Pre-Training Language Models for Identifying Patronizing and Condescending Language: An Analysis

Carla Pérez-Almendros,
Luis Espinosa-Anke
Steven Schockaert
Cardiff NLP, Cardiff University

Pre-Training Language Models for Identifying Patronizing and Condescending Language: An Analysis

1. A little bit of context: PCL
2. The data: the Don’t Patronize Me! dataset
3. Experiments and Results
4. Conclusions

Carla Pérez-Almendros
Cardiff NLP

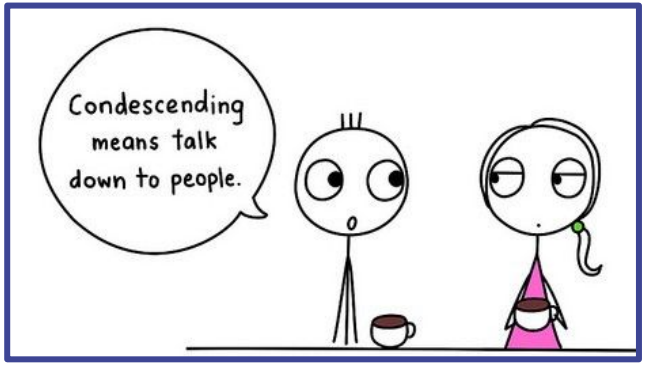
Cardiff NLP

A little bit of context:

Patronizing and Condescending Language

What is PCL

Someone engages in PCL when their language use shows a superior attitude towards others or depicts them in a compassionate way.



PCL ...

- is subtle and often unconscious
- has good intentions (e.g.., helping a community, raising funds, moving the audience to action)
- routinizes discrimination and makes it less visible (Ng, 2007)
- makes it more difficult to overcome difficulties and reach total inclusion (Nolan and Mikami, 2013)

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

How to identify PCL

When, referring to an underprivileged individual or community, we can identify one or several of the following traits:

- Clear differences between *us* and *them*.
- Feeling of pity towards an individual or a community.
- The author and the audience are presented as *saviours* of those in need.
  We can and must help them. *We know what they need*.
- The vulnerable community lacks the privileges the author's community enjoys or even the knowledge to overcome their own problems. *They need us*.
- The vulnerable community and its members are presented either as victims or as heroes.

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Why it is important to study

Research in sociolinguistics has suggested the following traits of PCL towards vulnerable communities:

- fuels discriminatory behaviour (Mendelsohn et al., 2020)
- creates and feeds stereotypes (Fiske, 1993)
- drives to greater exclusion, discrimination, rumour spreading and misinformation (Nolan and Mikami, 2013)
- it strengthens power-knowledge relationships (Foucault, 1980)
- communities in need are presented as passive receivers of help, waiting for a saviour to help them out of their situation (Bell, 2013; Straubhaar, 2015)
- it oversimplifies the wicked problems (Head et al., 2008) vulnerable communities face
- it proposes ephemeral and simple solutions (Chouliaraki, 2010)

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

The
Don’t Patronize Me! Dataset
(Pérez-Almendros et al, 2020*)

(*) Perez-Almendros, C., Espinosa-Anke, L., & Schockaert, S. (2020, December). Don’t Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5891-5902).

Cardiff NLP

Don’t Patronize Me! dataset (v_1.4.)

News stories

20 countries

10 keywords

10,469
paragraphs about vulnerable communities extracted from media in 20 English-speaking countries

Carla Pérez-Almendros
Cardiff NLP

(⁎) COLING 2020 paper: Perez-Almendros, C., Espinosa-Anke, L., & Schockaert, S. (2020, December). Don’t Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5891-5902).

(⁎) Dataset available at: https://github.com/Perez-AlmendrosC/dontpatronizeme

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

TAXONOMY of PCL categories

THE SAVIOUR

Unbalanced power relations

"[...] why not adopt poor families and help them break the cycle of poverty?"

Shallow solution

"Raise money to combat homelessness by curling up in sleeping bags for one night"

THE EXPERT

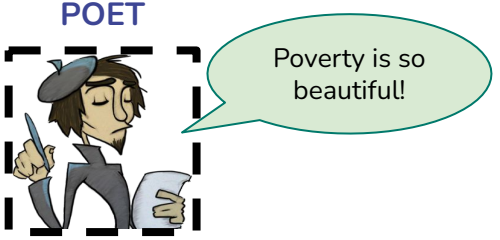
Presupposition

"[...] elderly or disabled people who are simply unable to evacuate due to physical limitations"

Authority voice

"Accepting their situation is the first step to having a normal life"

THE POET



Metaphors

"We have the opportunity to give the gift of love, to shine a light in the darkness of despair[...]"

