Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3902–3911
Marseille, 20-25 June 2022

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

Carla Pérez-Almendros, Luis Espinosa-Anke, Steven Schockaert
School of Computer Science & Informatics, Cardiff University, UK
{perezalmendrosc, espinosa-ankel, schockaerts1}@cardiff.ac.uk

Abstract

Patronizing and Condescending Language (PCL) is a subtle but harmful type of discourse, yet the task of recognizing PCL remains under-studied by the NLP community. Recognizing PCL is challenging because of its subtle nature, because available datasets are limited in size, and because this task often relies on some form of commonsense knowledge. In this paper, we study to what extent PCL detection models can be improved by pre-training them on other, more established NLP tasks. We find that performance gains are indeed possible in this way, in particular when pre-training on tasks focusing on sentiment, harmful language and commonsense morality. In contrast, for tasks focusing on political speech and social justice, no or only very small improvements were witnessed. These findings improve our understanding of the nature of PCL.

Keywords: Patronizing and Condescending Language, Pre-Training Strategies, Natural Language Processing

1. Introduction

The study of unfair, ideological, offensive or misleading discourse has become an important and well nourished topic of interest within the NLP research community. Most works on this topic address messages with a flagrant and clear intention of harming others, such as hate speech or offensive language detection (Zampieri et al., 2019; Zampieri et al., 2020; Basile et al., 2019). Others seek to address deceiving discourse, as in fake news (Conroy et al., 2015) or propaganda (Da San Martino et al., 2020). However, there are other kinds of discourse that, although equally harmful, present themselves in a more subtle way, making it more difficult, both for humans and NLP systems, to detect them. This is the case of Patronizing and Condescending Language (PCL), a type of discourse that presents a person or a community as superior to other(s). The use of PCL is often unconscious and well-intended, especially when referring to vulnerable communities (Wilson and Gutierrez, 1985; Merskin, 2011). This good will can make PCL especially harmful, as the audience receives this discriminatory language with low defense and is often unaware of its effects. PCL has been extensively studied in linguistics and social sciences (Margić, 2017; Giles et al., 1993; Huckin, 2002; Chouliaraki, 2006). Within NLP, however, only a few authors have addressed this topic. Two examples are Wang and Potts (2019), who compiled a corpus derived from Reddit comments, and our earlier work (Perez-Almendros et al., 2020), where we compiled a corpus from news stories about vulnerable communities. In addition, although not focusing on PCL, Sap et al. (2019) studied how certain uses of language indicate power relations, Mendelsohn et al. (2020) discussed the dehumanization of minorities through language and Zhou and Jurgens (2020) investigated the interplay between authoritative voices and expressions of condolences and empathy in online communities.

PCL detection presents an interesting challenge for NLP research, given its subtlety and subjectivity, as well as the fact that commonsense knowledge is often required to identify a message as being an instance of PCL, e.g., because they may refer to an implied *understanding* of human values and ethics. Although datasets that specifically address PCL are scarce, some of the associated challenges are also addressed in other tasks. For instance, Hendrycks et al. (2021) presents a number of tasks in which statements about human value judgements need to be assessed. While such tasks do not involve PCL, intuitively we would expect that modelling human value judgements plays an important role in PCL detection. In this paper, we analyse to what extent such tasks can improve the performance of a PCL detection system. While the idea of pre-training language models on auxiliary tasks is common practice (Ruder et al., 2019; Mao, 2020), the success of this strategy crucially depends on the relevance of these tasks (Poth et al., 2021).

The aim of this paper is to develop a better understanding of which types of pre-training tasks are most effective for PCL detection. In this way, we also aim to develop a better understanding of the nature of PCL itself. For instance, the performance of pre-training tasks that relate to human value judgements can be used to support or refute the hypothesis that such judgements are important for modelling PCL. As another example, we look at the performance of pre-training tasks that focus on more explicit forms of harmful language, such as hate speech. Given the subtle nature of PCL, and the fact that it is usually well-intended, it is unclear to what extent such tasks can be used for pre-training a PCL detection model.

