Original Article

Sociological Methods & Research
1-34
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241221134526
journals.sagepub.com/home/smr
SAGE

The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy

Salomé Do[1,2], Étienne Ollion[3], and Rubing Shen[2,3]

Abstract

The last decade witnessed a spectacular rise in the volume of available textual data. With this new abundance came the question of how to analyze it. In the social sciences, scholars mostly resorted to two well-established approaches, human annotation on sampled data on the one hand (either performed by the researcher, or outsourced to microworkers), and quantitative methods on the other. Each approach has its own merits - a potentially very fine-grained analysis for the former, a very scalable one for the latter - but the combination of these two properties has not yielded highly accurate results so far. Leveraging recent advances in sequential transfer learning, we demonstrate via an experiment that an expert can train a precise, efficient automatic classifier in a very limited amount of time. We also show

[1] ENS-Paris/PSL (LATTICE), Paris, France
[2] Sciences Po (Medialab), Paris, France
[3] Institut Polytechnique de Paris (CREST), Palaiseau, France

Corresponding Author:
Étienne Ollion, Institut Polytechnique de Paris, Palaiseau, France.
Email: etienne.ollion@polytechnique.edu

that, under certain conditions, expert-trained models produce better annotations than humans themselves. We demonstrate these points using a classic research question in the sociology of journalism, the rise of a “horse race” coverage of politics. We conclude that recent advances in transfer learning help us augment ourselves when analyzing unstructured data.

Keywords

computational social sciences, natural language processing, supervised machine learning, sequential transfer learning, digital humanities, text as data

The last decade witnessed a spectacular rise in the number of available textual data[1]. Whether born-digital, or digitized from other sources, text is now everywhere for researchers to use. Just like with digital data, this new abundance ushered in many changes in the human and social sciences. Using this trove, classic questions started to be analyzed anew, and new interrogations started to receive responses (Bearman 2015). New fields of research placing digital text as their central source emerged, such as digital humanities (Schreibman, Siemens, and Unsworth 2004). More recently, computational social sciences started making great use of it (Lazer et al. 2009; Mohr et al. 2020).

The availability of these volumes of text also gave way to the return of an old interrogation: what can the human and social sciences do with textual data when it comes *en masse*? To this question, these disciplines have long provided answers that fall into two broad categories. The first approach consists in using individuals to extract information from such a specific material. Resorting to human understanding is often seen as the best solution to capture meaning, especially when the tasks are complex. For a long time, researchers have thus annotated text according to a precise coding scheme. More often than never, they used assistants to help them. Starting with Lazarsfeld and Berelson’s (1948) hand annotation of swaths of newspaper articles in the immediate post-WWII era, they have long outsourced this grueling work.

This approach experienced a renewed popularity recently. The rise of online companies that match work providers with individuals willing to work from home (the most famous being Amazon Mechanical Turk, Appen, or TaskRabbit) provided a new and easy solution to handle this time-consuming activity (Wood et al. 2019). Still, this solution is not without problems. Critics often evoke the ethical problems raised by resorting to underpaid microworkers to conduct research (Casilli 2019; Fort 2016). Quality can also quickly become an issue as little trained and low-paid human annotators cannot be expected to perform well on annotation tasks that could be daunting, even for trained research assistants. Finally, the cost can still be a determining factor - even if the pay is low, the exponential rise in the volume can result in a hefty sum in the end.

The second classic way to analyze text, one that was also conceived to avoid spending too much time manually annotating, is to resort to statistical methods[2]. Here too, the approach is ancient, and this practice of a "distant reading" (Moretti 2013) is now a classic tool in the social and human sciences. Methods such as lexical statistics (Lebart, Salem, and Berry 1998), dictionary-based methods (Stone et al. 1966), or semantic networks (Leskovec, Backstrom, and Kleinberg 2009) are only a few classic approaches among a gamut of widely used techniques. Recent developments in machine learning further added powerful items to this collective toolbox. Latent Dirichlet allocation, an unsupervised method to detect topics in vast ensembles of documents, became extremely popular since it was introduced in the 2000s (Blei, Ng, and Jordan 2003).

In the last decade, scholars have turned to supervised machine learning to analyze vast corpora of texts[3]. The procedure consists in labeling a small sample of data in order to "train" an algorithm. Once this classifier is deemed to be good enough ("test"), it is used to annotate a much larger dataset ("prediction"). This procedure has demonstrated its merits in other areas: it was used to successfully extract information from raw data in areas so diverse as image recognition (Krizhevsky, Sutskever, and Hinton 2017), automatic transcription from audio signal to text (Hinton et al. 2012), or missing value imputation. In all these domains, machine learning demonstrated its "unreasonable effectiveness"[4] to automatically annotate data.

Alas, until recently, there was still a gap between the output of an automated text analysis and what a human could do. In a recently published article, Laura Nelson and her colleagues presented some recent advances in automatic text labeling for sociology. They convincingly demonstrated that human annotation could be replaced by supervised learning model, but also that it could do this mostly when tasks remained relatively simple. In their case, they successfully selected articles that dealt with inequality in a large corpus (Nelson et al. 2021). But no matter how useful supervised machine learning can be, the selection of articles based on a given topic is still a far cry from what human and social scientists may need for their analyses.

Detecting a subtle pattern in a text; annotating at the level of the sentence, or below; accurately identifying the subject and the object of a sentence… Many tasks that appear to be simple for a human are still complex, and sometimes simply off-limits, for a machine.

The recent development of large language models (LLMs) seems to offer a possibility to bridge this long-standing gap. Since their introduction after 2015, these models have demonstrated their effectiveness on a wide range of tasks that could be of great use to social scientists. Recent language models (Devlin et al. 2019; Brown et al. 2020) are generally built from *Transformers* (Vaswani et al. 2017), a neural network architecture able to process sequences of text on a wide range of tasks. Such contemporary language models leverage sequential transfer learning. This technique pre-trains models on massive corpora, before they are adjusted to the desired task. For the users, their main advantage is that they dramatically reduce of the amount of data needed to efficiently train very precise classifiers (Peters et al. 2018)[5].

These advances have nonetheless not been fully embraced by social scientists and digital humanists alike. The issue may have to do with the lack of familiarity with an approach that requires advanced technical skills, at least with respect to the disciplinary standards. But this avoidance has probably more to do with the idea that training an efficient classifier still necessitates large volumes of data in order to attain satisfying results. For a researcher who has to train the algorithm herself, the process could be very time-consuming, if not disheartening altogether.

In the light of these recent developments, this article raises two questions of crucial importance for the use of transfer learning in the human and social sciences:

- Q1: Can a social scientist, in a reasonable amount of time, train an algorithm that will carry out complex tasks of text annotation that are relevant to her research?
- Q2: Can certain tasks be outsourced to research assistants or to microworkers, or does expertise always play a crucial role in training efficient models?

To answer these two questions, we designed an experiment. We chose a topic relevant to social scientists: the rise of horse race journalism in political reporting. A classic line of inquiry in the sociology of journalism, this area has long resorted to the human annotation of texts. Using classic criteria from the literature, we had three groups of humans train an algorithm to detect patterns of interest in a large corpus of newspaper articles. We assessed the quality of the annotation produced by the said algorithm, and we measured the time required to reach a high-quality level of annotation. We also varied the level of expertise of the human annotators, to assess potential differences between groups of annotators.

The experiment yields a clear conclusion to both questions. Even on complex tasks, a well-trained supervised model using sequential learning can equal, and sometimes even surpass human annotation (Q1). The time spent annotating the text for the training part is small enough that even a time-constrained academic can perform it, thus limiting both the cost and the division of labor in research. What is more, resorting to research assistants and as is increasingly done nowadays, to microworkers can provide help in the case of relatively simple tasks, but does not surpass expert annotation (Q2).

Equipped with this technology, a social scientist can produce results that are at the same time tailored to her needs and more accurate than what research assistants or microworkers could produce, while avoiding a potentially harmful division of labor. Using these models, she can become this "augmented" individual the pioneers of computer science dreamed about, a human person enhanced with (not replaced by) technology to achieve more (Engelbart, 1962). This augmented social scientist is, we argue, a thrilling promise for the future of research.