Compassion

"From mother [...] who rejected him and a society that offered no respite, Siva was, in a nutshell, a hopeless street vagabond"

The poorer, the merrier

"[...] the disabled olympians, they have a genuine heart"

Cardiff NLP

EXPERIMENTS and RESULTS

PCL detection and categorization

Task 1 → Binary classification: PCL or not PCL (only 9% of paragraphs contain PCL)

Task 2 → Multilabel classification: What categories of PCL are present in the text

	F1-Score
Binary classification	53.27

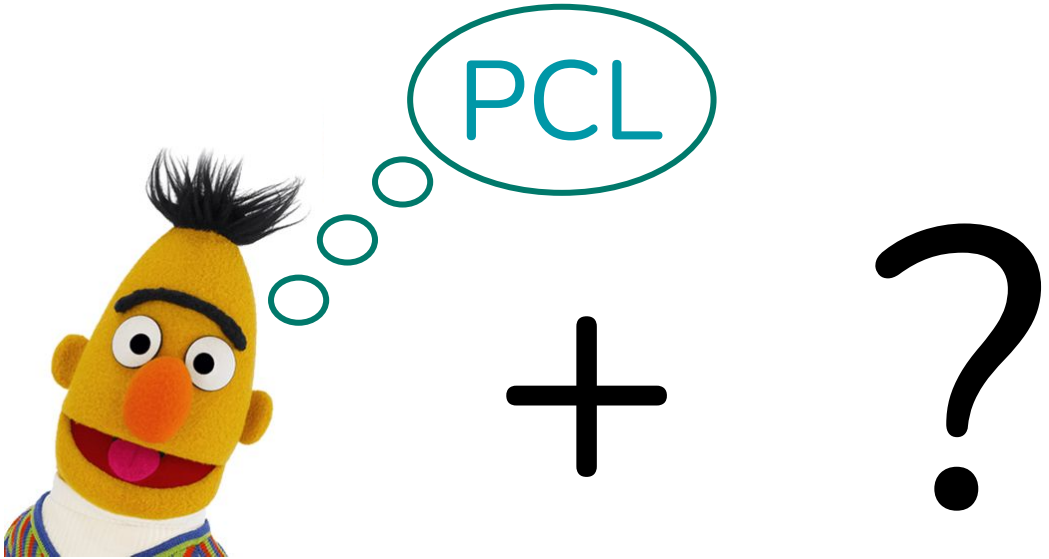
Category	F1-Score
Unb_pow_rel	51.81
Shallow solution	39.44
Presupposition	28.82
Authority voice	23.79
Metaphors	27.68
Compassion	46.48
Poorer, merrier	0.00

Baseline experiments with RoBERTa base

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

What does a NLP model need to know to better detect PCL towards vulnerable communities?



Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Auxiliary data

Human Values

Commonsense morality

Deontology

Social Justice
(Hendrycks et al., 2021)

StereoSet
(Nadeem et al., 2020)

Harmful Language

Offensive language
(Zampieri et al., 2020)

Hate Speech
(Basile et al., 2019)

Political Language

Democrats VS Republicans
(Pastor, 2018)

Hyperpartisan News Detection
(Kiesel et al., 2019)

Exploratory Data

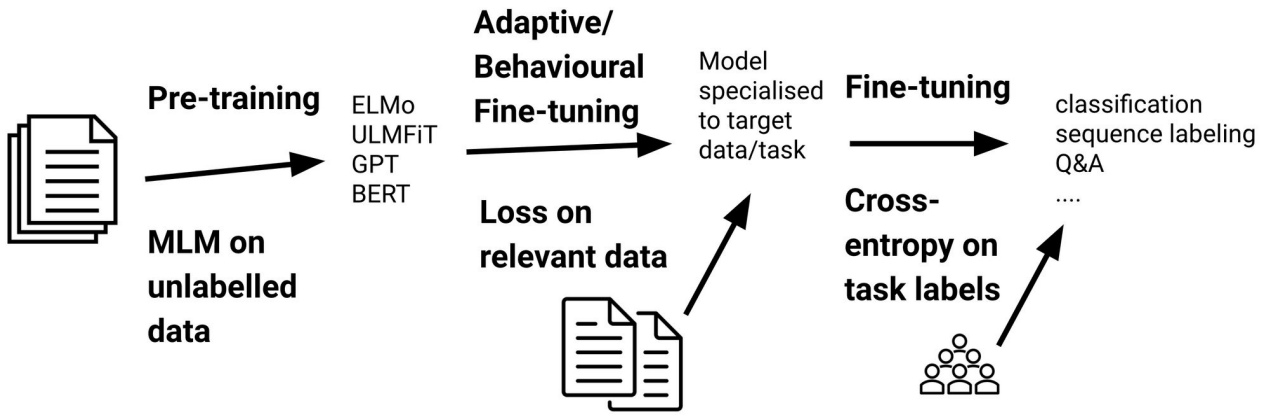
Irony
(Van Hee et al., 2018)

Sentiment
(Rosenthal et al., 2017)

Cardiff NLP

Pre-training strategies: From fine-tuning ...