2. PCL towards Vulnerable Communities

PCL towards vulnerable communities has been extensively studied in fields such as Social Sciences, Political Discourse, Psychology and Sociolinguistics (Margić, 2017; Giles et al., 1993; Huckin, 2002; Chouliaraki, 2006; Chouliaraki, 2010). These works present PCL as well-intended and sometimes unconscious. However, they also highlight the harmful effects it can have among those under-represented communities. PCL relies on subtle language, but can lead to discriminatory behaviour (Mendelsohn et al., 2020), as it creates stereotypes (Fiske, 1993), which drive to greater exclusion, discrimination, rumour spreading and misinformation (Nolan and Mikami, 2013). PCL also positions some communities as superior to others, which can be a tool for strengthening power-knowledge relationships (Foucault, 1980). Moreover, it praises and calls for charitable action versus cooperation, presenting powerful communities as *saviours* of more vulnerable ones (Bell, 2013; Straubhaar, 2015). PCL also tends to be shallow about the real, deep-rooted societal problems that lead to inequalities, simplifying situations, offering simple solutions (Chouliaraki, 2010) and even blaming the underprivileged communities or individuals for their situations. In summary, being the object of PCL makes it more difficult for vulnerable communities to overcome difficulties and reach total inclusion (Nolan and Mikami, 2013).

The Don't Patronize Me! dataset (Perez-Almendros et al., 2020) contains 10,637 paragraphs that refer to potentially vulnerable communities. The paragraphs were extracted from media sources in 20 English-speaking countries and mention, at least, one of a pre-defined set of keywords naming or referring to underrepresented groups. Each paragraph in the dataset is annotated with a label on a scale from 0 to 4, reflecting its level of condescension (0 being not condescending/patronizing and 4 being clearly condescending/patronizing). Following Perez-Almendros et al. (2020), we consider paragraphs labeled with 0 and 1 as negative examples and paragraphs labeled with 2, 3 or 4 as positive cases of PCL. Each positive example is furthermore labelled with one or more PCL categories. Among the positive examples of PCL, the distribution of categories is as follows: 73% UNB, 19% SHALL, 23.1% PRES, 23.6% AUTH, 48.7% COMP, 20.1% MET and 4.1% MERR. Note that most examples are labelled with more than one category. We briefly recall the meaning of these categories.

Unbalanced power relations (UNB): the author entitles themselves as being in a privileged situation, considering themselves as *saviours* of those in need (Bell, 2013; Straubhaar, 2015). For example, *"[...] why not adopt poor families and help them break the cycle of poverty?"*

Shallow solution (SHAL): a charitable, superficial and short-term action is presented as something which is life changing for the vulnerable community or individual. For example, *"Raise money to combat homelessness by curling up in sleeping bags for one night"*.

Presupposition (PRES): stereotypes and *clichés* are used to describe a community, relying on assumptions without having all the information. For example, *"[...] elderly or disabled people who are simply unable to evacuate due to physical limitations"*.

Authority voice (AUTH): the author stands as spokesperson and defendant of the community or individual and/or allows themselves to give expert advice about how to overcome underprivileged situations. For example, *"Accepting their situation is the first step to having a normal life"*.

Metaphor (MET): the author describes a difficult situation in a more poetic way through figures-of-speech such as metaphors and euphemisms. For example, *"We have the opportunity to give the gift of love, to shine a light in the darkness of despair[...]"*.

Compassion (COMP): the message uses flowery wording, and usually an abuse of adjectives, to reflect on the vulnerability or toughness of the situation, raising a feeling of pity among the audience. For example, *"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 (MERR): the author praises the vulnerability, granting positive values to all members of a vulnerable community and showing their admiration. For example, *"[...] the disabled olympians, they have a genuine heart"*.

3. Auxiliary Datasets

We consider four types of pre-training tasks for our experiments. First, we include tasks that involve modelling human value judgements. We consider three tasks from the ETHICS dataset. This dataset, introduced by Hendrycks et al. (2021), aggregates 5 tasks involving situations that need to be classified based on human values. We focus in particular on the Commonsense Morality, Social Justice and Deontology tasks, as they follow the same format as the *Don't Patronize Me!* dataset, i.e. they are binary text classification problems. For the three selected datasets, we combine the training and test splits to train our models. However, we discard the *test hard* partition, as it contains more ambiguous instances that could confound the model. In addition to the former, we also consider the StereoSet dataset (Nadeem et al., 2020) which measures stereotype bias in assumptions. We now describe the aforementioned tasks in more detail:

Commonsense Morality includes 17,795 assertions about specific scenarios, which need to be classified as acceptable or not based on commonsense moral judgements.

