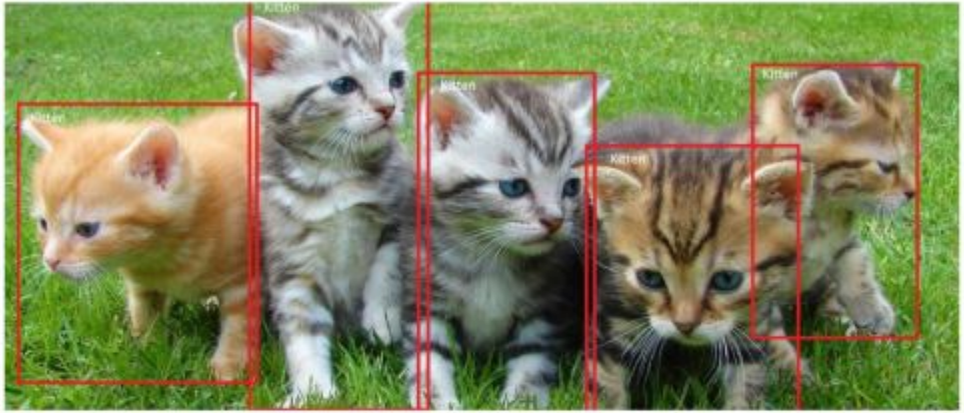
Keep in mind:

<u>you</u> get to design your indicators

Data & Methods

Method: Supervised Learning

Training a Model to Mimic Human Annotation

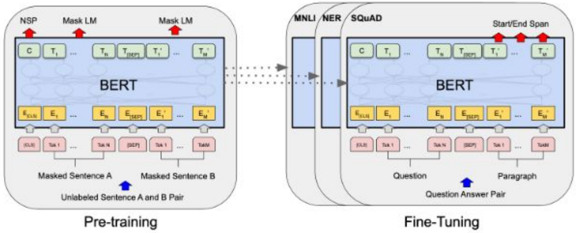


Data & Methods

Method: Supervised Transfer Learning

Using a Language Model to Drastically Reduce the Volume of Annotations

- Pretrained Language models : BERT (Devlin *et al.*, 2019) and CamemBERT for French (Martin *et al.*, 2020)
- Fine tuning: Adjusting the Model to Each of our two Tasks



The Experiment

3 types of Annotators

- **Social Scientist (SS)**
  - An expert in her field, often designs the indicators.
  - Time: limited

⇒ In this case: one of the authors of the paper

The Experiment

3 types of Annotators

- **Social Scientist (SS)**
- **Research Assistants (RA)**
  - Trained, qualified Students. Not experts in the field.
  - Interaction with the researcher

⇒ 3 Master's Level Students, carefully trained by us

The Experiment

3 types of Annotators

- **Social Scientist (SS)**
- **Research Assistants (RA)**
- **Microworkers (MW)**
  - Limited Training, no Connections to the Researcher

⇒ In our Case: 34 BA Students from a course (likely better than gig workers)

The Experiment

Against a carefully annotated test set ("gold standard"),

- RQ1: Compare the Performance of the Model Trained by the Social Scientist (ASS) against Human Annotations.

- RQ2: Compare the Performances of Models Trained by Annotators with Different Levels of Expertise (SS, RAs, MWs)

Results

Results

Research Question 1: Can a Model Achieve Human-Level Quality for Textual Annotation?

Results

	Policy vs. Politics	Unattributed
Human - Microworkers	0.65	0.7
Human - RAs	**0.80**	**0.86**

Table: F-1 Score for human annotation

Comparison to a Gold Standard annotated with care by experts

Results

	Policy vs. Politics	Unattributed
Human - Microworkers	0.65	0.70
Human - RAs	**0.80**	**0.86**
"Classic" supervised models	0.67	0.41
(SVM, LSTM)	[0.671, 0.673]	[0.390, 0.437]

Table: F-1 Score for human annotation vs. Model trained by the expert

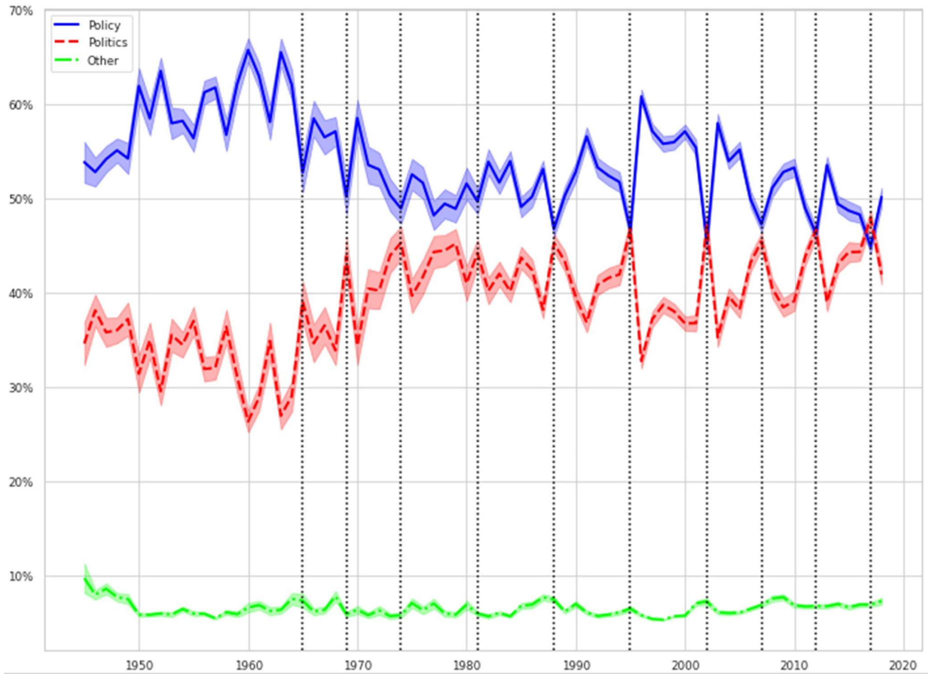
Results

	Policy vs. Politics	Unattributed
Human - Microworkers	0.65	0.70
Human - RAs	**0.80**	**0.86**
"Classic" supervised models (SVM, LSTM)	0.67 [0.671, 0.673]	0.41 [0.390, 0.437]
Augmented Social Scientist (camemBERT)	0.78 [0.781, 0.792]	0.82 [0.816, 0.834]

Table: F-1 Score for human annotation vs. Model trained by the expert

Results

Qualitative Assessment



Policy vs. politics in *Le Monde*

Results

Qualitative Assessment

Type	Frequency
(Quasi-) agreement	76%
Partial agreement	2%
In gold standard, but not predicted (= false negative)	10%
Predicted correctly by the algorithm, but not noticed by the expert	8%
Predicted incorrectly (= false positive)	4%

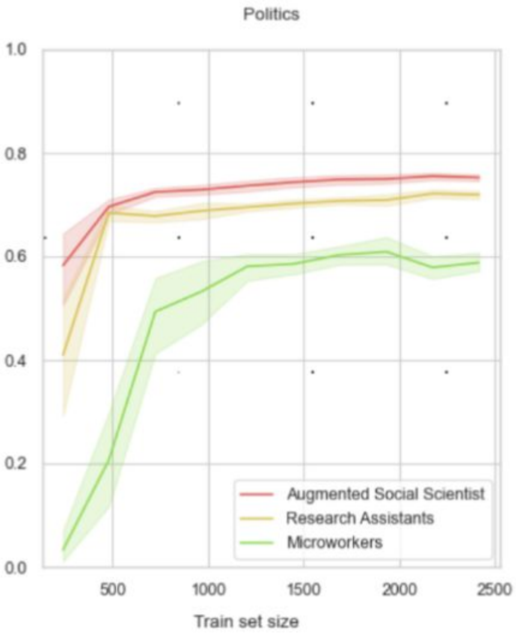
Manual evaluation
of the quality of prediction

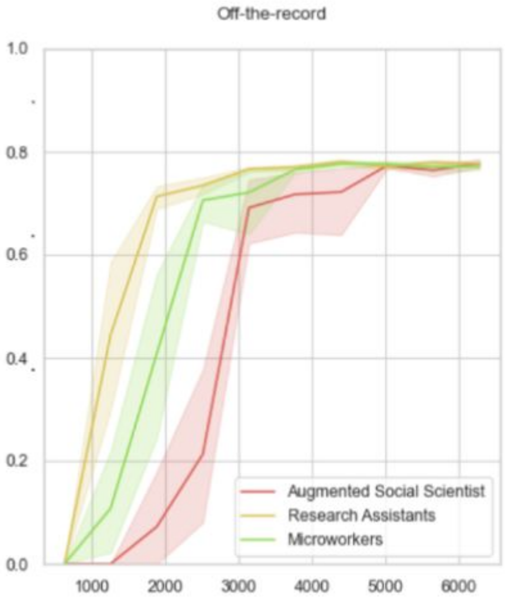
Results

Result1: **It is possible** to train a human-level quality classifier even for complex tasks

RQ 2: What Role does the Expertise of Annotators play?

