Microsoft

Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection

Rajkumar Pujari*, Erik Oveson+, Priyanka Kulkarni+ and Elnaz Nouri#

*Purdue University, +Microsoft Redmond, #Microsoft Research Redmond

1

Overview

Motivation

- Empirical success of large Pretrained Language Models led to them being ubiquitously used in daily-life applications that interact with humans. Unsupervised training on huge, uncurated datasets results in harmful text and societal bias creeping in their outputs

- This motivates a two-pronged solution:
  1) To diagnose and de-noise the bias in the PLMs
  2) To identify & regulate harmful text externally at the output

- This work focuses on the task of identifying *stereotypical associations* in text
- *Stereotypes* differ from other harmful text such as *hate speech, misogyny, abuse, threat, insult* etc., in two important ways:
  1) They could also express a positive sentiment towards the target
  2) We require knowledge of their existence in the society to identify them

Asians are good at math

My African-American friend loves watermelons

Contributions

- We devise a focused annotation effort for *"Stereotype Detection"* to construct a fine-grained evaluation dataset
- We leverage the existence of several correlated neighboring tasks to propose a *reinforcement-learning guided multitask framework* that identifies and leverages neighboring task data examples that are beneficial for the target task

2

Dataset

Existing Datasets

- There are two existing datasets for mitigating Stereotypical bias. Both of them are diagnostic in nature:
    1) *Stereoset (Nadeem et al. 2020* [1])
    2) *CrowS-Pairs (Nangia et al. 2020* [2])

- Blodgett et al. (2021) [3] demonstrate that both the datasets suffer from conceptual and operational issues

- In addition, diagnostic datasets, by nature, also suffer from lack of coverage of subtle manifestations of stereotypes in text

Annotation Approach

- We address the coverage issue by collecting potential data samples for annotation from two subreddits: */r/Jokes* (stereotype-rich) and */r/AskHistorians* (stereotype-poor)

- To avoid operational and conceptual pitfalls, we use *Cardwell 1996* [4]'s definition of *Stereotype*: *"a fixed, over-generalized belief about a particular group or class of people"*

- We ask the annotators to answer three questions for each sample:
  1) Is an over-simplified belief about a type of person "intentionally" expressed in the text?
  2) Is there an "unintentional", widely-known stereotypical association present in the text?
  3) Does the sentence seem made up (unlikely to occur in regular discourse)?

Our Dataset

➢ This focused annotation approach allows us to categorize the examples into three classes: *explicit stereotypes, implicit stereotypes* and *non-stereotypes*. We use *anti-stereotypes* from existing datasets.

1) Ethiopians like stew (*explicit stereotype*)
2) The lawyer misrepresented the situation and tricked the person (*implicit stereotype*)
3) Jews spend money lavishly (*anti-stereotype*)
4) There is an Asian family that lives down the street (*non-stereotype*)

Data Type	Size
Explicit Stereotypes	742
Implicit Stereotypes	282
Non-Stereotypes	1197

3

Model

Neighbor Tasks

- Several datasets for harmful language identification such as *hate speech detection, offensive language detection, misogyny detection* and *toxicity detection* are widely available. They often contain overlapping objectives. For example:

  1) She may or may not be a jew but, she's certainly cheap! (insult, stereotype)

  2) Burn in hell, you Asian bastard! (abuse, stereotype)