Social Justice includes 24,495 examples of the form "X deserves Y because Z", where the task is to predict whether the scenario is reasonable in terms of fairness.

Deontology contains 21,760 pairs of the form situation-assertion or petition-excuse, where the assertions and excuses need to be classified as being reasonable or not.

StereoSet includes 6,369 instances of the form context-assumption, where the task is to predict if (i) the assumption contains stereotypes; (ii) the assumption does not contain stereotypes; or (iii) the context and assumption are unrelated.

Second, we focus on tasks that involve detecting harmful language. We focus in particular on the Hate and the Offensive datasets (Basile et al., 2019; Zampieri et al., 2019), both of which are included in the TweetEval framework (Barbieri et al., 2020). The details of these pre-training tasks are as follows:

Offensive language is a collection of 14,100 tweets, where the task is to detect any kind of language that could offend either the target of the tweet or a general audience.

Hate speech contains 27,000 tweets, which need to be classified as containing hate speech or not.

We also consider two datasets that focus on political language. The interest in political discourse, in this context, stems from the fact that the way in which vulnerable communities are referred to plays an important role in such discourse. Indeed, PCL has been widely studied in relation to political discourse (Huckin, 2002). We focus in particular on Hyperpartisan News Detection (Kiesel et al., 2019) and Democrats vs Republicans Tweets[1]. The details are as follows:

Hyperpartisan News Detection is a small dataset with 645 news articles, which need to be classified as hyperpartisan or not.

Democrat vs Republican Tweets contains 86,460 tweets from US politicians, labelled as Democrat or Republican. The aim is to predict the political stance of the author of a given tweet.

Finally, as a more exploratory analysis, we also include two datasets from tasks that are intuitively less related to PCL detection, in particular the identification of irony (Van Hee et al., 2018) and sentiment analysis (Rosenthal et al., 2017), both also extracted from the TweetEval framework (Barbieri et al., 2020). Although the task of detecting irony may seem to have little in common with PCL detection, there are nonetheless some correspondences, such as the use of flowery and ornamented language and the prevalence of strongly opinionated inputs. Furthermore, we expect that some linguistic features that are related to the expression of sentiment might also help to detect PCL. The details of these tasks are as follows:

Irony consists of 4,601 tweets, where the task is to predict if they contain irony or not.

Sentiment consists of 59,899 tweets, where the task consists on classifying the sentiment of each input as negative, neutral or positive.

4. Experiments

The aim of this section is to explore the following research questions:

1. To what extent can the performance of PCL detection models be improved by pre-training these models on auxiliary datasets?
2. Which auxiliary tasks are most effective, and what does this tell us about the nature of Patronizing and Condescending Language?
3. How does the effectiveness of the pre-training strategies vary across different PCL categories?

After explaining our methodology in Section 4.1, we present our experimental results in Section 4.2. Finally, a qualitative analysis aimed at developing a better understanding of PCL is presented in Section 4.3. The code for our experiments can be found in https://github.com/Perez-AlmendrosC/pre-training_for_PCL_detection.

4.1. Methodology

We compare two standard strategies for pre-training a language model, namely full fine-tuning and the use of adapters (Houlsby et al., 2019). **Full fine-tuning** involves updating the parameters of a language model by training the model on some auxiliary task. Subsequently, the resulting model is trained on the target task. It is hoped that pre-training on this auxiliary task infuses some kind of knowledge or capability into the language model, which can then be exploited in the target task. However, an undesired consequence of pre-training in this way is the *catastrophic forgetting* of its previous knowledge that sometimes occurs (McCloskey and Cohen, 1989; Ratcliff, 1990). **Adapters** (Houlsby et al., 2019) are an alternative to full fine-tuning. In this case, new layers are added to the language model, which are trained on the auxiliary task, while the layers from the original model are frozen. The resulting model is then fine-tuned on the target task. Since the parameters from the original language model are not updated during pre-training, catastrophic

[1] www.kaggle.com/kapastor/democratvsrepublicantweets

forgetting should not occur. We consider two variants of the strategy with adapters: one in which the classification head for PCL detection is intialised based on the pre-training task (i.e. both the adapter layers and classification head are transferred) and one in which the classification head is randomly initialised (i.e. only the adapter layers are transferred). We will refer to these variants as *Adapters+Head* and *Adapters* respectively. We use the Simple Transformers library[2] for fine-tuning the models and Adapters-Hub (Pfeiffer et al., 2020b) for training the adapters, both of which are built over the Transformers library by Wolf et al. (2020).

