Identification and reduction of political information in algorithmic representation on recommender systems.

Tim Faverjon

SciencesPo
MÉDIALAB

LEARNING PLANET INSTITUTE

State of the art

- Recommender system predict new interactions from past interactions

Research suggests :

- Interactions are related to political attitudes
- Recommendations can have political bias

Regulation (DSA) :

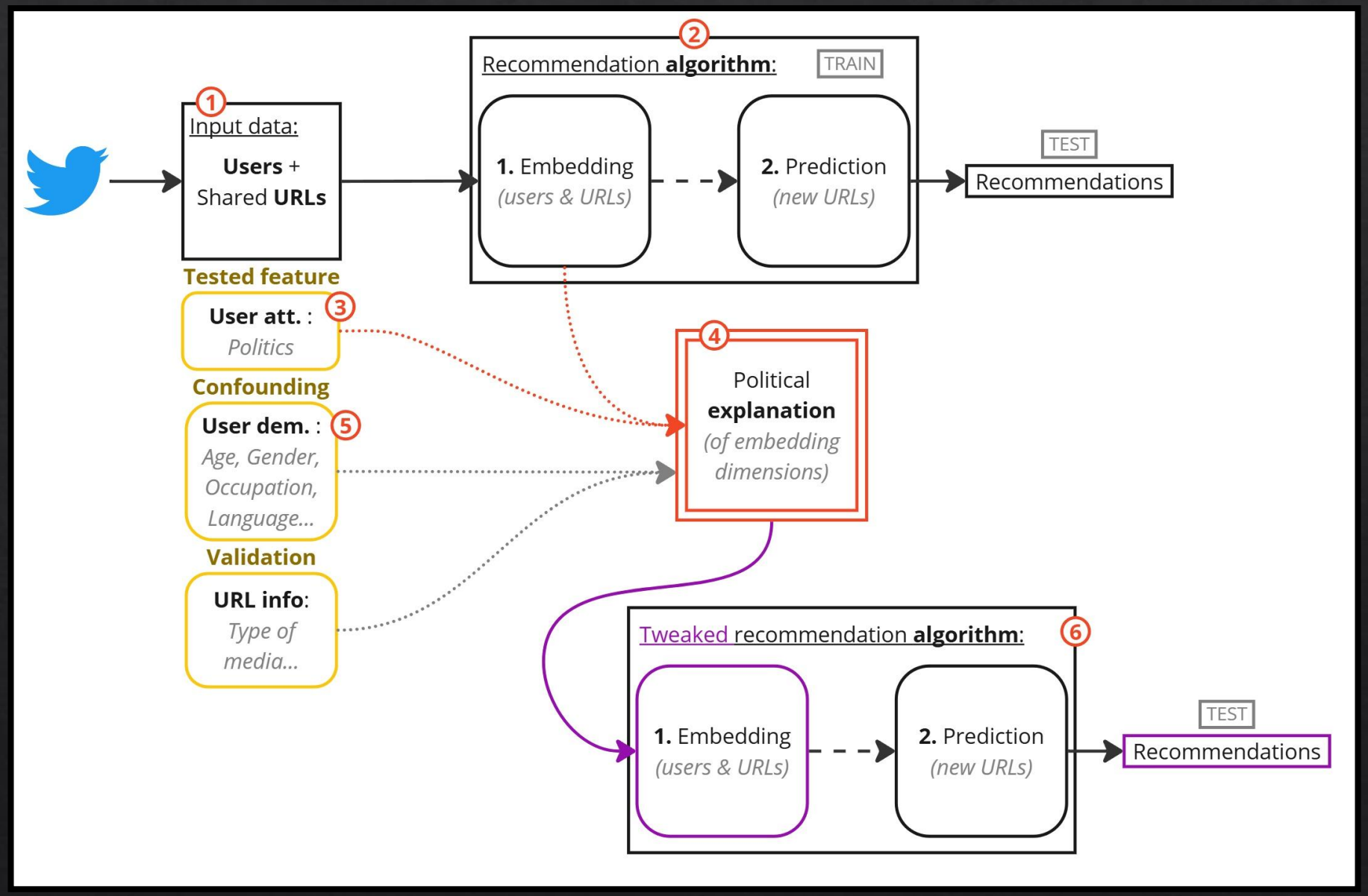
- Ask to reveal which features are important for recommendation
- Ask to not discriminate users by political opinion

State of the art

- Recommendation systems usually contain embedding layers

- We can use hidden semantic explanation :
  - Identify dimensions learning politic
  - Observe the result of embedding manipulation

Method


2
Recommendation **algorithm**:
TRAIN
1
Input data:
Users +
Shared **URLs**
1. Embedding
(users & URLs)
2. Prediction
(new URLs)
TEST
Recommendations
Tested feature
User att. :
Politics
3
4
Political
explanation
(of embedding dimensions)
Confounding
User dem. :
Age, Gender, Occupation, Language...
5
Validation
URL info:
Type of media...
Tweaked recommendation **algorithm**:
6
1. Embedding
(users & URLs)
2. Prediction
(new URLs)
TEST
Recommendations