(*)

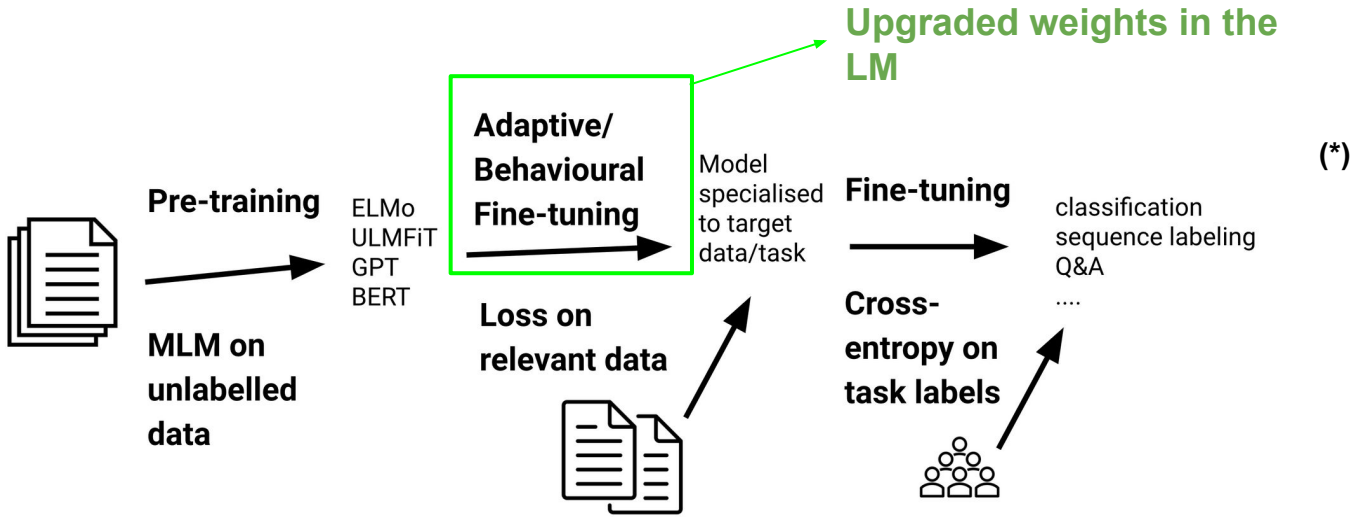


(*) https://ruder.io/recent-advances-lm-fine-tuning/

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Pre-training strategies: From fine-tuning ...

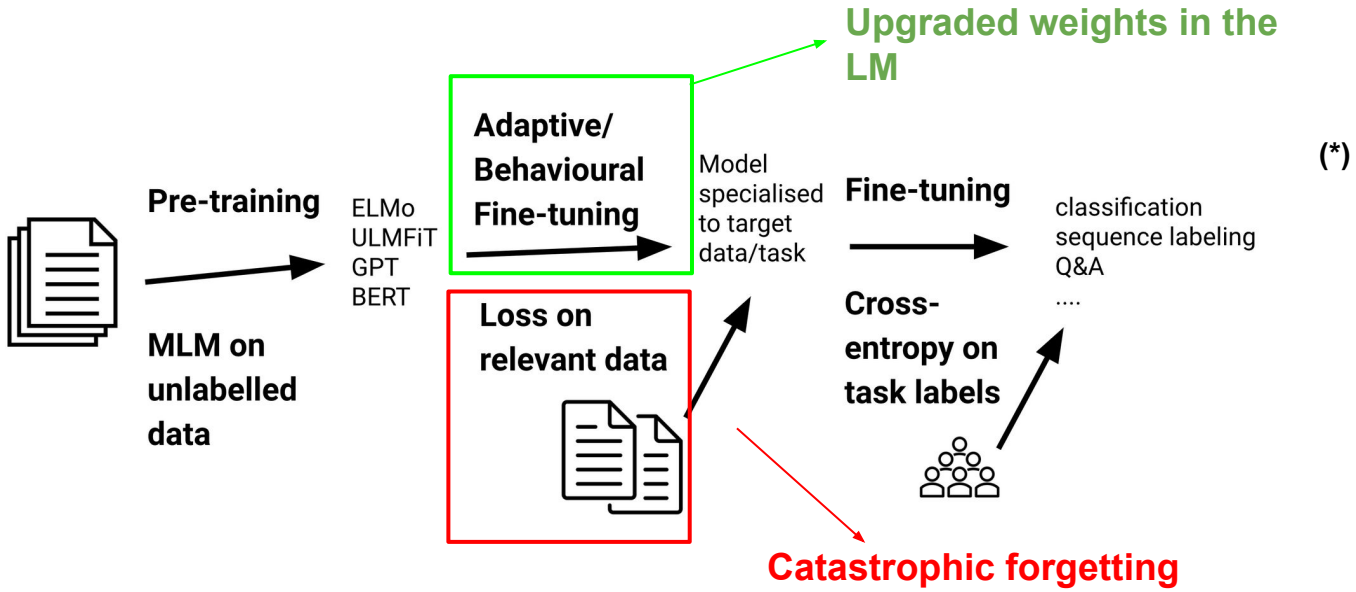


(*) https://ruder.io/recent-advances-lm-fine-tuning/

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Pre-training strategies: From fine-tuning ...