Investigating Political Narration by Journalists

A Classic Research Question

To investigate the relevance of these new large language models for the human and social sciences, we selected a classic line of study. In the sociology of journalism, the transformation of political reporting is a widely studied question. Across countries, scholars have consistently demonstrated that important changes took place in the narration of politics over the last decades. In particular, it was repeatedly shown that politics was increasingly described as a "horse race" (Broh 1980; Littlewood 1998; Farnsworth and Lichter 2011). Starting in the 1960s in the USA, reporters began to describe politics in a strategic fashion; they evoked the blows exchanged and the action of politicians - rather than the measures they passed or the effects of the latter (Patterson 1994). Journalists also increasingly considered it their duty to reveal deals and negotiations, and subsequently focused more on the backstage politics than on the measures they promoted, or on the debates they had. The timing, the cause of this shift are matters of scholarly discussions, but the rise of strategic news coverage is largely attested[6].

Data and Indicators

For this analysis, we used a dataset of articles published in the French daily *Le Monde*. Created in 1944, *Le Monde* is a highbrow daily that rapidly became the reference newspaper. We assumed it would be a good test case, as it has long resisted the use of this strategic language, but reluctantly adopted it at the close of the twentieth century (Kaciaf 2005). From the complete archives of the newspaper, we selected articles about politics, from 1945 to 2018. Our dataset has over 60,000 articles, and close to 39 million words. Taking cues from the literature in the sociology of journalism, we designed two indicators.

- **Task 1: Policy / Politics.** For each sentence of the article, we try to determine whether it evokes the content of a measure, its effects, and more broadly any action outside of the political sphere ("policy"), or if the sentence is *only* about the action of politicians, the strategy they use, the inner workings of the political field ("politics"). A third residual category was created to group all remaining sentences ("other").

  This task is arguably quite complex. There are many ambiguous situations that require humans to make an informed judgment call. The ambivalence is made worse by the fact that we aim to classify the text not at the level of the article, but at the level of the sentence. While the former is the most classic approach, the latter allows for a more nuanced indicator. We thus expect a trained human to be able to carry it out properly, but not without going through some reasoning, making it hard for untrained workers - and even more for a machine - to complete it successfully.
- **Task 2: "Off-the-record".** This task attempts to identify the parts of a sentence that introduce unattributed comments. These "off-the-record quotes" are conventionally seen as a possible proxy for measuring the extent to which journalists reveal backstage politics (strategies, internecine wars and alliances within the political field). They are regarded as a partial but good indicator of the phenomenon we are trying to investigate.

  This task is not trivial, since the wording is rarely the same across cases. But it is not hard either. A majority of prompts introducing "off-the-record" quotes display a clear and similar structure ("a source close to power expressed that", "a person who wants to remain anonymous confessed…”). We thus expect all groups to perform relatively well on this task[7].

The Experiment

To compare the quality of the annotation produced by a model trained by an expert (the social scientist) to other types of annotators, we designed an experiment that features three groups. They represent the main three options available when text annotation is carried out.

- **The Social Scientist.** She is an expert in her field. She has extensive knowledge about the topic at hand, she may have designed the indicators and written the coding scheme. Her time is limited, as she can at best dedicate a few dozens of hours to annotate a dataset. In this case, two of the authors carried out one task.
- **Research assistants.** Skilled, but not expert, individuals. They are often hired by the researcher who wants to do the analysis and may get to know her research. Their time is also limited, but less than that of the social scientist. In this case, we carefully trained three students with advanced qualifications (Master’s level). Before we gave them the sample of articles to annotate, we provided them with written guidelines, but also answered any questions they could have on a few representative texts.
- **Microworkers.** They are often hired on a gig economy platform or in universities. They have limited training, and they have little connection to the researcher or the project. In this case, we asked 34 French-speaking Bachelor students from one of our classes to participate in an experiment. We made it mandatory to validate the course, told them the entire experiment would not last more than 2 hours overall, and that we would offer them extra points for the course if performed well (we eventually gave the points to everyone who participated). They received written guidance, but no questions could be asked to the organizer of the experiment. Because of their educational background, we consider them to be at least as competent as low-paid gig workers one can hire through an online platform[8].

From the *Le Monde* corpus, we drew two separate sub-corpora: a “training set,” used to train the models, and a “test set,” used to evaluate them. In both sub-corpora, the articles were selected to ensure an appropriate distribution of articles over time. In order to provide an independent assessment of the quality of the model, the test set and the training set have no articles in common. The test set was first annotated with care by one of the authors, and then consolidated by contrasting these results with those of an independent hand-coding made by a second author. We use it as a benchmark to assess both the performance of human annotators and that of different models. We first asked the research assistants and the microworkers to annotate the sentences contained in the test set, in order to compare their performance to that of the "expert." Then we asked each group (1 social scientist, 3 research assistants, 34 microworkers) to annotate the training set in order to train a different model, whose performance is assessed on the same, expert annotated test set.[9]

The tasks take a different amount of time to complete, depending both on the type of annotator, and on the task (see Table 1 below). Because it is complex, the policy/politics task requires reflecting on virtually each sentence of the training set. In these circumstances, the expert is much faster than the two other groups, since she is attuned to the criteria and

Table 1. Summary of the Experiment.

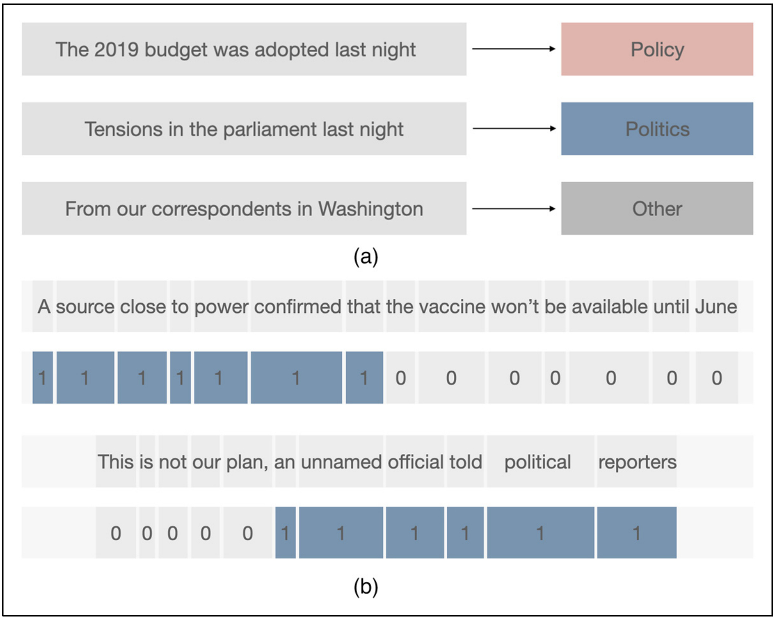
		Social Scientist	Research Assistants	Microworkers
Number of annotators		1 per task	3	34
Level of expertise		High	Moderate	Low
Training set	Policy/politics	2357 sentences from 63 articles 383,000 characters (0.12% of corpus) / 87,000 tokens		
	Off-the-record	6274 excerpts (3 sentences of an article) 3.16 million characters (1.00% of corpus) / 759,000 tokens		
Test set	Policy/politics	377 sentences from 11 articles[22] 60,000 characters / 14,000 tokens		
	Off-the-record	639 excerpts (3 sentences of an article) 317,000 characters / 77,000 tokens		
Time spent (in minutes)	Policy/politics task	480	1051	1243
	Off-the-record task	2220	1869	2654

can decide swiftly. As indicated above, the off-the-record task is simpler, as it just requires patient reading. The annotation time is more important because the share of "off-the-record" is low, thus requiring the various groups to read many texts to have positive examples, but the differences across groups are not so strong.

Methods

Translating Indicators to Supervised Learning Tasks

These two tasks can be presented as two classic natural language processing operations. The policy/politics task is a classic NLP operation called text classification. In this case, it is performed at the level of each sentence. Every article in the training set is first split in sentences with a sentence tokenizer, before it is manually labeled either as "Policy", "Politics" or "Other" by an annotator (Figure 1a).