Step 1:Auxiliary Task Pre-Training In each experiment, we start with a RoBERTa-base (Liu et al., 2019) language model, as this model obtained the best results in our earlier work (Perez-Almendros et al., 2020). The language model is then pre-trained on one of the auxiliary tasks described in Section 3, either using full fine-tuning or using adapters. For full fine-tuning, we use a learning rate of 1e-5, following Hendrycks et al. (2021). When using adapters, we use a learning rate of 1e-4, following Pfeiffer et al. (2020a). For both strategies, we use a batch size of 8 while training, which was the largest value we could fit into GPU memory. Furthermore, we fix the number of epochs depending on the size of the dataset, pre-training for 10 epochs on *Hyperpartisan* and *Irony* and for 2 epochs on the other tasks.

Step 2: PCL Fine-Tuning After pre-training on a given auxiliary task, we fine-tune the resulting model on the PCL dataset. We focus on the binary classification setting, i.e. determining whether a paragraph contains PCL or not. As a baseline, we directly train RoBERTa-base on the PCL dataset; this was the best-performing approach in (Perez-Almendros et al., 2020). Because the PCL dataset does not come with a fixed training-test split, we use 5-fold cross validation for all experiments. We train the models for 5 epochs, using a leaning rate of 1e-5 and a batch size of 8. Our main evaluation metric is the F1 score. However, to explore to what extent different categories of PCL are impacted, we also look at the recall per category, i.e. among all the paragraphs that are labelled with a particular category of PCL (e.g. UNB), we compute what percentage were predicted as positive examples by the model.

4.2. Experimental Results

Table 1 presents the results for each of the considered auxiliary tasks, for three pre-training strategies: adapters, adapters+head and full fine-tuning. Every experiment was repeated 5 times and we report the average F1 score across these 5 runs, as well as the standard deviation. One immediate conclusion is that using adapters outperforms full fine-tuned models in 7 out of 10 tasks. This suggests that catastrophic forgetting is indeed an issue in our setting, which could be related to the fact that the auxiliary tasks are only loosely related to the problem of PCL detection. In fact, *StereoSet*, *Hate Speech*, *Offensive Language* and *Sentiment* are the only tasks for which full fine-tuning outperforms the baseline.

Focusing now on the strategies with adapters, in most cases, *Adapters+Head* outperforms *Adapters*. For both of these variants, most configurations outperform the RoBERTa baseline, although the improvements in some cases are almost negligible. The strongest improvements are obtained for *Hate Speech*, *Commonsense Morality* and *Sentiment*, followed by *StereoSet* and *Offensive Language*. In contrast, the results for *Democrats vs Republicans*, *Hyperpartisan* and *Social Justice* are weak. The results for *Irony* are surprising: while these results are clearly below those of the best-performing configurations, they are nonetheless clearly above the baseline, despite the fact that irony detection is conceptually rather different from PCL detection. Similarly, the good results by pre-training on Sentiment are also unexpected.

In Table 2, we summarise the recall per category for the different auxiliary tasks. As before, all results are averaged across 5 runs. For this table, we considered the *Adapter* strategy. As can be seen, the improvement in recall across the categories is rather uneven. In general, the largest improvements are witnessed for UNB, AUTH and MERR. In contrast, for SHAL, PRES and MET any improvements that were obtained are smaller, and fewer auxiliary tasks are leading to any improvement at all. Comparing the performance of the different auxiliary tasks, we can see that *Commonsense Morality* achieves the most consistent results, outperforming the baseline across all categories. Somewhat surprisingly, given its rather disappointing overall performance, pre-training on *Deontology* leads to the best overall results for two categories: SHAL and MERR. Moreover, *Deontology* outperforms the baseline in all but one category. This is also the case for *Irony*. The worst results in Table 2 are for *Social Justice*, which is in line with its disappointing overall performance in Table 1.

