SemEval-2022 Task 4: Patronizing and Condescending Language Detection

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

Abstract

This paper presents an overview of Task 4 at SemEval-2022, which was focused on detecting Patronizing and Condescending Language (PCL) towards vulnerable communities. Two sub-tasks were considered: a binary classification task, where participants needed to classify a given paragraph as containing PCL or not, and a multi-label classification task, where participants needed to identify which types of PCL are present (if any). The task attracted 77 teams. We provide an overview of how the task was organized, discuss the techniques that were employed by the different participants, and summarize the main resulting insights about PCL detection and categorization.

1 Introduction

The study of unfair, misleading or offensive language has attracted the interest of many scholars from the NLP research community. Most relevant tasks in this context focus on explicit, aggressive and flagrant phenomena, such as fake news detection or fact-checking (Conroy et al., 2015; Nakov et al., 2018; Atanasova et al., 2019; Barrón-Cedeno et al., 2020); detecting propaganda techniques (Da San Martino et al., 2020); modeling offensive language (Zampieri et al., 2019, 2020) identifying hate speech (Basile et al., 2019); and rumour propagation (Derczynski et al., 2017). However, there also exist subtler but equally harmful types of language, which have received less attention by the NLP community, and which, due to their subtle nature, we can expect to be more difficult to detect. This is the case, among others, for Patronizing and Condescending Language (PCL), which was the focus of Task 4 at SemEval-2022.

An entity engages in patronizing or condescending communication when its use of language reveals a superior attitude towards others. These attitudes, when normalized, routinize discrimination and make it less visible (Ng, 2007). Furthermore, the use of PCL is often unconscious and well-intended, especially when referring to vulnerable communities (Wilson and Gutierrez, 1985; Merskin, 2011). This good will can make PCL especially harmful, as the audience receives this discriminatory language with low defense and is often unaware of its effects.

Research in sociolinguistics presents PCL as a subtle, often unconscious but harmful and discriminative kind of language (Mendelsohn et al., 2020). It creates and feeds stereotypes (Fiske, 1993), which result in greater exclusion, rumour spreading and misinformation (Nolan and Mikami, 2013). PCL also tends to strengthen power-knowledge relationships (Foucault, 1980), calling for charitable action instead of cooperation and presenting those who can help as *saviours* of those in a less privileged position (Bell, 2013; Straubhaar, 2015). Furthermore, PCL tends to conceal who is responsible for very deep-rooted societal problems, sometimes by implicitly or explicitly blaming the underprivileged communities or individuals for their situation, and often involves ephemeral and simple solutions. (Chouliaraki, 2010). The use of PCL by privileged communities has also been related to the so-called pornography of poverty (Nathanson, 2013), a communication style that depicts vulnerable situations with a pity discourse to move a target audience to charitable action and/or compassionate attitudes.

While the negative impact of PCL, both in social interactions and in corporate and political discourse, has been extensively studied in the social sciences, it still remains an under-explored phenomenon in NLP. Nonetheless, we believe that PCL detection offers a number of important challenges for NLP research, which warrant more work in this area, especially given the societal benefits that would result. For instance, given its subtle and subjective nature, we can expect PCL detection to be harder than tasks that are focused on more flagrant phenomena. Moreover, PCL detection often involves the need for an implied *understanding* of

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 298 - 307

human values and ethics, which requires a form of commonsense reasoning that NLP models are likely to struggle with. In this context, we have organized SemEval 2022 Task 4: Patronizing and Condescending Language (PCL) Detection. This task has attracted more than 300 participants, organized in 77 teams, during the official competition. The competition remains open on CodaLab to encourage further research on this topic[1].

2 Related Work

As already mentioned in the introduction, PCL has been extensively studied within the context of sociolinguistics (Margić, 2017; Giles et al., 1993; Huckin, 2002; Chouliaraki, 2006). Within NLP, however, modelling of patronizing discourse has only received limited attention. As a notable exception, Wang and Potts (2019) compiled a corpus of Reddit comments, which were annotated as using a condescending tone or not. Note that in contrast to our SemEval task, their work did not specifically focus on vulnerable communities. In our previous work (Perez-Almendros et al., 2020), we introduced *Don't Patronize Me!*, which is, to the best of our knowledge, the first annotated corpus of PCL towards vulnerable communities. This corpus was used as the training data for the SemEval task. Some other works have studied types of discourse that are highly related to condescension, including Sap et al. (2020), who studied how certain uses of language indicate power relations, Mendelsohn et al. (2020), who discussed the dehumanization of minorities through language and Zhou and Jurgens (2020), who investigated how some expressions of condolences and empathy interplay with authoritative voices in online communities.

3 Dataset

The seed material for this task is *Don't Patronize Me!* (DPM), an annotated dataset with Patronizing and Condescending Language towards vulnerable communities, which was introduced in our previous work (Perez-Almendros et al., 2020). This dataset contains 10,469 paragraphs, which were used as the training set for the SemEval task. To create the test set for this task, we annotated 3,898 additional paragraphs, following the same process. All paragraphs were extracted from news stories from media in 20 English speaking countries, originally provided by the News on Web (NoW) corpus[2] (Davies, 2013).

We used a keyword-based strategy to collect paragraphs, focusing on texts in which vulnerable communities are mentioned (e.g., refugees or homeless). The data was annotated by three annotators, with backgrounds in communication, media and data science. For the main dataset, two annotators annotated the instances with the following labels: 0 (not PCL), 1 (borderline), and 2 (PCL), achieving an inter-annotator agreement (IAA) of 41% for the raw annotations and 61% when removing borderline cases. For all the total disagreements (paragraphs labeled 0 by one annotator and 2 by the other), the third annotator acted as a referee, providing a final label. The final dataset uses a scale from 0 to 4, indicating the level of agreement between the annotators. Labels 0 and 4 correspond to clearly not condescending and clearly condescending (i.e. both annotators assigned 0 or both assigned 2), label 2 means that both annotators marked that paragraph as a borderline case (1-1), and labels 1 and 3 correspond to cases where either one of the annotators assigned the borderline label (0-1 or 1-2), or there was a disagreement that was resolved by the third annotator. Each positive example from the dataset is furthermore labelled with one or more PCL categories. We briefly recall the meaning of these categories.

Unbalanced power relations (UNB): the author entitles themselves as being in a privileged situation, considering themselves as *saviours* of those in need (Bell, 2013; Straubhaar, 2015).

Shallow solution (SHAL): a charitable, superficial and short-term action is presented as life changing.

Presupposition (PRES): stereotypes and *clichés* are used to describe a community, relying on assumptions without having all the information.

Authority voice (AUTH): the author stands as spokesperson and defendant of the community or individual and/or allows themselves to give expert advice about how to overcome underprivileged situations.

Metaphor (MET): the author describes a difficult situation in a more poetic way through

[1] https://competitions.codalab.org/competitions/34344

[2] Used with permission from the author.

Cat.	Examples
UNB	*They deserve another opportunity* or *You can make a difference in their lives.*
SHAL	*Raise money to combat homelessness by curling up in sleeping bags for one night.*
PRES	*Elderly or disabled people who are simply unable to evacuate due to physical limitations.*
AUTH	*Accepting their situation is the first step to having a normal life.*
MET	*Poor children might find more obstacles in their race to a worthy future.*
COMP	*[...] discarded in the streets of Europe [...]*
MERR	*Her mom is disabled and living with her gives her strength to face everyday's life* or *Refugees are wonderful people.*

Table 1: Examples of the different PCL categories.

figures-of-speech such as metaphors and euphemisms.

Compassion (COMP): the message uses flowery wording to reflect on the vulnerability or toughness of the situation, raising a feeling of pity among the audience.

The poorer, the merrier (MERR): the author praises the vulnerability, granting positive values to all members of a vulnerable community and showing their admiration.

Table 1 contains examples for each of these categories. The average of the IAA among categories is 57.43%[3]. It is worth mentioning that the distribution of labels is highly unbalanced in our dataset, with only around 9.5% of the inputs being labeled as containing PCL (positive cases). For the categories, the distribution is as follows: 73% UNB, 19% SHALL, 23.1% PRES, 23.6% AUTH, 48.7% COMP, 20.1% MET and 4.1% MERR.

4 Task Description

The aim of the proposed task is to identify the presence of PCL (Subtask 1), and to identify the categories of PCL that are present in a given paragraph (Subtask 2).

Training data The 10,469 annotated paragraphs from the DPM corpus were provided as training data. To frame Subtask 1 as a binary classification problem, paragraphs with labels 0 and 1 were considered as negative examples, while paragraphs with labels 3 and 4 were considered as positive examples of PCL. The original labels on the scale from 0 to 4 were also made available. The 993 positive examples in the training data are labelled with the corresponding PCL categories. Span annotations for these categories were also provided.

Test data A total of 3,898 paragraphs were released as test set, with the same format and meta-information as the training set, but without labels and span annotations. Paragraphs initially labelled as 2 were excluded from the test data, as these correspond to borderline cases.

External resources We welcomed the use of external resources in this task. Participants were encouraged to explore transfer learning or data augmentation techniques with a variety of source corpora and language resources.

Evaluation System submissions were ranked in the two subtasks as follows: **Subtask 1**: F1 score for the positive class. **Subtask 2**: Macro-averaged F1 over all categories.

4.1 Participation Framework

The task was hosted on CodaLab[4], with participants needing to register and submit their results through the platform. The competition involved the following three phases:

- **Practice phase:** The 10,469 paragraphs from the training data were split into 8,376 training paragraphs and 2,095 validation paragraphs. This was done to allow participants to compare their systems on a public leader board. The training-validation split respected the natural distribution of labels in the data.
- **Evaluation phase:** This was the official evaluation phase for the SemEval competition. The test data was released and the leader board for this phase remained hidden to prevent participants from fine-tuning their systems on the test data. Each participant was allowed two different submissions for each subtask.
- **Post-evaluation phase:** The learderboard for the evaluation phase and the official ranking for each subtask were published, as the SemEval competition ended. Participation in the SemEval task is no longer possible, but the

[3] See Perez-Almendros et al. (2020) for further details.

[4] https://competitions.codalab.org/competitions/34344

competition remains open on CodaLab to allow participants to re-test and further improve their systems.

5 Results and Discussion

A total of 77 different teams participated in the evaluation phase of our task, with 145 valid submissions for Task 1 and 84 for Task 2. For the competition, we allowed a maximum of 2 submissions per team. A total of 42 out of 77 teams outperformed the baseline for Subtask 1, while 37 out of 48 outperformed the baseline for Subtask 2. Tables 2 and 3 present the rankings for Subtask 1 and 2, respectively, where we have only listed the best performing system for each team. For Subtask 1, the best-performing systems used the following strategies:

Team PALI-NLP used an ensemble of pre-trained RoBERTa models (Liu et al., 2019). While training, they applied grouped Layer-Wise Learning Rate Decay, a variant of LLRD (Howard and Ruder, 2018), based on the idea that different layers capture different types of information (Yosinski et al., 2014). By optimizing the learning rate in different layers, the model captures more diverse and fine-grained linguistic features of PCL. To tackle the class imbalance in the dataset, they use weighted random samples (Hashemi and Karimi, 2018) to emphasize the positive instances.

Team STCE created adversarial examples to train an ensemble model of RoBERTa and DeBERTa (He et al., 2020). They also used weighted samples to address the class imbalance and explored different loss functions, establishing Cross Entropy and the contrastive loss algorithm NT-Xent introduced by Chen et al. (2020) as first and second loss function, respectively.

For Subtask 2, the best-performing systems used the following strategies:

Team BEIKE NLP participated with a system based on prompt learning (Petroni et al., 2019; Brown et al., 2020). They first reformulate PCL detection as a cloze prompt task and then fine-tune a pre-trained DeBERTa model.

Team PINGAN Omini-Sinitic proposed an ensemble model which used prompt training and a label attention mechanism, by adding a new label-wise attention layer ((Dong et al., 2021; Vu et al., 2021). Their system over-samples the positive examples. They also use a form of transfer learning from Subtask 1 to Subtask 2, by pre-training on Subtask 1 and using the resulting model as the starting point for training a model for Subtask 2.

For both sub-tasks, unsurprisingly, most systems rely on pre-trained language models, although a few teams have used CNN, LSTM, SVM or Logistic Regression based systems (XU, PC1, I2C, Ryan Wang, McRock, Amrita_CEN, SATLab and Team Lego, among others), or an ensemble of some of the above together with language models (UTSA_NLP, Taygete). Although the use of language models usually outperformed other systems in this task, some LSTM models, such as the one submitted by team Xu, achieved competitive results.

The ensembling of different models has also been a popular technique. Other strategies that proved successful include adversarial training, data augmentation and multitask learning. In the following, we summarize how these techniques have been used by the different systems.

Ensemble learning Ensembling different models has previously been found useful for text classification (Nozza et al., 2016; Kanakaraj and Guddeti, 2015; Fattahi and Mejri, 2021). Accordingly, ensembling was one of the most common strategies for improving on baseline PCL detection methods. Most of the teams combined different language models (e.g. PALI-NLP, STCE, PINGAN Omini-Sinitic,PAI_Team, LRL_NC, SSN_NLP_MLRG, ASRtrans, amsqr, UMass PCL). Considering the choice of language models, the most successful systems either used RoBERTa, DeBERTa or an ensemble which included the former ones and other models. For instance, these models were used by the best performing teams for both subtasks, i.e. PALI-NLP and STCE for Subtask 1 and BEIKE NLP and PINGAN Omini-Sintic for Subtask 2. To fine-tune the language models effectively, incorporating a contrastive loss function, in addition to the standard cross-entropy loss, has also proved useful. Finally, it should be noted that the combination of language models with different types of neural networks (Taygete, UTSA_NLP) has also proven useful.

	TEAM	P	R	F1		TEAM	P	R	F1		TEAM	P	R	F1
1	PALI-NLP	64.6	65.6	65.1	**27**	ML_LTU	58.0	51.4	54.5	**53**	RNRE NLP	39.0	50.2	43.2
2	stce	63.3	66.9	65.0	**28**	ZYBank-AI	54.8	53.9	54.4	**54**	SATLab	34.8	55.2	42.7
3	ymf924	63.8	65.6	64.7	**29**	Team LRL_NC	60.7	49.2	54.4	**55**	J.U.S.T-DL	49.0	37.5	42.5
4	BEIKE NLP	61.2	67.2	64.1	**30**	CS-UM6P & ESL	55.2	53.3	54.3	**56**	MaChAmp	58.8	32.8	42.1
5	holdon	60.3	67.5	63.7	**31**	Felix&Julia	40.1	77.3	52.8	**57**	I2C	61.1	31.2	41.3
6	cnxup	62.7	64.7	63.7	**32**	Stanford ACM	40.2	76.7	52.7	**58**	SMAZ	36.3	47.6	41.2
7	abcxyzw	58.8	68.5	63.3	**33**	UtrechtUni	44.6	62.5	52.0	**59**	MASZ	36.3	47.6	41.2
8	nowcoder	58.2	68.5	62.9	**34**	CSECU-DSG	59.0	46.4	51.9	**60**	Amrita_CEN	32.2	52.1	39.8
9	PINGAN Omini-Sinitic	61.8	63.7	62.7	**35**	Sapphire	59.4	46.1	51.9	**61**	Anonymus	27.6	59.9	37.8
10	bigemo	57.1	69.4	62.7	**36**	Ablimet	61.5	44.8	51.8	**62**	matan-bert	35.4	40.4	37.7
11	Leo_team	60.1	64.0	62.0	**37**	SSN_NLP_MLRG	42.3	66.6	51.7	**63**	Team LEGO	24.8	56.5	34.5
12	PAI-Team	66.3	57.7	61.7	**38**	Team PiCkLe	46.0	58.0	51.3	**64**	TüSoXi	38.8	29.3	33.4
13	Anonymus	53.5	70.4	60.8	**39**	sua	54.0	48.6	51.2	**65**	RNRE NLP RFC	30.0	36.9	33.1
14	BLING	63.5	55.5	59.3	**40**	UCL xNSI	41.5	65.3	50.7	**66**	jct_meir	25.3	47.0	32.9
15	Taygete	53.6	66.3	59.2	**41**	MS@IW	50.2	51.1	50.6	**67**	isys	22.4	59.3	32.5
16	NLP-Commonsense Reasoning team	61.2	56.8	58.9	**42**	University of Bucharest Team	49.1	50.8	49.9	**68**	AliEdalat team	18.4	87.1	30.3
17	GUTS	61.3	54.9	57.9	**43**	**RoBERTa Baseline**	**39.4**	**65.3**	**49.1**	**69**	ms_pa	23.4	39.1	29.3
18	DH-FBK	64.2	52.7	57.9	**44**	rematchka	44.5	53.9	48.8	**70**	Waad	64.0	18.0	28.1
19	ULFRI	56.4	58.7	57.5	**45**	fengxing	63.8	39.4	48.7	**71**	Ryan Wang	17.0	60.9	26.6
20	TUG-CIC	60.2	54.9	57.4	**46**	flerynn	67.2	38.2	48.7	**72**	PC1	37.8	18.6	25.0
21	amsqr	54.8	59.9	57.2	**47**	Team YNU-HPCC	65.9	36.6	47.1	**73**	UTSA_NLP	14.0	35.0	20.0
22	UMass PCL	52.9	58.4	55.5	**48**	niksss	51.8	42.0	46.3	**74**	yaakov	11.2	10.1	10.6
23	LastResort	51.5	59.9	55.4	**49**	JustTeam	55.0	39.8	46.2	**75**	ilan	14.5	6.0	8.5
24	Team Double_A	47.2	66.6	55.2	**50**	BWQ	51.0	41.3	45.6	**76**	Jiaaaaaa	8.2	6.3	7.1
25	thetundramanagainstpcl	54.3	55.5	54.9	**51**	Tesla	36.0	57.7	44.3	**77**	Anonymus	29.7	3.5	6.2
26	Xu	46.2	66.9	54.6	**52**	ASRtrans	35.6	58.4	44.2	**78**	Anonymus	10.6	2.8	4.5

Table 2: SemEval Task 4: Ranking by teams for Subtask 1: Binary Classification. The table reports Precision (%), Recall (%) and F1-Score (%) for the positive class.

	TEAM	UNB	SHAL	PRES	AUTH	MET	COMP	MERR	Avg
1	BEIKE NLP	65.6	52.9	36.9	40.7	35.9	49.2	47.1	46.9
2	PINGAN Omini-Sinitic	59.7	53.1	41.7	43.4	42.7	51.3	15.4	43.9
3	PAI_Team	57.6	45.2	35.2	39.4	38.4	44.5	26.7	41.0
4	stce	62.2	54.8	38.1	32.8	33.3	51.0	8.7	40.1
5	PALI-NLP	61.8	54.1	37.7	32.8	32.8	51.2	8.7	39.9
6	Leo_team	57.3	47.0	28.8	36.1	34.8	47.4	27.0	39.8
7	Anonymus	59.9	49.1	38.5	37.1	35.0	48.6	8.3	39.5
8	ymf924	61.6	54.1	36.8	31.3	33.3	50.0	8.7	39.4
9	bigemo	62.5	56.1	38.0	24.3	31.3	49.4	8.7	38.6
10	holdon	62.2	56.1	32.9	23.1	33.3	48.7	8.7	37.9
11	cnxup	60.2	53.3	30.6	24.1	40.0	48.1	8.7	37.8
12	Taygete	59.7	45.8	33.3	21.8	30.4	53.6	18.8	37.6
13	DH-FBK	52.5	36.2	27.0	37.7	31.9	46.0	30.3	37.4
14	abcxyzw	60.7	53.3	34.5	21.8	32.8	50.0	8.3	37.4
15	nowcoder	59.8	50.0	32.2	22.8	39.4	47.8	8.3	37.2
16	GUTS	55.6	47.4	24.0	34.3	25.6	44.4	27.6	37.0
17	BLING	55.1	38.9	23.4	29.0	31.5	50.9	26.7	36.5
18	UMass PCL	53.9	42.4	29.1	30.7	33.3	40.8	23.5	36.3
19	CS-UM6P & ESL	57.0	42.0	25.7	25.2	20.5	46.8	21.4	34.1
20	Fengxing	46.4	46.3	23.0	26.5	33.3	38.7	24.0	34.0
21	Team LRL_NC	52.1	42.7	25.2	30.4	28.8	43.3	14.8	33.9
22	thetundramanagainstpcl	50.5	50.0	18.4	16.5	20.3	41.5	24.0	31.6
23	Xu	55.0	48.4	28.0	24.0	13.6	49.0	0.0	31.1
24	SATLab	42.4	33.1	17.0	23.2	17.5	31.5	14.2	25.6
25	Felix&Julia	36.6	35.1	17.6	22.1	21.1	28.5	16.7	25.4
26	AliEdalat team	53.9	37.7	25.6	26.2	13.5	11.3	9.1	25.3
27	Tesla	43.7	38.3	16.3	19.2	17.9	35.7	0.0	24.5
28	Waad	36.9	33.3	17.5	15.3	16.5	28.7	19.5	24.0
29	ms_pa	32.3	32.9	19.2	20.6	22.2	26.4	7.1	23.0
30	rematchka	37.7	21.4	18.8	21.2	15.5	26.1	13.0	22.0
31	Team Double_A	33.5	31.9	18.4	19.1	23.4	24.5	0.0	21.5
32	SSN_NLP_MLRG	34.6	33.8	20.7	19.3	12.1	27.7	0.0	21.2
33	ASRtrans	18.6	8.8	8.3	19.8	13.2	27.8	35.7	18.9
34	MaChAmp	30.4	21.3	3.6	10.9	30.8	5.0	6.3	15.5
35	Team PiCkLe	10.9	22.5	14.4	21.0	19.2	6.5	11.5	15.2
36	LastResort	15.8	24.8	10.0	9.3	16.0	11.3	14.8	14.6
37	Ablimet	12.6	14.1	6.5	7.2	14.0	17.2	17.1	12.7
38	**RoBERTa Baseline**	**35.4**	**0.0**	**16.7**	**0.0**	**0.0**	**20.9**	**0.0**	**10.4**
39	BWQ	16.0	12.5	7.2	9.7	7.0	11.4	3.9	9.7
40	Stanford ACM	16.0	26.5	4.2	0.0	0.0	8.6	12.1	9.6
41	Team LEGO	11.8	20.6	1.9	6.4	6.5	10.2	0.0	8.2
42	CSECU-DSG	33.4	0.0	0.0	0.0	0.0	21.8	0.0	7.9
43	University of Bucharest Team	14.8	21.7	3.5	0.0	3.9	8.3	0.0	7.4
44	PC1	11.8	12.0	6.1	8.7	2.6	8.9	0.0	7.2
45	Team YNU-HPCC	10.9	0.8	3.5	3.3	0.0	5.8	0.0	3.5
46	NLP-Commonsense Reasoning team	9.7	0.2	0.0	3.2	3.2	4.4	1.1	3.1
47	Jiaaaaaa	2.8	1.9	0.0	2.0	0.0	4.8	6.9	2.6
48	Anonymus	5.9	8.3	0.0	2.4	0.0	1.4	0.0	2.6
49	niksss	0.0	1.0	0.0	0.0	0.0	0.0	1.1	0.3

Table 3: SemEval Task 4: Ranking by teams for Subtask 2: Categories Classification. The table reports F1-Score (%) for each one of the categories and the macro-averaged F1-score (%) for all categories. The categories stand for: Unbalanced Power Relations (UNB), Shallow Solution (SHAL), Presupposition (PRES), Authority Voice (AUTH), Metaphors (MET), Compassion (COMP) and The poorer, the merrier (MERR).

Balancing class distribution The class imbalance in the dataset has been addressed by participating teams in different ways. Some teams opted for downsampling the number of negative examples (Ryan Wang, LastResort, MS@IW), while others tried a cost-sensitive learning approach to address this issue (Amrita_CEN). However, the most popular approach to balance the class distribution has been through data augmentation (amsqr, Xu, Utrech Uni, UMass PCL, among others). To create new positive examples, participants have used strategies such as the use of large generative models like GPT3 (Brown et al., 2020) or T5 (Raffel et al., 2020) (MS@IW, PINGAN Omini-Sinitic and Tesla); back-translation (Taygete); the addition of synonymous sentences to the original data (I2C), or the application of the so-called Easy Data Augmentation methods, a set of simple but effective techniques such as synonym replacement, random insertion, random swap, and random deletion (AliEdalat) (Wei and Zou, 2019; Rastogi et al., 2020).

External resources Various types of external resources have been used. For example, lexical databases such as WordNet (Miller, 1995) have been used to augment, enrich and improve the training data (Ali Edalat). Datasets from related tasks, including TalkDown (Wang and Potts, 2019), and two metaphor detection datasets, namely MOH (Mohammad et al., 2016) and VUA (Steen et al., 2010), have been used both for pre-training and / or for data augmentation by different teams. PAWS, a dataset with Paraphrase Adversaries from Word Scambling, (Zhang et al., 2019) and xTREME, a bechmark for Cross-Lingual Transfer Evaluation of Multilingual Encoders (Hu et al., 2020), have also been used to improve several systems (Ali Edalat, ASRtrans, Tesla, MaChAmp). Other related NLP challenges have served as auxiliary tasks for pre-training PCL models (AliEdalat, UMass PCL), although such strategies have not always been successful (ULFRI). The MaChAmp team used 7 SemEval-2022 tasks, including ours, for training a model based on multi-task learning. The DH-FBK team also opted for multi-task learning, but they only used the data from the Don't Patronize Me dataset itself to create auxiliary tasks. For instance, they trained their model to predict the uncertainty of a label in Subtask 1, using the fine-grained set of labels (0-4); the agreement of the annotators in Subtask 2; the spans where the categories were present; or the country of origin of the news outlets. AliEdalat similarly used the meta-information from the Don't Patronize Me dataset as additional features for training their model.

Prompt learning Using prompts has also proven useful for PCL detection (BEIKE NLP, PINGAN Omini-Sintic, Ablimet). Specifically, the teams used prompts such as "*[paragraph]* is *[label]*", or "is *[paragraph]* *[label]*?" where *[paragraph]* is the original input. For Subtask 1, *[label]* is a natural language description of the binary class label (e.g. *"is (not) condescending or patronizing"*). For Subtask 2, *[label]* is the label of a given PCL category.

6 Conclusions

Patronizing and Condescending Language detection is a relatively new challenge for the NLP community. However, the high level of participation in this task has provided the community with valuable new insights about how to tackle this problem. A total of 42 out of 77 teams in Subtask 1 and 37 out of 48 for Subtask 2 outperformed the RoBERTa baseline. The performance of the best-performing systems shows that a judicious usage of state-of-the-art text classification techniques can bring significant benefits to PCL detection, especially when it comes to addressing the relative scarcity of the available training data and closely related external resources. However, there still remains considerable scope for further improvements. It is our expectation that further improvements may need to rely on techniques that are specifically targeted at PCL, e.g. by exploiting insights from linguistics about the linguistic features of PCL, or by building explicit models of stereotypes of vulnerable communities.

References

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: Check-worthiness. In *CEUR Workshop Proceedings, Lugano, Switzerland*.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer.

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).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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*, 13(2):107–126.

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.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. *Retrieved January*, 25:2019.

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. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728.

Jaouhar Fattahi and Mohamed Mejri. 2021. Spaml: a bimodal ensemble learning spam detector based on nlp techniques. In *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*, pages 107–112. IEEE.

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.

Mahdi Hashemi and Hassan Karimi. 2018. Weighted machine learning. *Statistics, Optimization and Information Computing*, 6(4):497–525.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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

Monisha Kanakaraj and Ram Mohana Reddy Guddeti. 2015. Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*, pages 169–170. IEEE.

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. *Frontiers in Artificial Intelligence*, 3:55.

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

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *International conference of the cross-language evaluation forum for european languages*, pages 372–387. Springer.

Janice Nathanson. 2013. The pornography of poverty: Reframing the discourse of international aid's representations of starving children. *Canadian Journal of Communication*, 38(1).

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.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Deep learning and ensemble methods for domain adaptation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 184–189. IEEE.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Chetanya Rastogi, Nikka Mofid, and Fang-I Hsiao. 2020. Can we achieve more with less? exploring data augmentation for toxic comment classification. *arXiv preprint arXiv:2007.00875*.

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*.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics*, 21(4).

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.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2021. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341.

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*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

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

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1298–1308.

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.