The 2019 budget was adopted last night → Policy

Tensions in the parliament last night → Politics

From our correspondents in Washington → Other

(a)

A	source	close	to	power	confirmed	that	the	vaccine	won't	be	available	until	June
1	1	1	1	1	1	1	0	0	0	0	0	0	0

This	is	not	our	plan,	an	unnamed	official	told	political	reporters
0	0	0	0	0	1	1	1	1	1	1

(b)

Figure 1. (a) illustration of a text classification task in NLP, (b) illustration of a sequence labeling task in NLP.

The second task, the detection of prompts introducing "off-the-record speech", is another classic NLP task called sequence labeling, a technique which consists in classifying every token[10] of an article according to a binary code (in this case, "off-the-record" or "not off-the-record"). Sequence labeling enables us to get a fine-grained detection of an indicator in precise spans of text - generally below the sentence level. For this reason, it is sometimes called "token classification" (Figure 1b). In our case, each token of the sentence has exactly one label: 1 if the token is part of a segment introducing "off-the-record", 0 if not.[11]

With a small, carefully hand coded sample, we aim to create a classifier that will replicate human behavior, and hopefully allow an automatic annotation of our vast corpus. By doing so, we resort to a supervised approach. Unlike many classic text analysis methods in the social sciences (such as clustering, semantic graphs or topic modeling), *supervised learning* consists in using human-annotated data. This small portion of a corpus serves as a supervisor for a model.

In mathematical terms, a model is a function $f{:}\Theta \times X \rightarrow Y$, where Θ is the parameter space, X the input space, and Y the output space. In the first task, X is the set of the sentences in the corpus, and $Y = \{Policy, Politics, Other\}$. In a supervised setting, we have a human-annotated training corpus (X_{train}, Y_{train}) and the test set (X_{test}, Y_{test}). The supervised training process is the following. Let $\theta_0 \in \Theta$ be the initial parameters. Let then

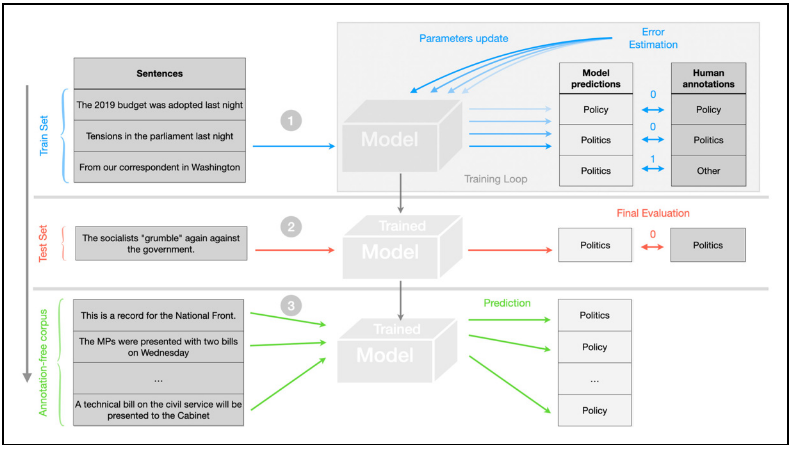

Train Set
Test Set
Annotation-free corpus
Sentences
The 2019 budget was adopted last night
Tensions in the parliament last night
From our correspondent in Washington
1
Model
Parameters update
Error Estimation
Model predictions
Human annotations
Policy
Politics
Politics
0
0
1
Policy
Politics
Other
Training Loop
2
Trained Model
The socialists "grumble" again against the government.
Final Evaluation
Politics
0
Politics
3
This is a record for the National Front.
The MPs were presented with two bills on Wednesday
...
A technical bill on the civil service will be presented to the Cabinet
Trained Model
Prediction
Politics
Policy
...
Policy

Figure 2. A supervised learning framework.

$(x_0, y_0) \in (X_{train}, Y_{train})$ be an input-output sample from the training set. The model predicts $\widehat{y_0} = f(\theta_0, x_0)$. Suppose that we have a loss function $l{:}Y \times Y \rightarrow \mathbb{R}^+$ which quantifies the error made by the model when the model predicts $\widehat{y_0}$ instead of y_0. We want to update the parameters so that the model makes the smallest error possible: we choose $\theta_1 \in \text{argmin}_{\theta \in \Theta} l(\widehat{y_0}, y_0)$. This process is repeated over and over with new samples from (X_{train}, Y_{train}). Once a stopping criterion is reached, the parameters are fixed ($\theta = \theta_{final}$) and the model $f_{\theta_{final}}{:}X \rightarrow Y$ is considered *trained*. $f_{\theta_{final}}$ can be used to predict the label of any sample of X: this process is called *prediction*.

Yet, we want to prevent the trained model from just replicating cases seen in the training sample (what is called "overfitting") and to be capable of *generalization*, at least to some extent. In order to test for generalization abilities, a performance metric is generally computed and averaged across the test set, which has never been seen by the model[12].

Figure 2 offers a schematic overview of this process. The training part comes first. During phase 1, the model relies on human annotations to try to learn the structure of the date. It iteratively corrects itself by comparing its predictions to the labels annotated by the human. Once the model has converged to the lowest error rate possible, it is evaluated (phase 2) on samples that were never seen during training, in order to test its generalization abilities ("Final Evaluation"). Finally, if the scientist deems the trained model performant enough, it is used to predict the rest of the (annotation-free) corpus ("Prediction").

Automated Text Labeling Using pre-Trained Language Models

In our case, we use a supervised framework but also draw on the recent developments of "sequential transfer learning" (Ruder 2019). The term refers to a machine learning technique which aims at *pre-training* a large neural network, generally on a "self-supervised" task, and then to *fine tune* the neural network on a so-called "downstream" task, different from the pre-training task. In our case, we built two models, each one with a downstream task. The first one classifies a sentence into three categories (politics, policy, or other), the second one labels each token ("off-the-record", or not).

Self-supervised learning tries to take advantage of massive, unlabeled textual corpora by designing training objectives which don't need human annotation, yet capture crucial linguistic features. One simple way of generating a supervised dataset without any human annotation is to hide a small part of the data from the model and to let it try to guess the right answer.

In BERT, an algorithm developed by a team at Google (Devlin et al. 2019) that became very popular and that we use here, the self-supervised task used for pre-training is the Masked Language Model (MLM) task (or Cloze test). It consists in masking 15% of a sentence's tokens at random, and in using the partially masked sentence as the input. The model has then to be trained to minimize its error while predicting the masked tokens. For instance, "The cat sits on the mat" could be transformed to "The [MASK] sits on the mat". The model would have to infer that the masked word is "cat."

Coupled with Transformers (Vaswani et al. 2017), a neural network architecture, masked language models are largely used these days. The reason for this appeal is that they are generally trained on massive corpora. BERT, for instance, was produced looking at two massive sets of textual data: the BooksCorpus and the English Wikipedia corpus (Zhu et al. 2015). Training these MLM on such corpora yielded major improvements on classic natural language processing benchmarks. Rather than learning core linguistic features (morphology, syntax, semantics) only through an idiosyncratic task (called a "downstream task"), the learning process is decomposed. First, the model learns a wide range of linguistic features on a large corpus, and then it is refined on the downstream task. An important feature of sequential transfer learning is that it efficiently reduces the amount of labeled data needed to correctly learn a downstream task (Peters et al. 2018).

Results

Assessing the Quality of the Predictions

In order to assess our model's performance, we compare its predictions to the one made by the expert on a smaller dataset. Using such a "test set" is a classic practice in machine learning. This dataset (sometimes also called "gold standard" or "ground truth") was never shown to the algorithm during the training phase. It is thus a good benchmark to assess the quality of the model's predictions.

We use the F1-score as an evaluation metric. F1-score is the harmonic mean of the precision (the positive predictive value, which responds to the question "how reliable is the model when it predicts label A") and the recall (the sensitivity of the model, which answers the question "which proportion of label A is correctly identified by the model?") for a given label[13]. F1-score values range from 0 to 1, with 1 being a prediction that is identical to the consolidated test set. This measure has an important feature. Averaged over all labels, it gives a quality estimate of the prediction that is not affected by unbalanced datasets. This is especially important in our case, as the *off-the-record* labels are rare in our corpus. They represent only 2.75% of the characters in the test set. In this case, using a basic accuracy measure would be misleading as a model predicting no *off-the-record* at all would yield a 97.25% accuracy.

The Superiority of Sequential Transfer Learning

Let us recall that our primary goal is to build a model that can correctly identify either the label of a sentence ("policy vs. politics"), or the occurrence of a prompt signaling the use of unattributed discourse ("off-the-record"). More specifically, our hope is that an algorithm trained with the annotations made by the social scientist could equal, or even surpass the quality of the annotations made by the two other groups on the control dataset. Should this be the case, we could avoid outsourcing this activity to a third party, be it a handful of student research assistants or a crowd of microworkers, to annotate a dataset for us.

Let us first compare the annotations made by the research assistants and the microworkers to those of the expert. Table 2 displays their respective performances on each of the two tasks[14]. The microworkers identified off-the-record prompts in 70% of the cases, while research assistants matched the annotations in the test set in 86% of the cases. These scores are pretty high, an indication that the task is, as we said, not too complex. This is not exactly the case for the policy/politics task. For this one, the research assistants concurred with the test set in 80% of the cases, but the microworkers reached only an agreement score of 65%. The difference is massive, and while we could certainly use annotations made by the research assistants without much hesitation, such is not the case for those of the microworkers - or only as a degraded, second-best option.

In the rest of the paper, we will take the score achieved by the research assistants as a baseline of what rigorous, well-trained humans can do in a reasonable amount of time. Since our goal is to replicate quality human annotation with a machine, our model trained by a human should come close, and potentially surpass, this threshold.

We now compare the labels predicted by the model, trained by the social scientist using pre-trained language models and the sequential transfer learning technique, to those annotated by the expert in the test set. The "Augmented Social Scientist" is a CamemBERT model (Martin et al. 2020), the BERT implementation for French language, fine-tuned on the training set annotated by the social scientist[15]. As indicated previously, this model was never shown any annotation present in the test set, since our goal is to assess whether we can produce a classifier that efficiently replicates the decisions of the expert. The performance of the model is estimated by averaging the F1 scores across multiple training runs with different random seeds (the confidence intervals is presented between brackets).

The third row in Table 2 provides a positive answer to our interrogation. In the case of the policy/politics task, the model trained by the expert almost equals the score reached by the research assistants (78% vs. 80%), a difference well within the margin of errors. The detailed results presented in Appendix B also show that this "augmented social scientist" was better at classifying "politics" sentences and lagged behind research assistants when it comes to the "policy" ones. Put otherwise, a model trained in a limited amount of time (8 h) by an expert was able to do as well as skilled research assistants (in double the time). But there are notable differences between these two annotation procedures. First, the expert was not removed from the material, and she may even have gained insights into it as she was annotating. Second, and more importantly, the now trained model can be used to annotate the entire corpus, a task that remains completely impossible in the second option, even with an army of diligent research assistants.

Table 2. Results for all groups of annotators on the two tasks.

	F1 – Policy vs. Politics	F1 – Off the record
Human – Microworkers	0.65	0.70
Human – Research assistants	**0.80**	**0.86**
Model without pre-training	0.67 [0.671, 0.673]	0.41 [0.390, 0.437]
Augmented social scientist (model with pre-training)	0.78 [0.781, 0.792]	0.82 [0.816, 0.834]

On both tasks, the research assistants performed better than microworkers and models without pretraining did. However, they only surpassed the expert trained model by very few percentage points.

The value in each cell reports the F1-score for each model or group of human annotators on the test set. The F1-scores range between 0 and 1: a higher F1-score indicates a better reconstruction of the test set labels. The numbers between brackets provide the 95% confidence intervals of the model scores, computed using the results of multiple training runs with different random seeds. Bold fond indicates the best performance on each task.

The same remark can be made about the off-the-record task. In this case, the annotation performed by the research assistants achieved an F1-score of 86%, while the annotation performed by the microworkers achieved a score of 70%. The model trained by the social scientist reached 82%, a score slightly lower than that of the research assistants[16], but still very high given the complexity of the task.

To assess the benefits of using the sequential transfer learning technique, we also trained classic supervised models that do not benefit from pre-training. The latter were until recently the method of choice in quantitative textual analysis[17]. The results are not catastrophic, since the model is right in 67% of the cases for policy/politics, but only in 41% for off-the-record (Table 2, third row). This is still a far cry from the results using a pre-trained, large language model.

The main result of this experiment and the response to our first research question is that a model trained by the social scientist using the sequential transfer learning technique, can produce reliable predictions on a sample of text it was never exposed to. On this test set, it yields results that are comparable to the annotations made by skilled research assistants. It also does so without the financial cost or time requirements (for hiring, for training).

When Should an Expert Annotate (and for how long?)

But could we have cut down on this training phase? And could a model trained either by research assistants, or by microworkers, perform as well as the augmented social scientist? To these interrogations, we can also provide responses. To compare the models trained by annotators with different levels of expertise, Figure 3 shows the sample-efficiency curves of the training of each model using the annotations of each group (augmented social scientist, research assistants and microworkers). These plots illustrate a trade-off between the size of the training set size and the quality of the prediction quality. The x-axis represents the volume of data in the training set, and the y-axis the quality of the overall prediction.

They are rich with information. First, we see that we could have reduced the annotation time for all three groups, since the curves reach a plateau, sometimes with little training data. This is clearly the case for the policy/politics task. The F1-score starts to plateau when the model is provided with just one-third to half of the samples. This means that our social scientist could have annotated for less than 4 h and still trained a similarly efficient classifier. For the off-the-record, at least 50% of the annotations were required to reach such a plateau, which is equivalent to about 18 h of annotation for the

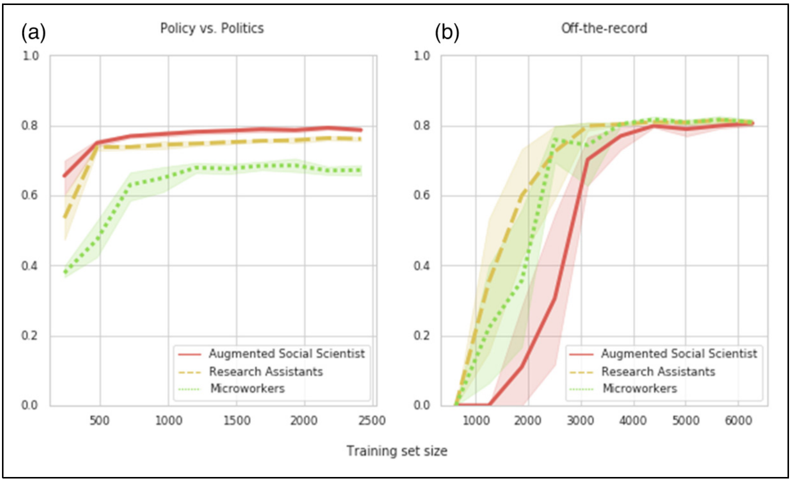


Figure 3. (a, b): sample-efficiency curves for the models trained by different annotators. The expert trained model (augmented social scientist) performs better than the two other groups on the policy vs. politics tasks no matter the size of the training sample. This is not the case for the “off-the-record task”. Each line represents the average F1-score reached with a given amount of training data and the shaded areas provide the associated 95% confidence intervals.[23].

expert[18]. The reason for such a difference between the tasks does not have much to do with their relative difficulty. As indicated above, both are complex, but the off-the-record task is probably easier than the policy/politics task. The surplus in time rather stems from the sparsity of the instances of “off-the-record” in the dataset (2.75%). With such a low prevalence, a single annotator is forced to read thousands of texts if she wants to find enough examples to teach the algorithm a correct representation, but she eventually succeeds.

This sample efficiency curve for the off-the-record task also shows that for annotation jobs that are simple (and potentially time consuming), it can be outsourced to other types of annotators. Past a certain volume of data, the models for all three groups converge at the same (high) level of performance. The sample efficiency curve even displays a surprising pattern: it shows the model trained by the social scientist lagging behind the two others. It turns out that another factor played a role here: annotation fatigue. Upon inspection of the results, we realized that on this quite draining task (one must read thousands of sentences), the expert missed prompts introducing off-the-record more than just a few times. Since her annotation load was much bigger (respectively 3 and 34 times larger than for the two other groups), she likely experienced a classic “fatigue effect” (Rousson, Gasser, and Seifert 2002), which made her miss some instances. We shall return to this point in more detail soon, but in those occasions where the volume of annotation is high and the task not too complex, it may be worth outsourcing the annotation job to other persons, or at least part of it.