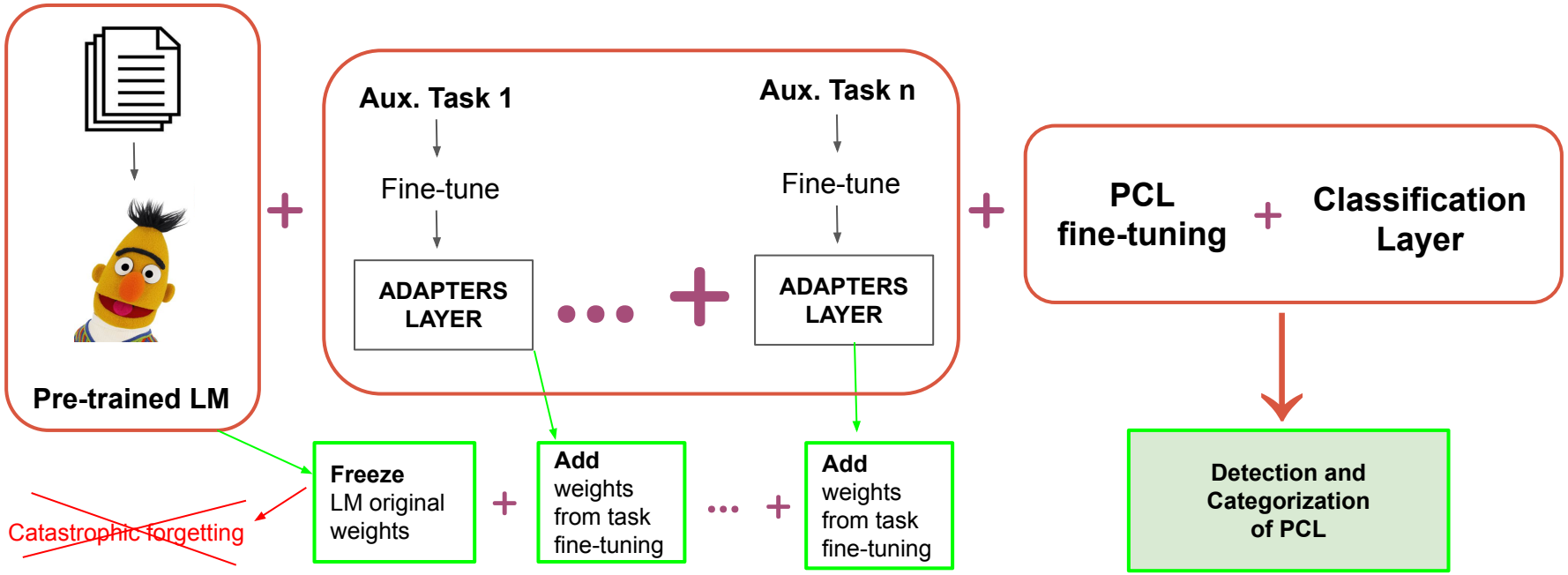
(*) https://ruder.io/recent-advances-lm-fine-tuning/

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies: ... to Adapters (*)



(*) Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzkebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In Proceedings of ICML 2019.

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 1

	Adapters	Adapters+Head	Fine-Tuning
RoBERTa baseline	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$
Commonsense Morality	**$54.14_{\pm 0.36}$**	**$55.24_{\pm 0.79}$**	$53.23_{\pm 0.20}$
Deontology	**$53.66_{\pm 0.48}$**	**$53.49_{\pm 0.31}$**	$52.22_{\pm 0.35}$
Social Justice	$53.06_{\pm 0.13}$	$53.04_{\pm 0.30}$	$51.45_{\pm 0.25}$
StereoSet	**$53.82_{\pm 0.54}$**	-	**$54.42_{\pm 0.54}$**
Hate Speech	**$54.16_{\pm 0.32}$**	**$55.37_{\pm 0.23}$**	**$53.59_{\pm 0.20}$**
Offensive Language	**$53.89_{\pm 0.33}$**	**$54.35_{\pm 0.52}$**	**$54.43_{\pm 0.43}$**
Democrat vs Republican	**$53.39_{\pm 0.40}$**	$53.08_{\pm 0.46}$	$51.61_{\pm 0.20}$
Hyperpartisan	**$53.47_{\pm 0.34}$**	**$53.72_{\pm 0.56}$**	$52.59_{\pm 0.41}$
Irony	**$53.76_{\pm 0.65}$**	**$54.18_{\pm 0.42}$**	$53.05_{\pm 0.18}$
Sentiment	**$54.50_{\pm 0.50}$**	-	**$54.50_{\pm 0.57}$**