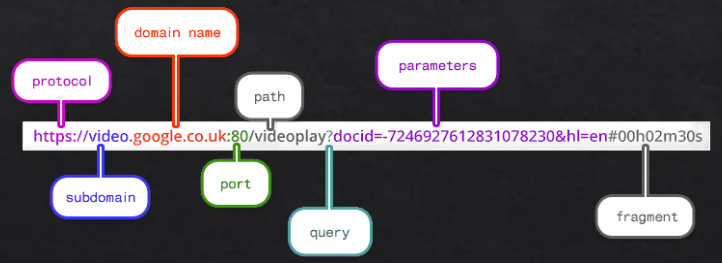
1. Data collection

Users : ***Twitter users***

- Active[1] users
- following French MPs[2]
- $\approx 10^4$: randomly selected

Items : ***URLs***

- Shared on Twitter
- $\approx 10^5$: shared by at least 10 users



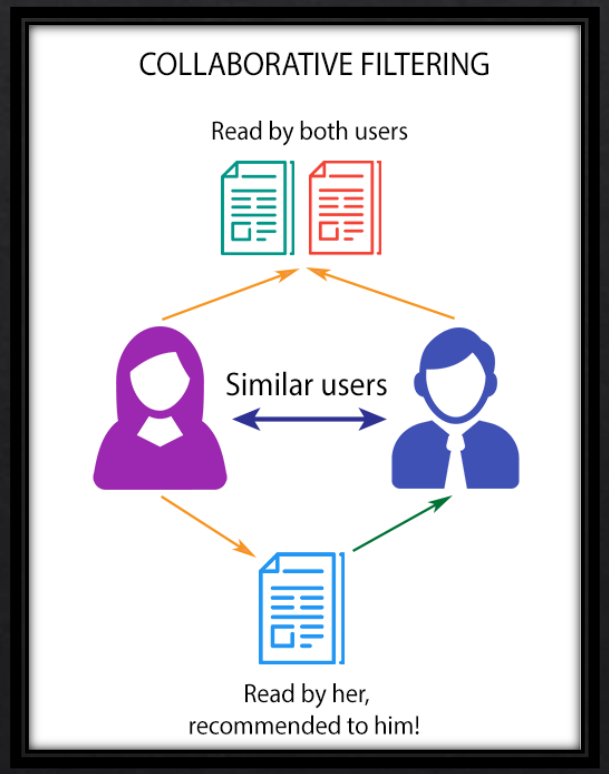
[1] : Sharing links, following at least 3 MPs and 25 users and with at least 25 followers (to avoid bots)
[2] : Members of Parliament

2. Recommendation algorithm

2.1. Representation Space :

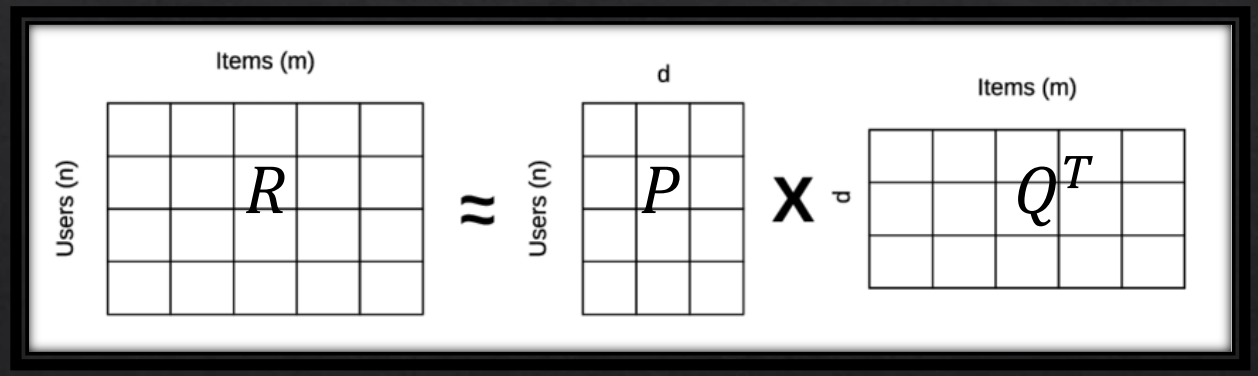
From : interaction data - *To* : vector space

The idea :



The method :

Non-negative Matrix Factorization



The Loss-function :

$$\mathcal{L}(P,Q) = \frac{1}{2} \cdot \|R - PQ^T\|_2^2$$

$$+ Regularisation \ldots$$

2. Recommendation algorithm

2.2. Prediction :

From : vector space - *To* : prediction of new Items for each User

The predition :

- Predicted rating for user u and item i is a **scalar product** : $\widehat{r_{ui}} = P_u \cdot Q_i$
- We recommend the best rated new items

The accuracy metric : Hits@10

The results : Hits@10 = 0.34

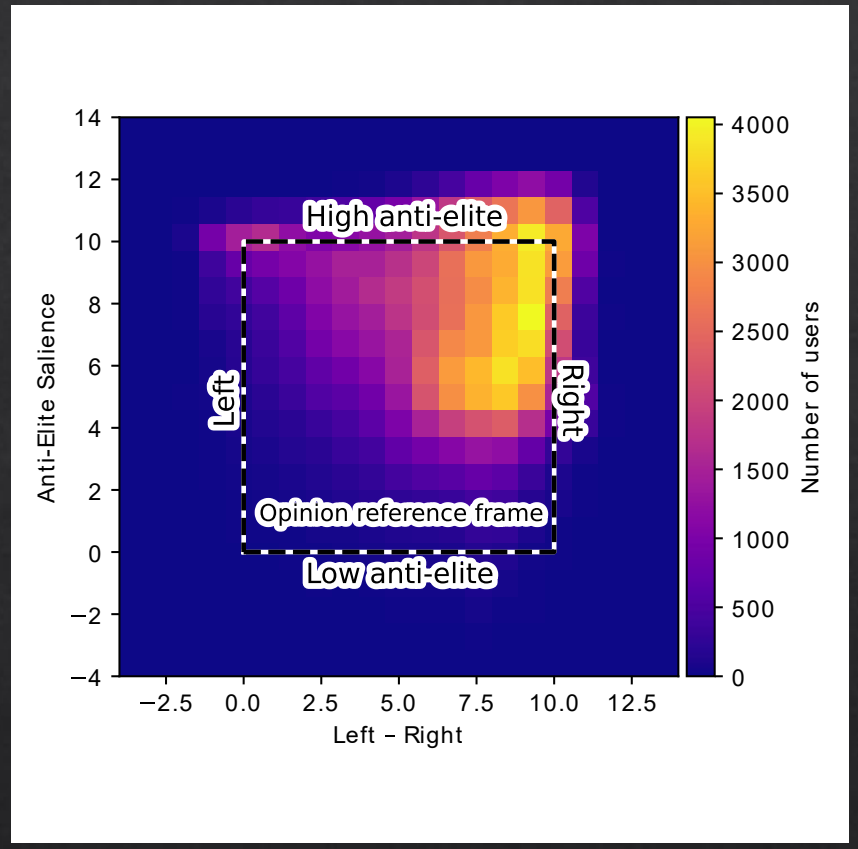
3. Users political attitudes

- Spatialisation **following network of French MPs**
- Scaling based on CHES survey

France - 2 most explicative dimensions

Left – Right : classical left-right

Anti-Elite Salience : negative attitude toward élites and insitutions



5. Explanation design choices

- Latent embedding explanation
  - Generalization (model agnostic)
  - Statistical significance
  - Safety engineering
- Post training
  - Governance
- User based
  - Politicized content
- Global
  - Systemic phenomena evaluation
- Multi-indicators
  - Account for specific learning

Explanation

Global indicator :

- **Explanation Power** : average sensibility and bias

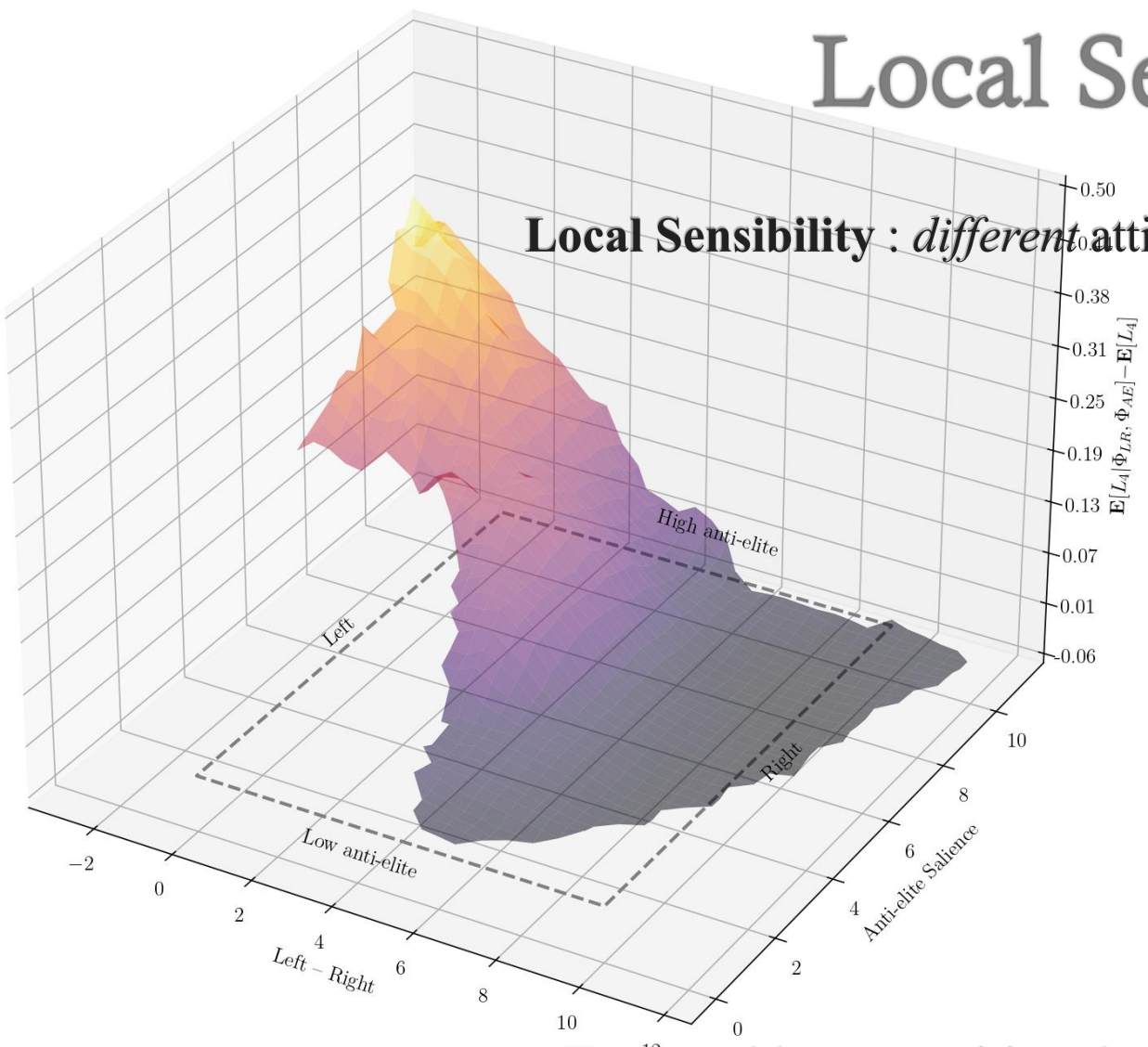
Local indicators :

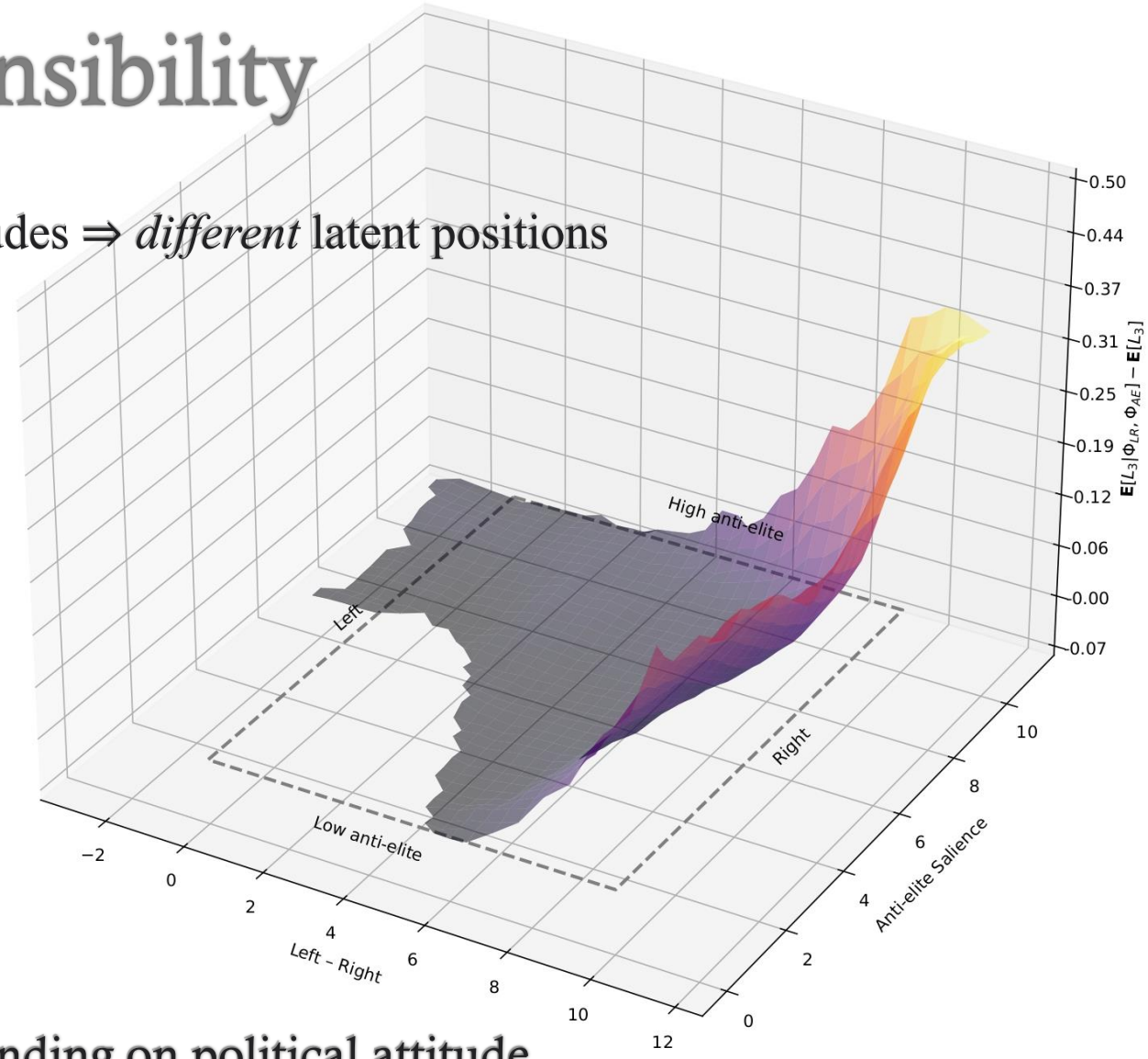
- **Local Sensibility** : *different* attitudes ⇒ *different* latent positions
- **Local Dispersion** : *similar* attitudes ⇒ *similar* latent positions

- **Latent Bias** : *different* latent positions ⇒ *different* attitudes
- **Latent Diversity** : *similar* latent positions ⇒ *similar* attitudes

