Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities

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

Abstract

In this paper, we introduce a new annotated dataset which is aimed at supporting the development of NLP models to identify and categorize language that is patronizing or condescending towards vulnerable communities (e.g. refugees, homeless people, poor families). While the prevalence of such language in the general media has long been shown to have harmful effects, it differs from other types of harmful language, in that it is generally used unconsciously and with good intentions. We furthermore believe that the often subtle nature of patronizing and condescending language (PCL) presents an interesting technical challenge for the NLP community. Our analysis of the proposed dataset shows that identifying PCL is hard for standard NLP models, with language models such as BERT achieving the best results.

1   Introduction

In this paper, we analyze the use of Patronizing and Condescendig Language (PCL) towards vulnerable communities in the media. An entity engages in PCL when its language use shows a superior attitude towards others or depicts them in a compassionate way. This effect is not always conscious and the intention of the author is often to help the person or group they refer to (e.g. by raising awareness or funds, or moving the audience to action). However, these superior attitudes and a discourse of pity can routinize discrimination and make it less visible (Ng, 2007). Moreover, general media publications reach a large audience and we believe that unfair treatment of vulnerable groups in such media might lead to greater exclusion and inequalities.

While there has been substantial work on modelling language that purposefully undermines others, e.g. offensive language or hate speech (Zampieri et al., 2019; Basile et al., 2019), the modelling of PCL is still an emergent area of study in NLP. Some reasons for this might include that the use of PCL in the media is commonly unconscious, subtler and more subjective than the types of discourse that are typically targeted in NLP. Specifically, a special focus in PCL towards vulnerable communities has not been yet considered, to the best of our knowledge.

Within a broader setting, there has been some work on PCL which is concerned with the communication between two parties, where one is patronized by the other, such as in social media interactions. In particular, Wang and Potts (2019) recently published the Talkdown corpus for condescension detection in comment-reply pairs from Reddit. In this work, the authors highlight the difficulty of the task and the need for a high-quality dataset annotated by experts, which is the approach we take for studying PCL towards vulnerable communities.

To encourage more research on detecting PCL language, we introduce the *Don't Patronize Me! dataset*[1]. This dataset contains more than 10,000 paragraphs extracted from news stories, which have been annotated to indicate the presence of PCL at the text span level. The paragraphs were selected to cover English language news sources from 20 different countries, covering different types of vulnerable communities (e.g. homeless people, immigrants and poor families). We furthermore propose a taxonomy

---

[1]Available at `https://github.com/Perez-AlmendrosC/dontpatronizeme`.

of PCL categories, focused on PCL towards vulnerable communities. Each of the PCL text spans from our dataset has been annotated with a category label from this taxonomy. Finally, we also provide some analysis of the dataset. Among others, we find that even simple baselines are able to detect PCL to some extent, which suggests that this task is feasible for NLP systems, despite the subtle nature of PCL. On the other hand, we also find that the considered models, including approaches based on BERT (Devlin et al., 2019), struggle to detect certain categories of PCL, suggesting that there is still considerable room for improvement. In particular, while some forms of PCL can be detected by identifying relatively simple linguistic patterns, many other cases seem to require a non-trivial amount of world knowledge.

2 Related Work

Condescending and patronizing treatment has been widely studied in various fields, such as language studies (Margić, 2017), sociolinguistics (Giles et al., 1993), politics (Huckin, 2002) or medicine (Komrad, 1983). Within NLP, there has been extensive work on several forms of harmful language, but this work has generally focused on explicit, aggressive and flagrant phenomena such as fake news detection (Conroy et al., 2015); trust-worthiness prediction and fact-checking (Atanasova et al., 2018; Atanasova et al., 2019); modeling offensive language, both generic (Zampieri et al., 2019), and geared towards specific communities (Basile et al., 2019); or rumour propagation (Derczynski et al., 2017). Recently, however, some work on condescending language has started to appear. For instance, Wang and Potts (2019) introduced the task of modelling condescension in direct communication from an NLP perspective, and developed a dataset with annotated social media messages. In the same year, Sap et al. (2019) discussed the social and power implications behind certain uses of language, an important concept in the unbalanced power relations that are often present in condescending treatment. Also related to unfair treatment of underprivileged groups, Mendelsohn et al. (2020) analyzed, from a computational linguistics point of view, how language has dehumanized minorities in news media over time.

