ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Evaluating Concept Discovery Methods for Sensitive Attributes in Language Models

Sarah Schröder[1] and Alexander Schulz[1] and Barbara Hammer[1] *

1- Bielefeld University - AG Machine Learning
Inspiration 1, 33615 Bielefeld - Germany

Abstract. This paper examines how to improve interpretability of language models in the context of fairness. While traditional concept learning focuses on identifying the most important concepts for a task, this study explores how to locate the representation of sensitive attributes in pretrained language models. We address challenges such as the potential low importance and sparsity of sensitive attributes in training data, and the limited amount of labeled data for this purpose. Our experiments evaluate potential methods to obtain such identity concepts, considering factors like label sparsity, generalizability, and the influence of different language models on the representation of sensitive attributes.

1 Introduction

Concept learning has become a popular approach for interpretability as it allows to attribute model predictions to abstract concepts rather than the input space. Usually concept learning is applied to identify task-specific concepts that are most relevant to explain model predictions [1, 2, 3]. In the scope of fairness however, model decisions need to be put in context of sensitive attributes. This requires additional cost for labeling. Especially in language models self-supervised learning and automated labeling of datasets are popular to handle large amounts of data at minimal labeling costs. By identifying how sensitive attributes are represented in language models, one could improve transparency and allow attribution of model decisions to sensitive attributes without respective labels. This paper explores the usability of existing concept learning methods as well as methods from the fairness literature to solve this problem. Specifically, we aim to predict "identity concepts", which activate at mentioning of protected groups or when dialect or language style hints at the author's identity. Compared to the usual use-cases for concept learning, learning these identity concepts is presumably more challenging in the sense that (i) sensitive attributes should not be among the most important concepts for task X and (ii) might be sparse in the training data. Another challenge of concept learning is to find concepts that generalize across datasets [4]. Bias subspaces [5, 6, 7] from the fairness literature are modeled from small sets of words or phrases representing the sensitive attributes, instead of optimizing these to some dataset and thus might be more suited in terms of generalizability.

We conduct experiments to evaluate the potential of bias subspaces and concept learning to discover identity concepts. Given the challenges mentioned

*Funded by the Ministry of Culture and Science of North-Rhine-Westphalia in the frame of the project SAIL, NW21-059A.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

before, we specifically investigate the influence of sparsity of identity labels and the generalizability of learned concepts between datasets. Another aspect is the influence of different language models, which might represent sensitive attributes in different ways.

2 Foundations

In our experiments we consider two concept learning approaches: Concept Activation Vectors and Concept Bottleneck Models. An alternative approach that does not require learning are bias subspaces. All methods assume that concepts are represented linearly in the embeddings of language models and provide user-defined concepts. We do not consider unsupervised concept learning since it cannot guarantee to deliver the identity concepts.

Concept Activation Vectors Kim et al. [3] propose Concept Activation Vectors (CAV) to learn user-defined concepts. They use a linear classifier to learn the activation vectors of concept X based on training samples that include concept X and random counter examples.

Concept Bottleneck Models Concept Bottleneck Models (CBM) [1, 2] realize concept learning by introducing an additional concept bottleneck layer before the prediction head. Using a contrastive loss $l = \lambda l_c + l_p$, the model is guided to optimize the prediction loss l_p while aligning the concept representation with user-defined concepts using concept loss l_c.

Bias subspaces Under the hypothesis that word embeddings encode semantic similarity by similarity in the vector space, researchers proposed to measure social bias by the relation of terms describing protected groups in the embedding space[5, 6, 7]. By contrasting over embeddings for terms of different protected groups they obtain bias directions or bias subspaces. We follow the approach of SAME [7], which allows us to construct bias spaces using an arbitrary amount of protected groups (other than [6]) while maintaining interpretability of the bias space (as opposed to [5]). Precisely, given n protected groups for some sensitive attributes, we define the concept for group i by $\mathbf{a_i} = \hat{\mathbf{e}}_\mathbf{i} - \mu$ where $\hat{\mathbf{e}}_\mathbf{i}$ is the mean of embeddings chosen to represent group i and $\mu = \frac{1}{n}\sum_j^n \hat{\mathbf{e}}_\mathbf{j}$. For the sake of interpretability we do not enforce the bias space to be orthonormal (as in [7]).

3 Experiments

3.1 Setup

3.1.1 Datasets

In the experiments, we use four dataset with labels for protected groups: (i) A supervised version of the BIOS dataset [8, 9] for occupation classification on biographies with binary gender labels, (ii) the TwitterAAE dataset [10] with tweets labeled as either african american or 'white' english, (iii) the CrowS-Pairs dataset [11], a collection of stereotypical phrases for bias evaluation in masked language modeling, and (iv) the Jigsaw Unintended Bias dataset [12]

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

for toxicity detection in comments. While BIOS and TwitterAAE provide single binary identity labels, CrowS-Pairs and Jigsaw follow a multi-label approach. In CrowS-Pairs every sample is assigned at least one identity label, though the groups are highly imbalanced. Similarly in Jigsaw identity labels are imbalanced, but in addition a majority of samples does not have any positive label at all. In Jigsaw and CrowS-Pairs we only considered protected groups with a minimum amount of positive samples to avoid random effects. For further details, see our implementation on Github[1].

3.1.2 Language Models

We consider encoder and decoder language models from Huggingface and a state-of-art embedding model from OpenAI (for details see [1]). All language models were used in their pretrained state to obtain text embeddings. For the Huggingface models we used both mean and cls pooling. In models without [CLS] token, the cls pooling was mimicked by adding a [CLS] token at the end of the input.

3.1.3 Experiment Design

To evaluate the robustness to label sparsity and availability, we evaluate CAV, CBM, and bias subspaces on a variety of datasets and protected attributes (see 3.1.1). For CAV and CBM we further distinguish cross-dataset transfer against training and testing on splits of the same dataset. The bias subspaces are dataset-independent, so we do not need to make the distinction. The BIOS dataset provides only gender concepts, TwitterAAE race concepts (in the form of dialect), while Jigsaw and CrowS-Pairs include concept labels for gender, ethnicity, religion and disability. Hence, cross-dataset transfer can be done between two or three datasets per protected attribute. We report Pearson correlations of concept activations with the ground-truth labels of protected groups. For CAV and CBM, we can easily derive binary concept labels from the concept activations, and thus report F1 scores. For bias subspaces determining the threshold is not trivial. The implementation details for the concept methods follow. For further details, see our implementation[1].

CAV: The CAVs are obtained by training a Logistic Regression in a one vs. rest scheme for each protected group. Concept activations of unseen inputs are then calculated using the dot product of CAVs and the input. CAV can be applied to any dataset where concept labels are available.

CBM: We realize the CBMs by two parallel concept layers: One for the identity concepts, which is optimized by the concept loss, and another one that is unconstrained. The stacked concepts are then fed into a 2-layer MLP for predictions. We select $\lambda = 0.5$. The output of the identity concept layer is taken as concept activations. Since concepts and predictions are optimized jointly, CBMs were only trained on the BIOS and Jigsaw dataset, where class labels are available.

Bias subspaces: We construct simple sets of defining terms for the protected groups used in the training of CAV and CBM (e.g. binary gender, ethnicity

[1] https://github.com/HammerLabML/PreSeCoLM/tree/esann25

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

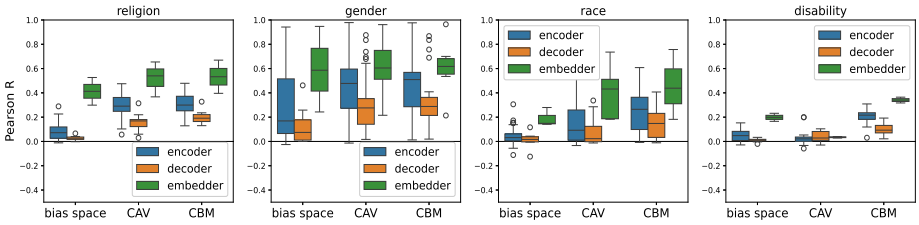

Fig. 1: Boxplot of Pearson correlations of concepts with ground truth labels. R values are aggregated over different datasets, models and pooling strategies.

from a US perspective). While we could easily extend the subspaces (e.g. non-binary gender) this might harm the comparability. Similarly to CAV, concept activations are computed by the dot product with the new input.

3.2 Results

Figure 1 gives an overview of Pearson correlations, aggregated over different datasets and models including both the transfer and non-transfer scenarios. We observe that correlations vary strongly between the different model types and sensitive attributes. Particularly, we report stronger correlations for the gender and religion concepts compared to race and disability. Overall, the concept learning methods deliver slightly higher correlations. Yet, there are cases where the bias subspaces perform equal or almost equally well. When closer inspecting the results, we find strong variations depending on the datasets, cross-dataset transfer, the protected groups (of the same attribute) and the models and pooling methods. Thus, we investigate these aspects in more detail.

3.2.1 Concept learning methods struggle with concept transfer

Table 1 shows the F1-scores per dataset (mean over different sensitive attributes) for CAV and CBM when trained on the same dataset or being transferred from another dataset (T). The transfer results are further aggregated over all possible transfer cases. We observe a significant drop of F1-scores in transfer scenarios. This is most obvious for BIOS. While both methods almost perfectly predict gender when trained and tested on BIOS, they fail when transferred from or to other datasets. The low correlations on TwitterAAE can be explained by the fact that concepts were transferred from 'mentioning of racial attributes' (Jigsaw/ CrowSPairs) to dialect. Since CBM could not be trained on TwitterAAE, we only report transfer results for CAV, too.

3.2.2 Concept predictability varies between datasets

Figure 2 shows the Pearson correlations for gender concepts separated by dataset. This emphasizes that some datasets are more challenging than others. In particular, the concepts are easy to learn on BIOS compared to CrowS-Pairs and

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

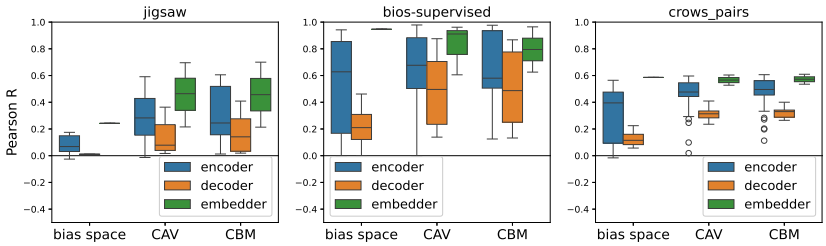

Fig. 2: Pearson correlations of gender concepts with ground truth labels.

Jigsaw. This can be explained by the multi-labels, different density and imbalance of identity labels (see Section 3.1.1). Despite using positive class weights for the concepts in the training process, the F1-scores and correlations remain limited. While bias subspaces are not affected by label imbalanced, the simultaneous presence of two groups (e.g. male and female) is an issue since the gender concepts were derived by contrasting between these two groups.

3.2.3 Concept learning favors frequent labels; bias subspaces marginalized groups

To further highlight the strengths of bias subspaces and concept learning, we look into the concept correlations for specific protected groups. Unsurprisingly CAV and CBM perform best on those groups that are more frequent in the training data. On the other hand, bias subspaces give better estimates for marginalized groups compared to the majority group. This accounts for a lot of variance in Figures 1 and 2.

3.2.4 Concept quality varies between language models and pooling methods

We observe differences in the performance of encoder, decoder and embedding models. In general, the concepts of the embedding model achieve the highest correlations. This is not surprising, considering that the embeddings are much larger (factor 1.5 to 2 compared to Huggingface models) and thus are more likely to represent these concepts linearly. Other than that, concepts of encoder outperform the decoder models, and concepts of mean pooled embeddings outperform those of cls pooled embeddings.

Table 1: Concept F1-scores for CAV and CBM for the given dataset and sensitive attribute. (T) indicates that concepts were learned on a different dataset.

dataset	CAV	CAV (T)	CBM	CBM (T)
BIOS	0.91 ± 0.023	0.45 ± 0.111	0.92 ± 0.024	0.11 ± 0.004
TwitterAAE	n.a.	0.08 ± 0.016	n.a.	$0.005 \pm 5e-5$
jigsaw	0.38 ± 0.034	0.23 ± 0.025	0.43 ± 0.051	0.22 ± 0.004
CrowS-Pairs	n.a.	0.37 ± 0.043	n.a.	0.45 ± 0.060

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

4 Discussion

The experiments show that concept learning methods highly depend on the availability and density of identity labels for a specific downstream task. While bias subspaces are not dataset-specific and might generalize better, none of the methods produced convincing results out-of-the-box. One central aspect that needs to be considered is the language models' influence and whether non-linear methods are more suitable to discover identity concepts. Based on the findings in this paper, future work could explore several directions: (i) addressing the label imbalance issue in concept learning methods, (ii) comparing linear and non-linear methods for concept retrieval, (iii) investigating how bias subspaces could be improved to identify majority groups and handle simultaneous presence of same-attribute groups, (iv) focusing more on large models or models optimized for embeddings.

References

[1] P.W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *Proceedings of the 37th ICML*, volume 119 of *PMLR*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[2] Z. Tan, L. Cheng, S. Wang, B. Yuan, J. Li, and H. Liu. Interpreting pretrained language models via concept bottlenecks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 56–74. Springer, 2024.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th ICML*, volume 80 of *PMLR*, pages 2668–2677. PMLR, 10–15 Jul 2018.

[4] Anwar et al. Foundational challenges in assuring alignment and safety of large language models. *TMLR*, 2024.

[5] T. Bolukbasi, K.W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NIPS*, 29, 2016.

[6] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[7] S. Schröder, A. Schulz, and B. Hammer. The same score: Improved cosine based measure for semantic bias. In *proceedings of IJCNN*, pages 1–8, 2024.

[8] De-Arteaga et al. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the FAccT*, pages 120–128, 2019.

[9] S. Schröder, A. Schulz, I. Tarakanov, R. Feldhans, and B. Hammer. Measuring fairness with biased data: A case study on the effects of unsupervised data in fairness evaluation. In *proceedings of IWANN*, pages 134–145. Springer, 2023.

[10] S.L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on EMNLP*, pages 1119–1130. ACL, November 2016.

[11] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[12] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.