Local Sensibility

Local Sensibility : *different* attitudes ⇒ *different* latent positions

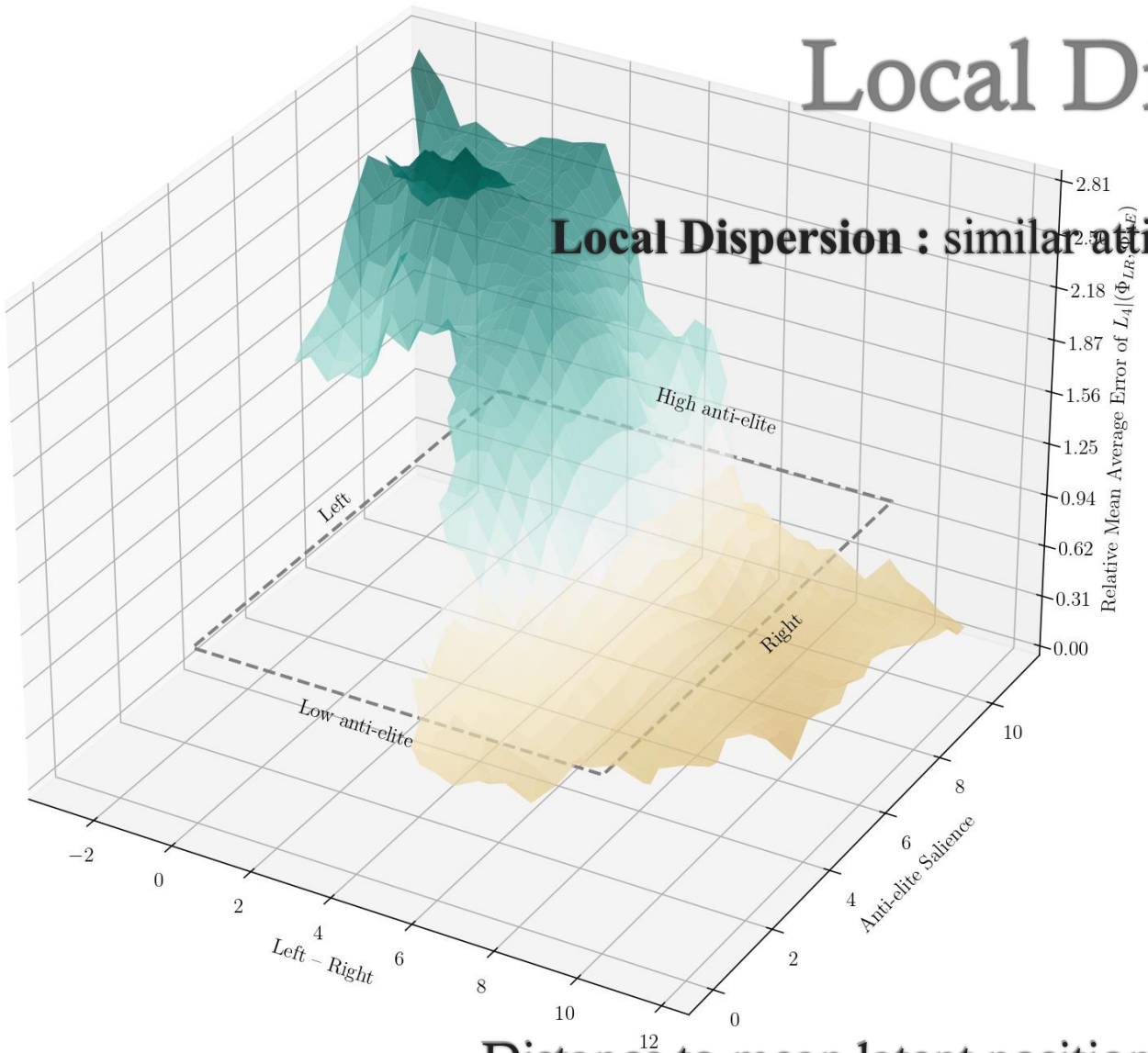


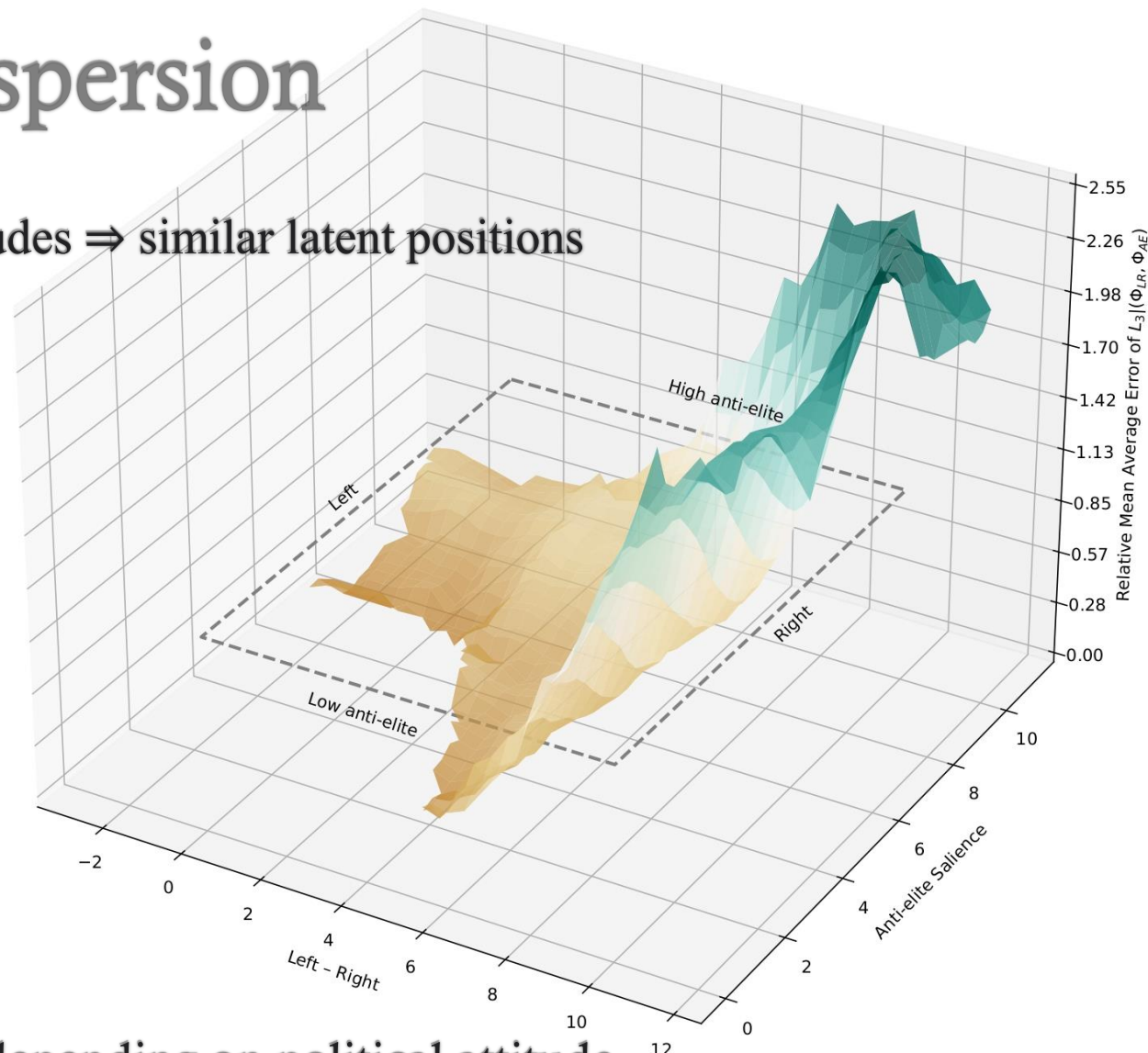


Expected latent position depending on political attitude

Local Dispersion

Local Dispersion : similar attitudes ⇒ similar latent positions



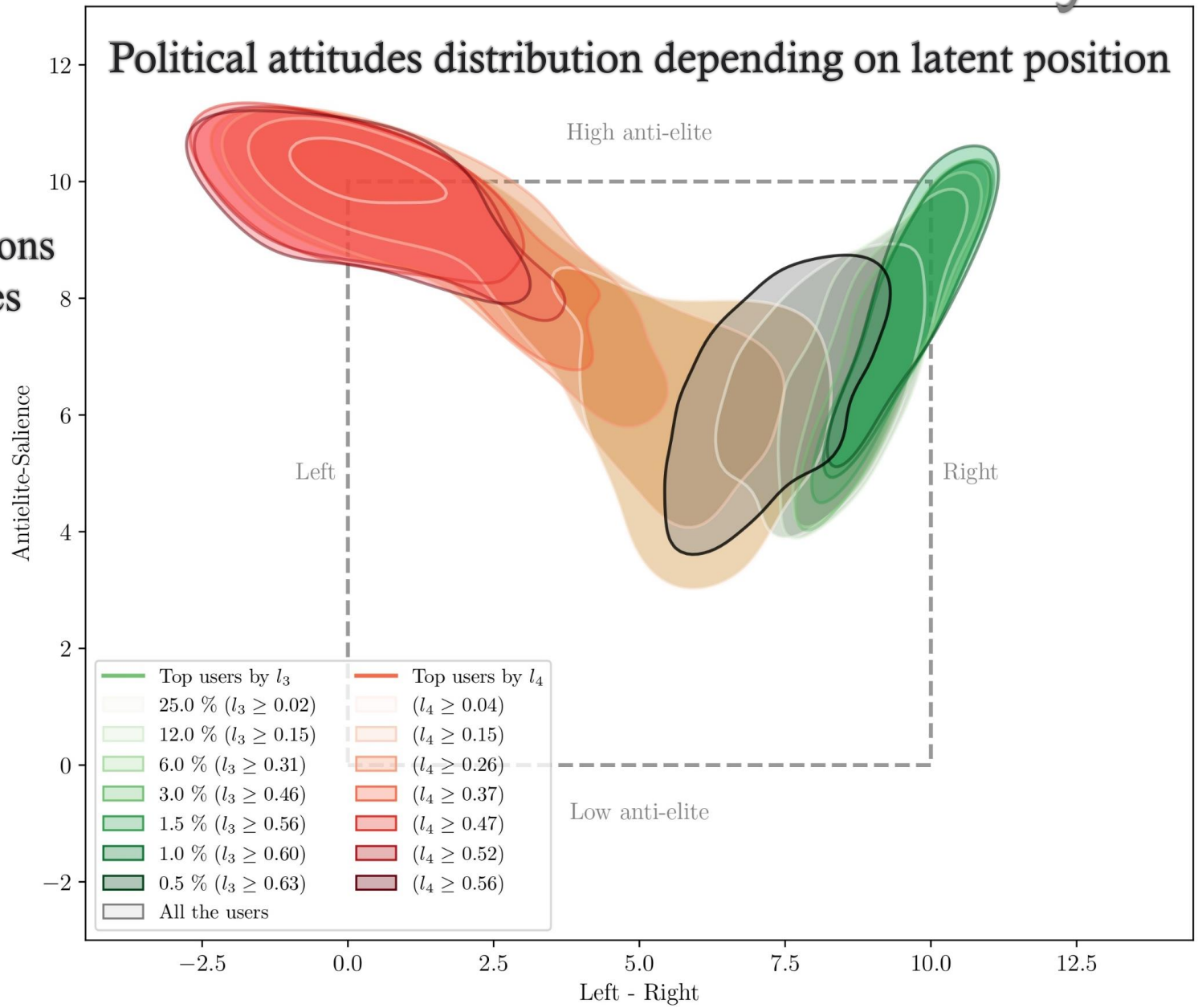


Distance to mean latent position depending on political attitude

Latent Bias and Diversity

Latent bias :

different latent positions
⇒ different attitudes

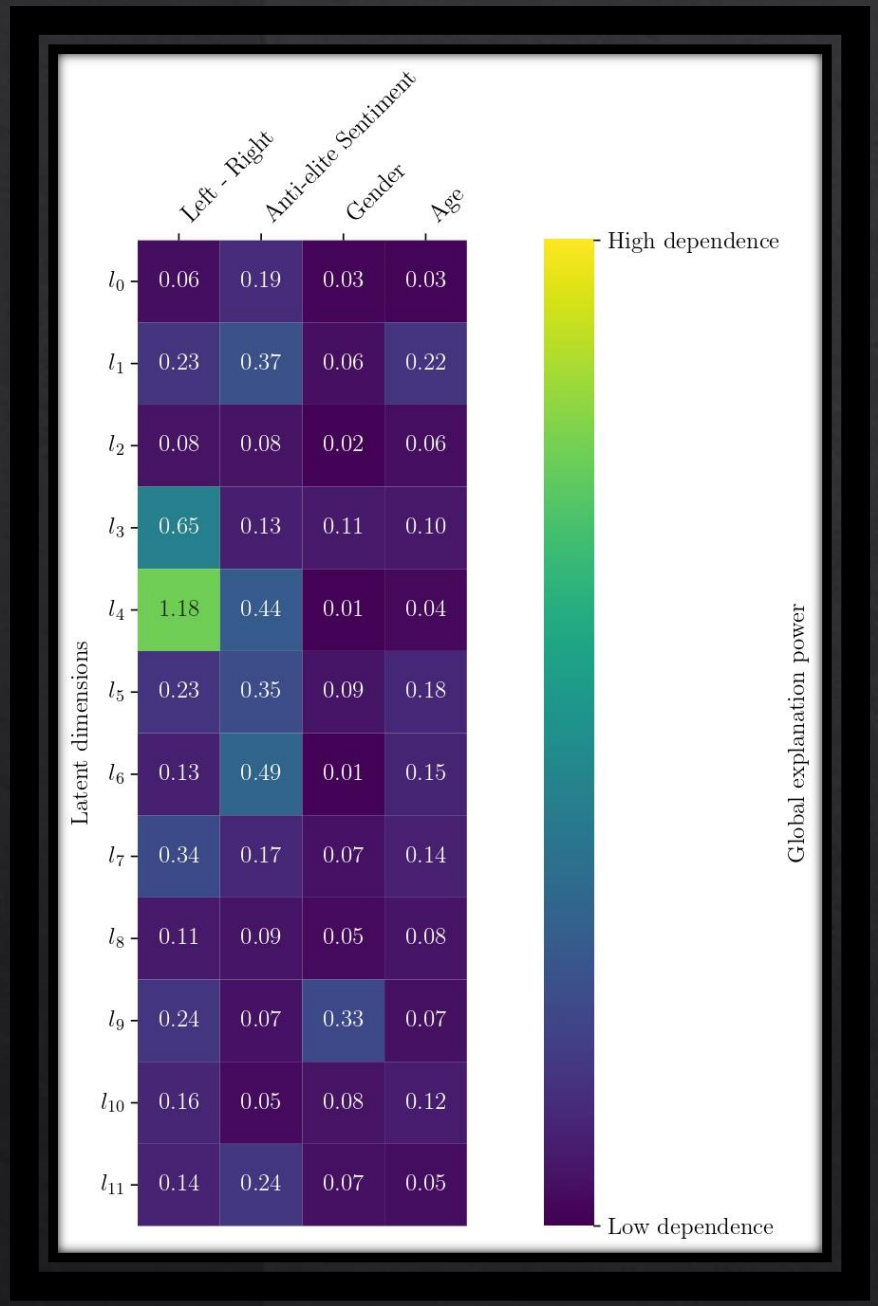


Latent diversity :

