Automating text annotation when doing interpretative social science: what state of the art supervised machine learning can do?

Andrey Indukaev (University of Helsinki)

Presentation at the **Research Seminar in Language Technology**

5.11.2020 Helsinki

Presentation outline

Not a new technique, but more reflection on applying existing techniques to a new kind of problems

Theory

- Introducing the interpretative component in social science and humanities research
- Interpretative research and machine learning: overview of existing practices
- Supervised machine learning and interpretative research

Case

- Justification Theory
- Annotation setup
- Interpreting agreement
- Interpreting classifier output

Interpretative research

It such a fundamental category in social sciences and humanities epistemology that authors do not often provide definition

Interpretation is “the action of explaining the meaning of something” (Oxford English Dictionary)

- Many debated features of interpretative research practice:

Impossibility of: generalization and quantification (Williams 2000), systematic formalization of research practice, coding (Biernacki 2012).

- Essential feature of the interpretative component of research:

To understand their object of study (predict it behavior, manipulate it) such research require understanding what meaning the studied realities have for those who are essential to study object

Interpretative research

The centrality of meaning to research practice have logical consequences:

- Sensitivity to context is essential
- Open ended, bottom-up character: the analytical categories and even the object definition have an emergent component

What could that mean for the disciplines?

- Fundamental divide: interpretative component of research is what is true research in social science and humanities
- More pragmatic vision: most research activities in social science and humanities either have, either should have this component

Interpretative research an ML in text analysis

What ML can bring to interpretative research working with texts (but not necessarily vise-versa).

Discussion superimposed on existing methodological debates. How to improve interpretative analysis?

Problems to solve (Pääkkönen and Ylikoski 2020):

- Lack of transparency: trust me, I am a humanist.
- Underdetermination: why this interpretation and not another?
- Limited scale

Coding of counting?

Coding is bad!!(?)

- Biernacki (2012): following a formal coding procedures does not solve any issue but kills the open-ended character of interpretative analyses, forces arbitrary interpretation.

Is counting a solution?

- When it comes to formal analyses, we might say that bad sociologists code, and good sociologists count. The reason is that the former disguises the interpretation and moves it backstage, while **the latter delays the interpretation**, and then presents the reader with the same data on which to make an interpretation that the researcher herself uses. (Lee and Martin 2015, 24)

Unsupervised ML is the solution?

Could this "good sociologist count" logic be even reinforced by the use unsupervised ML?

- Known model applied to known data as a starting point for interpretation, not and individual reading (or coding) of a text, idiosyncratic and prone to cherry-picking. (Pääkkönen and Ylikoski 2020)

My estimate is that in SSH unsupervised ML is way more popular

It is working, but not perfectly:

- Quite naïve vision of counting as interpretation-free activity by Lee and Martin (2015)
- Limited evidential role of the models in research, limited potential for such a role (analyzed for topic modeling by Pääkkönen and Ylikoski 2020)

Basically, the main problem solved is that of the scale (that might be ok if we are not much concerned by the methodological weaknesses, and we do not have to be)

Supervised ML for scaling-up interpretative research

Supervised ML might be used for scaling up research in SHS without compromising the interpretative component if:

- The creation of training data is reflexive, iterative process where the coding schema and the underlying theory are open to be revised. Gradual move from close reading to large-scale annotation.
- The classification results are not immediately seen as more or less certain application of coding schema to large-volume data, but are triangulated with other techniques requiring interpretation (ex. selective close reading of labelled data)

Supervised ML and transparency and underdetermination issues

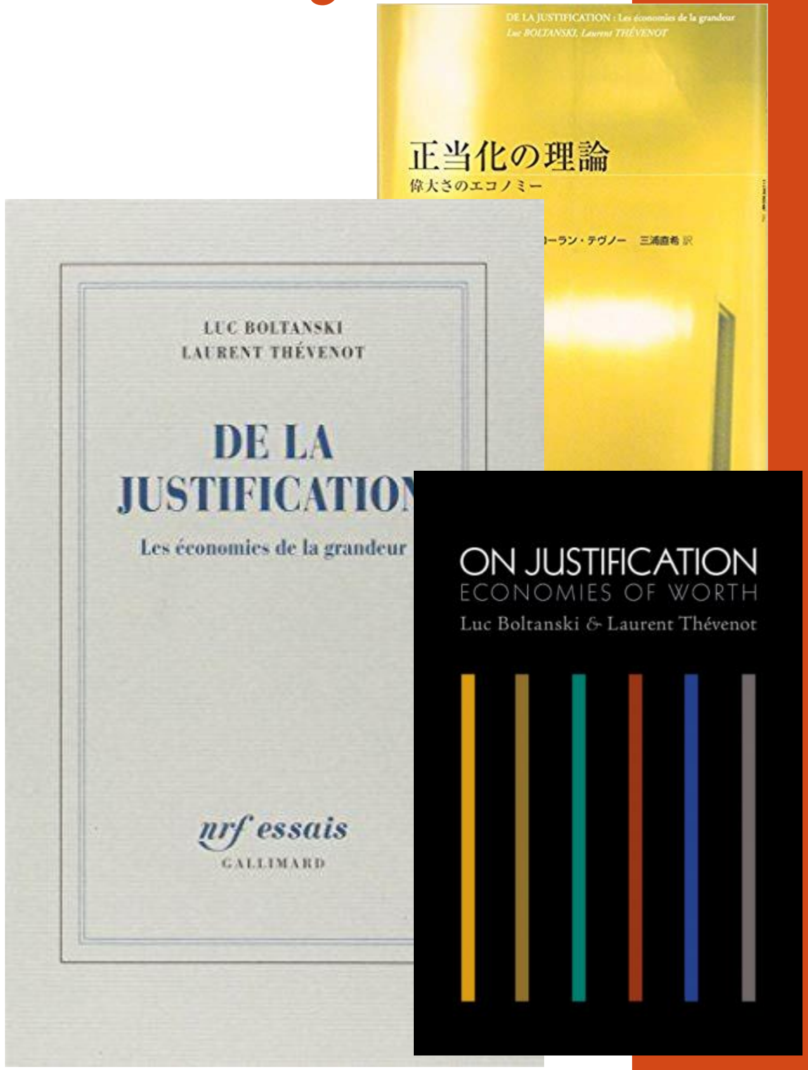
Supervised ML use in research practice might be used for solving more problems than that of the scale:

- Coding (annotation) practices have high standards
- Agreement measures: not only as a measure of quality of annotation process, but as a source of information for the interpretative and open-ended adjustment of the coding schema.
- Classification results itself could be as well contain provide material for interpretative analysis: class-level accuracy measures; features contributing the most to classification.

Automating justification analysis

Luc Boltanski and Laurent Thévenot influential theory

The power of an argument in almost any debate is conditioned by its capacity to mobilize one of a limited set of justification principles – orders of worth – that reflect collectively shared representations of the common good in a given society



Orders of worth

6 detected (in French context):

- Inspired
- Domestic
- Fame
- Civic
- Market
- Industrial

Further addition:

- Ecological

Project-specific additions

Subcategories for Civic and Domestic justifications

Other ways to claim that an issue is important:

- Individual Interest (Eranti 2018)

NOT IN MY BACKYARD!!!

- Familiarity (Ylä-Anttila 2017)

Data

Petitions and “reasons to sing” from Change.org (hundreds of thousands)

Petition to Ban Petitions

Reasons for signing

R C · 8 years ago

Petitions are dumb!

0 · Report

R C started this petition to United States Supreme Court and 1 other

Remember that time that one petition saved the world?

Yea, me neither. Which is why we at the Lottery Party are demanding that the United States Supreme Court take measures to ban all petitions, regardless of content or aim. No more vegetarian lesbians shoving clipboards in our faces while we're at the coffee shoppe- we don't care how hot she is. No more electronic petitions clogging up our email spam filters or online

Petition Closed

This petition had 5 supporters

Petition to Ban Petitions

Share on Facebook

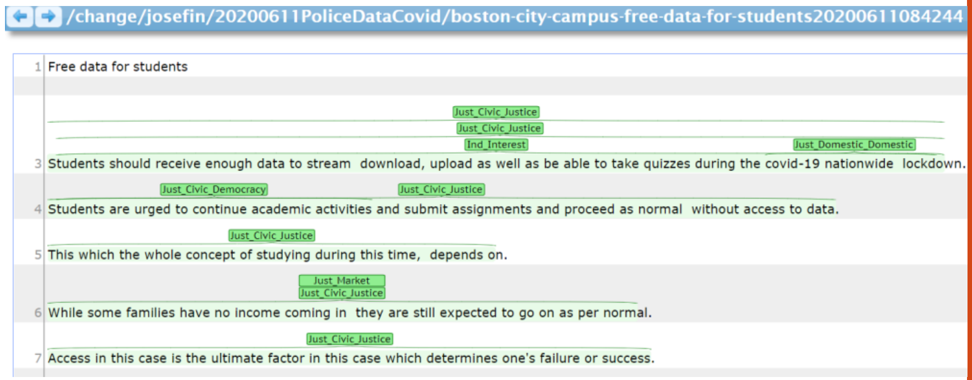
Send a Facebook message

Send an email to friends

Tweet to your followers

Copy link

Annotation setup



- Brat annotation tool
- Discussion board (GitLab)
- Dashboard with some agreement information (not used much)
- Bi-weekly meetings (Zoom☹) 1-2h
- Codebook on 7 pages
- Readings and discussions before coding
- Reimbursed Master students as annotators

	Project	Phrases	Phrases Coded	A all
	All			
1	08042020Wednesday	96	22	0.365
2	20020424FridayCovid	172	43	0.308
3	20020429covidmigrants	289	48	0.454
4	20020504covidprivacy	427	82	0.375

4000 sentences with multi-label annotations

Low agreement of interpretations

- Low agreement, Krippendorff's alpha (with MASI distance) fluctuating around 0.4, 03 if one exclude NONE
- Lots of discussions and codebook edits with decreasing improvement
- Quite possible that there is a limitation due to data diversity and theory complexity

Interpretation of agreement?

- Some subcategories make things complicated, some do not
- Some new categories seem to work well (Individual Interest) some not as well (Familiarity)

Familiarity	0.23
Ind_Interest	0.33
Just_Civic_Democracy	0.22
Just_Civic_Justice	0.25
Just_Civic_Legal	0.30
Just_Domestic_Domestic	0.21
Just_Domestic_Heritage	0.19
Just_Ecological	0.45
Just_Fame	0.22
Just_Industrial	0.29
Just_Inspired	0.25
Just_Market	0.30
NONE	0.30

Familiarity	0.23
Ind_Interest	0.33
Just_Civic	0.28
Just_Domestic	0.25
Just_Ecological	0.45
Just_Fame	0.22
Just_Industrial	0.29
Just_Inspired	0.25
Just_Market	0.30
NONE	0.30

Classifier

- The one used in (Öhman et al. 2020) 🙏
- Really provisory results

BERT Macro F1: 031

A lot of confusion in the model, but NONE and Civic Justice being classified quite well.

Linear SVC one-vs-rest Macro F1: 0.616

Accuracy variation from 0.68 to 0.996 (largely explained by class size)

References

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry*. New York: Palgrave Macmillan US. https://doi.org/10.1057/9781137007285.

Eranti, Veikko. 2018. "Engagements, Grammars, and the Public: From the Liberal Grammar to Individual Interests." *European Journal of Cultural and Political Sociology* 5 (1–2): 42–65. https://doi.org/10.1080/23254823.2018.1442733.

Lee, Monica, and John Levi Martin. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3 (1): 1–33. https://doi.org/10.1057/ajcs.2014.13.

Öhman, Emily, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. "XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection." In *The 28th International Conference on Computational Linguistics: COLING 2020*. https://researchportal.helsinki.fi/en/publications/xed-a-multilingual-dataset-for-sentiment-analysis-and-emotion-det.

Pääkkönen, Juho, and Petri Ylikoski. 2020. "Humanistic Interpretation and Machine Learning." *Synthese*, September. https://doi.org/10.1007/s11229-020-02806-w.

Williams, Malcolm. 2000. "Interpretivism and Generalisation." *Sociology* 34 (2): 209–24. https://doi.org/10.1177/S0038038500000146.

Ylä-Anttila, Tuukka. 2017. "Familiarity as a Tool of Populism: Political Appropriation of Shared Experiences and the Case of *Suvivirsi*." *Acta Sociologica* 60 (4): 342–57. https://doi.org/10.1177/0001699316679490.