3 Background on PCL

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

- it fuels discriminatory behaviour by relying on subtle language (Mendelsohn et al., 2020);

- it creates and feeds stereotypes (Fiske, 1993), which drive to greater exclusion, discrimination, rumour spreading and misinformation (Nolan and Mikami, 2013);

- it strengthens power-knowledge relationships (Foucault, 1980), positioning one community as superior to others;

- it usually calls for charitable action instead of cooperation, so communities in need are presented as passive receivers of help, unable to solve their own problems and waiting for a *saviour* to help them out of their situation (Bell, 2013; Straubhaar, 2015);

- it tends to avoid stating the reasons for very deep-rooted societal problems, by concealing those responsible or even, in some cases, by apportioning blame to the underprivileged communities or individuals themselves;

- it proposes ephemeral and simple solutions (Chouliaraki, 2010), which oversimplify the wicked problems (Head and others, 2008) vulnerable communities face.

The use of PCL makes it more difficult for vulnerable communities to overcome difficulties and reach total inclusion (Nolan and Mikami, 2013).

3.1 How to identify PCL?

In this work, we analyze discourse on vulnerable communities. We will consider a piece of text as containing PCL when, referring to an underprivileged individual or community, we can identify one or several of the following traits:

- The use of the language states the differences between the *'us'* and the *'them'*. The vulnerable community is depicted as different to *us*, with other experiences and life stories. This discourse establishes an invisible distance between the two communities.

- The language raises a feeling of pity towards the vulnerable community, for example by using (or abusing) adjectives or by recurring to flowery words to depict a certain situation in a literary way (i.e., metaphors, euphemisms or hyperboles).

- The author and the community they belong to are presented as *saviours* of those in need. Not only do they have the capacity to solve their problems, but also a moral responsibility to do so. The superior or privileged community is also presented as having the knowledge and experience to face and solve the problems of the vulnerable ones.

- In the opposite direction, the members of the vulnerable community are described as lacking the privileges the author's community enjoys, or even the knowledge or experience to overcome their own problems. They will need, therefore, the help of others to improve their situation.

- The vulnerable community and its members are presented either as victims (i.e. overwhelmed, victimized or pitied) or as heroes just because of the situation they face.

3.2 What is not PCL?

Precisely because we are studying the discourse towards vulnerable communities, it can be easy to classify a piece of text as condescending mistakenly. We want to highlight, in particular, the following two situations where the language that is used to talk about unprivileged groups is not condescending.

- Because they are experiencing vulnerability, the news about them often depicts rough situations. The description of an extreme situation can be harsh and stark and leave the reader with a feeling of sadness and helplessness, while not necessarily being condescending.

- With PCL, the superiority of the author is concealed behind a friendly or compassionate approach towards the situation of vulnerable communities. Thus, a message which is openly offensive, aggressive or containing prejudiced, discriminatory or hate speech is not considered to be PCL for the purpose of our dataset.

4 The Don't Patronize Me! dataset

The Don't Patronize Me! dataset currently contains 10,637 paragraphs about potentially vulnerable social groups. These paragraphs have been selected from general news stories and have been annotated with labels that indicate the type of PCL language that is present, if any. The paragraphs have been extracted from the News on Web (NoW) corpus[2] (Davies, 2013). To this end, we first selected ten keywords related to potentially vulnerable communities widely covered in the media and susceptible of receiving a condescending or patronizing treatment: *disabled*, *homeless*, *hopeless*, *immigrant*, *in need*, *migrant*, *poor families*, *refugee*, *vulnerable* and *women*. Next, we retrieved paragraphs in which these keywords are mentioned, choosing a similar number of paragraphs for each of the 10 keywords and each of the 20 English speaking countries that are covered in the NoW corpus. An overview of the number of paragraphs for each keyword-country combination can be found in Table 1. All the selected paragraphs come from news stories that were published between 2010 and 2018. The data was annotated by three expert annotators, with backgrounds in communication, media and data science. Two annotators annotated the whole dataset (*ann1* and *ann2*), while the third one (*ann3*) acted as a referee to provide a final label in case of disagreements. An extended data statement (Bender and Friedman, 2018) about the corpus will be published together with the dataset.