F1 score (for the positive class) on PCL Detection with different auxiliary tasks and pre-training strategies. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 1

	Adapters	Adapters+Head	Fine-Tuning
RoBERTa baseline	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$
Commonsense Morality	**$54.14_{\pm 0.36}$**	**$55.24_{\pm 0.79}$**	$53.23_{\pm 0.20}$
Deontology	**$53.66_{\pm 0.48}$**	**$53.49_{\pm 0.31}$**	$52.22_{\pm 0.35}$
Social Justice	$53.06_{\pm 0.13}$	$53.04_{\pm 0.30}$	$51.45_{\pm 0.25}$
StereoSet	**$53.82_{\pm 0.54}$**	-	**$54.42_{\pm 0.54}$**
Hate Speech	**$54.16_{\pm 0.32}$**	**$55.37_{\pm 0.23}$**	**$53.59_{\pm 0.20}$**
Offensive Language	**$53.89_{\pm 0.33}$**	**$54.35_{\pm 0.52}$**	**$54.43_{\pm 0.43}$**
Democrat vs Republican	**$53.39_{\pm 0.40}$**	$53.08_{\pm 0.46}$	$51.61_{\pm 0.20}$
Hyperpartisan	**$53.47_{\pm 0.34}$**	**$53.72_{\pm 0.56}$**	$52.59_{\pm 0.41}$
Irony	**$53.76_{\pm 0.65}$**	**$54.18_{\pm 0.42}$**	$53.05_{\pm 0.18}$
Sentiment	**$54.50_{\pm 0.50}$**	-	**$54.50_{\pm 0.57}$**

F1 score (for the positive class) on PCL Detection with different auxiliary tasks and pre-training strategies. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 1

	Adapters	Adapters+Head	Fine-Tuning
RoBERTa baseline	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$	$53.27_{\pm 0.36}$
Commonsense Morality	**54.14**$_{\pm 0.36}$	**55.24**$_{\pm 0.79}$	$53.23_{\pm 0.20}$
Deontology	**53.66**$_{\pm 0.48}$	**53.49**$_{\pm 0.31}$	$52.22_{\pm 0.35}$
Social Justice	$53.06_{\pm 0.13}$	$53.04_{\pm 0.30}$	$51.45_{\pm 0.25}$
StereoSet	**53.82**$_{\pm 0.54}$	-	**54.42**$_{\pm 0.54}$
Hate Speech	**54.16**$_{\pm 0.32}$	**55.37**$_{\pm 0.23}$	**53.59**$_{\pm 0.20}$
Offensive Language	**53.89**$_{\pm 0.33}$	**54.35**$_{\pm 0.52}$	**54.43**$_{\pm 0.43}$
Democrat vs Republican	**53.39**$_{\pm 0.40}$	$53.08_{\pm 0.46}$	$51.61_{\pm 0.20}$
Hyperpartisan	**53.47**$_{\pm 0.34}$	**53.72**$_{\pm 0.56}$	$52.59_{\pm 0.41}$
Irony	**53.76**$_{\pm 0.65}$	**54.18**$_{\pm 0.42}$	$53.05_{\pm 0.18}$
Sentiment	**54.50**$_{\pm 0.50}$	-	**54.50**$_{\pm 0.57}$

F1 score (for the positive class) on PCL Detection with different auxiliary tasks and pre-training strategies. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 2

