Explainable AI and Bias in Word Embeddings

Jelke Bloem & Giovanni Colavizza

Text Mining
Amsterdam University College
With materials from AI BSc course AI for Society

May 16, 2022

Announcements

- Thursday 19/05: Project update 2

Stochastic Parrots: Questions

- How can we train language models more efficiently while still retaining their high accuracy?
- How do socio-technical systems work?
- What is the response of the industry or academia after the publication of this paper?
- To what extent are researchers actually thinking about the environmental impact of their models and the wider societal impact?
- Are the currently any established metrics or measurements in use for social or environmental impact of machine learning models, or technology in general?

Bias in Word Embeddings: Questions

- Will a solution, like the one adopted in the paper, help to fix models like Tay(bot) which became racist through word association and stereotypes?
- Could the Debiasing of word embeddings be generalized and used for other biases?
- Asides from word embedding fixes, what other fixes have there been for word debiasing?
- Are ten crowd-workers really a sufficient amount to produce statistically significant assessments of the resulting debiased word embeddings?
- Why were only lower case letter words used from the Google News corpus?
- I wonder if this method of reducing gender bias and thus tackle inherent sexism could do the same for racial biases/stereotypes and reduce inherent racism machine models might pick up from human training data.

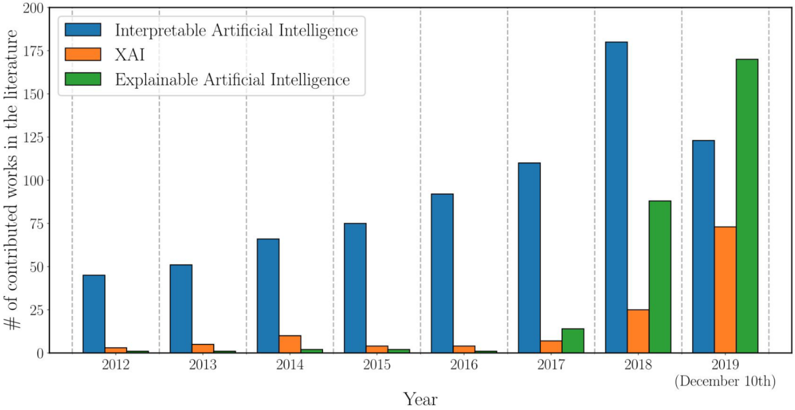
Bias in Word Embeddings: Questions

- How one would approach such a task in languages like German where gender neutral words don't exist?
  - `https://aclanthology.org/2021.trustnlp-1.6.pdf`
- Will the gender bias in word embedding be more or less significant in languages with grammatical gender, where most noun definitions incorporate a gender marker?
- How does the indirect bias removal work?
- I still did not fully understand how the pair bias gets removed (mathematically) when applying the hard de-biasing algorithm (what the intuitive idea is behind this process).
- What other algorithms do we have to debiasing word embeddings?
  - Training methods and post-processing algorithms
  - Project into a semantic space that is orthogonal to gender-specific words
  - Maximize distance betwen masculine and feminine words
  - Separate gender component and semantic component
  - Causal inference methods
  - Overview: `https://www.aaai.org/AAAI22Papers/AISI-6900.DingL.pdf`

Bias in Word Embeddings: Questions

- When doing sentiment analysis, would a very large and biased dataset still be useful if the labelling is done in a diverse manner such that the biases are taken into consideration, but as something negative?
- The question for me is what if one piece of text is supposed to be sexist and full of discrimination. How do we deal with this issue, wouldn't de-biasing the text hinder the initial effect?
- would it be realistic to create laws around the use of biased models so that we can follow a baseline to check whether a model is too biased to use in real-life settings?
- Fully fix issue? It seems like trying to fix the bias in our algorithms is a good way to start, but I'm wondering if it will ever be enough that we can see a society without any bias toward gender?

A recent trend...



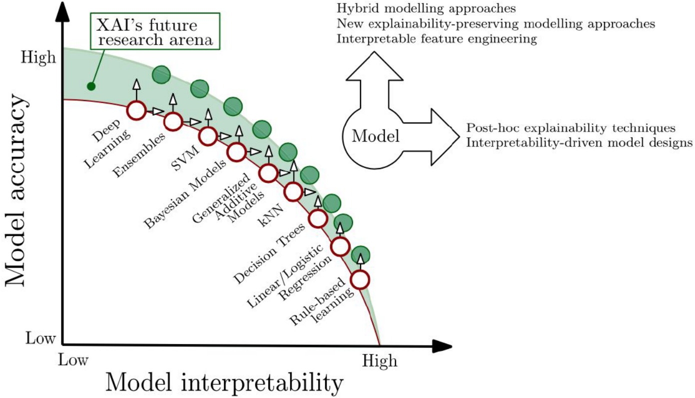
`https://arxiv.org/abs/1910.10045`

Explainable and Interpretable AI

- The definitions are not very clear yet, as it is an emerging field
- Interpretation: how does a model work? (model transparency)
  - Allow human to grasp the mechanism used to come up with a decision
- Explanation: what can a model tell me? (post-hoc reasoning)
  - Deconstruct steps that were used in making a decision

Explain to whom?

