Learning to Analyze, Learning to Produce Data

Using AI to revisit the History of Political Journalism (in France)

Étienne Ollion (CNRS – Ecole polytechnique)

Joint Paper with **Salomé Do** (Sciences Po - ENS)

What can (Social) Sciences do with Machine Learning ?

Machine learning “*follows the same procedure [*as classic statistics*] of generating, testing, and discarding or refining hypotheses. But while a scientist may spend her whole life [*doing so*], machine learning can do it in a fraction of second*”

P. Domingos, *The Master Algorithm*, 2015, p. 13

Machine learning “*follows the same procedure* [as classic statistics] *of generating, testing, and discarding or refining hypotheses. But while a scientist may spend her whole life* [doing so]*, machine learning can do it in a fraction of second*”

Machine learning is “***the scientific method on steroids***". It is thus "*no surprise that* ***it is revolutionizing science***”

P. Domingos, *The Master Algorithm*, 2015, p. 13

What can (Social) Sciences do with Machine Learning ?

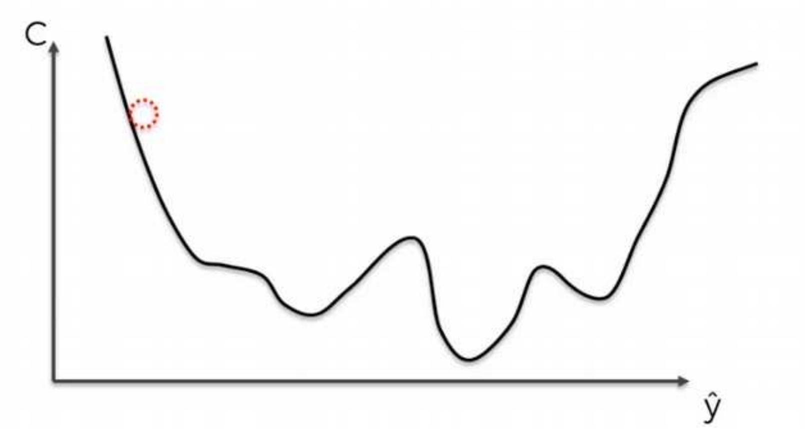
Promises

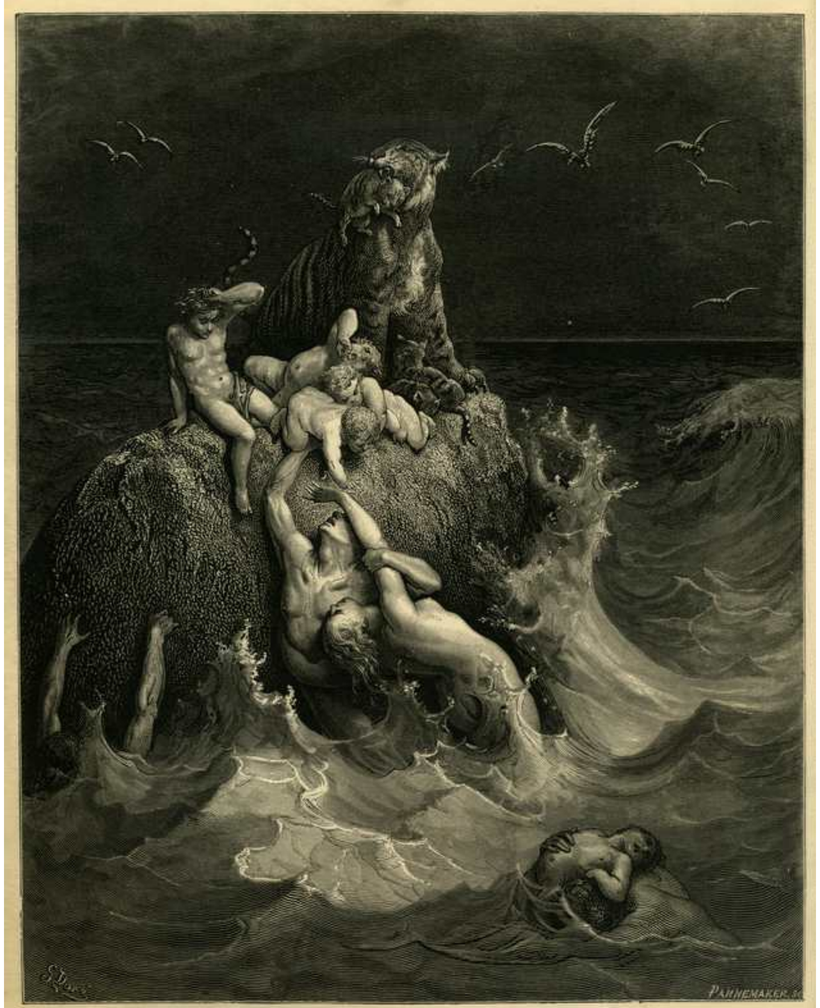
- More flexible than standard methods
  - More fine-grained analyzes
  - More contextual
- Universal approximation

What can (Social) Sciences do with Machine Learning ?

And known criticisms

- No clear mathematical demonstration
- Uncertain optimality
- Lack of interpretability
- Prediction



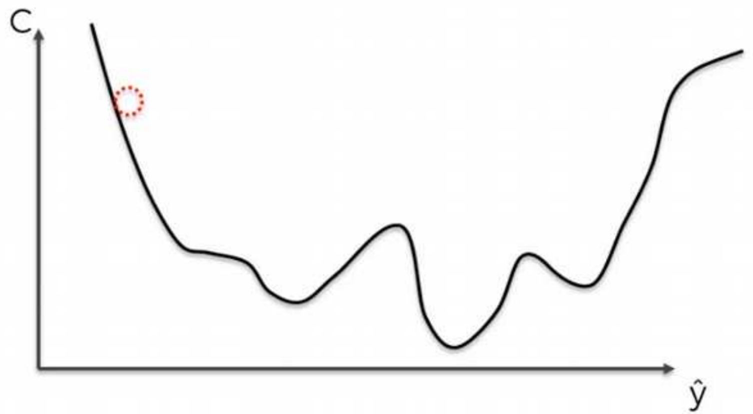


THE DELUGE.