---

[2]The corpus is used with the permission of its author.

	dis	hom	hop	imm	need	mig	poor	ref	vul	wom	**Total**
Australia	56	51	52	56	57	57	54	54	60	55	552
Bangladesh	51	57	46	50	51	56	46	52	55	53	517
Canada	53	53	52	51	52	47	55	56	61	52	532
Ghana	62	55	57	56	51	58	25	53	54	55	526
Hong Kong	60	58	32	53	55	59	22	49	52	61	501
Ireland	61	49	55	58	58	58	36	58	48	55	536
India	53	52	62	60	57	52	52	58	59	50	555
Jamaica	53	62	47	56	58	51	11	54	50	51	493
Kenya	52	51	55	56	51	54	55	49	57	61	541
Sri Lanka	53	57	57	59	48	53	32	56	49	50	514
Malaysia	58	48	47	54	62	58	53	58	60	56	554
Nigeria	55	60	49	52	53	56	49	56	60	55	545
New Zealand	63	49	61	51	50	56	51	49	49	47	526
Philipines	61	56	56	48	54	59	53	51	55	52	545
Pakistan	50	55	51	51	58	57	57	56	54	56	545
Singapore	51	56	53	57	59	59	54	45	54	50	538
Tanzania	51	55	18	53	50	51	38	48	53	51	468
UK	55	50	47	55	56	53	59	58	58	51	542
United States	58	60	54	51	54	55	53	61	47	58	551
South Africa	60	54	63	58	55	54	59	50	47	56	556
Total	1116	1088	1014	1085	1089	1103	914	1071	1082	1075	10637

Table 1: Number of paragraphs per keyword and country in the dataset. The considered keywords are disabled (dis), homeless (hom), hopeless (hop), immigrant (imm), in-need (need), migrant (mig), poor-families (poor), refugees (ref), vulnerable (vul) and women (wom).

4.1 Categories of PCL towards vulnerable communities

For all text spans that were annotated as containing PCL, the annotators also provided a category label. This allows us to analyze at a finer-grained level to what extent NLP models are able to recognize the different traits of PCL. These labels might also make it easier to train NLP models for detecting PCL, for instance by treating them as privileged information during training (Vapnik and Vashist, 2009). Inspired by the characteristics of PCL discussed in Section 3, we have used the following seven categories, which we grouped into three higher-level categories.

- **The saviour.** The community which the author and the majority of the audience belong to is presented in some way as *saviours* of those vulnerable or in need. The language used subtly positions the author in a better, more privileged situation than the vulnerable community. They express the will to help them, from their superior and advantageous position. There is a clear difference between the *we* and the *they*. As part of this category, we can find examples of the following subcategories:

  - **Unbalanced power relations.** By means of the language, the author distances themselves from the community or the situation they are talking about, and expresses the will, capacity or responsibility to help them. It is also present when the author entitles themselves to give something positive to others in a more vulnerable situation, especially when what the author *concedes* is a right which they do not have any authority to decide to give.

    (i.e. *'You can make a difference in their lives'* or *'They come back in with nothing and we need to outfit them again'* or *'They deserve another opportunity'* or *'They also have the right to love'*).

  - **Shallow solution.** A simple and superficial charitable action by the privileged community is presented either as life-saving/life-changing for the unprivileged one, or as a solution for a deep-rooted problem.

    (i.e. *'Raise money to combat homelessness by curling up in sleeping bags for one night'* or *'If every supporter on Facebook donated just one box each it would make a real difference to many poor families'*).

- **The expert.** The underlying message is that the privileged community, which the author and their

audience belong to, knows better what the vulnerable community needs, how they are or what they should do to overcome their situation. We consider the following subcategories:

- **Presupposition**, when the author assumes a situation as certain without having all the information, or generalises their or somebody else's experience as a categorical truth without presenting a valid, trustworthy source for it (e.g. a research work or survey). The use of stereotypes or clichés are also considered to be examples of presupposition.

  (i.e. *'[...] elderly or disabled people who are simply unable to evacuate due to physical limitations'* or *'If the economy fills with women, it will develop beautifully'*);

- **Authority voice**, when the author stands themselves as a spokesperson of the group, or explains or advises the members of a community about the community itself or a specific situation they are living.

  (i.e. *'Accepting their situation is the first step to having a normal life'* or *'We also know that they can benefit by receiving counseling from someone who can help them understand.'*);

- **The poet.** The focus is not on the *we* (author and audience), but on the *they* (the individual or community referred to). The author uses a literary style to describe people or situations. They might, for example, use (or abuse) adjectives or rhetorical devices to either present a difficult situation as somehow beautiful, something to admire and learn from, or they might carefully detail its roughness to touch the heart of their audience. The subcategories we establish are:

  - **Metaphor**. They can conceal PCL, as they cast an idea in another light, making a comparison between unrelated concepts, often with the objective of depicting a certain situation in a softer way. For the annotation of this dataset, euphemisms are considered as an example of metaphors.

    (i.e. *'Poor children might find more obstacles in their race to a worthy future'* or *'those who cling to boats to reach a shore of survival'*);

  - **Compassion**. The author presents the vulnerable individual or community as needy, raising a feeling of pity and compassion from the audience towards them. It is commonly characterized by the use of flowery wording that does not provide information, but the author enjoys the detailed and poetic description of the vulnerability;

    (i.e. *'Some are lured by corrupt "agents", smuggled across the searing Sahara and discarded in the streets of Europe, resigned to selling fake designer bags as undocumented immigrants'* or *'For the roughly 2,000 migrants who call it home, the broken windows and decaying walls of the decrepit warehouse offer scant respite from the harsh blizzard conditions currently striking Serbia'*).

  - **The poorer, the merrier**. The text is focused on the community, especially on how the vulnerability makes them better (e.g. stronger, happier or more resilient) or how they share a positive attribute just for being part of a vulnerable community. People living vulnerable situations have values to admire and learn from. The message expresses the idea of vulnerability as something beautiful or poetic. We can think of the typical example of 'poor people are happier because they don't have material goods'.

    (i.e. *'He is reminded of the true meaning of hope by people living in situations the world would see as hopeless'* or *'her mom is disabled and living with her gives her strength to face everyday's life'* or *'refugees are wonderful people'*)

Finally, in the dataset, we also included an "Other" category, to classify all the text spans which the annotators considered to contain PCL, but which they could not assign to any of the previous categories. However, the annotators did not need to use this label for any instance.

4.2 Annotation

To annotate the dataset, a two-step process has been followed. In the first step, annotators determined which paragraphs contain PCL. Subsequently, in the second step, the annotators indicated which text spans within these paragraphs contain PCL and they labelled each of these text spans with a particular PCL category. We now discuss these two steps in more detail.

4.2.1 Step 1: Paragraph-Level Identification of PCL

The aim of this annotation step is to decide for each paragraph whether or not it contains PCL. This annotation step proved more difficult than expected, stemming from the often subtle and subjective nature of PCL. To mitigate this, we decided to annotate the paragraphs with three possible labels: 0, meaning that the paragraph does not contain PCL, 1, meaning that it is considered to be a borderline case, or 2, meaning that it clearly contains PCL. We computed the Kappa Inter-Annotator Agreement (IAA) between two main annotators (*ann1* and *ann2*) across the three labels, obtaining a moderate agreement of 41%. If we omit all paragraphs which were marked as borderline by at least one annotator, the IAA reaches a substantial 61% (Landis and Koch, 1977).

Overall, ann1 and ann2 agreed in 9,182 paragraphs and disagreed in 1457. Among the disagreements, 590 were total disagreements (0 vs 2) and 867 cases included borderline cases. To maximize the amount of information captured by the annotations, and in particular obtain a finer-grained assessment about borderline cases, we combined the labels provided by the two annotators into a 5-point scale, as follows:

- Label 0: both annotators assigned the label 0 (0 + 0).

- Label 1: one annotator assigned the label 0 and the other assigned the label 1 (0 + 1).

- Label 2: both annotators assigned the label 1 (1 + 1).

- Label 3: one annotator assigned the label 2 and the other assigned the label 1 (2 + 1).

- Label 4: both annotators assigned the label 2 (2 + 2).

Note how partial disagreement between the annotators is thus reflected in the final label. The cases of total disagreement, where one annotator labeled the instance as clearly not containing PCL and the other annotated it as clearly containing PCL (0 + 2), were annotated by *ann3*. After this supplementary annotation, the paragraph is either labelled as 1, if the third annotator considered the paragraph not to contain PCL, as 2, if they considered it to be a borderline case, or as 3, if they considered the paragraph to clearly contain PCL. In this way, the labels 0 and 4 remain reserved for clear-cut cases. For the experimental analysis presented in this paper, we treated paragraphs with final labels 0 and 1 as negative examples (i.e. as instances not containing PCL) and paragraphs with final labels 2, 3 and 4 as positive examples (i.e. as instances containing PCL). In total, interpreted in this way, the dataset contains 995 positive examples of PCL.

4.2.2 Step 2: Identifying Span-Level PCL Categories

Those paragraphs labelled as containing PCL in Step 1 are collected for further annotation. The aim of this second step is to specify which text spans within these paragraphs contain PCL and to identify which PCL categories these text spans belong to. For this step, we used the BRAT rapid annotation tool (Stenetorp et al., 2012)[3]. Note that each paragraph might contain one or more text spans with PCL, which may be assigned to the same or to different categories. Table 2 shows how many spans have been labelled with each of the categories.

In Task 2, we compute the IAA for each category, reaching the following agreements: *Unbalanced power relations*: 58.43%; *Authority voice*: 48.34%; *Shallow solution*: 56.50%; *Presupposition*: 52.94%; *Compassion*: 66.40%; *Metaphor*: 52.72%, and *The poorer, the merrier*: 66.72%. When computing the agreement for the three higher-level categories, we obtain a IAA of 63.02% for *The Saviour* (*Unbalanced power relations* and *Shallow solution*), 57.21% for *The Expert* (*Presupposition* and *Authority voice*), and 66.99% for *The Poet* (*Compassion*, *Metaphor* and *The poorer, the merrier*).

---

[3]https://brat.nlplab.org/

	unb	com	pre	aut	sha	met	merr	Total
Disabled	96	55	26	23	21	17	12	250
Homeless	231	154	38	31	84	56	6	600
Hopeless	105	224	95	60	6	59	6	555
Immigrant	29	32	21	7	4	5	4	102
In-need	347	85	17	42	85	36	6	618
Migrant	40	45	9	14	4	10	4	126
Poor-families	185	131	63	59	41	67	11	557
Refugee	93	78	22	17	33	20	5	268
Vulnerable	130	54	22	41	12	36	1	296
Women	51	30	34	33	12	13	9	182
Total	1307	888	347	327	302	319	64	3554

Table 2: Number of text spans that have been labelled with each of the PCL categories, per keyword. The considered categories are unbalanced power relations (unb), compassion (comp), presupposition(pres), authority voice (auth), shallow solution (shal), metaphor (met), and the poorer, the merrier (merr).

5 Experiments

We experiment with a number of different methods to provide baselines for further research in modeling PCL. We consider two settings: predicting the presence of PCL, viewed as a binary classification task (Task 1), and predicting PCL categories, viewed as a multi-label classification task (Task 2). We evaluate the following methods:

- **SVM-WV.** We use paragraphs embeddings as the input for a Support Vector Machine implemented with SciKit-Learn. To create the paragraphs embeddings, we use the average of the standard 300 dimensional Word2Vec Skip-gram word embeddings trained on the Google News corpus (Mikolov et al., 2013). For Task 1, the parameters that were selected after hyper-parameter tuning were C=10, gamma='scale', kernel='poly', while for Task 2 we found that C=100, gamma='scale', kernel='rbf' yielded the best results on the validation data.

- **SVM-BoW.** We use a TF-IDF weighted Bag-of-Words representation of the paragraphs as input to an SVM, also implemented with SciKit-Learn. In this case, the hyperparameters that were selected are C=10, gamma= 'scale', kernel= 'rbf' for Task 1 and C=100, gamme='scale', kernel= 'linear' for Task 2.

- **BiLSTM.** We used a bidirectional LSTM, using the same Word2Vec embeddings as SVM-WV to represent the individual words. As hyper-parameters, we used 20 units for each LSTM layer and a dropout rate of 0.25% at both the LSTM and classification layers. We trained for 300 epochs, using the Adam optimizer, with early stopping and a patience of 10 epochs.

- **Fine-tuned Language Models.** We fine-tune a BERT language model (Devlin et al., 2018) for sequence classification. We considered two variants of this method, were we respectively used the BERT-large-cased and BERT-base-cased pre-trained models. To further explore the performance of language models, we also fine-tuned a RoBERTa-base (Liu et al., 2019) model, which can be viewed as an optimized version of BERT, and a DistilBERT (Sanh et al., 2019) model, which is a lighter and faster variant of BERT. In all cases, we trained the model for 10 epochs with a batch size of 32. For reproducibility, we fixed the random seeds at 1 in all cases.

- **Random.** To put the results in context, we include a classifier that relies on random guessing, choosing the positive class with 50% probability in Task 1, and independently selecting each label with a probability of 50% in Task 2.

For both Task 1 and Task 2 we used 10-fold cross validation for all the experiments. For the BiLSTM models, we used 10% of the training data in each fold as a validation set for early stopping. For the SVM models, we instead tuned the hyper-parameters using Grid Search Cross-Validation. As mentioned

	P	R	F1
SVM-WV	46.53	57.80	47.10
SVM-BoW	49.95	40.48	40.59
BiLSTM	62.61	54.43	57.75
Random	24.82	50.65	33.31
BERT-base	72.39	63.27	67.33
RoBERTa	73.08	68.51	**70.63**
DistilBERT	70.73	66.17	68.24
BERT-large	57.06	51.50	53.91

Table 3: Results for the problem of detecting PCL, viewed as a binary classification problem (Task 1).

before, for Task 1 we view paragraphs labelled with 0 or 1 as negative examples, and the remaining paragraphs, labelled with 2, 3 or 4, as positive examples. The results are reported in terms of the precision, recall and F1 score of the positive class. Task 2 is viewed as a paragraph-level multi-label classification problem, where each paragraph is assigned a subset of the PCL category labels. Therefore, in these baselines, span boundaries are not used as part of the training data. We report the precision, recall and F1 score of each of the individual category labels.

	SVM-WV			**SVM-BoW**			**BiLSTM**			**Random**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Unb. power rel.	82.51	85.37	83.82	80.02	78.91	79.21	83.94	84.58	83.92	71.63	49.89	58.54
Authority voice.	40.85	37.27	37.96	33.01	37.94	34.83	42.54	21.27	25.71	21.33	42.01	28.07
Shallow solu.	57.86	50.72	53.49	43.20	39.05	40.39	64.06	31.84	40.46	22.27	55.53	31.67
Presupposition	46.88	42.87	44.28	40.01	39.73	38.12	52.44	36.02	41.15	22.53	49.81	30.86
Compassion	68.31	70.41	69.13	62.25	62.25	60.92	74.48	69.86	71.34	49.03	52.64	50.56
Metaphor	37.93	32.80	34.71	29.53	29.03	28.63	7.83	1.99	3.14	20.12	48.36	28.29
The p., the mer.	40.00	12.17	17.89	5.00	1.43	2.22	0.00	0.00	0.00	4.40	55.58	8.08
	BERT-large			**BERT-base**			**RoBERTa**			**DistilBERT**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Unb. power rel.	84.28	93.35	88.55	84.47	93.53	88.70	85.84	93.34	**89.4**	84.11	92.44	88.01
Authority voice.	54.24	52.65	**53.06**	54.52	43.60	47.43	56.34	48.00	50.9	51.73	37.16	41.71
Shallow solu.	70.93	52.59	59.67	71.08	49.64	57.47	69.09	55.62	**61.0**	72.80	45.21	54.89
Presupposition	60.42	59.71	59.61	59.94	55.92	57.22	60.95	58.90	**59.7**	60.32	49.04	53.60
Compassion	78.56	76.66	77.46	77.85	76.29	76.92	78.83	77.67	**78.1**	74.17	74.80	74.37
Metaphor	58.51	31.48	40.09	62.81	27.93	38.21	59.36	35.74	**43.4**	65.15	27.60	37.93
The p., the mer.	23.33	8.50	11.90	0.00	0.00	0.00	40.83	15.00	**20.5**	0.00	0.00	0.00

Table 4: Results for the problem of categorizing PCL, viewed as a paragraph-level multi-label classification problem (Task 2).

The results of Task 1 are summarized in Table 3. As can be seen, all of the considered methods clearly outperform the random baseline. Unsurprisingly, the BERT-based methods achieve the best results, with RoBERTa performing slightly better than DistilBERT and BERT-base. The performance of BERT-large is surprisingly weak compared with the other BERT-based models, performing worse than the BiLSTM. This suggests that BERT-large is more prone to over-fitting, given the relatively small number of training examples. Table 4 shows the results we obtained in Task 2. RoBERTa outperforms the rest of the models in all the categories except for Authority voice, where BERT-large gets the best results. We can also notice the fairly good performance of the SVM methods. In some categories, such as *Methaphors*, the SVM-WV model performs almost on par with DistilBERT and BERT-base and outperforms the BiLSTM results. For *The poorer, the merrier* it outperforms all the models except for RoBERTa.

Comparing the results for different categories, we can see that *Unbalanced power relations* appear relatively easy to detect. This is not unexpected, given that the presence of words such as *us*, *they*, *must* or *help* are strong and common indicators of such language. For similar reasons, instances of *Compassion* appear relatively easy to detect. *The poorer, the merrier* is the least represented category in the entire dataset, with just 64 samples, which can explain the poor results for this category. However, the poor performance for the *Metaphor* category cannot be explained in this way, given that the number of

Pred.	Paragraph	Gold
pos.	After Vatican controversy, McDonald's helps feed homeless in Rome.	pos.
pos.	From his personal story and real-life encounters with poor families, manpower correspondent Toh Yong Chuan suggested shifting the focus from poor parents who repeatedly make bad decisions to their children (Lifting families Out of poverty: Focus on the children; last Thursday).	pos.
pos.	He said their efforts should not stop only at creating many graduates but also extended to students from poor Families so that they could break away from the cycle of poverty.	pos.
neg.	"The biggest challenge is the no work policy. I think that refugees who come here, or asylum seekers, they're unable to work and they have kids here - their kids are stateless. That's really the cause of a lot of stress in the community."	pos.
neg.	"The people of Khyber Pakhtunkhwa are resilient. I did not see hopelessness on any face," he said.	pos.
neg.	Teach kids to give back: When Kang runs summer camps with kids, she includes "Contribution Fridays" - the kids work together as a team to make sandwiches for the homeless and dole out the food in shelters.	pos.
pos.	These shocking failures will continue to happen unless the Government tackles the heart of the problem - the chronic underfunding of social care which is piling excruciating pressure on the NHS, leaving vulnerable patients without a lifeline.	neg.
pos.	Lilly-Hue: His ability to make sure our family is never in need - his sacrificial self.	neg.
pos.	Any Kenyan small-scale farmer with such an income could not be said to be hopelessly mired in agrarian destitution. But of course, nothing in life is ever so simple as to allow for neat and precise answers.	neg.
pos.	Selective kindness: In Europe, some refugees are more equal than others.	neg.

Table 5: Examples of incorrect predictions made by RoBERTa in Task 1.

training examples for this category is higher than the number of examples for *Shallow solution* and very similar to the number of examples for *Authority voice*. More generally, while some of the differences in performance are due to variations in the number of training examples, the categories with the weakest performance also tend to be those that require some forms of world knowledge. For instance, to detect presuppositions, we need to determine whether the assumption which is made is reasonable or not. Similarly, detecting shallow solutions requires assessing the quality of the proposed solution, which can clearly be challenging.