4.3. Qualitative Analysis

To better understand how pre-training can benefit the PCL detection model, Tables 3, 4 and 5 show some examples of paragraphs containing PCL, which were incorrectly classified by the baseline model but correctly classified by one of the pre-trained models.

The **Commonsense Morality** dataset contains various sentences about situations where the author is helping others. Some examples of such sentences are as follows:

- *I received admiration for helping an elderly woman cross the street.*
- *A troubled friend came to me for help so I listened to her.*

[2] github.com/ThilinaRajapakse/simpletransformers

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

Table 1: 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. Note that for *StereoSet* and for *Sentiment*, the classification head of the Adapter can not be used, as the number of labels in the auxiliary task and the main task is different.

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

Table 2: 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.

- *I served leftover meat to the homeless.*
- *I put a dollar bill in the homeless man's empty cup.*

Comparing these sentences to the first block of examples in Table 3 reveals some clear similarities. For example, in both cases, the format of the text fragments centers on personal experiences or personal actions told in first person. By pre-training on *commonsense morality*, the model thus becomes exposed to language that is characteristic of PCL, especially when it comes to recognising *authority voice* (i.e. *I know what to do*, *I know what it means*, *I know what they think*), *unbalanced power relations* (i.e. the unbalanced relation between those who give and those who receive), and shallow solutions.

The examples from the **Deontology** dataset tell us what a person in a certain situation should morally or ethically do. Positive examples from this dataset mainly talk about helping and protecting others, which is a clear trait of PCL, especially for unbalanced power relations. Some examples are as follows:

- *I am a real estate agent in a large city. So I should make good deals to people who need homes.*
- *I am a high school English teacher. So I should offer tutoring to students who need help.*
- *I am a nun in the Catholic church. So I should pray for people who need it.*
- *I am a camp counselor at a small camp. So I should make sure children are kept safe.*
- *I am a babysitter that works on weekends. So I should ensure the children's safety.*

Model	Text	Categories
C. Morality	"I have a lot of sympathy for folks who are in need in the city," Mr Edmonds-Waters said. "This has become an extremely expensive city to live in. The divide between those who have and those who don't is ridiculously ginormous."	UNB, COMP
	We have done a great deal of work but the biggest thing we have done is to make the ministry a powerful tool for the protection of women... over the last four years, every time somebody comes up with a good idea or proposes an area in which women need help, we have adopted it immediately. We have been doing two things – policymaking, which also pushes other ministries, and looking at individual cases of suffering. For me, both components are equally important.	UNB, COMP
	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
	It wasn't uncommon for Johannah to hear through the media of an individual in need of assistance, and she would immediately make it her objective to get that person what they often desperately needed, and rarely without success.	UNB
Deontology	"We are in a state of mourning but we have to carry on because we have a duty to those people who are so in need of our support." Mr Bird said thousands of people worldwide had paid their respects to the two men, with tributes from as far afield as Australia.	UNB
	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

Table 3: PCL paragraphs correctly classified by models pre-trained on *C.Morality* and *Deontology* and missed by the baseline model

Model	Text	Categories
Hate	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
	School for the blind, deaf and dumb, Isulo, Anambra State, which parades a number of beautiful structures, is one of the schools battling with lack of facilities to meet the special educational needs of the children. According to Felix Nwaochi, President-General of Isulo Community, the school is seriously in need of water supply as many of the blind students have to fetch water from a stream to survive in the school.	UNB, SHAL, MET
	"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
	As Maas put it, "the loss of this organisation could unleash an uncontrollable chain reaction. "Kids would be pushed from Unrwa classrooms onto the streets, where they would be more vulnerable to dangerous scenarios such as recruitment efforts by terrorists, who will surely jump at the chance to argue that if we can't keep our aid promises, peaceful coexistence with the West is impossible. Child marriage, child labour, and child trafficking would rise. A generation of children and young people would be lost, in a region more unstable than ever.	UNB, AUTH, MET, COMP

Table 4: PCL paragraphs correctly classified by the model pre-trained on *Hate* and missed by the baseline model

The model pre-trained on *deontology*, therefore, learns about what is the right or wrong thing to do in different situations. Examples of PCL often have a similar message, as can be seen in the examples in Table 3 for the *deontology* pre-trained model.