What can (Social) Sciences do with Machine Learning ?

And known criticisms

- No clear mathematical demonstration
- Uncertain optimality
- Lack of interpretability
- Prediction



J. Boelaert & E. Ollion, ‘The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences’, in *Revue Française de Sociologie*, 2018.

What can (Social) Sciences do with Machine Learning ?

Changing our goals

⇒ From ML to analyze data to ML to Produce data

What can (Social) Sciences do with Machine Learning ?

Changing our goals

⇒ From ML to analyze data to ML to produce data

Producing data as a way out of the classic conundrum

- Prediction is not a problem anymore
- Neither is the ‘black box’ aspect
- Validation is easy
- Amenable to any type of (social) scientific research

Case : History of Political Journalism (in France)

Many studies about the transformations of journalism

Case : History of Political Journalism (in France)

Many studies about the transformations of journalism

One of them : ‘How has the narration of politics changed?’

→ Rise of ‘strategic news coverage’ (SNC)

- Change in what is reported (issues → strategies)
- Change in how this is reported (winners & losers, backstage stories)

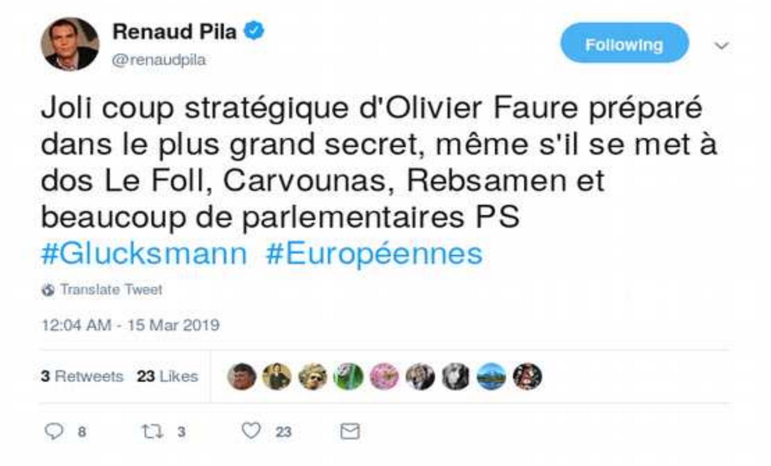
Case : History of Political Journalism (in France)

Many studies about the transformations of journalism

One of them : ‘How has the narration changed?’

→ Rise of ‘strategic news coverage’ (SNC)

- Change in what is reported (issues → strategies)
- Change in how this is reported (winners & losers, backstage stories)



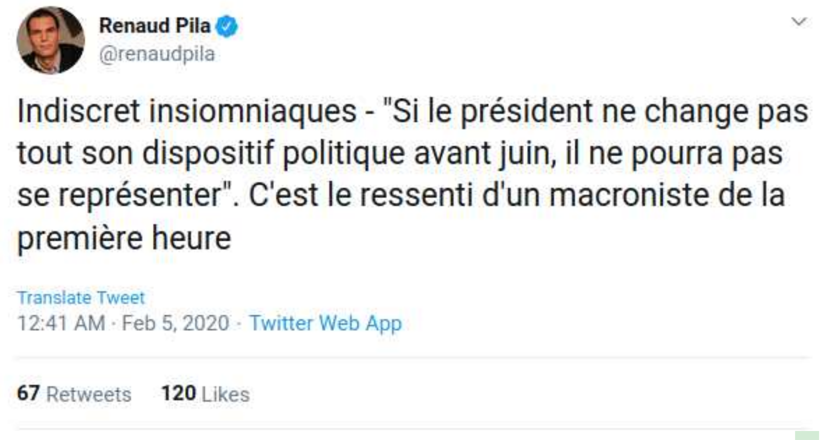
Renaud Pila
@renaudpila

Joli coup stratégique d'Olivier Faure préparé dans le plus grand secret, même s'il se met à dos Le Foll, Carvounas, Rebsamen et beaucoup de parlementaires PS #Glucksmann #Européennes

Translate Tweet

12:04 AM - 15 Mar 2019

3 Retweets 23 Likes



Renaud Pila
@renaudpila

Indiscret insiomniaques - "Si le président ne change pas tout son dispositif politique avant juin, il ne pourra pas se représenter". C'est le ressenti d'un macroniste de la première heure