similar latent positions
⇒ similar attitudes

Explanation Power

Explanation Power : average sensibility and bias



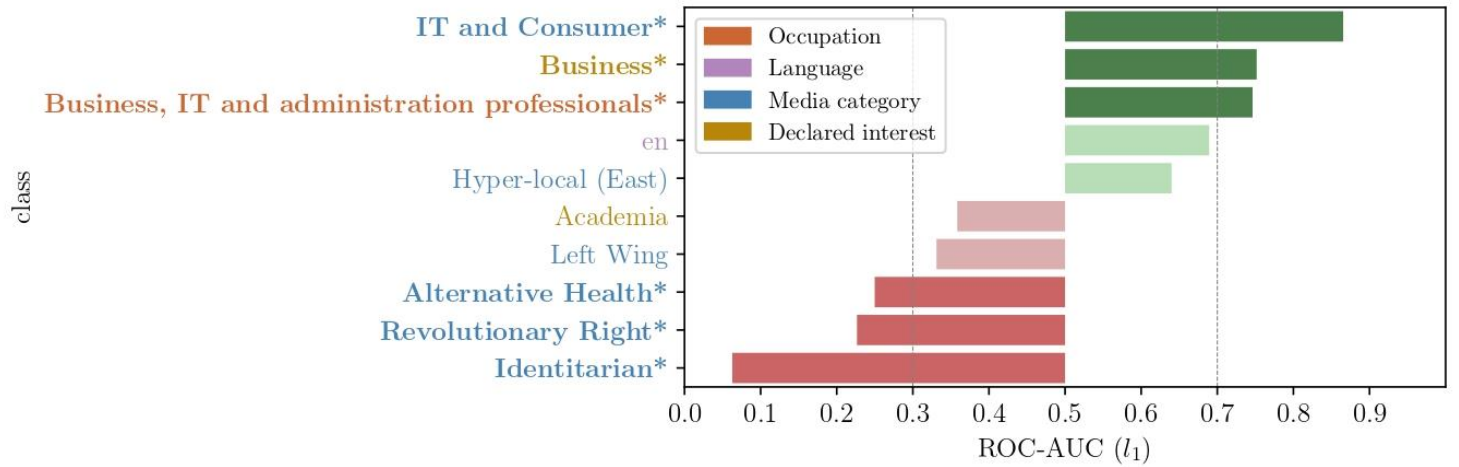
5. Confounding factors

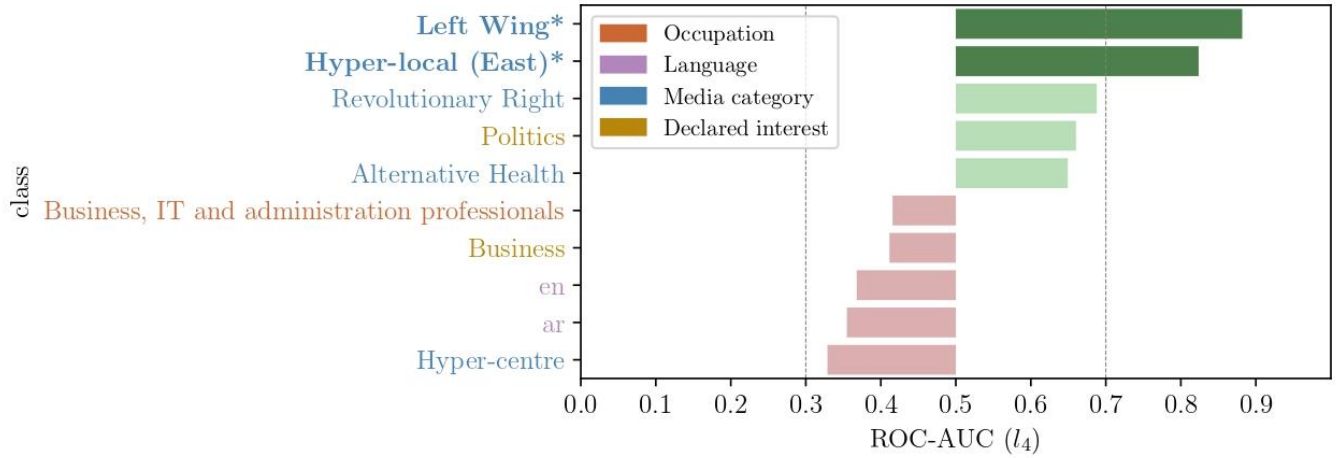
Confounding factors considered :

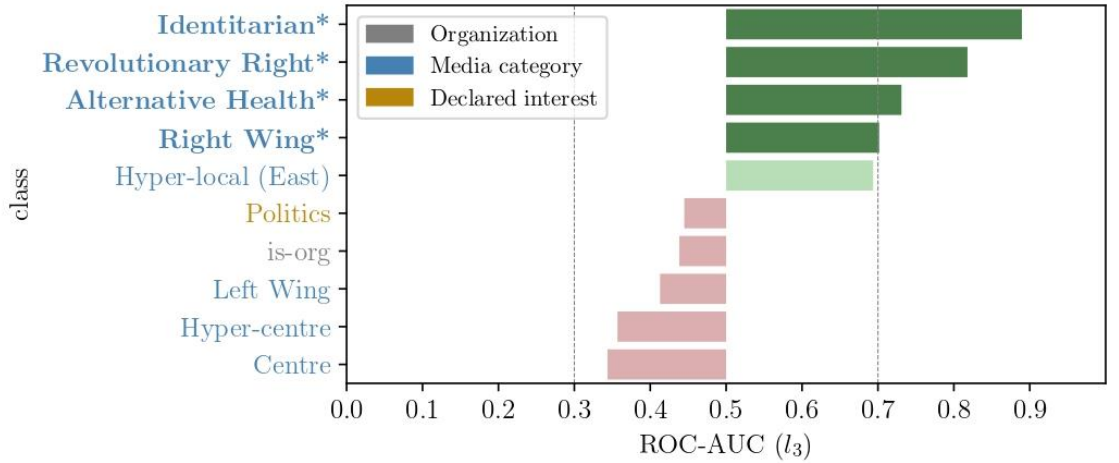
- Age
- Gender
- Organisation
- Language
- Occupation
- Declared interest

Validation :