	UNB	SHAL	PRES	AUTH	MET	COMP	MERR
RoBERTa baseline	$69.80_{\pm 0.60}$	$69.08_{\pm 1.12}$	$68.04_{\pm 1.47}$	$63.30_{\pm 0.79}$	$77.26_{\pm 1.21}$	$71.26_{\pm 1.15}$	$70.50_{\pm 3.71}$
Commonsense Morality	**$72.46_{\pm 0.81}$**	**$69.49_{\pm 2.29}$**	**$69.38_{\pm 1.99}$**	**$66.09_{\pm 0.31}$**	**$77.77_{\pm 0.98}$**	**$72.15_{\pm 0.78}$**	**$74.00_{\pm 2.24}$**
Deontology	**$71.76_{\pm 0.51}$**	**$70.00_{\pm 1.05}$**	$68.04_{\pm 1.32}$	**$64.09_{\pm 1.64}$**	**$78.48_{\pm 1.71}$**	**$72.92_{\pm 1.13}$**	**$76.00_{\pm 4.18}$**
Social Justice	**$69.83_{\pm 0.72}$**	$66.73_{\pm 2.12}$	$65.89_{\pm 1.80}$	$63.13_{\pm 1.72}$	$74.52_{\pm 1.50}$	$69.55_{\pm 1.17}$	$69.50_{\pm 2.74}$
StereoSet	**$71.56_{\pm 0.82}$**	**$69.49_{\pm 1.50}$**	$66.61_{\pm 1.28}$	$62.70_{\pm 1.13}$	$76.04_{\pm 0.83}$	$71.17_{\pm 0.71}$	**$74.00_{\pm 4.18}$**
Hate Speech	**$72.07_{\pm 0.66}$**	$68.88_{\pm 1.80}$	**$68.93_{\pm 1.16}$**	**$66.70_{\pm 1.56}$**	**$77.66_{\pm 0.36}$**	**$72.71_{\pm 0.69}$**	**$72.00_{\pm 3.26}$**
Offensive Language	**$70.31_{\pm 1.21}$**	$67.35_{\pm 2.04}$	$67.68_{\pm 1.47}$	**$64.52_{\pm 1.91}$**	**$80.41_{\pm 0.99}$**	**$73.65_{\pm 0.94}$**	$70.00_{\pm 3.06}$
Democrat vs Republican	**$70.11_{\pm 0.68}$**	**$69.18_{\pm 2.06}$**	$66.34_{\pm 1.94}$	**$63.39_{\pm 1.95}$**	$75.63_{\pm 1.24}$	$70.92_{\pm 1.44}$	**$75.00_{\pm 2.50}$**
Hyperpartisan	**$69.92_{\pm 0.82}$**	$67.14_{\pm 2.21}$	**$68.57_{\pm 1.32}$**	$63.13_{\pm 0.57}$	$76.35_{\pm 1.67}$	**$71.90_{\pm 1.14}$**	**$72.50_{\pm 5.86}$**
Irony	**$70.89_{\pm 1.63}$**	$68.57_{\pm 1.64}$	**$68.57_{\pm 1.36}$**	**$65.22_{\pm 2.20}$**	**$77.46_{\pm 1.59}$**	**$72.15_{\pm 1.45}$**	**$72.00_{\pm 2.74}$**
Sentiment	**$71.96_{\pm 0.55}$**	**$69.90_{\pm 0.62}$**	$67.32_{\pm 0.58}$	**$65.13_{\pm 0.89}$**	$76.24_{\pm 0.23}$	$71.26_{\pm 0.82}$	**$74.00_{\pm 2.85}$**

Recall per category for models that were pre-trained using adapters. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 2

	UNB	SHAL	PRES	AUTH	MET	COMP	MERR
RoBERTa baseline	$69.80_{\pm 0.60}$	$69.08_{\pm 1.12}$	$68.04_{\pm 1.47}$	$63.30_{\pm 0.79}$	$77.26_{\pm 1.21}$	$71.26_{\pm 1.15}$	$70.50_{\pm 3.71}$
Commonsense Morality	**72.46**$_{\pm 0.81}$	**69.49**$_{\pm 2.29}$	**69.38**$_{\pm 1.99}$	**66.09**$_{\pm 0.31}$	**77.77**$_{\pm 0.98}$	**72.15**$_{\pm 0.78}$	**74.00**$_{\pm 2.24}$
Deontology	**71.76**$_{\pm 0.51}$	**70.00**$_{\pm 1.05}$	$68.04_{\pm 1.32}$	**64.09**$_{\pm 1.64}$	**78.48**$_{\pm 1.71}$	**72.92**$_{\pm 1.13}$	**76.00**$_{\pm 4.18}$
Social Justice	**69.83**$_{\pm 0.72}$	$66.73_{\pm 2.12}$	$65.89_{\pm 1.80}$	$63.13_{\pm 1.72}$	$74.52_{\pm 1.50}$	$69.55_{\pm 1.17}$	$69.50_{\pm 2.74}$
StereoSet	**71.56**$_{\pm 0.82}$	**69.49**$_{\pm 1.50}$	$66.61_{\pm 1.28}$	$62.70_{\pm 1.13}$	$76.04_{\pm 0.83}$	$71.17_{\pm 0.71}$	**74.00**$_{\pm 4.18}$
Hate Speech	**72.07**$_{\pm 0.66}$	$68.88_{\pm 1.80}$	**68.93**$_{\pm 1.16}$	**66.70**$_{\pm 1.56}$	**77.66**$_{\pm 0.36}$	**72.71**$_{\pm 0.69}$	**72.00**$_{\pm 3.26}$
Offensive Language	**70.31**$_{\pm 1.21}$	$67.35_{\pm 2.04}$	$67.68_{\pm 1.47}$	**64.52**$_{\pm 1.91}$	**80.41**$_{\pm 0.99}$	**73.65**$_{\pm 0.94}$	$70.00_{\pm 3.06}$
Democrat vs Republican	**70.11**$_{\pm 0.68}$	**69.18**$_{\pm 2.06}$	$66.34_{\pm 1.94}$	**63.39**$_{\pm 1.95}$	$75.63_{\pm 1.24}$	$70.92_{\pm 1.44}$	**75.00**$_{\pm 2.50}$
Hyperpartisan	**69.92**$_{\pm 0.82}$	$67.14_{\pm 2.21}$	**68.57**$_{\pm 1.32}$	$63.13_{\pm 0.57}$	$76.35_{\pm 1.67}$	**71.90**$_{\pm 1.14}$	**72.50**$_{\pm 5.86}$
Irony	**70.89**$_{\pm 1.63}$	$68.57_{\pm 1.64}$	**68.57**$_{\pm 1.36}$	**65.22**$_{\pm 2.20}$	**77.46**$_{\pm 1.59}$	**72.15**$_{\pm 1.45}$	**72.00**$_{\pm 2.74}$
Sentiment	**71.96**$_{\pm 0.55}$	**69.90**$_{\pm 0.62}$	$67.32_{\pm 0.58}$	**65.13**$_{\pm 0.89}$	$76.24_{\pm 0.23}$	$71.26_{\pm 0.82}$	**74.00**$_{\pm 2.85}$