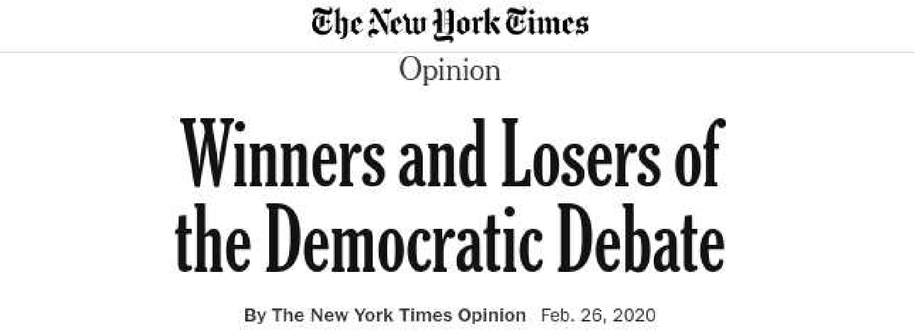
Translate Tweet

12:41 AM · Feb 5, 2020 · Twitter Web App

67 Retweets 120 Likes

Case : History of Political Journalism

> Same pattern exists in other countries



The New York Times

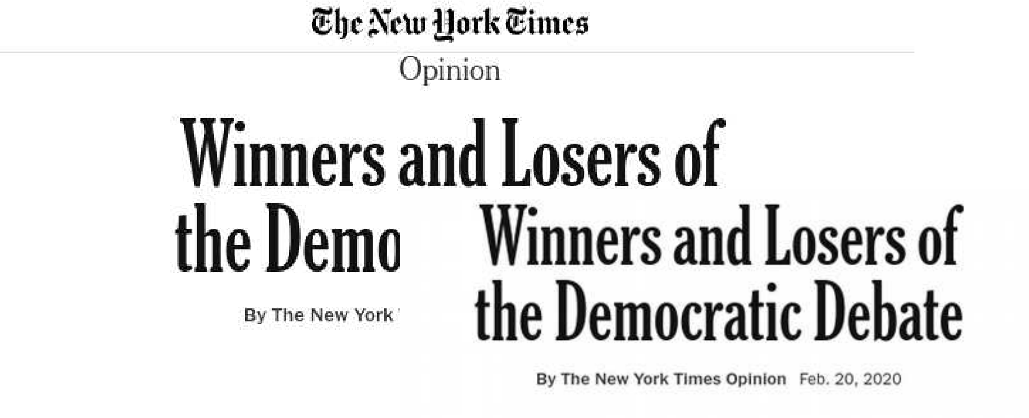
Opinion

Winners and Losers of the Democratic Debate

By The New York Times Opinion Feb. 26, 2020

Case : History of Political Journalism

> Same pattern exists in other countries



The New York Times

Opinion

Winners and Losers of the Demo

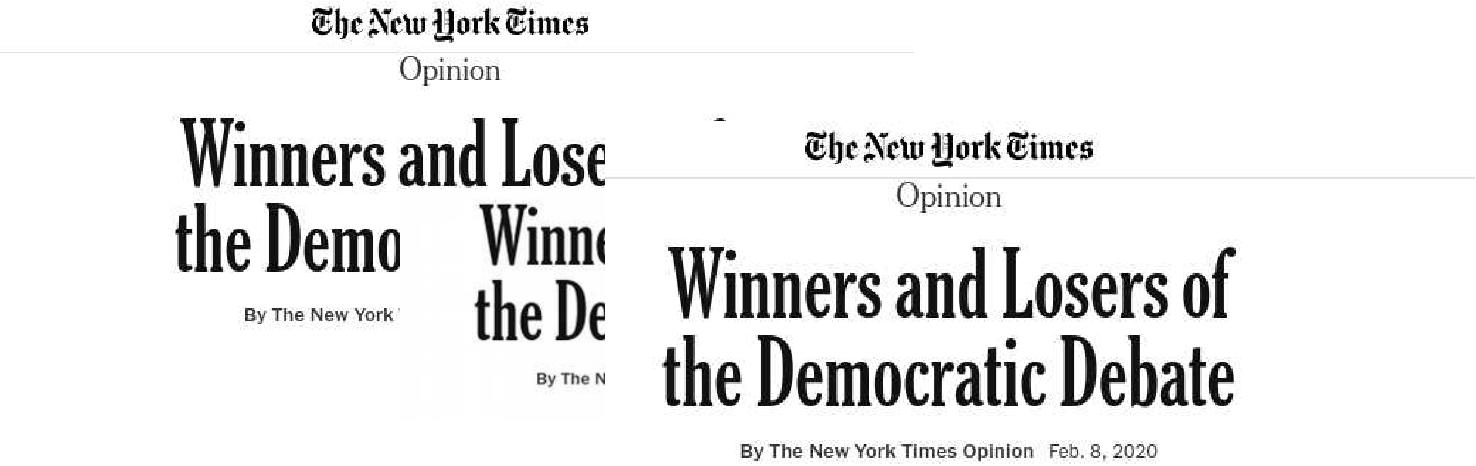
By The New York

Winners and Losers of the Democratic Debate

By The New York Times Opinion Feb. 20, 2020

Case : History of Political Journalism

> Same pattern exists in other countries



The New York Times

Opinion

Winners and Lose

the Demo

By The New York

Winn

the De

By The N

The New York Times

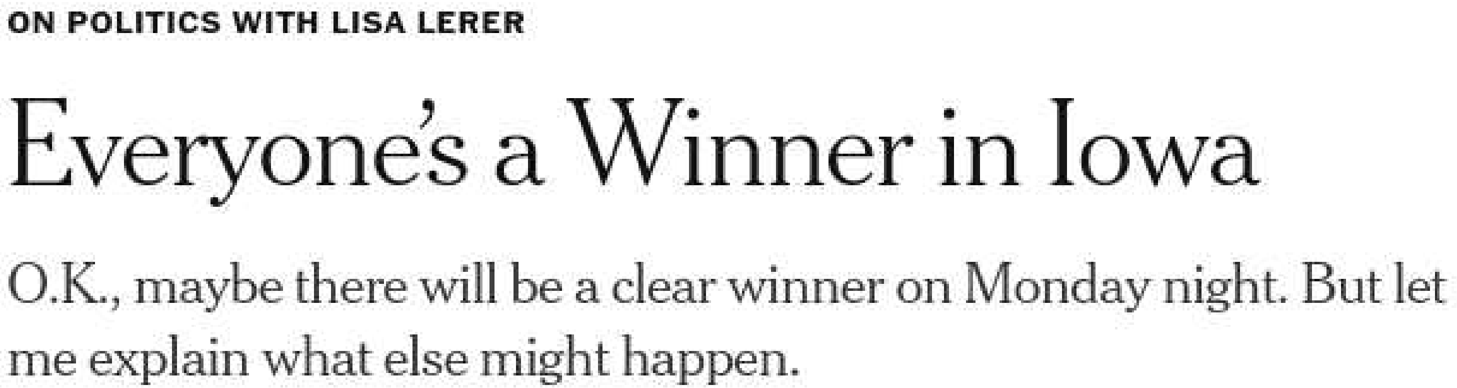
Opinion

Winners and Losers of the Democratic Debate

By The New York Times Opinion Feb. 8, 2020

Case : History of Political Journalism

> Same pattern exists in other countries



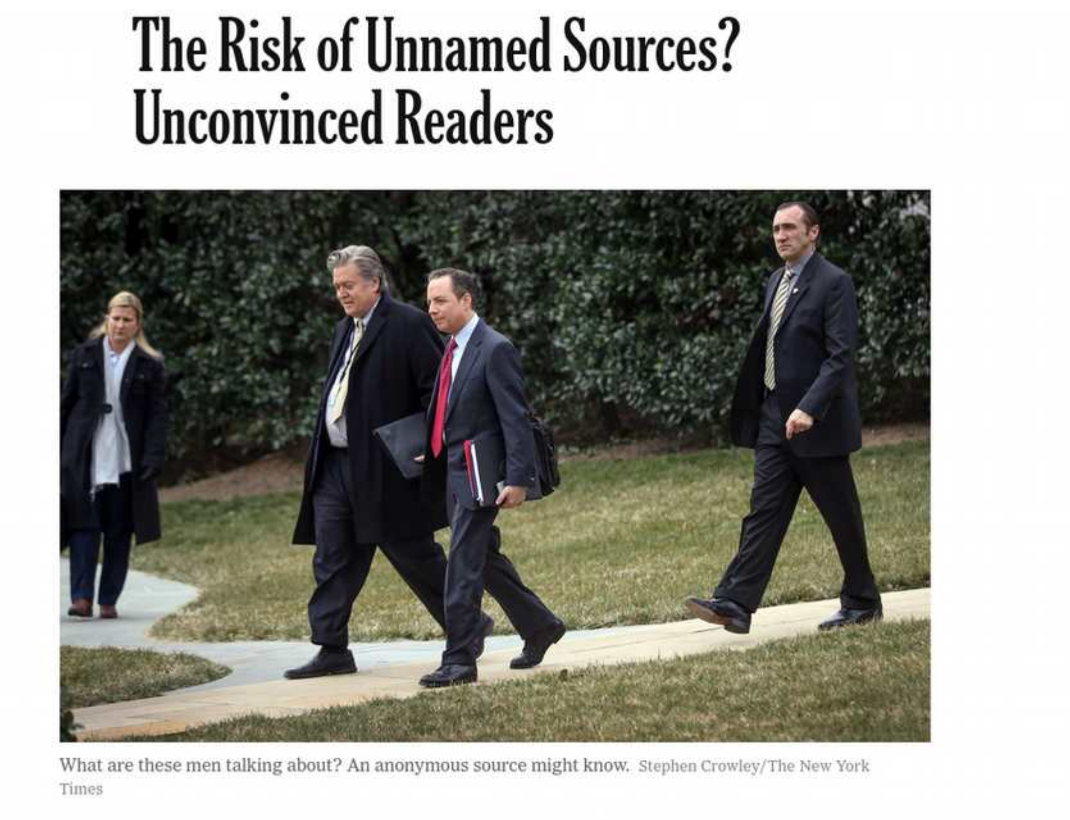
ON POLITICS WITH LISA LERER

Everyone's a Winner in Iowa

O.K., maybe there will be a clear winner on Monday night. But let me explain what else might happen.

Case : History of Political Journalism

> Same pattern exists in other countries



The Risk of Unnamed Sources? Unconvinced Readers

What are these men talking about? An anonymous source might know. Stephen Crowley/The New York Times

Case : History of Political Journalism

A story about the media is a story about the public sphere (resp. the State, polarization, public trust, war, etc.)

Zoizner, ‘The Consequences of Strategic News Coverage for Democracy: A Meta-Analysis’, *Communication Research*, 2018

Leads to

- Lack of public trust
- Increased polical cynicism amongst citizen
- Sharper elites/ citizen divide
- ...

Case : History of Political Journalism (in France)

Many studies about the transformations of journalism

One of them : ‘How has the narration changed?’

Well-documented, though a few shortcomings

- Precise timeline?
- Causes?
- Segregated or widespread?
- Only prevalent during election periods ? (‘horseracism’)

Case : History of Political Journalism (in France)

Many studies about the transformations of journalism

One of them : ‘How has the narration changed?’

More specifically, competing theories

- Around the 1960s (& new political regime)
- Around 1980s (& rise of investigative journalism)
- Around 2000s (& cable news)

Goal: Revisit this question computationally

- Determine a series of indicators pointing to this shift in political journalism.
- Then train an algorithm to outsource the job to it
- Measure evolution over time, in each article

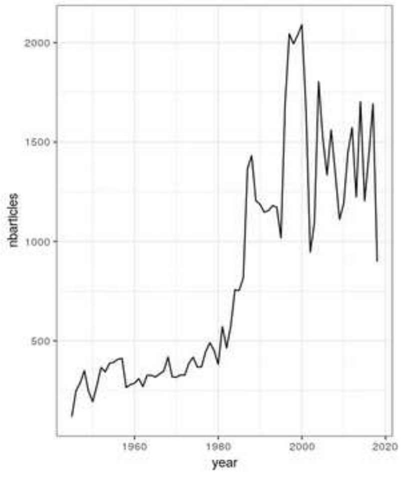
⇒ In other words: do not do away with the qualitative approach, but rather outsource the repetitive aspect of it to a machine.

Corpus: All articles by political journalists in *Le Monde* since 1946

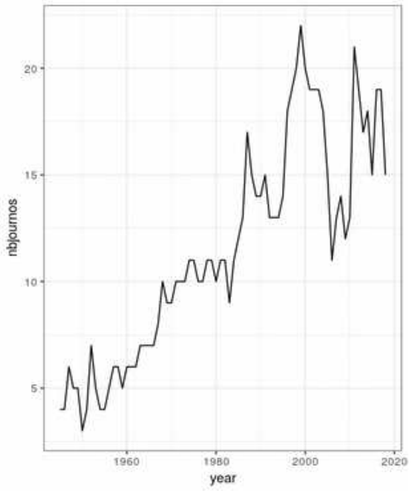
- *Le Monde*: Reference newspaper in France
- (hand) curated list of political journalists
- Selection of all all articles they published as political journalists