Performance vs Interpretability tradeoff


XAI's future research arena
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

Social aspects of the explanation/interpretation

- Confidence: grows when the rationale of a decision is close to the thought processes of the user
- Trust: grows when decisions do not require validation to be acted upon
- Safety: the system is consistent and relible, displays uncertainty or confidence level, is robust to outliers etc.
- Ethics: the system does not violate a certain well-defined code of principles

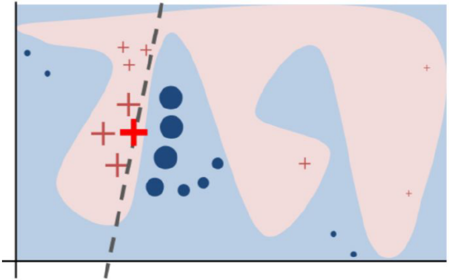
`https://arxiv.org/abs/2004.14545`

Contextual aspects of the explanation/interpretation

- Contrastive: identify elements unique to this decision
- Selective: provide the most relevant causes
- Provide causes: humans are bad at interpreting probabilities
- Social context: may call for different kind of explanation

LIME: Local Interpretable Model-agnostic Explanations

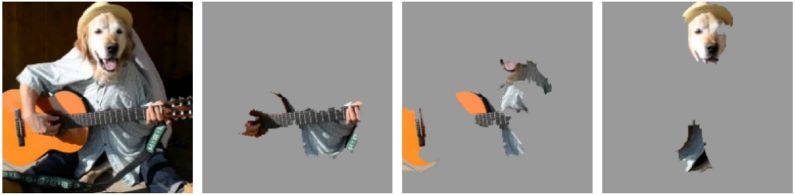
- Algorithm that explains predictions of a classifier by approximating it locally (in the vicinity of the predicted data point) with an interpretable model
- Treat original model as black box
- Train simple interpretable linear classifier on input features and classification decision



```
https://arxiv.org/abs/1602.04938
```

LIME: Example

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \;\; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

+ SP-LIME: Method to select representative examples of a classification problem to show to the user
`https://arxiv.org/abs/1602.04938`

Explainable AI

- Can we have explanation without interpretability?
- Can people accurately explain how they make decisions?

Links on Explainable AI

- List of libraries to explain black-box models: https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets
- LIME implementation in Python: https://github.com/marcotcr/lime
- SHAP unifies LIME and many more methods: https://github.com/slundberg/shap
- AIX360: https://github.com/Trusted-AI/AIX360
- Language Interpretability Tool (from UvA): https://github.com/pair-code/lit

Measuring (gender) bias in word embeddings

- Define a set of "definitional word pairs" that capture the gender dimension (e.g., he/she, man/woman, etc.)
- Measure bias by how differently a word *w* projects onto word pairs.
    - *x_he = cos("politician", "he")*
    - *x_she = cos("politician", "she")*
    - *x_he* - *x_she* = measure of bias towards the masculine gender

Bolukbasi et al. (2016): `https://arxiv.org/abs/1607.06520`

Measuring (gender) bias in word embeddings

Identify the gender subspace:

- Consider the pairwise differences among the set of “definitional word pairs” that capture the gender dimension (he-she, etc.)
- Apply dimensionality reduction on them (e.g. PCA), and find the gender subspace.
- Use the cosine between any word and this gender subspace to quantify its bias. This bias can be averaged over a set of words.
- If you take masculine - feminine, a positive cosine might be indicative of bias towards the masculine gender, vice versa for a negative one.

Dealing with (gender) bias in word embeddings

- Neutralize and equalize (**hard de-biasing**): enforces that any gender neutral word is set to zero onto the gender subspace.
- Soften (**soft de-biasing**): Ensures that neutral words are equidistant from equality sets. For example, it ensures that brother, sister and husband, wife are both equidistant from babysitting, although probably the latter set will still be closer than the former.

Approaches to bias in word embeddings

1. Work on data (e.g. filtering the training corpus)
2. Work on the algorithm (loss, bias mitigation via a constrained optimization objective)
3. Post-hoc methods (transforming the embeddings in some way)

https://www.aclweb.org/anthology/P19-1159

https://www.aaai.org/AAAI22Papers/AISI-6900.DingL.pdf