Results

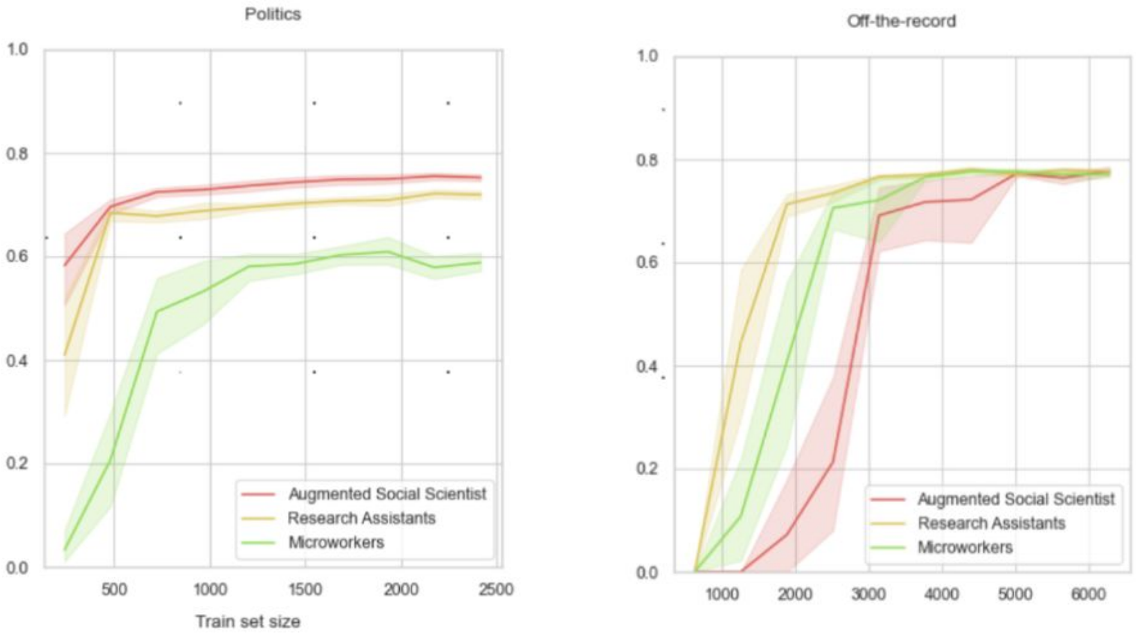




Sample efficiency curves (F1 Scores)

Results

- In both cases, we could have massively cut down on annotation time
- And this is before doing active learning
- But quality matters for complex tasks (*politics*)



Sample efficiency curves (F1 Scores)

Results

Result 1: It is possible to train a human-level quality classifier even for complex tasks

Result 2: Annotation plays an important role for complex tasks

Discussion

Discussion

1. An Immense Promise
    - An improvement with respect to alternative methods
        - Expert hand annotation, at scale
        - More tailored to research needs than non-supervised models
        - No outsourcing

Discussion

1. An Immense Promise
    - An improvement with respect to alternative methods
    - Ability to fully annotate an entire data set
        - Comprehensive AND fine-grained Analysis
        - Forces conceptual clarification (Bonikowski *et al.*, 2022)
        - Avoid “Fatigue effect”, “Learning effect” (Rousson *et al.*, 2002)

Discussion

2. Limitation, Challenges, and Future Developments
    - Computer time and hardware
        - Hard Without a GPU
        - Colab and its Problems

Discussion

2. Limitation, Challenges, and Future Developments

- And what about non-foundational LLMs? (chatGPT)
  - A Recent Experiment (July 2023)
  - Easiest task of the two: detecting “unnamed sources”

Discussion

	F1-Score (“unattributed”)
Microworkers	0.7
Research Assistants	0.86
Augmented (Expert + BERT)	0.82 - (.94)
chatGPT (gpt 3.5-turbo) Few-shot learning Zero-shot learning	

Discussion

	F1-Score ("unattributed")
Microworkers	0.7
Research Assistants	0.86
Augmented (Expert + BERT)	0.82 - (.94)
chatGPT (gpt 3.5-turbo) Zero-shot learning Few-shot learning	 ~.37 .48 - .6

Discussion

2. Limitation, Challenges, and Future Developments

- And what about non-foundational LLMs? (chatGPT)
    - Result? Not anywhere close to expert annotation + BERT
        - Sure, we could better engineer prompts; use a foundational model (Llama), do this in English, etc…
        - Yet: for research purposes, <u>as of now</u>, <u>in our experience</u>, what is relevant is hard to extract with zero or few-shot learning.

Discussion

2. Limitation, Challenges, and Future Developments

- Computer time and hardware
- When to use it, when to not use it?

- And what about non-foundational LLMs? (chatGPT)
  - As of today, clearly better on complex tasks
  - Also: are you sure you <u>don't want</u> to read your corpus??!

Conclusion

The return of an old debate: replacement, or augmentation

Creating an in-silico replica of ourselves

> Doug Engelbart, the Internet and "The Augmentation Research Lab"

> "Increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems." (Engelbart, *Augmenting Human Intellect*, 1962).

Resources

- Do, Ollion and Shen, “The Augmented Social Scientist”, *Sociological Methods and Research*, 2022.
- GitHub repository
- Google Colab

Schedule

Augmented: General Presentation

Lab 1: Using the Augmented Package

BREAK

Lab 2: Conducting A Real Life Project

Augmented: Tips & Tricks