Dates	Articles	Number of words	Number of pol. journalists
1946-2018	65,100	~ 3,9 M	849

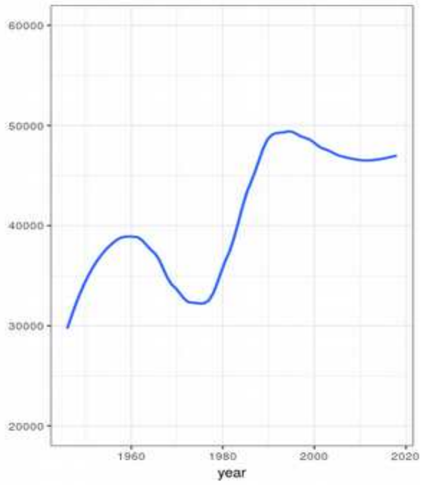
Corpus: All articles by political journalists in *Le Monde* since 1946



Nb articles



Nb pol. journalists



Nb. words per journalist

Indicators of a new register to evoke politics

Indicators of a new register to evoke politics

1. Mention of Opinion Polls [‘sondage’]

Polls as a classic vector of ‘horse-racism’

⇒ Measure the number of articles mentioning ‘polls’

Difficulty: disambiguating multiple meanings

‘Bloomberg rose again in the polls’

‘Having people to wait in line to vote is a poll tax’

‘Heavy attendance at the poll this morning’

Does anyone know how to do this ?

Indicators of a new register to evoke politics

2. Focus on the political field only, or else?

Is the text referring to events purely endogeneous to the political field ? ⇒ Looking at the title

- Endogeneous: Only about the PF
e.g. 'Tensions in the majority group'

- Exogeneous: Mention of external event/actor
e.g. 'Tensions in the majority over the healthcare bill

- Margins: Pol. institutions, electorate, unions…
+ 'Ambiguous', + 'Uncategorizable'

Difficulty: 65,000+ articles is a lot to annotate

Indicators of a new register to evoke politics

3. ‘Off the record’ quotes

‘Off the record’ as a way for journalists to describe backstage aspects of politics, from a source that cannot be named.

- ‘According to a source close to power,...’
- ‘An unnamed official suggested that...’
- ‘People in the X campaign expressed...’
- ‘An anonymous source...’

<u>Difficulty :</u> 65,000+ articles is a lot to annotate
Countless ways of introducting ‘*off the record*’ speech

Method: Supervised Machine Learning
> Train an algorithm to recognize a pattern

Method: Supervised Machine Learning
> Train an algorithm to recognize a pattern

1. Human annotation (**very careful**) [Small N]

Task	**Volume of annotations**
Title classification	3000 titles
Off the record	2100 sequences (out of 5400 articles)

Articles sampled at random, by decade

Method: Supervised Machine Learning
> Train an algorithm to recognize a pattern

1. Human annotation (**very careful**) [Small N]

(2. Active learning and human determination [Selected N])

3. Prediction on a limited (holdout) set and quality assessment

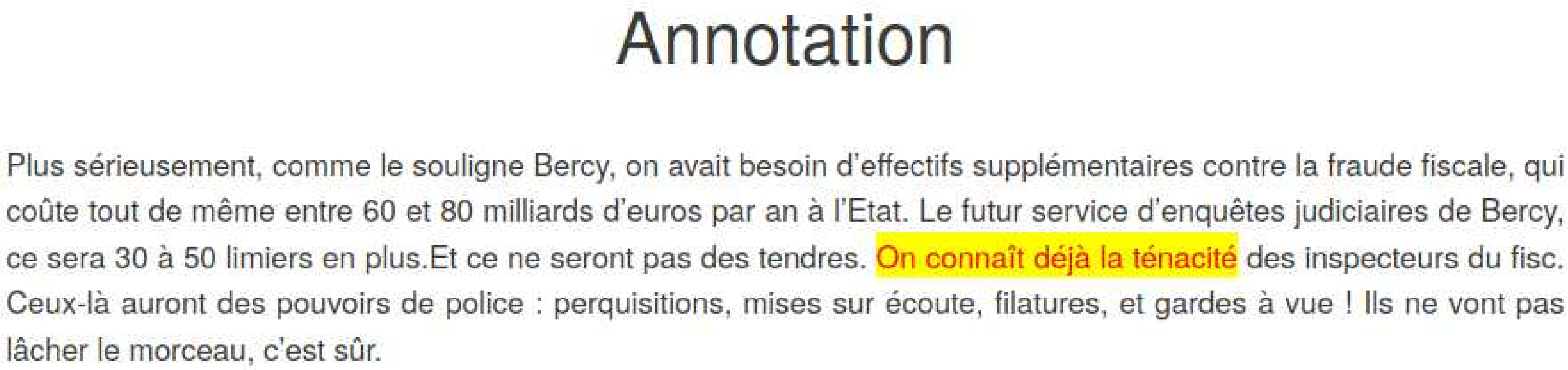
4. Prediction on the whole data set

Method: Supervised Machine Learning
> Train an algorithm to recognize a pattern

1. Human annotation (**very careful**) [Small N]
(2. Active learning and human determination [Selected N])
3. Prediction on a limited set and quality assessment
4. Prediction on the whole data set.

Specifically:
1. Regex for now...
2. Text classifier (CamemBERT + Prediction)
3. Word embedding + Sequence labelling (bi-LSTM + CRF)

Method: Supervised Machine Learning
> Train an algorithm to recognize a pattern

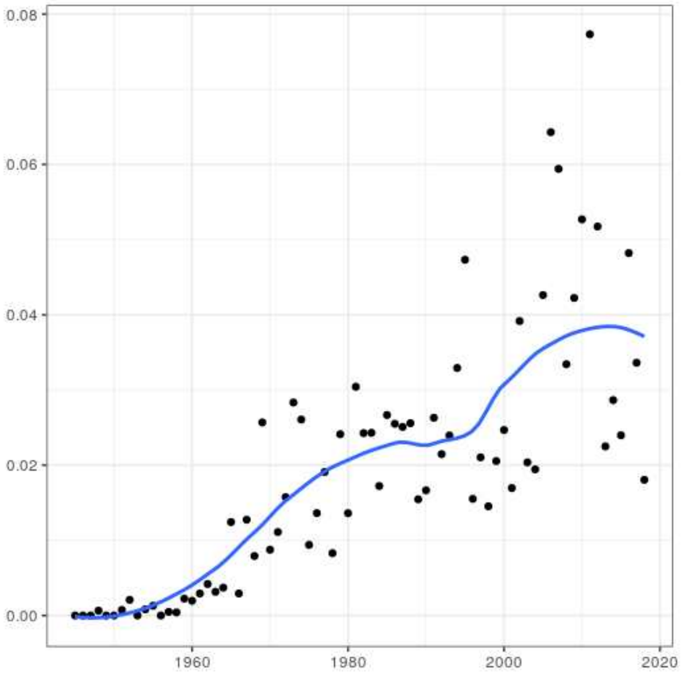


Annotation

Plus sérieusement, comme le souligne Bercy, on avait besoin d'effectifs supplémentaires contre la fraude fiscale, qui coûte tout de même entre 60 et 80 milliards d'euros par an à l'Etat. Le futur service d'enquêtes judiciaires de Bercy, ce sera 30 à 50 limiers en plus.Et ce ne seront pas des tendres. On connaît déjà la ténacité des inspecteurs du fisc. Ceux-là auront des pouvoirs de police : perquisitions, mises sur écoute, filatures, et gardes à vue ! Ils ne vont pas lâcher le morceau, c'est sûr.

Results

1. Opinion polls



Mention of opinion polls

Mean (per article, relative to the nb of words, expressed as %)

Results

2. Text Classification

Goal: Determine whether the title only mentions the political field

Tensions within the Democratic Party (purely ENDOGENEOUS)

vs.

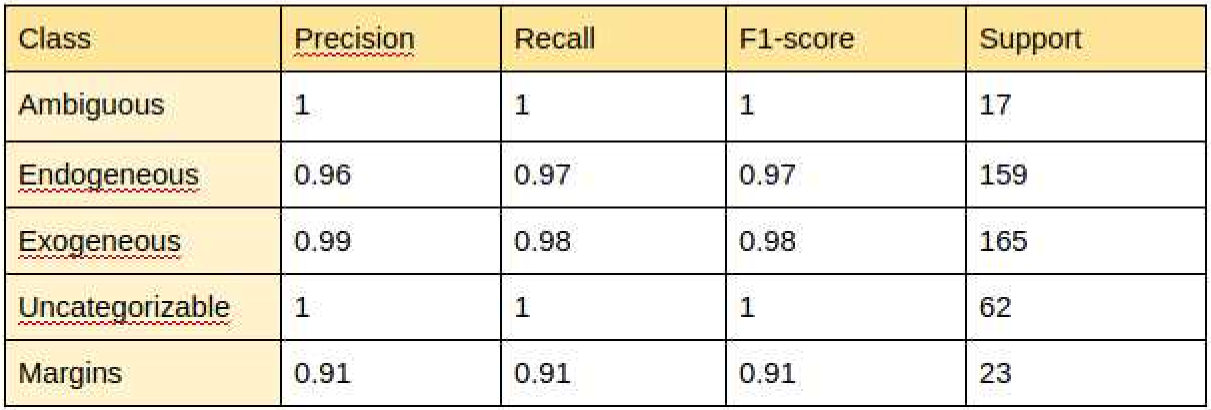
Medicare for all: Tensions within the Democratic Party (EXO)

Results

2. Text Classification

Goal: Determine whether the title only mentions the political field

Overall F1-score: 0.97 (max = 1)

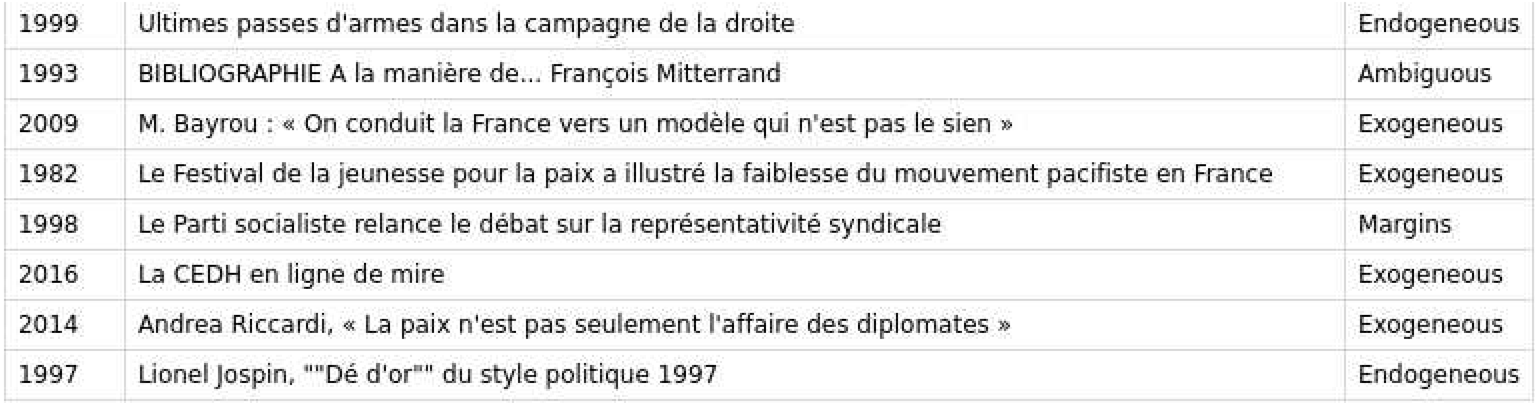


Class	Precision	Recall	F1-score	Support
Ambiguous	1	1	1	17
Endogeneous	0.96	0.97	0.97	159
Exogeneous	0.99	0.98	0.98	165
Uncategorizable	1	1	1	62
Margins	0.91	0.91	0.91	23

Results

2. Text Classification

Goal: Determine whether the title only mentions the political field.