A Qualitative Assessment of the Predictions

To assess the quality of the predictions in a more qualitative fashion, one can look at the results. Visualizing the prevalence of each category and comparing their time trends to established results yields supplementary information about the quality of our classifier. Figure 4 shows these results[19].

Figure 4 confirms what historians of journalism have already demonstrated, namely that starting in the 1970s, French journalists increasingly focused on the actions of politicians, at the expense of the narration of

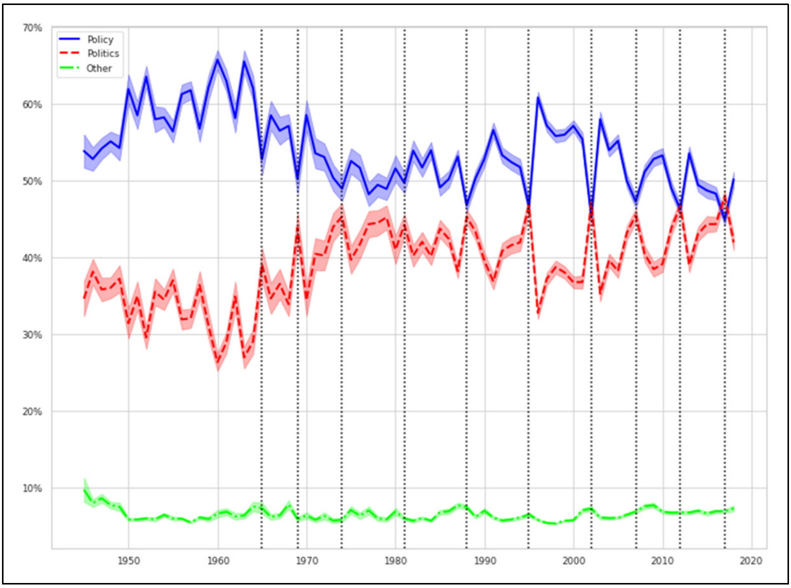


Figure 4. Evolution of politics and policy indicators in *Le Monde* since 1945. In 2000, out of all the articles written by the political reporters in Le Monde, 57% evoked policy measures, 37% dealt with politics, and 6% with other aspects. Each line indicates the estimated prevalence of each category for each year. The shaded areas indicate the associated 95% confidence intervals.

policy measures (Kaciaf 2013). In the data, the change is particularly visible between the 1960s and the 1980s. During these two decades, the share of sentences only focusing on “politics” rose sharply - from 27% to 45%. During those years, an important transformation in the French political journalism scene also happened. The once revered specialty of “parliamentary journalism” progressively disappeared, only to be replaced with the more generic position of “political journalist”. The change was not just in name. It coincided with the disappearance of some practices, one of them being the publication of the almost unabridged minutes of the parliamentary debates (Kaciaf 2005). This disappearance, in turn, drastically lowered the space allocated to discussing policy measures. At the same time, the dissemination of the new norm of journalistic objectivity pushed journalists to demonstrate their distance towards politicians. One manifestation of this was an increased attention to the struggles and tensions within the political field, which became central objects of attention in their daily reporting.

Figure 4 offers yet another sign of the quality of our automatic coding procedure. The dotted vertical lines indicate years of presidential election in France. They also precisely correspond to local peaks in the “politics” coverage. For virtually all these years, the rates of “politics” and “policy” reach respectively their local maximum and minimum, a fact that has long been established in the research on horse-race journalism (Patterson 1994). In the years following the election, the discrepancy between the two lines increased, before a new cycle started again. The method thus confirms results from the literature, but it also goes beyond, as one can investigate the narration of politics besides these very specific moments that electoral campaigns are. Due to obvious limitations in the volume of texts they could annotate, scholars have so far focused their attention on these moments of high attention. But equipped with this technique, we can now analyze the entire corpus, or compare these results with other periods or other newspapers.

We shall return to these advantages in a moment but let us consider the results provided by the second classifier, the one that looks for ‘off-the-record’ prompts. Since there are no established results on the topic for France, it is hard to use this external evidence to validate the prediction. But one can qualitatively inspect the predictions produced by the algorithm once it annotates the full corpus. To do so, we compared its predictions on the entire test set (639 segments of texts). We sorted the annotations on these segments in five categories: (quasi -) agreement between the annotators and the algorithm (1); partial agreement between the expert annotator and the algorithm (2); annotation in the test set, but not by the algorithm (known in the literature as a false negative, 3); correct prediction by the algorithm, but not by the human (4), partial overlap between the algorithm and the human (4), and misclassification by the algorithm (known in the literature as a false positive, 5).

Table 3 illustrates the results of this classification. The quasi-agreement between the expert annotators and the algorithm is the rule, this category representing 76% of the total annotations on all the text segments (a quasi-agreement means that there is an overlap of at least half of the tokens of the longest phrase), to which we could add 2% of partial agreement (some annotation in common, but less than 50% of the tokens). The rest, namely 22%, are broken down accordingly: 10% were annotated by the human, but not by the algorithm - a classic case of false negatives. Conversely, 4% are incorrectly predicted by the algorithm.

Much more interestingly, 8% of all annotations were correctly labeled as "off-the-record" by the algorithm, but not by the human. Upon close inspection, we were forced to recognize that… the algorithm was right. Despite our attention to producing a rigorous test set, the two expert annotators had missed these prompts introducing "off-the-record". This result dovetails with the above remarks mentioning a possible "fatigue effect" from human annotators on long and tedious tasks. It also shows that the above-mentioned metrics were conservative, since the algorithm was in fact correctly annotating in 86% of the cases (agreement, partial agreement, and correct prediction not noticed by experts), thus matching research assistants.

Table 3. Manual Classification of Positive Predictions.

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

Out of the 639 excerpts in the test set sample for the "off-the-record" task, 314 had a positive label. Of these 314, 76% saw a complete agreement between the algorithm and the human, 10% were labelled by humans but not seen by the algorithm. More interestingly, 8% were correctly annotated by the algorithm, but initially not seen by the human who had missed them.

Discussion

An Important Promise

For the human and social sciences, the promise of these methods is massive. Equipped with such algorithms, researchers can craft their own indicators. This stands in contrast to the output of unsupervised models (such as topic models, or more recently word embeddings), arguably the most frequently used tool nowadays in our disciplines. With the latter, the researcher is provided with a result she then needs to interpret and confirm using external sources, thus leaving lots of room to subjectivity (Chang et al. 2009). With supervised models, the validation is also made much simpler, as one can assess the quality of the generalization by comparing, on a holdout sample, the predicted labels to the classification made by the expert. These two properties (self-designed indicators, and measurable error) are the reason why supervised models are often regarded as a method of choice.

But so far, supervised models remained too general to effectively replace human annotation, unless the task was basic. Not anymore, since the introduction of transfer learning coupled to transformer architectures made it possible to achieve quasi-human precision in the automatic coding of texts. The promise is thus nothing shy of bridging the long-standing gap between two established approaches to text analysis, the qualitative approach and the quantitative one. This is the reason why these models start being used in various disciplines that rely on text. For instance, Luo, Card and Jurafsky (2020) successfully leveraged the same algorithms to detect argumentation strategies about climate change. Analyzing tweets about immigration, Mendelsohn, Budak and Jurgens (2021) managed to classify messages according to their framing about immigration, thus operationalizing classic (qualitative) theories about political communication. The list is likely to grow, as in addition to being accurate, the method is very cost effective. In our case, we were able to train our two classifiers in respectively 8 and 30 h, but our results show that we could have cut down on the annotation time by at least half.

Annotating an entire corpus is not just a matter of increasing the size. We contend that it also changes the nature of the analysis, in at least three ways. First, by increasing the volume of text they analyze, researchers are in a position to offer much finer grained analyzes. In our case, we can now compare newspapers, analyze the difference between election years and other periods. We can also zoom in on journalists, and track evolutions in style over the course of their careers - as they get experienced or change organizations. We can link this newly produced information to external data (circulation, ownership, size of the newsroom, socio-demographic variables of the journalists) in order to respond to still disputed questions - such as causes for a change in the narration of politics.

Using this method has another important advantage for (social) scientists: it can, in some cases, make us more certain about the validity of our claim. So far, social and human scientists working on text had to limit themselves to a fraction of the corpus, leaving room for misinterpretation due to hidden patterns or to outliers. With these models, a classifier can label a portion or the whole corpus in virtually the same amount of time. The nagging question of the representativity of the sample and its correlates (selection strategies, interpolation) thus tends to disappear, at least when the entire corpus is available

One last merit of using this set of methods is that they can be easily replicated. Another researcher can, using the same dataset, train her own indicator or annotate according to another coding scheme. She can thus, in a limited amount of time, check for the robustness of the coding. By doing so, we can effectively move towards a "more comprehensive scientific replication standard in which the mandate is to replicate data production, not just data analysis," as (Benoit et al. 2016) invited us to do a few years ago.

Limitations

This method is of course no silver bullet, and limitations exist. First, our estimates of time do not take into account the countless hours (and, in some cases, days) spent preparing for the annotation. This time includes accessing a corpus and making sure it is reliable and can be used to answer the question at hand[20]. It also includes the search for the appropriate indicator, and all preliminary attempts to operationalize it. This latter task can, in fact, be very time consuming. But such a procedure is mandatory for any researcher willing to analyze text, irrespective of the method. Whether one wants to label the text herself or to outsource it to research assistants or to microworkers, these phases still need to take place. We would even argue that this moment is a fruitful one for research and would encourage social scientists to carry out the annotation phase themselves - especially now that it is not so time-consuming.

There are other types of limitations to using language models. One has to do with the need for computational skills. To collect the texts, to clean and format them, but also to use state-of-the-art NLP methods requires a good knowledge in at least one computer programming language. These skills, nonetheless, disseminate rapidly and packaged solutions make the use of sequential transfer learning more democratic. The Transformers package (Wolf et al., 2019) democratizes the use of language models such as BERT (Devlin et al. 2019) or CamemBERT

(Martin et al. 2020). Moreover, using supervised learning algorithms requires a relatively user-friendly annotation interface. We used Doccano (Nakayama et al. 2018), an open-source annotation interface, which is easy to roll out and to maintain, even for people who don't have a developer background. Still, reading the documentation and creating users and tasks is time-consuming.

Yet another issue has to do with the computational cost implied. Arguably, most of the work is done by the computer scientists that pre-trained the model. This phase can take over 80 hours on massive computers[21]. In comparison, the fine-tuning part is relatively limited. Depending on the downstream task and the available data, it took us from 30 min to a few hours on V100 GPUs. Still, and even though many have tried to reduce the computational cost of transformer-based models (Sanh et al. 2019; Jiao et al. 2020), carrying out these tasks is still more easily done on servers with high computation capacities, which may be available in the cloud, but ends up being costly and have a non-negligible environmental impact.

Despite this success, one must also bear in mind that supervised learning models may fail at tasks relying on a high level of textual understanding. Although they perform better than most tools used so far in the domain, supervised language models do not integrate the communicative intent which is inherent to human language and intelligence (Bender and Koller 2020). Instead, they approach it through statistical correlations, with only basic compositionality. Supervised learning models based on pre-trained language models can thus only automate tasks that were already defined but could in no way replace the analysis by the social scientist.

Finally, it should be noted that while we showed that a high performing classifier can accurately label texts, the social scientist's ultimate goal is often the prevalence estimation, also called *quantification* in statistical learning. Our naïve "classify and count" approach does not always provide the best prevalence estimation when compared to quantification-specific methods (Schumacher et al. 2021). Models specifically designed for prevalence estimation are increasingly explored from a statistical learning point of view (see González et al. 2018 for a review).

Conclusion

The experiment demonstrates that using a supervised learning algorithm to fully annotate a textual database might not only be the cheapest, but also the best option available to researchers who want to analyze large scale text corpora. The time spent annotating the text for the training part is small enough that even a time-constrained academic can perform it. The scholar will, in addition, gain crucial insights into her material, and more often than not improve her coding scheme. But even if one were to outsource the annotating part to other humans, she would still gain from using such a tool, as the quality of the annotation is often increased, rather than decreased, by resorting to a model.

Equipped with this technology, social scientists can yield more accurate results than the currently used alternatives. With the help of a pre-trained language model, fine-tuned on their own research questions, they can in many situations extend themselves, while preserving the control they have over the annotation of the data. The same remarks apply to images and films, which can also be investigated using video data analysis with deep learning (Mazières et al. 2021). Put otherwise, the researcher equipped with such a tool could become this augmented social scientist that computer scientists once fantasized about. In the 1960s, the pioneers of the internet discussed how technology could be mobilized to enhance, and not replace, humans. Just like these individuals could be cognitively amplified by this nascent system, researchers could extend themselves, and parse immense corpora with an expert eye.

Author's Note

- The code and data for this paper are available at: https://github.com/Language-Power/Replication_Augmented
- A short tutorial showing how to use the package is available at: https://colab.research.google.com/drive/132_oDik-SOWve31tZ8D1VOx1Sj_Cyzn7?usp=sharing

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Svenska Vetenskapsrådet, Labex, Agence Nationale de la Recherche, (grant number Mining for Meaning 2018-05170, ANR-11-LABX-0047, ANR-19-P3IA-0001).

ORCID iDs

Salomé Do https://orcid.org/0000-0002-6095-6253
Étienne Ollion https://orcid.org/0000-0003-3099-5240
Rubing Shen https://orcid.org/0000-0002-5504-6108

Notes

1. The authors are grateful to Jean-Philippe Cointet, Marion Fourcade, Miriam Hurtado Bodell, Marc Keuschnigg, Thierry Poibeau, Patrick Präg and the three anonymous reviewers for their insightful remarks. They also want to acknowledge the generous comments made by the participants to the Institute for Advanced Studies conference on AI and Social Sciences (March 2020), to the Stockholm University Computational Social Science workshop and to the CREST seminar at ENSAE. This research is supported by the Huma-Num infrastructure, by a grant from the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ ANR-11-LABX-0047, and ANR-19-P3IA-0001) and by the Swedish Vetenskapsrådet (Project Mining for Meaning, grant 2018-05170).
2. For a recent overview, see (Grimmer and Stewart 2013; Cointet and Parasie 2018; Gentzkow, Kelly, and Taddy 2019).
3. See among others (Barberá et al. 2021; King, Pan, and Roberts 2013).
4. According to the oft-cited review by Karpathy (2015).
5. Early transfer learning models reached – with only 1% of the data in the training samples – the same performances as the baseline model with 10% of the training set.
6. See (Aalberg, Strömbäck, and Vreese 2012) for a review.
7. It can be noted that both tasks attempt to move "beyond keyword" or analyzes premised on word frequency. The practice is common in "classic text analysis" but has also generated criticisms (see for instance (Houghton et al. 2019)).
8. The annotation quality of gig workers is often reported to be uneven, and platforms such as Amazon Mechanical Turk cannot ensure that annotators are native speakers (see for instance Tratz and Hovy 2010). The B.A. students here are native French speakers, have a good knowledge of French politics, and are familiar with the writing style found in newspapers. It is thus plausible that their annotations will be of a better quality than those of microworkers recruited on crowdsourcing platforms.
9. Thus, to answer our first research question, we shall compare the performance score of the model trained by the social scientist to that of human annotations. And for the second question, we shall compare the models trained by each group of annotators differing in their level of expertise.