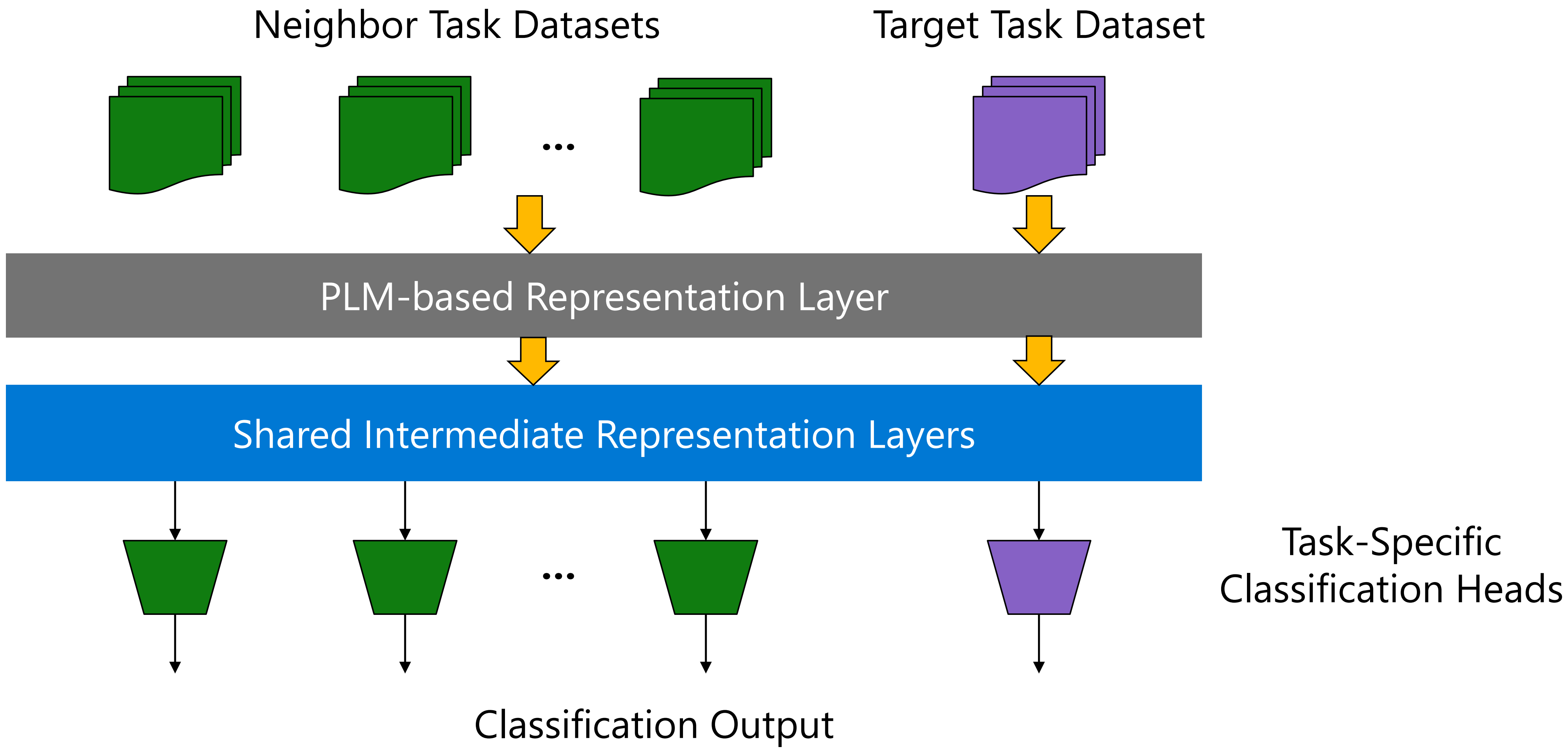
- We hypothesize that solving these tasks require understanding largely similar linguistic characteristics of the text. We call these tasks *"neighbor tasks"*.

Multi-Task Learning Model

- **Motivation:** Leverage the transfer learning gains from the neighbor tasks to improve the target task.

- As the tasks have "*overlapping objectives*" and largely require "*understanding similar linguistic characteristics*" of text, leveraging the *intermediate representations* from the neighbor tasks should benefit the target task.

Multi-Task Learning Architecture



RL-Guided Multi-Task Learning Model

- **Intuition:** *Not all examples from the neighbor task are equally useful in learning the target task*

- We train an RL-agent on top of the MTL model to identify examples from neighbor tasks, which are beneficial for the target task

  - **Step 1:** For each example in the neighbor task, RL-actor makes a *select*/*reject* decision
  - **Step 2:** MTL model is trained on the selected examples from the neighbor task
  - **Step 3:** The RL-actor is rewarded based on the *change in performance on the target task*
  - **Step 4:** The loss between *RL-actor's actual reward* and *RL-critic's expected reward* is used to train the RL-agent

4 Experiments

Experimental Setup

- We perform experiments using *six* datasets in *three* phases:
  - Phase 1: Fine-tune PLM-based classifier
  - Phase 2: Train a multi-task learning (MTL) model for all the datasets
  - Phase 3: Train RL-guided MTL model for each task as target task