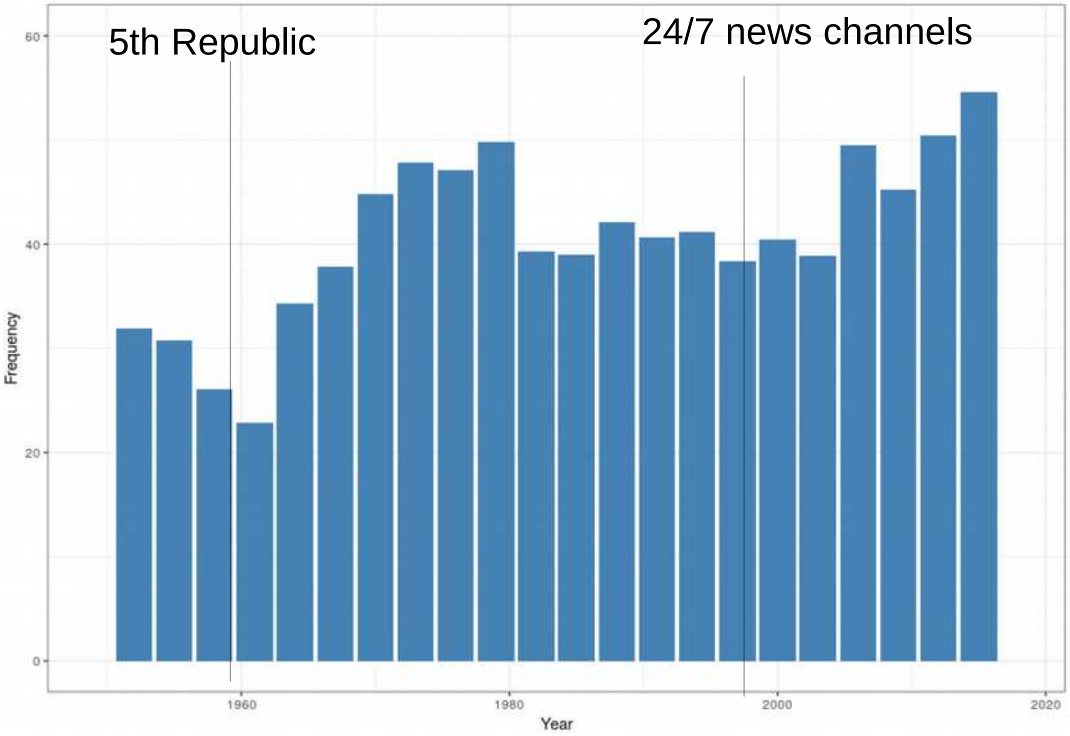


1999	Ultimes passes d'armes dans la campagne de la droite	Endogeneous
1993	BIBLIOGRAPHIE A la manière de... François Mitterrand	Ambiguous
2009	M. Bayrou : « On conduit la France vers un modèle qui n'est pas le sien »	Exogeneous
1982	Le Festival de la jeunesse pour la paix a illustré la faiblesse du mouvement pacifiste en France	Exogeneous
1998	Le Parti socialiste relance le débat sur la représentativité syndicale	Margins
2016	La CEDH en ligne de mire	Exogeneous
2014	Andrea Riccardi, « La paix n'est pas seulement l'affaire des diplomates »	Exogeneous
1997	Lionel Jospin, ""Dé d'or"" du style politique 1997	Endogeneous

Results

2. Text Classification

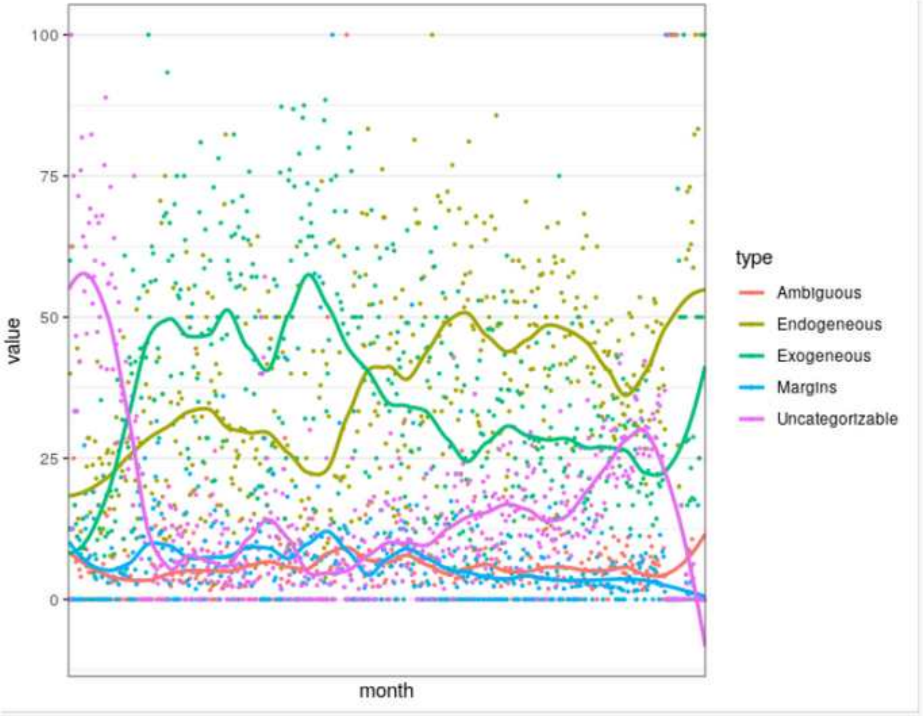
Frequency of ‘endogeneous’ label



Results

2. Text Classification

Goal: Determine whether the title only mentions the political field.



