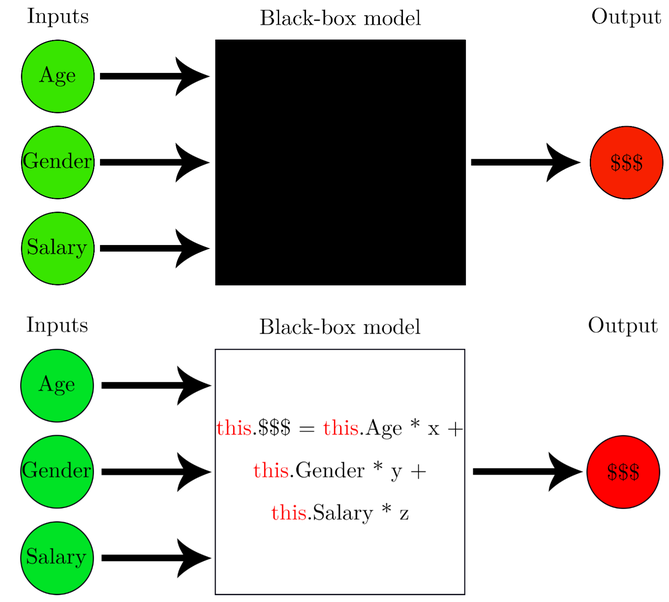
“Honey, Tell Me What's Wrong”, Global Explainability of NLP Models through Cooperative Generation

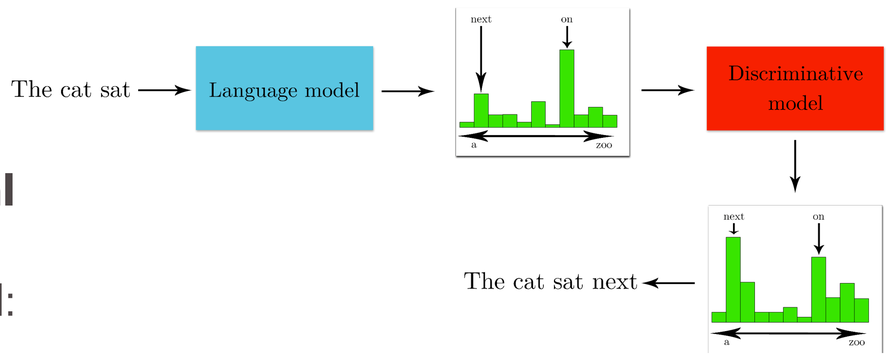
1. Explainability

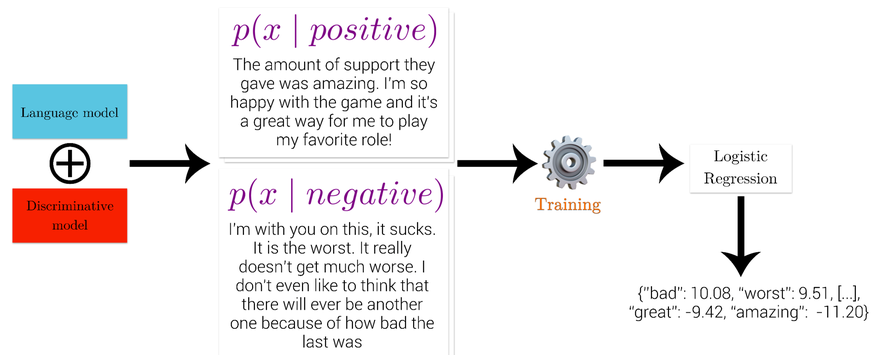
- Neural networks learn a **very complex mapping** between inputs and outputs
- Explainability try to give **insights about the decisions**
- **Model-agnostic** (not relying on internal functioning)
- **Local explanations**: differences in the model decisions for small variations of the input
- Several drawbacks:
  1. Requires data **(confidentiality/privacy)**
  2. Selecting **representative data** is hard
  3. Explain the decision for **this input and this input only**



2. Therapy

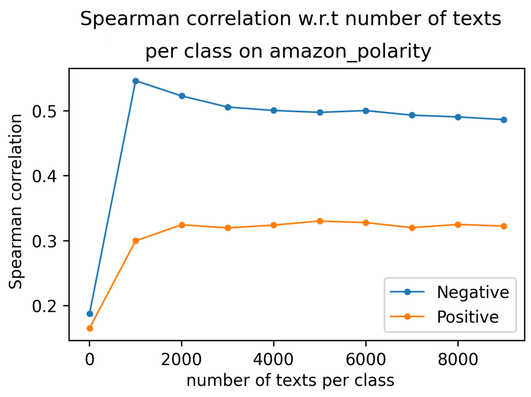
- Use a **pre-trained generative language model to guide the search**
- **Cooperative generation** to generate texts following the **conditional distribution**
- Use the distribution of cooperatively generated texts to explain the model: **words with high frequencies are likely to be important**
- Logistic regression trained to predict classes of generated texts
- **Weights associated to each word can be returned as explanation**
- Tf-idf to get **words frequent for a given class**