Recall per category for models that were pre-trained using adapters. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Pre-training strategies results - Task 2

	UNB	SHAL	PRES	AUTH	MET	COMP	MERR
RoBERTa baseline	$69.80_{\pm 0.60}$	$69.08_{\pm 1.12}$	$68.04_{\pm 1.47}$	$63.30_{\pm 0.79}$	$77.26_{\pm 1.21}$	$71.26_{\pm 1.15}$	$70.50_{\pm 3.71}$
Commonsense Morality	**72.46**$_{\pm 0.81}$	**69.49**$_{\pm 2.29}$	**69.38**$_{\pm 1.99}$	**66.09**$_{\pm 0.31}$	**77.77**$_{\pm 0.98}$	**72.15**$_{\pm 0.78}$	**74.00**$_{\pm 2.24}$
Deontology	**71.76**$_{\pm 0.51}$	**70.00**$_{\pm 1.05}$	$68.04_{\pm 1.32}$	**64.09**$_{\pm 1.64}$	**78.48**$_{\pm 1.71}$	**72.92**$_{\pm 1.13}$	**76.00**$_{\pm 4.18}$
Social Justice	**69.83**$_{\pm 0.72}$	$66.73_{\pm 2.12}$	$65.89_{\pm 1.80}$	$63.13_{\pm 1.72}$	$74.52_{\pm 1.50}$	$69.55_{\pm 1.17}$	$69.50_{\pm 2.74}$
StereoSet	**71.56**$_{\pm 0.82}$	**69.49**$_{\pm 1.50}$	$66.61_{\pm 1.28}$	$62.70_{\pm 1.13}$	$76.04_{\pm 0.83}$	$71.17_{\pm 0.71}$	**74.00**$_{\pm 4.18}$
Hate Speech	**72.07**$_{\pm 0.66}$	$68.88_{\pm 1.80}$	**68.93**$_{\pm 1.16}$	**66.70**$_{\pm 1.56}$	**77.66**$_{\pm 0.36}$	**72.71**$_{\pm 0.69}$	**72.00**$_{\pm 3.26}$
Offensive Language	**70.31**$_{\pm 1.21}$	$67.35_{\pm 2.04}$	$67.68_{\pm 1.47}$	**64.52**$_{\pm 1.91}$	**80.41**$_{\pm 0.99}$	**73.65**$_{\pm 0.94}$	$70.00_{\pm 3.06}$
Democrat vs Republican	**70.11**$_{\pm 0.68}$	**69.18**$_{\pm 2.06}$	$66.34_{\pm 1.94}$	**63.39**$_{\pm 1.95}$	$75.63_{\pm 1.24}$	$70.92_{\pm 1.44}$	**75.00**$_{\pm 2.50}$
Hyperpartisan	**69.92**$_{\pm 0.82}$	$67.14_{\pm 2.21}$	**68.57**$_{\pm 1.32}$	$63.13_{\pm 0.57}$	$76.35_{\pm 1.67}$	**71.90**$_{\pm 1.14}$	**72.50**$_{\pm 5.86}$
Irony	**70.89**$_{\pm 1.63}$	$68.57_{\pm 1.64}$	**68.57**$_{\pm 1.36}$	**65.22**$_{\pm 2.20}$	**77.46**$_{\pm 1.59}$	**72.15**$_{\pm 1.45}$	**72.00**$_{\pm 2.74}$
Sentiment	**71.96**$_{\pm 0.55}$	**69.90**$_{\pm 0.62}$	$67.32_{\pm 0.58}$	**65.13**$_{\pm 0.89}$	$76.24_{\pm 0.23}$	$71.26_{\pm 0.82}$	**74.00**$_{\pm 2.85}$

Recall per category for models that were pre-trained using adapters. Configurations that outperform the RoBERTa baseline are shown in bold. For all results, we show the average across five runs, as well as the standard deviation.