- We experiment with *four* PLMs as base-classifiers: BERT-base, BERT-large (Devlin et al., 2019 [5]), BART-large (Lewis et al., 2020 [6]) and XLNet-large (Yang et al., 2019 [7])

- We use the following *six* datasets for our experiments:

1) Hate Speech Detection (de Gilbert et al., 2018 [8])
2) Offensive Language Detection (Davidson et al., 2017 [9])
3) Misogyny Detection (Fersini et al., 2018 [10])
4) Coarse-Grained Stereotype Detection: combination of *StereoSet* and *CrowS-Pairs* datasets
5) Fine-Grained Stereotype Detection (*our dataset*)
6) Jigsaw Toxicity Dataset [11] (*used only for training*)

Results

Model	Hate Speech Detection	Offense Detection	Misogyny Detection	Coarse-grained Stereotypes	Fine-grained Stereotypes
BERT-base	66.47	66.13	74.16	65.71	61.36
BERT-large	67.05	63.90	72.13	59.63	55.42
BART-large	68.91	65.86	73.12	63.40	54.64
XLNet-large	59.14	48.33	63.16	63.71	53.80
Multi-Task Learning					
BERT-base + MTL	69.21	68.57	73.48	68.29	65.00
BERT-large + MTL	69.78	65.14	73.94	61.96	61.65
BART-large + MTL	67.79	68.03	74.40	65.77	64.90
XLNet-large + MTL	61.68	46.35	64.42	65.21	57.00
RL-guided Multi-Task Learning					
BERT-base + RL-MTL	**72.06**	**68.97**	74.48	**74.18**	65.72
BERT-large + RL-MTL	69.82	65.97	**75.21**	70.88	64.74
BART-large + RL-MTL	69.60	66.76	75.14	74.11	**67.94**
XLNet-large + RL-MTL	61.97	47.60	63.21	67.98	56.37

5 Analysis

Impact of MTL Prior on RL-MTL

- In our experiments, we initialize the parameters of RL-MTL model with trained parameters from the MTL model.

- In this ablation, we initialize the RL-MTL model randomly and observe the difference in performance

Task	MTL Initialization	Random Initialization
Hate Speech Detection	72.06	70.23
Offense Detection	68.97	67.23
Misogyny Detection	74.78	71.10
Coarse-grained Stereotypes	74.18	60.42
Fine-grained Stereotypes	65.72	57.32

Neighbor Task Impact

- In this ablation, we study the impact of each neighbor task with each task as a target task

- It is interesting to note that *coarse-grained stereotype* data doesn't contribute as significantly to the performance improvement on *fine-grained stereotype detection task*. This might be due to the presence of anti-stereotypes and several other issues pointed out in Blodgett et al. (2021) [3].

Target \ Neighbor	Hate Speech Detection	Offense Detection	Misogyny Detection	Coarse-grained Stereotype
Hate Speech	-	69.69	70.07	**71.10**
Offensive Language	66.71	-	66.56	**67.39**
Misogyny	70.98	**75.87**	-	73.89
Coarse Stereotype	66.15	**67.40**	63.82	-
Fine Stereotype	**63.80**	63.65	59.94	56.12

Conclusion

- We tackle the problem of *Stereotype Detection* from *data annotation* and *low-resource computational framework* perspectives

- We devise a *focused annotation task* in conjunction with selective data candidate collection to create a fine-grained evaluation set for the task

- We utilize neighbor tasks with abundance of high-quality gold data in our *multi-task learning model*. We further propose an *RL-guided multi-task learning model* that learns to select examples from the neighbor tasks which benefit the target task.


Microsoft
Thank you

References

[1] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. arXiv.

[2] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. arXiv.

[3] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In ACL-IJCNLP 2021.

[4] Mike Cardwell. 1996. Dictionary of psychology. Routledge.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of NAACL-HLT, Volume 1 (Long and Short Papers), pages 4171–4186.

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880

[7] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5754–5764.

[8] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

[9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM '17, pages 512–515.

[10] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In IberEval@ SEPLN, pages 214–228.

[11] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification