Identifying Condescending Language: A Tale of Two Distinct Phenomena?

Carla Perez-Almendros and Steven Schockaert
School of Computer Science & Informatics, Cardiff University, UK
{perezalmendrosc, schockaerts1}@cardiff.ac.uk

Abstract

Patronizing and condescending language is characterized, among others, by its subtle nature. It thus seems reasonable to assume that detecting condescending language in text would be harder than detecting more explicitly harmful language, such as hate speech. However, the results of a SemEval-2022 Task devoted to this topic paint a different picture, with the top-performing systems achieving remarkably strong results. In this paper, we analyse the surprising effectiveness of standard text classification methods in more detail. In particular, we highlight the presence of two rather different types of condescending language in the dataset from the SemEval task. Some inputs are condescending because of the way they talk about a particular subject, i.e. condescending language in this case is a linguistic phenomenon, which can, in principle, be learned from training examples. However, other inputs are condescending because of the nature of what is said, rather than the way in which it is expressed, e.g. by emphasizing stereotypes about a given community. In such cases, our ability to detect condescending language, with current methods, largely depends on the presence of similar examples in the training data.

1 Introduction

Patronizing and Condescending Language (PCL) has been a topic of interest across a wide range of disciplines, including Politics, Journalism and Medicine (Huckin, 2002a; Chouliaraki, 2006; Draper, 2005; Oldenburg et al., 2015). The use of PCL implies a position of superiority of the author regarding the person or community they are referring to, suggesting an imbalance in terms of power or privilege (Foucault, 1980). Especially when directed towards vulnerable communities, PCL fuels discrimination and perpetuates inequalities (Ng, 2007; Mendelsohn et al., 2020), feeds stereotypes and misinformation (Fiske, 1993), and makes it more difficult for underrepresented groups to overcome social difficulties (Nolan and Mikami, 2013).

The NLP community has recently also turned its attention to the study of PCL, focusing on the task of detecting and categorizing this kind of harmful discourse. For instance, Wang and Potts (2019) introduced the *Talk Down* dataset, which is focused on condescending language in social media, while Perez-Almendros et al. (2020) introduced the *Don't Patronize Me!* (DPM) dataset, which is focused on the way in which vulnerable communities are described in news stories. From the NLP point of view, the study of PCL is interesting because it is more subtle, and therefore presumably harder to detect, than other forms of harmful language, such as hate speech (Basile et al., 2019) and offensive language (Zampieri et al., 2019, 2020). Moreover, identifying PCL often seems to require a deep commonsense understanding of human values (Pérez-Almendros et al., 2022). Consider the following example from the DPM dataset:

> "People across Australia ordered pizzas to be delivered on Saturday night, with the ample leftovers donated to local homeless shelters."

We can understand that, although donating food can be socially valuable, the impact of this particular action is painted in an excessively positive light (e.g. as evident in the phrase *ample leftovers*). Moreover, this seems to refer to a campaign to increase the consumption of pizzas with the excuse to help homeless people, which as humans we might also find condescending. However, an NLP model might struggle to infer such connotations.

Based on the premise that PCL detection would present unique challenges, SemEval-2022 featured a task devoted to PCL detection and categorization (Perez-Almendros et al., 2022). The top-ranked submissions for this task achieved a remarkably strong performance, which seems to somewhat

undermine the assumption that the subtle nature of PCL would make its detection inherently hard. Moreover, even the best systems (Deng et al., 2022; Wang et al., 2022; Hu et al., 2022), relied on a judicious use of more or less generic text classification techniques, improving on the RoBERTa (Liu et al., 2019) baseline by addressing the class imbalance, adding a contrastive learning loss, using ensembles of language models, etc. In particular, there was little evidence of the presumed need to focus on commonsense understanding of human values.

In this paper, we present an analysis of the SemEval-2022 PCL detection dataset, in light of the aforementioned observations. Our central argument is that the dataset contains examples of two rather distinct types of condescending language, and that the difference between the two is fundamental to understanding why the task, as it has been formulated, might be significantly easier than the task of detecting condescending language in general. We argue that a deeper understanding of these two phenomena might lead to a better performance on PCL detection, which in turn can mitigate the discourse of condescension towards vulnerable communities. We will refer to these two types as *Linguistic PCL* and *Thematic PCL*.

Linguistic PCL Some instances of PCL are related to the way in which a given claim is expressed. Consider the following example:

> "...we must rally together as humans, understanding that we have a responsibility to help the world's most vulnerable to survive and rebuild their lives [...]"

In this sentence, we can see two common aspects of PCL. First, expressions such as *we must* or *we have a responsibility*, indicate an authority voice and attitude (Simpson, 2003). Second, the sentence evokes the idea of a *saviour* and a *victim*. Note how the condescending tone of the sentence is related to linguistic aspects that are relatively easy to identify (e.g. the presence of modal verbs such as *must*) and largely independent of the community being referred to. We will refer to such cases as *linguistic PCL*. Our hypothesis is that detecting linguistic PCL is relatively straightforward for language models, as this is ultimately about learning to detect a particular writing style (Iyer and Vosoughi, 2020).

Thematic PCL There are also examples of PCL where the message itself is condescending, irrespective of how it is formulated. We will refer to such cases as *Thematic PCL*. Consider the following example:

> "The problem of what to do about the Dreamers, as the immigrants are known[...]"

Calling young immigrants *Dreamers* has condescending connotations, as it implies that the author is in a privileged position which the immigrants aspire to reach. To recognize this, we need a deeper understanding about the nature of condescending language, and we need access to particular world knowledge. For instance, we need to know that the author refers to the DREAM Act[1] and that this tries to protect young immigrants brought to the US as children and fulfill their aspiration to live in America as a *dreamed life*. Our hypothesis is that detecting themed PCL often requires a level of understanding about human values, and the world in general, that goes above what we can expect to be captured by standard language models. However, the training and test data from the SemEval task is focused on a small number of vulnerable communities, with the same communities being covered in the training and test data. As such, the model may detect instances of PCL by identifying that they express a similar argument as some training example, rather than by developing an understanding of the underlying reasons why a given example is condescending. In this case, we can expect the model to fail to detect PCL towards communities that are not seen in the training set. Similarly, the model may struggle to adapt when the themes appearing in PCL towards previously seen communities change.

Overview In this paper, we present an analysis of the SemEval-2022 dataset, aimed at testing the aforementioned hypotheses about linguistic and themed PCL. First, we carry out two experiments in which models are trained such that they are prevented, to some extent, from learning about condescending themes associated with individual communities. Our experiments show that there are some communities for which this leads to a dramatic drop in performance, while for other communities there was no negative impact at all. This suggests that there is indeed considerable overlap in the kinds of themed PCL that can be found in the training and test sets of the SemEval dataset, but only for some communities. We then complement

[1] www.americanimmigrationcouncil.org/research/dream-act-overview

these results with a qualitative analysis based on ideas from critical Discourse Analysis (CDA), a technique which emerged from Critical Linguistics in the 1970s (Fowler et al., 2018; Fairclough and Chouliaraki, 1999; Fairclough, 2013; Wodak, 2004; Van Dijk, 2015; Huckin et al., 2012). CDA looks at the relation between power and language, and how discourse expresses social hierarchy and inequalities. This qualitative analysis provides further support for the idea that (i) PCL detection models can identify Linguistic PCL even if they have not seen similar cases during training while (ii) their ability to detect instances of themed PCL is much more dependent on the training examples.

2 Related Work

The Study of PCL The discourse of condescension has been widely studied in disciplines such as Sociolinguistics, Politics, Psychology, Medicine, Cultural Studies, Public Relations, Journalism and International Cooperation (Huckin, 2002a,b; Giles et al., 1993; Margić, 2017; Chouliaraki, 2006, 2010). Within the NLP community, the study of PCL is more recent, although there is a longer tradition of looking at harmful language more generally (Basile et al., 2019; Zampieri et al., 2020; Conroy et al., 2015; Da San Martino et al., 2020; Feng et al., 2021; Farha et al., 2022). As already mentioned in the introduction, Wang and Potts (2019) and Perez-Almendros et al. (2020) addressed condescension in different types of discourse, while other recent works addressed some closely related aspects, such as how language conceals power relations (Sap et al., 2020), expresses authoritarian voices as empathy (Zhou and Jurgens, 2020) or dehumanizes minorities (Mendelsohn et al., 2020).

PCL towards vulnerable communities is a subtle and subjective kind of language, often unconscious and well intended (Wilson and Gutierrez, 1985; Merskin, 2011). An author might use PCL while trying to help a community or individual, raise their voice for them or move the audience to action. However, PCL can be very harmful, as it routinizes discrimination (Ng, 2007), creates stereotypes (Fiske, 1993) and reinforces inequalities (Nolan and Mikami, 2013; Chouliaraki, 2006, 2010), feeding the dichotomy of a *saviour* (Bell, 2013; Straubhaar, 2015) and a *helpless victim*. PCL contributes to the "distorted and stereotyped representation" (Caspi and Elias, 2011) that vulnerable communities or underrepresented groups frequently receive in the media.

The Coverage of Minorities in the Media Our emphasis on the distinction between *thematic* and *linguistic* PCL draws from previous analysis of the relation between discourse and power, and how language can reinforce inequalities and exclusion. Such studies are mainly based on Critical Discourse Analysis (CDA), which is concerned with the analysis of unbalanced power relations and privilege in public discourse and the construction of identities in the media. It also draws our attention to the influence (voluntary or not) that the author of a public discourse has over the construction of an image in the mind of the audience by, for instance, their selection of words, use of linguistic structures and omissions when depicting a specific community or situation (Huckin, 2002a). Huckin (2002a) suggests that, in the critical study of a discourse, an analyst should look for certain linguistic or stylistic features in a text, such as the use of modal verbs (modality) or the identity of the subject and the object of an action (transitivity), to find expressions of power imbalance and inequality. He also suggests to look at recurrent themes and stereotypes in the media coverage of minorities. Along this direction, the same author studied the treatment of homelessness in the US in 1999 (Huckin, 2002b). He collected a corpus of 163 newspapers articles and editorials which mentioned the keyword *homeless* and analyzed, among others, the more recurrent themes and stereotypes related to this community. For instance, he shows that the analyzed data includes "desire of independence" or "lack of life skills" as common themes when referring to causes of homelessness. Also, the theme "bad grooming" is highlighted as one effect of homelessness. "Religious support", "food donation" and "donated clothes" are common themes in the discussion of public responses, which represent shallow and ephemeral solutions for a structural, deep-rooted problem, and thus again reinforce the charitable, *saviour-victim* treatment of a community. Using a similar approach, Díaz-Rico (2012) analyzed 93 articles about Mexican immigrants from the Los Angeles Times, published in 2010. She claims that the selection of topics and themes is the most important aspect of Journalism and that newspapers use the drama of a story to gain attention from their audience. Although the language and topics she analyses in this work are often openly discriminatory and offensive, she also finds expres-

	Neg.Inst.	Pos.Inst.	%Pos.Inst.
Migrant	1052	36	3.3
Immigrant	1031	30	2.8
Refugee	981	86	8.1
In need	906	176	16.3
Poor fam.	759	150	16.5
Vulnerable	1000	80	7.4
Women	1018	52	4.9
Disabled	947	81	7.9
Homeless	899	178	16.5
Hopeless	881	124	12.3
All data	9474	993	9.5

Table 1: Number of negative and positive training examples per community. We also report the percentage of positive instances.

	Neg.Inst.	Pos.Inst.	%Pos.Inst.
Migrant	359	12	3.2
Immigrant	383	17	4.3
Refugee	390	26	6.3
In need	357	42	10.5
Poor fam.	267	56	17.3
Vulnerable	382	18	4.5
Women	390	22	5.3
Disabled	308	24	7.2
Homeless	337	57	14.5
Hopeless	342	43	11.2
All data	3515	317	8.3

Table 2: Number of negative and positive test examples per community. We also report the percentage of positive instances.

sions that, through rhetorical figures, connotation and semantic selection, reinforce power relations and inequalities (e.g.“help new arrivals get on their feet”, or "ballot crusade").

3 Methodology

In Sections 4 and 5, we describe experiments in which PCL classifiers are trained in a way that (partially) prevents them from learning about community-specific thematic PCL. This will allow us to better characterise the abilities of fine-tuned language models, as the overlap between the themes covered by the training and test sets is reduced. In this section, we first describe the basic experimental setup that we rely on throughout the paper (Section 3.1). Subsequently, we describe a simple strategy for characterizing topics or themes that are strongly associated with particular vulnerable communities or groups (Section 3.2).

3.1 Experimental Setup

Dataset We use the dataset that was provided for the Patronizing and Condescending Language Detection Task at SemEval-2022 (Perez-Almendros et al., 2022). This dataset consists of 14,299 annotated paragraphs (10,467 for training and 3,832 for testing). The paragraphs were extracted from English news stories and cover traditionally vulnerable communities and underrepresented groups. In particular, each paragraph mentions at least one of the following vulnerability-related keywords: *immigrants*, *migrants*, *refugees*, *poor families*, *in need*, *hopeless*, *homeless*, *disabled*, *women* and *vulnerable*. We only use the binary labels from the dataset, i.e. whether a paragraph is considered to contain PCL or not[2]. We show the number of positive and negative instances for each community for the training data in Table 1 and for the test data in Table 2.

Training Details For our experiments, we fine-tune RoBERTa-base (Liu et al., 2019) on different versions of the training set. While better results have been reported for RoBERTa-large and DeBERTa (Hu et al., 2022; Deng et al., 2022; Wang et al., 2022), we found the results with RoBERTa-base to be more stable across different runs, which is more important than the absolute level of performance for the analysis in this paper. We train our models for 5 epochs, using the Transformers library (Wolf et al., 2020). We use AdamW with a learning rate of 1e-5 and a batch size of 4. All the reported results have been averaged over 5 runs. As can be seen in Tables 1 and 2, the SemEval dataset is highly imbalanced, with 9,474 negative and 993 positive cases of PCL. For this reason, when training the language model, we down-sample the negative cases to 5,000 and over-sample the positive cases five times.

3.2 Community-Related Terms

We associate each of the vulnerable communities from the SemEval dataset with a set of terms, which essentially describe the topics or themes that are specific to, or at least strongly related to, that community. To associate terms with a given community, we compare the set of paragraphs, from the SemEval dataset, in which the keyword associated

[2]The dataset also includes a categorisation of positive examples according to the taxonomy from Perez-Almendros et al. (2020), as well as labels which indicate the level of inter-annotator agreement for a given example.

Community	Associated terms
Immigrants	First-generation, resentment, cultures, foreign-born, undocumented, sentiment, spouses, applicant
Migrants	Hatred, incoming, dreamers, coast, trafficking, racism, protections, deported, gangs, rescued
Refugees	Repatriation, offshore, queer, seekers, resettlement, camps, fled, abuses, mercy, forget
In need	Donor, desperately, Christ, drought, kindness, foster, budgets, compassionate, humanitarian, blankets
Poor families	Diapers, nutritious, scholarship, rice, poverty, expenses, savings, malnutrition, babies, orphans
Vulnerable	Droughts, prey, strategies, hub, resilience, crop, proactive, exploitation, fragile, hazards
Women	Feminist, maternity, abortions, husbands, beauty, fertility, unsafe, empowering, motivated, honour
Disabled	Assistive, pension, impaired, heroes, integrating, consideration, allowance, disadvantaged, begging
Homeless	Downpour, jobless, addicts, evicted, shelters, hungry, streets, rough, roofs, soup

Table 3: Selection of terms found for the different communities, with $k = 100$.

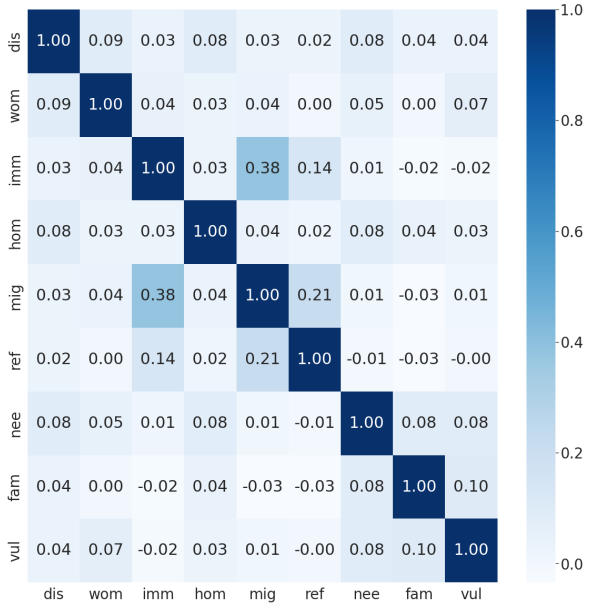

Figure 1: Similarity between the different communities from the SemEval dataset. The communities are identified by the following keywords: disabled (dis), women (wom), immigrants (imm), homeless (hom), migrants (mig), refugees (ref), in need (nee), poor families (fam) and vulnerable (vul).

with that community is mentioned (e.g. *homeless*) with the remaining paragraphs. We first select those terms that are mentioned in at least five paragraphs for the considered community. Then we rank these terms according to Pointwise Mutual Information (PMI), i.e. by comparing how strongly the presence of a given term (e.g. *addicts*) is associated with the presence of the community keyword (e.g. *homeless*). Finally, we select the top-k highest ranked terms for each community, where we have considered $k = 100$ and $k = 500$ in our experiments. Note that the selected terms are not necessarily indicative of PCL. However, even for $k = 100$ we observed that many of the selected terms reflect stereotypes and condescending attitudes. Table 3 shows a selection of terms that were found for $k = 100$.

Finally, we analyse to what extent the ten keywords from the SemEval dataset refer to distinct communities. To this end, we represent each keyword/community as a PMI-weighted bag-of-words vector. Figure 1 displays the cosine similarities between the vectors we obtained for the different communities. As can be seen, and somewhat unsurprisingly, there is a high degree of overlap between *migrants* and *immigrants*. For this reason, these two communities/keywords will be merged for the analyses in this paper. We can furthermore see that *migrants* and *refugees* are also somewhat similar in the dataset, but since the similarity between *immigrants* and *refugees* is much lower, we keep *refugees* as a separate community. Note that we omitted the keyword *hopeless* in Figure 1, as we found this keyword to be too generic to be viewed as describing a particular community. For this reason, we will not consider this keyword in our community-specific experiments and analysis.

4 Omitting Community-Specific Training Data

Our main hypothesis, as outlined in the introduction, is that the SemEval PCL detection task is easier than one might expect because it involves a combination of linguistic PCL, which is easier to detect, and thematic PCL. While we believe that thematic PCL can be hard to detect in general, our hypothesis is that it is simplified, in the context of the SemEval dataset, because of the overlap between the themes covered in the training and test data. If a language model is truly able to recognize PCL, then it should be capable of identifying (thematic) PCL about communities it has not seen during training. In this section, we report the results of an experiment where we test the performance of the model per community in two settings. First, we consider the standard setting, where the model

has had access to the entire training set. Second, we consider the setting where all examples about the community being tested were removed from the training set. Note that for the latter case, we need to train a separate model for every community, each time omitting the corresponding training examples.

	Full Training	Comm. Omitted
Migr. + Imm.	**43.6**$_{\pm 7.89}$	25.3$_{\pm 3.27}$
Refugees	50.4$_{\pm 8.36}$	**54.0**$_{\pm 5.12}$
In need	**55.3**$_{\pm 3.12}$	51.2$_{\pm 1.04}$
Poor families	52.7$_{\pm 6.34}$	**53.7**$_{\pm 7.18}$
Vulnerable	**54.7**$_{\pm 3.75}$	51.6$_{\pm 3.29}$
Women	31.5$_{\pm 8.79}$	**41.7**$_{\pm 7.53}$
Disabled	**54.6**$_{\pm 5.52}$	52.4$_{\pm 3.85}$
Homeless	**60.2**$_{\pm 1.85}$	54.4$_{\pm 2.49}$
All communities	53.2$_{\pm 2.54}$	-

Table 4: Performance of RoBERTa-base models fine-tuned with (Full Training) and without (Comm. Omitted) training examples about the test community. Result are reported in terms of F1-score % and are averaged over 5 runs. We also report the standard deviation.

The results are summarized in Table 4. We can make a number of clear observations. First, the performance of the model that was trained on the full training set varies substantially across the different communities. For instance, the F1 score for *homeless* is almost twice as high as that for *women*. Second, excluding training examples about the test community has a substantial impact on the results for some communities, but not for others. For *migrants + immigrants*, we can see a particularly large drop in performance, which suggests that PCL towards this community is more likely to be thematic than for the other communities. For some of the other communities, we also see drops, although these are much smaller. Surprisingly, for some communities, the performance improves when omitting training examples from that community, which is most pronounced for *women*. This suggests that PCL towards women is more likely to be linguistic (and thus community-independent), while the model may have learned incorrect associations from the themes that are present in the training examples about women. This will be further explored in the qualitative analysis.

5 Masking Community-Specific Terms

We now present a variant of the experiment from the previous section, where no training examples are removed, but we instead mask (some) occurrences of community-related terms, as identified in Section 3.2, in the training data. Note that we mask occurrences of such terms regardless of the community a training example is about (e.g. a term that was identified for *refugees* would still be masked in examples about *immigrants*). This setup has the advantage that the number of training examples remains constant. Moreover, the model may now also be prevented from learning thematic PCL by training on related communities. For instance, in the setting from Section 4, the model may be able to learn condescending themes about the *homeless* community from training examples mentioning the *vulnerable* keyword.

The results are reported in Table 5, where the masking probability for mentions of community-related terms is varied from 0% to 100%. The main findings from Section 4 are confirmed by this experiment. In particular, for *migrants + immigrants*, we find that masking community-related terms leads to a substantial drop in performance (especially when 100% of the mentions are masked). This again suggests that the classifier, in the standard setting, heavily relies on the fact that condescending themes from the test set are also present in the training set. For *women*, we can see that masking can improve the results, which again suggests that the type of PCL for this community is mostly linguistic. In fact, for all but two communities, the best overall results are obtained with some degree of masking. This suggests that linguistic PCL is prevalent across the dataset, and that the fine-tuned RoBERTa-base model is susceptible to lean incorrect associations between thematic terms and the presence of PCL.

6 Qualitative Analysis

The experiments in Sections 4 and 5 have revealed stark differences in the robustness of PCL detection models across different communities, when the model is (partially) prevented from learning community-specific themes during training. In particular, our results suggest that PCL examples for *migrants + immigrants* are often thematic in nature, with the same themes recurring in both the training and test sets. Conversely, the results for *women* suggest that PCL towards that community is more likely to be linguistic in nature. In this section, we supplement our findings with a qualitative analysis, where we focus on these two communities.

	Top-100 community-based terms					Top-500 community-based terms					Baseline
	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	0%
Migr. + Imm.	27.7	38.0	31.6	35.7	40.0	25.2	34.3	42.3	36.0	34.9	**43.6**
Refugees	49.9	50.1	47.1	**52.2**	**53.0**	49.6	49.5	48.1	48.5	**53.5**	50.4
In need	**55.6**	55.2	**55.8**	**56.5**	**58.6**	**56.9**	54.7	**58.6**	**57.1**	55.1	55.3
Poor families	**55.9**	**57.5**	52.0	47.8	52.7	51.7	52.2	52.1	50.2	46.6	52.7
Vulnerable	54.3	**56.8**	52.7	**57.5**	**55.8**	48.4	47.5	**56.3**	54.1	52.3	54.7
Women	31.0	**37.6**	**39.3**	**41.0**	**39.7**	**38.2**	**39.8**	**39.9**	**39.5**	**35.9**	31.5
Disabled	51.8	49.3	52.4	48.7	48.0	45.8	46.3	54.4	52.1	53.0	**54.6**
Homeless	58.5	58.4	57.8	57.7	**62.1**	54.6	54.9	**61.3**	60.0	57.9	60.2
All communities	52.3	**53.4**	51.6	52.5	**53.9**	51.2	50.7	**54.6**	52.9	52.3	53.2

Table 5: Performance of RoBERTa-base models fine-tuned on variants of the training set in which community-related terms are masked. Results are shown with the $k = 100$ and the $k = 500$ top terms from each community, and with varying masking probabilities. Configurations which outperform the baseline (i.e. the setting where the original training set is used) are shown in bold, while the best overall result for each community is underlined. Result are reported in terms of F1-score % and are averaged over 5 runs. The standard deviation is reported in Appendix A

.

Migrants + Immigrants In Table 6, we can see examples of PCL which were consistently[3] classified correctly when including the community in the training set, but where the model was unable to recognise the PCL when trained without examples from the test community. Therefore, these are paragraphs where we would expect to see community-related themes that make the message condescending. Note that the word *Dreamer* is present in all the examples from this table. It thus seems safe to infer that the model has learned that this term is highly predictive of the presence of PCL, when such examples are included in the training data. The use of other terms such as *deportation*, *undocumented* or *citizenship* are also strongly related to the community and might help the model to identify the presence of PCL.

In contrast, the examples of PCL in Table 7 were consistently identified correctly, whether the training examples for *migrant + immigrant* were included or not. As expected, we can indeed think of these examples as being primarily *linguistic PCL*, in the sense that what makes them condescending is *how* the message is expressed, more than *what* is being expressed. For instance, in the first example we can see an excess of flowery wording and adjectives to express a message, the use of metaphors and an almost poetic style to describe a vulnerable situation, which are common features of PCL (Perez-Almendros et al., 2020). The second and third examples also show clear differences in power and privilege, for instance, through the use of expressions such as *we have a moral responsibility*, *show them solidarity* or *permitting them to work and study without fear*. The last example conveys a distance between the author and the community (*breaking through the barrier of migrant communities*) and expresses presuppositions and an authority voice based on the idea of a *saviour-victim* relation (*I grapple with this*, *I'm trying to help, to make things better, but many women find comfort in the norms and the way things are*). These examples of *linguistic PCL* are independent of the community they are addressing, which is why the model still recognises them even when no training examples for the *migrants + immigrants* community are provided.

Women Table 8 shows examples of PCL that were missed when using the full training set, but consistently classified correctly when omitting *women* examples. In the first paragraph, the phrase *their shame continues*, a community-independent value judgement, makes the text condescending. The second and third example express a *saviour-victim* relation, where the differences between power and vulnerability, as well as an admiration towards the *saviour*, are explicitly stated. As these examples are clearly linguistic, we can expect that a model which has not seen *women* examples should be able to classify them correctly. Surprisingly, all three paragraphs were missed by the model that was trained on the full training data. To understand why this is the case, note that 95% of the training examples for *women* are negative. As a result, several of the terms that are associated

[3]We focus on cases where the classification is consistent across different runs of our experiments, i.e. with different random seeds, to reduce the influence of instances that were classified correctly or incorrectly by chance.

Classified correctly only with full training set
On the campaign trail, **Trump** promised to **deport** all **undocumented** migrants. Since taking office, he appeared to soften on **dreamers**, a relatively well-educated and industrious group who he described as "incredible kids"
But without resolution, the centrists warn they will have enough petition signatures by Tuesday to force House votes later this month, including on their preferred bill which provides young **"Dreamer"** immigrants **protection** from **deportation** and a chance to apply for **citizenship**.
Passage of the measure came over the opposition of Democratic leaders who demanded the promise of a vote to protect **"Dreamer"** immigrants brought to the country illegally as children. A band of tea party Republicans was also against the legislation over what it sees as spiralling spending levels.
The New York senator said he was hopeful about talks on so-called **Dreamers**, more than 700,000 young immigrants brought to the US as children who were **protected** under the Obama-era Deferred Action for Childhood Arrivals **(Daca)** programme.

Table 6: Examples of PCL for *migrants + immigrants*, which are consistently classified correctly when the model is trained on the full training set, but consistently misclassified when training examples about this community are excluded from the training set. In bold, we highlight some community-specific themes that are common in examples of PCL, which the model is unable to learn when not presented with similar examples during training.

Classified correctly even without community-specific training examples
The Irish famine led to a massive influx of Irish immigrants to New York during the late 1840s and 1850s. As the **downtrodden** Irish escaped the famine in their home country, however, **they came to a place where life was just as tough**. Disembarking from **coffin ships**, Irish newcomers were **greeted with a new life of hardship, slums and tough, endless labor**.
Vatican City: As record numbers of people flee conflict, persecution and poverty, governments, citizens and the Church **have a moral obligation** to safeguard migrants and **show solidarity** with them, the Pope has said.
Barack Obama implemented the DACA program five years ago to **help bring** the children of undocumented immigrants **out of the shadows** of illegality, **permitting them** to study and work **without fear**.
It's been hard **breaking through the barrier of migrant communities**. Many women from my own community do not take my work seriously and do not support it, and **I grapple with this**. **I'm trying to help, to make things better, but many women find comfort in the norms and the way things are**.

Table 7: Examples of PCL for *migrants + immigrants*, which are consistently classified correctly both when including or excluding the community from the training set. in bold, we highlight the presence of some common linguistic features of PCL.

Classified correctly only without community-specific training examples
Many of these women now lie in unmarked graves, a situation that is slowly being rectified by the work of the voluntary Justice for Magdalenes Group. **Their shame continues**.
However, "when **a major male rock star who could do anything at all with his life decides to focus on the rights of women and girls worldwide** - well, all **that's worth celebrating**. We're proud to name that rock star, Bono, our first Man of the Year," it said.
A Cosmopolitan spokesperson says with a focus on empowerment, **the magazine is "proud of all that the brand has achieved for women around the world"**.

Table 8: Examples of PCL for *women*, which are classified correctly only when excluding the community from the training set. In bold, we highlight the presence of some common linguistic features of PCL.

with women (almost) exclusively appear in negative training examples. This can lead the model to believe that these words are indicative of a lack of PCL. By masking community-related terms, or omitting training examples from this community entirely, we can prevent the model from learning such coincidental associations.

7 Conclusions

We have studied the challenge of detecting Patronizing and Condescending Language (PCL), with the aim of improving our understanding of its na-

Classified correctly only with partial masking
"Eleven months into his administration, the country is showing signs of progress in most sectors of the economy. With the implementation of the free senior high school programme, most students, **especially those from poor families**, **who hitherto would not have progressed to the senior high school**, **have the opportunity now to receive** secondary education **to make them better and more functional in society**", Dr Nyarko said.
Today, Brooklyn is home to people of all races, most struggling to make ends meet. Council flats continue to degrade as the population swells – **unemployment and homelessness sees people of different races** lining up side-by-side **for a plate of free food**. It's a representation of **the rainbow nation in trauma**, **with its colours dulled and blended together by suffering**.
Helping refugee children fit in a bonus for Juventus football camp.
Swimming superstar Adam Peaty is set to unveil **a new motorbike for charity** in memory of schoolgirl Imogen Evans, who used the service. The Shropshire and Staffordshire Blood Bikes is a charity which **saves lives** by delivering vital blood supplies to those in need.
RADIO Veritas, the leading faith-based AM station in Mega Manila, continues **its commitment to charity and public service** through an initiative dubbed as "**Good Samaritan**". Since it was launched last June 2017 (airing every Monday to Friday from 1-2 p.m.), Radio Veritas has listed 182 cases of pleads and requests that have been fulfilled through this program. It serves as a platform **for those in need to make on-air appeals** for legal, spiritual, medical, material and financial assistance, and **link them to "Good Samaritans" who are willing to share**.

Table 9: Examples of PCL for different communities which are consistently classified correctly when partially masking community-related terms, but that are missed when training either on all data or removing all the community-specific training examples.

ture. We highlighted the distinction between two types of PCL. On the one hand, linguistic PCL is concerned with how the message is expressed and is largely community-independent. On the other hand, thematic PCL is more concerned with the message itself, and often relates to aspects that are highly community-specific. Our analysis suggests that for some communities, instances of PCL are mostly linguistic, while for other communities, thematic PCL is more prevalent. Moreover, detecting thematic PCL remains highly challenging in settings where the training data does not include examples covering similar themes. A better understanding of these phenomena can help future work to improve the detection of PCL and, eventually, contribute to more responsible and inclusive communication. As a first step, we envisage that a more fine-grained annotation of PCL detection datasets will be needed, distinguishing between (sub-categories of) linguistic and thematic PCL, to help us train better models and allow for a more insightful evaluation.

8 Ethical and societal implications

With our study of Patronizing and Condescending Language towards vulnerable communities we aim at contributing to more ethical communication. PCL is more subtle and subjective than other kinds of harmful language, such as hate speech or offensive language, but equally damaging, especially when spread by the media. Crucially, the use of PCL is often unintentional, hence developing tools that flag instances of PCL, which could work similarly to spelling and grammar checkers, can bring about meaningful change. This makes PCL detection an important social challenge that should be addressed by the NLP community. Although recent works have shown that fine-tuned language models can identify PCL to some extent, this paper tries to deepen our understanding of the nature of this kind of language,and of the fundamental challenges that still remain to be solved in this area. Among the limitations of this work, we include the small size of the analyzed dataset, as well as the limited number of communities that are covered.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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.

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

Dan Caspi and Nelly Elias. 2011. Don't patronize me: media-by and media-for minorities. *Ethnic and Racial Studies*, 34(1):62–82.

Lilie Chouliaraki. 2006. *The spectatorship of suffering*. Sage.