To get further insights into the dataset, Table 5 shows some examples of paragraphs from Task 1, their gold labels and the predictions by RoBERTa. There are three correctly classified instances and seven misclassified examples (i.e. three false negatives and four false positives). In many cases, we can see words and phrases that are often used in PCL, but which are not actually used in a condescending context, causing the model to predict false positives. For instance, in the seventh example, excess of adjectives and flowery wording, e.g. *shocking failures* and *excruciating pressure*, are often used in PCL fragments from the *Compassion* category. In this example, however, it is used in a political context, without being condescending towards any particular group. In the fifth example, the model misclassifies the paragraph as not containing PCL. In this case, we have an example of the category *The poorer, the merrier*, which all models struggle to detect. Surprisingly, this category has the highest inter-annotator agreement in the annotation of the dataset. This suggest that, while for human annotators it is very easy to identify cases of this category, the models struggle to detect such cases. In Table 6, some incorrect predictions from Task 2 are presented. Among others, these examples illustrate how RoBERTa struggles to distinguish between presuppositions and authority voices, which are often incorrectly predicted together. Shallow solutions are also often neglected by RoBERTa. A particularly clear case is the last example, where recognizing the presuppositions and shallow solutions in the text will require external knowledge of the situation and the needs of those affected. We can also see examples where the occurrence of a particular structure of language appears to mislead RoBERTa, e.g. *to open the doors wider for [...]*, in the fourth example, seems to lead the model to bet for a shallow solution. *Methaphors*, as in this same example, are also difficult to identify for RoBERTa in this context.

Paragraph	Gold	Pred
[...] The blacks want all our farmland without compensation. Give it to them. Let the farmers flock into the cities and make a new life for themselves . With their resilience I am sure it will not be so difficult for them to establish a new, happy and productive life. They will have no money but the clothes on their back to start off with , but that is what so many immigrant Americans had to face. Through guts, determination and sheer will power, they rose above it all, and look what America is today.	unba, pres, comp, merr	unba, auth, pres, comp, meta
According to the foundation , a number of children between the ages of six and 14 homeless and roaming the streets is becoming alarming.	comp	unba, comp
The photo of a Hyderabad traffic policeman feeding an elderly homeless woman has gone viral , earning him accolades from social media users [...].	unba, shal	unba
Practical ways to open the doors wider for our disabled	unba, meta	unba, shal
He could have also taken his condition to mean he must be disabled from seeking to live for others. He could have degenerated into self pity as many do, wallowing in the muddy fields of self-obsession and low self esteem. Yusuf did not; everything was not about his immediate interests, but a social impact that touched even the lives of strangers [...].	unba, comp, meta, merr	auth, pres, comp
She called on the general public to volunteer to donate blood and that way rescue the lives of patients in need of blood transfusion.	unba, auth	unba, auth, meta
For now the families are staying with friends and family. During the day they clean up the debris left by the fire, hoping that someone will come to their rescue. They received emergency relief packs, but they are still in need of clothes, beds, blankets and kitchen appliances.	unba, shal, pres, comp	unba, comp

Table 6: Examples of incorrect predictions made by RoBERTa in Task 2.

6 Conclusions and Future Work

We have introduced the Don't Patronize Me! dataset, which is aimed at introducing the NLP community to the challenge of identifying and categorizing Patronizing and Condescending Language (PCL) towards vulnerable communities. As another contribution of this paper, we also introduced a two-level taxonomy of PCL categories, which was used for annotating the dataset. Our exploratory analysis shows that identifying condescending or patronizing texts is a difficult challenge, both for human judges and for NLP systems. Apart from the subtle and subjective nature of PCL, a particular challenge comes from the fact that accurately modelling such language often requires knowledge of the world and common sense (e.g. to assess whether a proposed solution is shallow, or whether a particular presupposition is warranted). Nonetheless, we found that both identifying PCL (Task 1) and categorizing occurrences of PCL (Task 2) is feasible, in the sense that non-trivial results can be achieved, with BERT-based approaches outperforming simpler methods. Future work will include the development of new models for both detecting and categorizing PCL. In addition, we plan to continue to extend the Don't Patronize Me! dataset with more paragraphs from news stories, as well as text fragments from different sources, such as social media or NGO campaigns, to create a useful and updated resource for the community.

Acknowledgements

References

Pepa Atanasova, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv preprint arXiv:1808.05542.*

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019.

Overview of the clef-2019 checkthat! lab on automatic identification and verification of claims. task 1: Checkworthiness. In *CEUR Workshop Proceedings, Lugano, Switzerland*.

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).

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. Available online at `https://www.english-corpora.org/now/`.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

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.

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

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

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

Mark S Komrad. 1983. A defence of medical paternalism: maximising patients' autonomy. *Journal of medical ethics*, 9(1):38–44.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

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.

Vladimir Vapnik and Akshay Vashist. 2009. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557.

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*.

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). *arXiv preprint arXiv:1903.08983*.