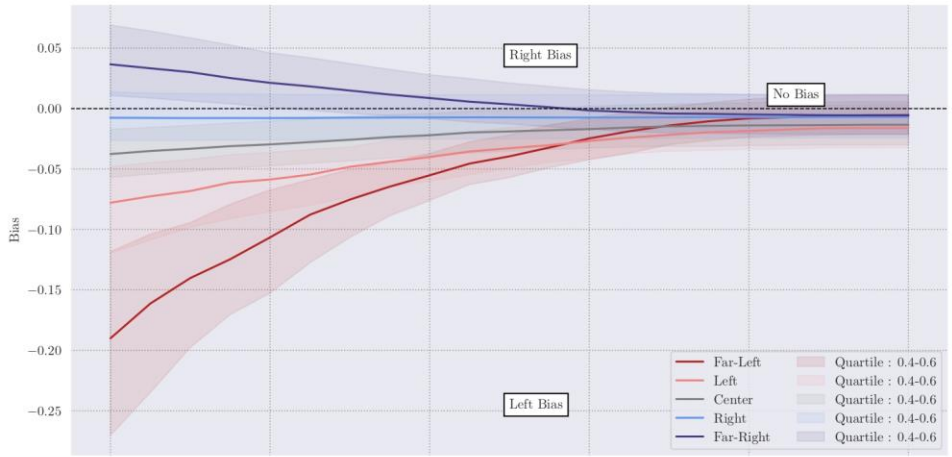
- Media category

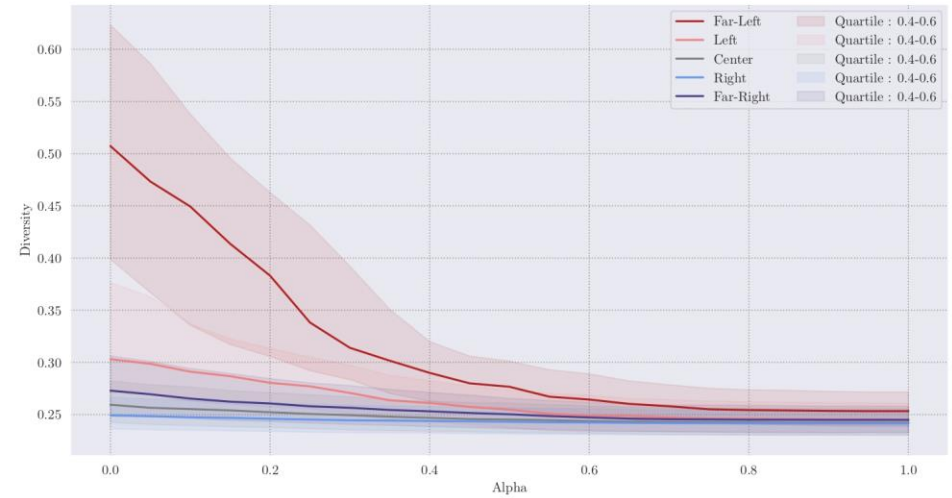






6. Political information reduction



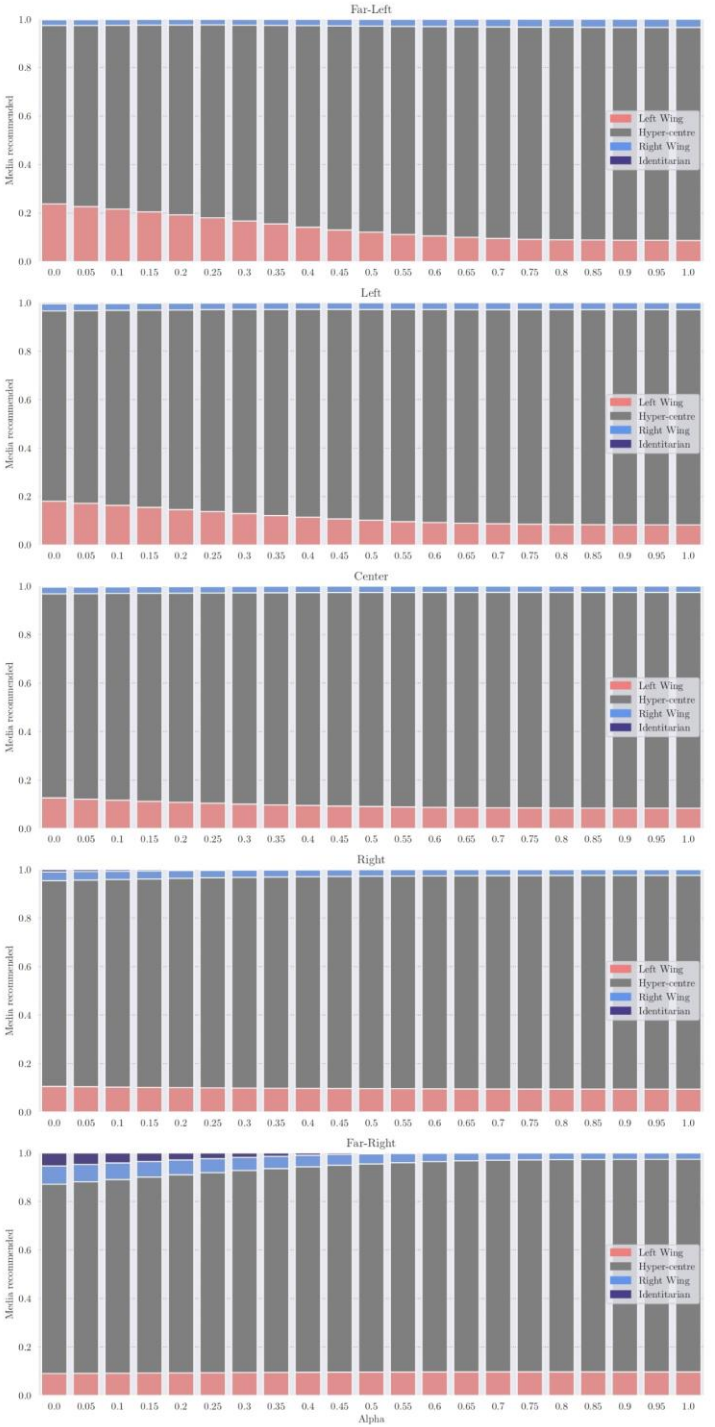


Reducing political dimensions :

- Reduced bias
- Reduced diversity

Mechanism :

- Mainstream media recommendation



Conclusion

Method :

- Political explanation
- Political impact reduction

Case study :

- Some dimensions learn specific political features
- Reducing political information in models reduce bias while reducing also diversity