10. A token is defined as a word or subword unit. In our case, tokens are subword units encoded with WordPiece (Wu et al. 2016).
11. If our goal had only been to measure the prevalence of unattributed speech, it would have been easier to formulate the task as a problem of text classification, thus asking the algorithm to determine whether each sentence contains an "off-the-record" quote. Since our purpose here is also to showcase the potentialities offered by sequential transfer learning, we chose to use another technique. The results offered by a sentence classifier are, in fact, even better.
12. One can further use a cross-validation procedure to train and assess the model with different training and test sets, ensuring a more robust performance. In our case, we choose to use a single hold-out test set, since we want to compare the performance of different actors (models, humans) on the exact same benchmark.
13. Formally, $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN}$, $F1 = 2\frac{precision*recall}{precision+recall}$, where TP, FP, FN denote the number of true positive, false positive and false negative predictions.
14. Detailed results are reported in Appendix B.
15. The training of the models requires two learning parameters: the learning rate and the number of epochs, which we determined through a quick manual grid-search. They are then fixed for the training of all models.
16. For the off-the-record task, we considered that the model was predicting correctly when there was an overlap of at least 50% of the tokens annotated in the test set with the prediction of the algorithm. This idiosyncratic measure was designed in order to avoid penalizing the model because it would miss one word out of an entire sentence.
17. Linear SVM classifier with TF-IDF encoding for the text classification task (policy vs. politics), LSTM network with one-hot encoding for the sequence labeling task (off-the-record).
18. This has a negative downside: past a certain level, it is not possible to further improve the quality of the annotation. But with 85% of correct guesses, our classifiers are on a par with what skilled human annotators would do.
19. The estimation procedure is the following: for each category, the actual prevalence μ_y can be linked to the prevalence of positive predictions $\mu_{\hat{y}}$ using two performance metrics: *precision* and *false omission rate* (*FOR*). Their relationship is the following $\mu_y = Precision^*\mu_{\hat{y}} + FOR^*(1-\mu_{\hat{y}}) = FOR + (Precision - FOR)^*\mu_{\hat{y}}$. (See Appendix C for a formal derivation.) Therefore, we first use the predicted data to produce the classic empirical mean estimator of $\mu_{\hat{y}}$ and the associated confidence interval with Student's t-distribution. We then transform these estimations into estimations of the actual prevalence μ_y using the previous relationship. We assume here that the performance metrics are constant over time. One can produce a more realistic estimation by assessing the model performance year by year.

20. On this question, see (Hurtado Bodell, Magnusson, and Mützel 2022).
21. You et al. (2020) estimate it at around 80 hours on 16 TPUs for BERT.
22. We sampled at article-level because we wanted the model to carry out the very same task research assistants are conventionally asked to complete, namely, to annotate an entire article (and not sentences drawn at random). However, as a reviewer aptly points out, this choice might introduce a correlation between the sentences in the test set within each article. But as our goal is to compare performances, any bias which might exist would be the same for all actors (humans, models), thus does not substantially impact our results.
23. For each given training set size, we randomly sample a subset in the entire training set. We use this sub-training set to train the model, the F1-score of which is then measured. We repeat this procedure multiple times with different subsets drawn at random, in order to compute the means and the confidence intervals.

References

Aalberg, Toril, Jesper Strömbäck, and Claes H. de Vreese. 2012. "The Framing of Politics as Strategy and Game: a Review of Concepts, Operationalizations and key Findings." *Journalism* 13(2):162–78.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: a Practical Guide." *Political Analysis* 29(1):19–42.

Bearman, Peter. 2015. "Big Data and Historical Social Science." *Big Data & Society* 2(2): 1–5.

Bender, Emily M. and Alexander Koller. 2020. "Climbing Towards NLU: on Meaning, Form, and Understanding in the Age of Data." Pp. 5185–98 in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and agile Production of Political Data." *American Political Science Review* 110(2):278–95.

Berelson, Bernard and Paul F. Lazarsfeld. 1948. *The Analysis of Communication Content*. Chicago: University of Chicago Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Broh, C. Anthony. 1980. "Horse-Race Journalism: reporting the Polls in the 1976 Presidential Election." *Public Opinion Quarterly* 44(4):514–29.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, and Clemens Winter,… Dario Amodei. 2020. “Language Models are Few-Shot Learners.” *Advances in Neural Information Processing Systems* 33:1877–901.

Casilli, Antonio. 2019. *En Attendant les Robots: Enquêtes sur le travail du clic*. Paris: Seuil.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. “Reading Tea Leaves: how Humans Interpret Topic Model.” Pp. 288–96 in Proceedings of the 22nd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc.

Cointet, Jean-Philippe and Sylvain Parasie. 2018. “Ce que le big Data Fait à l’analyse Sociologique des Textes: un Panorama Critique des Recherches Contemporaines.” *Revue Française de Sociologie* 59(3):533–57.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: pre-Training of Deep Bidirectional Transformers for Language Understanding.” Pp. 4171–86 in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics.

Engelbart, Douglas C. 1962. “Augmenting Human Intellect: A Conceptual Framework.” SRI Summary Report AFOSR-3223.

Farnsworth, Stephen J. and Samuel R. Lichter. 2011. *The Nightly News Nightmare: Media Coverage of U.S. Presidential Elections. 1988–2008*. Lanham, Maryland: Rowman & Littlefield.

Fort, Karën. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. London: ISTE Ltd.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature* 57(3):535–74.

González, Pablo, Alberto Castaño, N. Chawla, and Juan José del Coz. 2018. “A Review on Quantification Learning.” *ACM Computing Surveys* 50(5):1–40.

Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–97.

Houghton, James P., Michael Siegel, Stuart Madnick, Nobuaki Tounaka, Kazutaka Nakamura, Takaaki Sugiyama, Daisuke Nakagawa, and Buyanjargal Shirnen. 2019. “Beyond Keywords: tracking the Evolution of Conversational Clusters in Social Media.” *Sociological Methods and Research* 48(3):588–607.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. “Deep Neural Networks for Acoustic

Modeling in Speech Recognition: the Shared Views of Four Research Groups." *IEEE Signal Processing Magazine* 29(6):82–97.

Hurtado Bodell, Miriam, Måns Magnusson, and Sophie Mützel. 2022. "From Documents to Data: A Framework for Total Corpus Quality." Retrieved August 1, 2022 (https://osf.io/preprints/socarxiv/ft84u/).

Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. "TinyBERT: Distilling BERT for Natural Language Understanding."

Kaciaf, Nicolas. 2005. Les métamorphoses des pages politiques dans la presse française (1945–2000). Doctoral thesis, Université Paris-1 Panthéon Sorbonne, Paris, France.

Kaciaf, Nicolas. 2013. *Les pages "Politique": Histoire du journalisme politique dans la presse française (1945–2006).* Rennes, France: Presses universitaires de Rennes.

Karpathy, Andrej. 2015. "The Unreasonable Effectiveness of Recurrent Neural Networks", blog post (http://karpathy.github.io/2015/05/21/rnn-effectiveness/).

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2):1–18.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60(6):84–90.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science (New York, N.Y.)* 323(5915):721–3.

Lebart, Ludovic, André Salem, and Lisette Berry. 1998. *Exploring Textual Data.* Dordrecht: Kluwer Academic Publishers.

Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. "Meme-Tracking and the Dynamics of the News Cycle." Pp. 497–509 in Proceeding of the 15thACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery.

Littlewood, Thomas. 1998. *Calling Elections: The History of Horse-Race Journalism.* Notre Dame, Indiana: University of Notre Dame Press.

Luo, Yiwei, Dallas Card, and Dan Jurafsky. 2020. "Detecting Stance in Media on Global Warming." Pp. 3296–315 in Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. "CamemBERT: a Tasty French Language Model." Pp. 7203–19 in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics.

Mazières, Antoine, Telmo Menezes, and Camille Roth. 2021. “Computational Appraisal of Gender Representativeness in Popular Movies.” *Humanities and Social Sciences Communications* 8:137. (https://doi.org/10.1057/s41599-021-00815-9).

Mendelsohn, Julia, Ceren Budak, and David Jurgens. 2021. “Modeling Framing in Immigration Discourse on Social Media.” Pp. 2219–63 in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics.

Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terence E. McDonnel, Ann Mische, Iddo Tavory, and Frederich F. Wherry. 2020. *Measuring Culture*. New York: Columbia University Press.

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Nakayama, Hiroki, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. “Doccano: Text Annotation Tool for Human,” software available from https://github.com/doccano/doccano.

Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. “The Future of Coding: a Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods.” *Sociological Methods & Research* 50(1):202–37.

Patterson, Thomas. 1994. *Out of Order: An Incisive and Boldly Original Critique of the News Media’s Domination of America’s Political Process*. New York: Vintage.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” Pp. 2227–37 in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). New Orleans, Louisiana: Association for Computational Linguistics.

Rousson, Valentin, Theo Gasser, and Burkhardt Seifert. 2002. “Assessing Intrarater, Interrater and Test–Retest Reliability of Continuous Measurements.” *Statistics in Medicine* 21(22):3431–46.

Ruder, Sebastian. 2019. Neural Transfer Learning for Natural Language Processing. Doctoral thesis, National University of Ireland, Galway, Ireland.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.”

Schreibman, Susan, Ray Siemens, and John Unsworth (eds.). 2004. *A Companion to Digital Humanities*. Oxford: Blackwell Publishing Ltd.

Schumacher, Tobias, Markus Strohmaier, and Florian Lemmerich. 2021. "A Comparative Evaluation of Quantification Methods" Retrieved February 15, 2021 (https://arxiv.org/abs/2103.03223).

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, Massachusetts: MIT Press.

Tratz, Stephen and Eduard Hovy. 2010. "A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation." Pp. 678–87 in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All you Need." Pp. 6000–10 in Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California: Curran Associates Inc.

Wolf, Thomas, Lysandre Debut and Victor Sanh. 2019. Huggingface's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771.*

Wood, Alex J., Mark Graham, Vili Lehdonvirta, and Isis Hjorth. 2019. "Networked but Commodified: the (Dis)Embeddedness of Digital Labour in the Gig Economy." *Sociology* 53(5):931–50.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, and Hideto Kazawa, … Jeffrey Dean. 2016. "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation." Retrieved June 12, 2021 (https://arxiv.org/abs/1609.08144).

You, Yang, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. "Large Batch Optimization for Deep Learning: Training BERT in 76 min." Retrieved June 12, 2021 (https://arxiv.org/abs/1904.00962).

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. "Aligning Books and Movies: towards Story-Like Visual Explanations by Watching Movies and Reading Books." Pp. 19–27 in 2015 IEEE International Conference on Computer Vision (ICCV). Los Alamitos, California: IEEE Computer Society.

Appendix

Appendix A: Details about the dataset

See Table A1.

Table A1. Label Distributions (Expressed in %).

	Social Scientist Training set	Research Assistants Training set	Microworkers Training set	Test set
"Politics"	39	36.2	33.3	37.6
"Policy"	55.3	52.2	56	56.6
"Other"	5.7	11.6	10.6	5.8
"Off-the-record"	2.47	3.3	4.27	2.75

In the training set annotated by the expert social scientist, 39% of the sentences were labeled as "politics", and 2.47% of the tokens as introducing "off-the-record.

Appendix B: Detailed Performance Scores

For the policy/politics task, the metric used is a classic F1-score. For the off-the-record task, we considered that the model was correct when there was an overlap of at least 50% of the tokens annotated in the test set with the prediction of the algorithm. This idiosyncratic measure was designed in

Table B1. F1-Score Performances.

	F1 - Off	F1 - Policy	F1 - Politics	F1 - Other
Human - Microworkers	0.70 *(0.55)*	0.72	0.57	0.56
Human - Research assistants	0.86 *(0.79)*	0.83	0.77	0.66
Model without pre-training	0.41 (*0.46)* [0.390, 0.437] *([0.448, 0.480])*	0.74 [0.743, 0.744]	0.60 [0.600, 0.602]	0.09 [0.089, 0.091]
Augmented social scientist (model with pre-training)	0.82 (*0.80)* [0.816, 0.834] *([0.796, 0.815])*	0.82 [0.815, 0.826]	0.75 [0.745, 0.759]	0.34 [0.318, 0.3653]

The model trained by the social scientist predicted correctly 82% of the instances of "Off-the-record" in the test set (80% at the level of the character).

Table B2. Precision Performances.

	Precision - Off	Precision - Policy	Precision - Politics	Precision - Other
Human - Microworkers	0.67 *(0.44)*	0.74	0.57	0.47
Human - Research assistants	0.86 *(0.75)*	0.87	0.77	0.52
Model without pre-training	0.46 *(0.70)* [0.439, 0.477] *([0.687, 0.718])*	0.71 [0.708, 0.709]	0.60 [0.603, 0.605]	1.00 [1.00, 1.00]
Augmented social scientist (model with pre-training)	0.78 *(0.79)* [0.764, 0.797] *([0.779, 0.807])*	0.80 [0.796, 0.811]	0.75 [0.744, 0.759]	0.48 [0.448, 0.515]

The model trained by the social scientist had a precision of 78% when detecting “Off-the-record” (79% at the level of the character).

Table B3. Recall Performances.

	Recall - Off	Recall - Policy	Recall - Politics	Recall - Other
Human - Microworkers	0.73 *(0.71)*	0.69	0.58	0.70
Human - Research assistants	0.86 *(0.83)*	0.80	0.77	0.91
Model without pre-training	0.38 (*0.35*) [0.350, 0.406] *([0.330, 0.367])*	0.78 [0.780, 0.783]	0.60 [0.596, 0.601]	0.05 [0.047, 0.048]
Augmented social scientist (model with pre-training)	0.89 (*0.81*) [0.877, 0.911] *([0.800, 0.838])*	0.83 [0.833, 0.845]	0.75 [0.742, 0.763]	0.26 [0.244, 0.292]

The model trained by the social scientist had a recall of 89% when detecting “Off-the-record” (81% at the level of the character).

order to avoid penalizing the model just because it missed one word out of a sentence. A measure at the level of the character is also mentioned (italicized). (Tables B1–B3)

Appendix C: Estimation of the Actual Prevalence of a Classification Category, Using Predicted Data and Performance Metrics

For a given classification category, let y_i, $\widehat{y_i}$ be the binary random variables respectively indicating if the sentence i truly belongs to the category and if it is predicted by the classifier to belong to the category. We have

$$y_i \sim Bernoulli(\mu_y),$$

$$\widehat{y_i} \sim Bernoulli(\mu_{\hat{y}}).$$

Let's introduce two classification performance metrics, *precision* and *false omission rate (FOR)*, given by

$$Precision = p(y_i = 1|\widehat{y_i} = 1),$$

$$FOR = p(y_i = 1|\widehat{y_i} = 0).$$

Using these performance metrics, the means μ_y and $\mu_{\hat{y}}$ can be linked to each other using the following formula

$$\mu_y = FOR + (Precision - FOR) * \mu_{\hat{y}}. \tag{1}$$

Indeed, we have

$$\begin{aligned}\mu_y =& p(y_i = 1)\\ =& p(y_i = 1|\widehat{y_i} = 1)p(\widehat{y_i} = 1) + p(y_i = 1|\widehat{y_i} = 0)p(\widehat{y_i} = 0)\\ =& Precision * \mu_{\hat{y}} + FOR * (1 - \mu_{\hat{y}})\\ =& FOR + (Precision - FOR) * \mu_{\hat{y}}.\end{aligned}$$

Thus, to estimate the actual average μ_y and its associated confidence interval, one can first estimate those of $\mu_{\hat{y}}$ using the predicted data, then linearly transform them using Equation 1. Typically, if $\overline{\mu_{\hat{y}}}$ is the classic empirical mean estimator of $\mu_{\hat{y}}$ and if the 95% confidence interval given by Student's t-distribution is noted $\mu_{\hat{y}} \in [\overline{\mu_{\hat{y}}} - \eta, \overline{\mu_{\hat{y}}} + \eta]$, then $\overline{\mu_y} = FOR + (Precision - FOR) * \overline{\mu_{\hat{y}}}$ is an unbiased estimator of μ_y (assuming that *Precision* and *FOR* are correctly estimated) and the associated 95% confidence interval is given by $\mu_y \in [\overline{\mu_y} - (Precision - FOR) * \eta, \overline{\mu_y} + (Precision - FOR) * \eta]$. One can also choose other estimation methods for $\mu_{\hat{y}}$ better fitted to the predicted data structure.

The performance metrics *Precision* and *FOR* can be estimated using prediction results. This is done by counting the number of true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*): $Precision = \frac{TP}{TP+FP}$, $FOR = \frac{FN}{TN+FN}$. However, the estimation may have uncertainties, due to training set sampling, test set sampling, stochasticity of the training or hand-coding errors. One way to estimate these uncertainties is to use cross-validation and/or bootstrapping. Then, the uncertainties of the performance metrics can be taken into account when using Equation (1), by considering appropriate upper and lower bounds of the performance metrics.