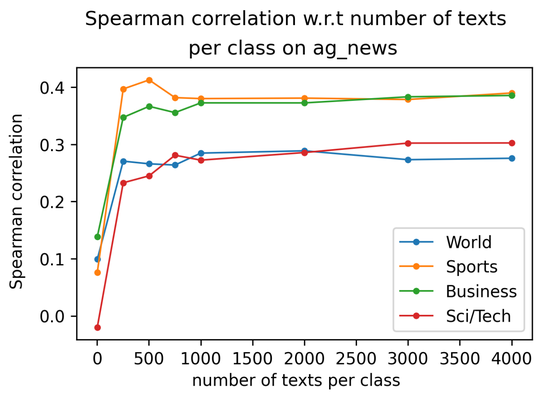




3. Experiments

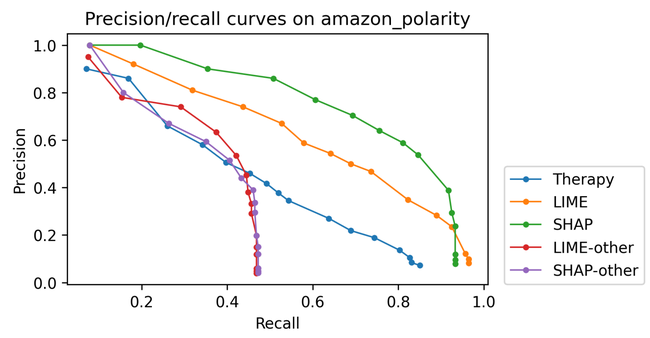
- Two classification datasets**: amazon_polarity & ag_news**
- No ground truth available: **glass-box model**
- **Spearman correlation**: assign similar (relative) weights
- **Precision/recall**: returned features are important and most important features are found
- **Insertion/deletion**: returned features affect the model predictions

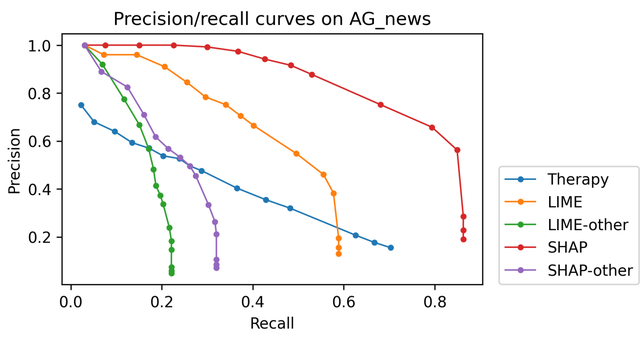


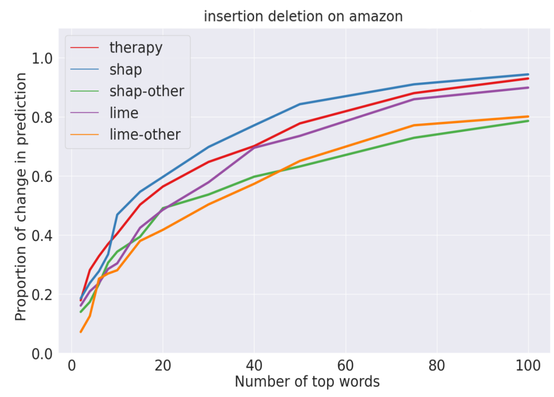


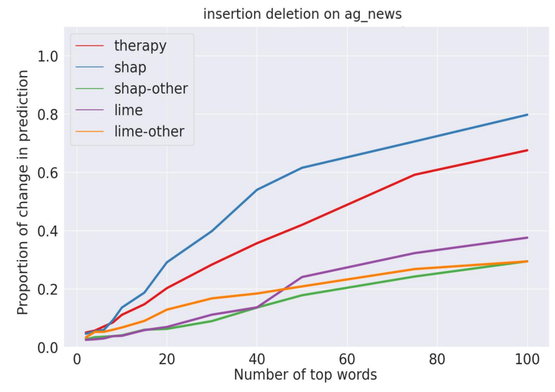
Dataset	AMAZON_POLARITY		AG_NEWS			
Class	Positive	Negative	World	Sports	Business	Sci/Tech
LIME	0.64 (5.0e-7)	0.44 (1.5e-3)	0.09 (0.53)	0.16 (0.27)	0.20 (0.16)	0.19 (0.19)
LIME-other	0.21 (0.14)	0.18 (0.21)	-0.03 (0.85)	0.23 (0.12)	0.09 (0.52)	0.29 (0.04)
SHAP	0.71 (7.6e-9)	0.76 (1.6e-10)	0.47 (6.2e-4)	0.62 (1.7e-06)	0.53 (8.0e-5)	0.61 (2.4e-6)
SHAP-other	0.02 (0.87)	0.26 (0.06)	-0.05 (0.71)	0.04 (0.77)	0.15 (0.31)	0.12 (0.41)
Therapy	0.49 (3.3e-08)	0.31 (1.0e-4)	0.27 (1.6e-07)	0.37 (4.0e-12)	0.38 (5.6e-13)	0.3 (8.9e-09)

Table 1: Spearman correlation (p-value) between the top words of a logistic regression glass box and the four explanation methods. Results are shown per class and dataset. ‘other’ indicate that the explanations are generated using the other dataset.









5. CONCLUSION

- **Model-agnostic global explanations working without input data**
- **Competitive results against usual methods**
- Substantially better **when no data or not very specific data** is available
- Code available on Github
  - Experiments with other type of model, e.g **CLIP (cross-modal regression)**

Antoine Chaffin
IMATAG, IRISA, France
Julien Delaunay
INRIA, France

Contacts
antoine.chaffin@irisa.fr
julien.delaunay@irisa.fr

References
"Why Should I Trust You?": Explaining the Predictions of Any Classifier.
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
A Unified Approach to Interpreting Model Predictions.
Scott Lundberg, Su-In Lee