Results

3. ‘Off the Record’

Goal: determine sequences introducing ‘off the record’ quotes

Results

3. ‘Off the Record’

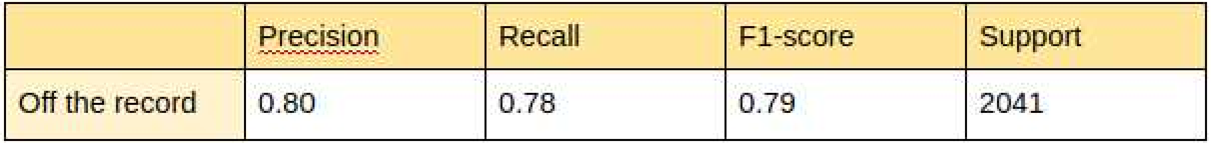
Goal: determine sequences introducing ‘off the record’ quotes

Chez [OFF] les [OFF] proches [OFF] de [OFF] Martine [OFF] Aubry [OFF] , [OFF] on [OFF] estime [OFF] que [OFF] la balle est dans le camp du chef de l'État.

Results

3. ‘Off the Record’

Goal: determine sequences introducing ‘off the record’ quotes



	Precision	Recall	F1-score	Support
Off the record	0.80	0.78	0.79	2041

Results

3. ‘Off the Record’

```
SELON nos [OFF]  informations [OFF]  obtenues [OFF]  dans [OFF]  l
[OFF] ' [OFF] entourage [OFF]  du [OFF]  ministre [OFF]  de [OFF]  l
[OFF] ' [OFF] économie [OFF]  et [OFF]  des [OFF]  finances [OFF]  ,
c'est le vendredi 3 septembre - et non le jeudi 2, comme cela
l'avait été dit jusqu'à présent - que Nicolas Sarkozy annoncera sa
décision sur son avenir au sein de l'Union pour un mouvement
populaire (UMP). Il devrait confirmer sa candidature, que [OFF]  l
[OFF] ' [OFF] un [OFF]  de [OFF]  ses [OFF]  proches [OFF]  présente
[OFF]  comme [OFF]  « [OFF]  plus que probable », à la présidence de
l'UMP, vacante depuis la démission d'Alain Juppé. Toujours selon
[OFF]  les [OFF]  mêmes [OFF]  sources [OFF]  , cette annonce devrait
prendre la forme d'un simple communiqué.
```

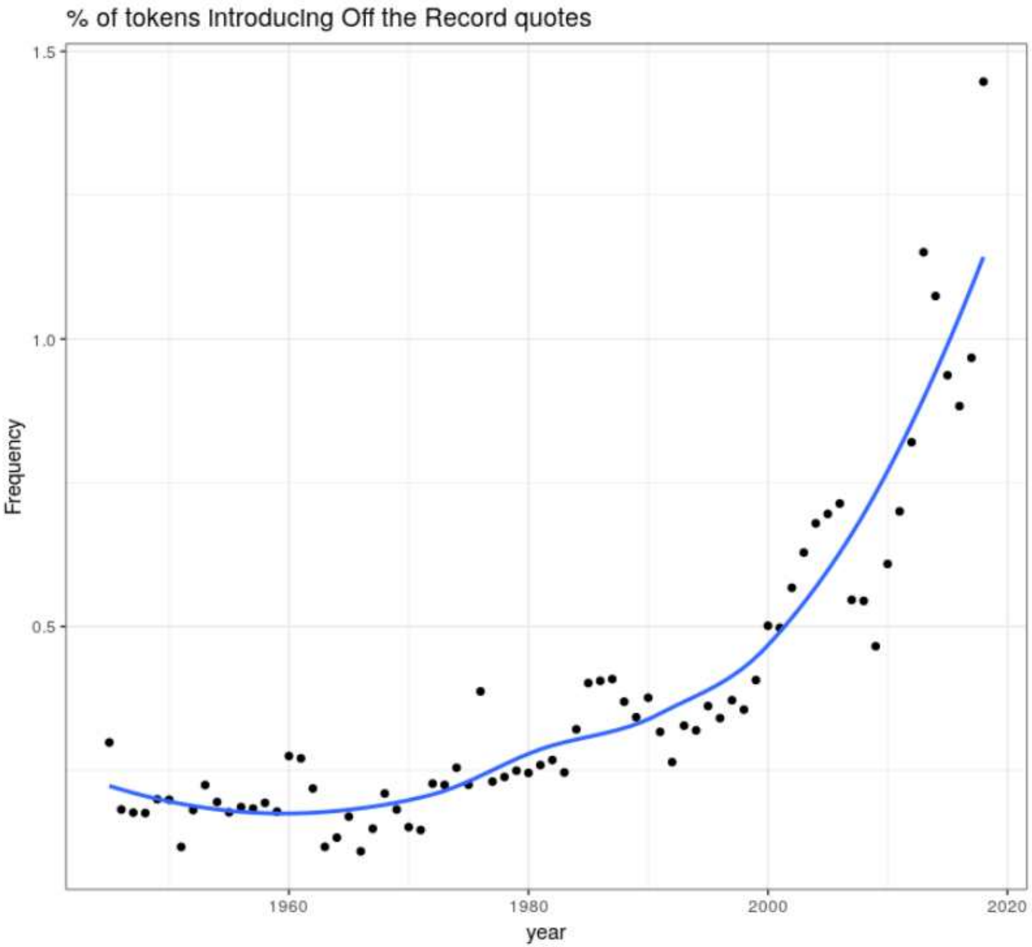
3 sequences of ‘off the record’

32 tokens (out of 101)

2 misclassified tokens (1FP, 1FN)

Results

3. ‘Off the Record’



Conclusion

- Another way to use ML: ML to produce data

  - Transcribing interviews, speeches

  - Imputing missing values

  - Finding individuals in pictures, videos

  - Measuring incoming flows of goods/persons through satellite videos

  - Detect writing styles (Gothic handwriting *Kurrentschrift*, any type of handwriting) and convert it.

  - ...

Conclusion

- Machine Learning and Quantitative Methods in the Social Sciences
  - Another form of quantification: check our results, limit monoculture
  - A radically inductive approach
  ⇒ towards quantitative description ?