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”

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”

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
What can (Social) Sciences do with Machine Learning?

And known criticisms
- No clear mathematical demonstration
- Uncertain optimality
- Lack of interpretability
- Prediction

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)
Case: History of Political Journalism

> Same pattern exists in other countries

"Winners and Losers of the Democratic Debate"

Case: History of Political Journalism

> Same pattern exists in other countries
Case: History of Political Journalism

> Same pattern exists in other countries
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.)


Leads to

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

20
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

<table>
<thead>
<tr>
<th>Dates</th>
<th>Articles</th>
<th>Number of words</th>
<th>Number of pol. journalists</th>
</tr>
</thead>
<tbody>
<tr>
<td>1946-2018</td>
<td>65,100</td>
<td>~ 3,9 M</td>
<td>849</td>
</tr>
</tbody>
</table>
Corpus: All articles by political journalists in *Le Monde* since 1946

**Le Monde**

- 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…’

Difficulty: 65,000+ articles is a lot to annotate
Countless ways of introducing ‘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]

<table>
<thead>
<tr>
<th>Task</th>
<th>Volume of annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title classification</td>
<td>3000 titles</td>
</tr>
<tr>
<td>Off the record</td>
<td>2100 sequences (out of 5400 articles)</td>
</tr>
</tbody>
</table>

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

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’État. 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)

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Endogeneous</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
<td>159</td>
</tr>
<tr>
<td>Exogeneous</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>165</td>
</tr>
<tr>
<td>Uncategorizable</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>Margins</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>23</td>
</tr>
</tbody>
</table>
## Results

### 2. Text Classification

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

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

2. Text Classification

Frequency of ‘endogeneous’ label

![Bar chart showing frequency over years for 5th Republic and 24/7 news channels.](chart.png)
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

Results

3. ‘Off the Record’
Goal: determine sequences introducing ‘off the record’ quotes

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off the record</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79</td>
<td>2041</td>
</tr>
</tbody>
</table>
3. ‘Off the Record’

SELON nos informations obtenues dans l'entourage du ministre de l'économie et des finances, 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 un de ses proches présente comme « plus que probable », à la présidence de l'UMP, vacante depuis la démission d'Alain Juppé. Toujours selon les mêmes sources, 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?