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Brief qualitative analysis

(Misclassified by baseline model and correctly classified by pre-trained models)

Model	Text	Categories
C. Morality	There are also angels who get together and help a larger group of people in need.	UNB, MET
	At a ceremony held in Accra, she said the presence of disable persons begging on the streets and the absence of ambulance to aid in the transfer of patients in need of critical help, moved her to donate the items.	UNB, SHAL, COMP
Deontology	But the goal isn't only to get the reality of homelessness onto social media.	SHAL, AUTH
	"The people of Khyber Pakhtunkhwa are resilient. I did not see hopelessness on any face," he said.	PRES, MERR

Model	Text	Categories
Hate	"I and my daughter Monica are excited about providing a space for disabled people to be able to get together and earn fair prices for their work," Mr. Rogers said.	UNB
	Apparently in Dr. Ablow's eyes, people who undergo the transgendered process are broken individuals, in need of repair. There are no transgendered people – only people who are confused and in need of treatment to alleviate their condition.	PRES, MET, COMP

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Carla Pérez-Almendros
Cardiff NLP

Brief qualitative analysis

(Misclassified by baseline model and correctly classified by pre-trained models)

Model	Text	Categories
Irony	As a matter of life views, migrants generally see opportunities where locals don't. They see how their home society has handled different problems and they can draw from that experience to simply copy and paste amazing solutions that change a society. These innovations are what an economy needs to grow and solve its own issues in dynamic ways.	PRES, MERR
	"It 's not just a matter of income poverty. What matters is children in very poor families in crowded, cold and damp houses. There is an income issue, there is a housing supply issue and there is a housing quality issue."	AUTH, COMP
Sentiment	The boxers were from poor families and had nothing. I was trying to feed them in my own home, and I wasn't thinking about my own family. All I knew was I had food in my house and I had to feed the boxers.	UNB, AUTH, COMP
	A kind-hearted woman has rescued a 11-year-old girl fleeing from her home in the Sri Lankan refugee camp near Madurai and re-united her with her family with the help of police in Tiruchi.	UNB

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

CONCLUSIONS

Cardiff NLP

Conclusions:

- Pre-training on auxiliary tasks helps LM to better identify and classify PCL
- Using Adapters is a better strategy in this setting, as it avoids the catastrophic forgetting.
- PCL detection improves especially by pre-training on:
  - → Harmful language (Hate speech)
  - → Human values (Common-sense Morality)
  - → More distant datasets (Sentiment, Irony)
- Different auxiliary tasks help the model to identify different categories of PCL
- This findings help us to understand more about the nature of PCL and the potential approaches for future work on the detection and categorization of PCL.

Where to find out more about this work:

https://github.com/Perez-AlmendrosC/pre-training_for_PCL_detection

https://carlaperezalmendros.medium.com/

Cardiff NLP

References:

Katherine M Bell. 2013. Raising Africa?: Celebrity and the rhetoric of the white saviour. PORTAL Journal of Multidisciplinary International Studies, 10(1).

Lilie Chouliaraki. 2010. Post-humanitarianism: Humanitarian communication beyond a politics of pity. International journal of cultural studies, 13(2):107–126.

Susan T Fiske. 1993. Controlling other people: The impact of power on stereotyping. American psychologist, 48(6):621.

Michel Foucault. 1980. Power/knowledge: Selected interviews and other writings, 1972-1977. Vintage.

Brian W Head et al. 2008. Wicked problems in public policy. Public policy, 3(2):101.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. Frontiers in Artificial Intelligence, 3:55.

Sik Hung Ng. 2007. Language-based discrimination: Blatant and subtle forms. Journal of Language and Social Psychology, 26(2):106–122

David Nolan and Akina Mikami. 2013. 'the things that we have to do': Ethics and instrumentality in humanitarian communication. Global Media and Communication, 9(1):53–70.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5891–5902.

Rolf Straubhaar. 2015. The stark reality of the 'white saviour'complex and the need for critical consciousness: A document analysis of the early journals of a freirean educator. Compare: A Journal of Comparative and International Education, 45(3):381–400.

Auxiliary Datasets:

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt, 2021. Aligning AI with shared human values. Proceedings of the International Conference on Learning Representations (ICLR).

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., ... & Potthast, M. (2019, June). Semeval-2019 task 4: Hyperpartisan news detection. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 829-839).

Kyle Pastor, 2018. Democrats Vs Republicans Tweets (version 4). Published at kaggle.com. URL: https://www.kaggle.com/kapastor/democratvsrepublicantweets

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and C¸ oltekin, C¸ . (2020). Semeval-2020 task ¨ 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63..

Van Hee, C., Lefever, E., and Hoste, V. (2018). Semeval-2018 task 3: Irony detection in english tweets. In Proceedings of The 12th International Workshop on Semantic Evaluation, pages 39–50.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada, August. Association for Computational Linguistics.

Nadeem, M., Bethke, A., and Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456.

Carla Pérez-Almendros
Cardiff NLP

Cardiff NLP

Thanks for your attention!