The strong results for **Hate** are to some extent surprising, as the style of the tweets in this dataset, which is often about insulting and aggressively addressing people, is very different from PCL, which is more about praising and pitying individuals or communities. However, the vulnerable communities from the PCL dataset are commonly targeted in hate speech. A model which is pre-trained on hate speech can thus learn about what kind of attitudes towards these communities are acceptable. Moreover, the authoritarian or aggressive tone, the hyperboles and the abuse of adjectives that can be

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
	Bombarded by schizophrenia, addiction and homelessness, you might say that Eoghan O'Driscoll has been to hell and back. but he is finding a new balance through painting. Interview: Michael Lanigan	MET, COMP
	Many celebrities wore blue ribbons to support the American Civil Liberties Union, which is seeking to shed light on the plight of young immigrants facing the potential of being deported.	UNB, SHAL, MET, COMP
Sentiment	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
	The actor, who will be seen later this month in Avengers: Infinity War, found himself called upon to make the day of a young fan in need. On Wednesday, he hung out with Jacob Monday, who is a 16-year-old from upstate New York who has terminal cancer. The teen, who has a rare form of bone cancer, has a bucket list he's working through and it included meeting his favorite movie star.	UNB, SHAL
	Discrimination of the disabled by society is one of the major problems undermining the progress of democratic practice in the country. It is always the dream of people with disabilities that so long as the disability bill is passed, their position in society will be influenced positively.	PRES, AUTH
	He said the victims who are currently rendered homeless can now be relieved of troubles as the 5,000 iron sheets from Mwanza had arrived, with 1,200 already distributed to victims in Bukoba Municipality.	UNB, SHAL, AUTH
	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

Table 5: PCL paragraphs correctly classified by models pre-trained on *Irony* and *Sentiment* and missed by the baseline model

found in hate speech are also common traits of PCL, especially for the categories AUTH, COMP and MET. Some examples of sentences from the Hate dataset are as follows, with the first two being positive examples of hate speech and the last two being negative examples.

- *@user Coward Cameron go on welcome migrants with housing etc while destroying disabled peoples benefits its not a secret ur no good.*
- *Prevent new refugee crisis? You can stop doing the lies n propagandas bullshit.You can't even take care of your poor ppl at home. Space Force is too expensive for the ppl w 2 jobs.You can't even take care of Puerto Rico. Good night millions of homeless on the streets of US.*
- *Why we need to protect refugees from the ideas designed to save them.*
- *Lots of events coming up next week. Sign up to take action! On Aug 15th call Governor Wolf and demand he take action to protect immigrant families. Stop being complicit with Trump/ICE. Governor Tom Wolf...*

Some parallels with the examples for *hate* in Table 4 can be observed. First, in the examples above, we see how vulnerable communities are presented as being in need of protection and attention, which is similar to the examples from the PCL dataset in Table 4. The authoritarian and aggressive tone from the two positive examples above also resembles the last example for *hate* in Table 4.

The relevance of **Irony** may seem less clear than that of the other datasets. However, our experimental results nonetheless show that pre-training on this dataset is beneficial. To understand why this is the case, it is worth pointing out that instances from this dataset often contain strongly opinionated language and value judgments, which are related to the *AUTH* and *PRES* categories, as well as generalizations and hyperboles,

which are relevant for the *MERR* and *PRES* categories. Moreover, a speaker using irony often decorates their language with unnecessary, flowery wording, which is relevant for the *MET* category. The following examples from the *Irony* dataset illustrate these points:

- *"Now that i can seem to afford good things, material things in life ... its the simple things that i need and really want ... of my life"*
- *@user try having no internet for a month. Now I know how Ethiopians feel.*
- *"so, sane peoples would talk to themselves in twitter because they can't find other sane humans to talk to. that #retweet #ifagree"*
- *"@user I don't think, I know x"*

The model seems to learn from the assumptions, exaggerations and generalizations in the *irony* dataset, as we can also find them in the *irony* examples in Table 5, for instance in the generalization and assumption that migrants see what locals do not. In other examples from Table 5, we can see a dichotomy between a dramatic situation and a shallow solution (e.g. painting or wearing blue ribbons), which is reminiscent of the dichotomies that often appear in ironic language. The authoritarian, confident tone of the last two examples extracted from the *irony* dataset is also a common feature of PCL.