Lilie Chouliaraki. 2010. Post-humanitarianism : Humanitarian communication beyond a politics of pity. *International Journal of Cultural Studies*.

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

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414.

Yong Deng, Chenxiao Dou, Liangyu Chen, Deqiang Miao, Xianghui Sun, Baochang Ma, and Xiangang Li. 2022. Beike nlp at semeval-2022 task 4: Prompt-based paragraph classification for patronizing and condescending language detection. *arXiv preprint arXiv:2208.01312*.

Lynne Díaz-Rico. 2012. *Tools for Discourse Analysis*, pages 149–159. SensePublishers, Rotterdam.

Peter Draper. 2005. Patronizing speech to older patients: A literature review. *Reviews in Clinical Gerontology*, 15(3-4):273–279.

Norman Fairclough. 2013. *Language and power*. Routledge.

Norman Fairclough and Lilie Chouliaraki. 1999. Discourse in late modernity.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at semeval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104.

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.

Roger Fowler, Bob Hodge, Gunther Kress, and Tony Trew. 2018. *Language and control*. Routledge.

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

Dou Hu, Zhou Mengyuan, Xiyang Du, Mengfei Yuan, Jin Zhi, Lianxin Jiang, Mo Yang, and Xiaofeng Shi. 2022. PALI-NLP at SemEval-2022 task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 335–343, Seattle, United States. Association for Computational Linguistics.

Thomas Huckin. 2002a. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, 155:176.

Thomas Huckin. 2002b. Textual silence and the discourse of homelessness. *Discourse & Society*, 13(3):347–372.

Thomas Huckin, Jennifer Andrus, and Jennifer Clary-Lemon. 2012. Critical discourse analysis and rhetoric and composition. *College composition and communication*, pages 107–129.

Aarish Iyer and Soroush Vosoughi. 2020. Style change detection using bert. In *CLEF (Working Notes)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Branka Drljača Margić. 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.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization.

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

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.

Jan Oldenburg, Jorge Aparicio, Jörg Beyer, Gabriella Cohn-Cedermark, M Cullen, T Gilligan, U De Giorgi, Maria De Santis, Ronald de Wit, SD Fosså, et al. 2015. Personalizing, not patronizing: the case for patient autonomy by unbiased presentation of management options in stage i testicular cancer. *Annals of Oncology*, 26(5):833–838.

Carla Perez-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.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. Pre-training language models for identifying patronizing and condescending language: An analysis. *LREC*.

Carla Perez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 298–307, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Association for Computational Linguistics*.

Paul Simpson. 2003. *Language, ideology and point of view*. Routledge.

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.

Teun A Van Dijk. 2015. Critical discourse analysis. *The handbook of discourse analysis*, pages 466–485.

Ye Wang, Yanmeng Wang, Baishun Ling, Zexiang Liao, Shaojun Wang, and Jing Xiao. 2022. PINGAN omini-sinitic at SemEval-2022 task 4: Multi-prompt training for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 313–318, Seattle, United States. Association for Computational Linguistics.

Zijian Wang and Christopher Potts. 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*.

Clint C Wilson and Felix Gutierrez. 1985. Minorities and the media. *Beverly Hills, CA, London: Sage*.

Ruth Wodak. 2004. Critical discourse analysis. *Qualitative research practice*, 185:185–204.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, 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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

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

A Appendix: Standard deviation for Table 3.

	Top-100 community-based terms					**Top-500 community-based terms**					**Baseline**
	100%	80%	60%	40%	20%	100%	80%	60%	40%	20%	0%
Migr. + Imm.	±4.76	±8.78	±8.56	±6.83	±3.02	±6.82	±6.89	±6.47	±7.24	±6.77	±7.89
Refugees	±3.17	±6.11	±2.81	±3.76	±3.61	±7, 72	±1, 60	±2, 56	±5, 30	±4, 85	±8.36
In need	±1.19	±1.21	±1.87	±1.45	±3.77	±1.10	±1.65	±2.44	±2.80	±3.46	±3.12
Poor families	±3.31	±03.05	±6.93	±6.24	±4.89	±3.90	±5.92	±5.60	±4.44	±2.62	±6.34
Vulnerable	±6.28	±4.80	±7.90	±3.70	±6.27	±3.33	±2.26	±6.12	±5.35	±2.47	±3.75
Women	±9.92	±5.44	±4.91	±2.74	±3.97	±2.60	±06.05	±8.62	±4.76	±7.10	±8.79
Disabled	±2.81	±5.59	±5.23	±2.42	±4.52	±3.15	±02.06	±4.43	±6.51	±4.54	±5.52
Homeless	±0.79	±2.94	±2.64	±5.22	±1.95	±1.86	±2.63	±03.01	±1.84	±5.74	±1.85
All communities	±1.49	±2.15	±1.70	±1.39	±0.87	±3.59	±1.15	±2.59	±2.59	±1.97	±2.54

Table 10: Standard deviation for Table5 over 5 runs.