Pre-training on **Sentiment** also improves the model's performance on PCL detection. There are several features in the *sentiment* dataset which can help the model to detect condescension. For instance, the inputs from this dataset often contain a confident, strongly opinionated tone, characteristic of tweets, which is also a feature of the *AUTH* and *PRES* categories in PCL. To express sentiment, the texts also contain a fair number of adjectives, which can be easily linked to the *COMP* category in PCL. If we look at some examples from the dataset, we can also see a recursive structure of content, where someone does something for another person, a structure also shared by the *UNB* and *SHAL* categories of PCL. Some of these features can be observed in the examples below, extracted from the *sentiment* dataset:

- *We've got the info on how YOU can help those in need in SLC w/ @user & @user #ad*
- *Support CEO Keith Bradshaw as he spends a night sleeping at Adelaide Oval on THURSDAY raising money for the homeless*
- *'Knock Knock: Live:' David Beckham Surprises Family In Need: Tuesday marked the debut of "Knock Knock:... #family"*
- *"Jeff Foxworthy leads a Bible study with homeless guys on Tuesday mornings, and has for years. How cool is that?"*
- *"In the Oregon experiment, 10,000 previously-excluded people (poor & childless) were given access to Medicaid for the first time"*

The language in the above examples clearly shows unbalanced relations between those who can help and those who are helped, a highly indicative feature of PCL. Furthermore, in these examples those in a more powerful situation are praised by their charitable actions, which is as well a common theme in PCL, as shown in most of the examples of *Sentiment* in Table 5. There, the individuals who help and their actions take center stage in the paragraph, above the community or individuals who receive the action. By pre-training on *sentiment* the model seems to learn associations between some communities and their positions of power and need, and that helping others is considered an action with positive sentiment. This knowledge helps the model to better identify PCL.

5. Conclusions

In this paper, we studied which tasks can be used for pre-training a PCL detection model. Perhaps unsurprisingly, our findings confirm that pre-training the model on other forms of harmful language, such as hate speech, can be beneficial. However, we also identified several tasks whose success is more surprising. Most notably, we obtained clear improvements when using a dataset focused on commonsense morality, which supports the idea that PCL detection requires an assessment of human values. We also found irony detection to be a useful pre-training task. While this task is conceptually rather different from PCL detection, we found several similarities in the underlying discourse, such as the use of hyperboles and strongly opinionated language. Apart from comparing different pre-training tasks, we also compared different pre-training strategies, where we found that the use of adapters generally leads to the best results.

6. Bibliographical References

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., and Neves, L. (2020). TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

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.

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

Chouliaraki, L. (2006). *The spectatorship of suffering*. Sage.

Chouliaraki, L. (2010). Post-humanitarianism : Humanitarian communication beyond a politics of pity. *International Journal of Cultural Studies*.

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Da San Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., and Nakov, P. (2020). Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Giles, H., Fox, S., and Smith, E. (1993). Patronizing the elderly: Intergenerational evaluations. *Research on Language and Social Interaction*, 26(2):129–149.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. (2021). Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Huckin, T. (2002). Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, 155:176.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mao, H. H. (2020). A survey on self-supervised pretraining for sequential transfer learning in neural networks. *arXiv preprint arXiv:2007.00800*.

Margić, B. D. (2017). Communication courtesy or condescension? linguistic accommodation of native to non-native speakers of english. *Journal of English as a lingua franca*, 6(1):29–55.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Mendelsohn, J., Tsvetkov, Y., and Jurafsky, D. (2020). A framework for the computational linguistic analysis of dehumanization.

Merskin, D. L. (2011). *Media, minorities, and meaning: A critical introduction*. Peter Lang.

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

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

Perez-Almendros, C., Espinosa-Anke, L., and Schockaert, S. (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.

Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2020a). Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020b). Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Poth, C., Pfeiffer, J., Rücklé, A., and Gurevych, I. (2021). What to pre-train on? efficient intermediate task selection. *arXiv preprint arXiv:2104.08247*.

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

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.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2019). Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Straubhaar, R. (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.

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.

Wang, Z. and Potts, C. (2019). Talkdown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Wilson, C. C. and Gutierrez, F. (1985). Minorities and the media. *Beverly Hills, CA, London: Sage*.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing: System Demonstrations, pages 38–45.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

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

Zhou, N. and Jurgens, D. (2020). Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626.