DH-FBK at SemEval-2022 Task 4: Leveraging Annotators' Disagreement and Multiple Data Views for Patronizing Language Detection

Alan Ramponi
Fondazione Bruno Kessler
Trento, Italy
alramponi@fbk.eu

Elisa Leonardelli
Fondazione Bruno Kessler
Trento, Italy
eleonardelli@fbk.eu

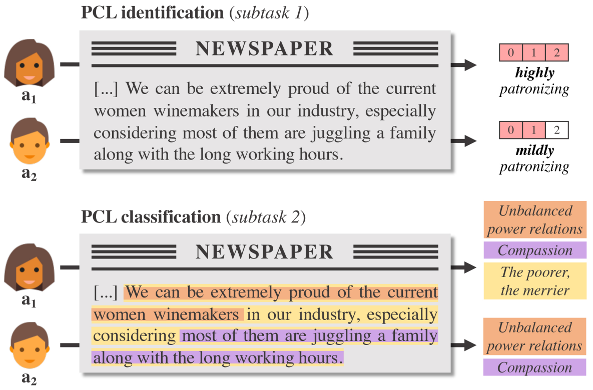
Abstract

The subtle and typically unconscious use of patronizing and condescending language (PCL) in large-audience media outlets undesirably feeds stereotypes and strengthens power-knowledge relationships, perpetuating discrimination towards vulnerable communities. Due to its subjective and subtle nature, PCL detection is an open and challenging problem, both for computational methods and human annotators. In this paper we describe the systems submitted by the DH-FBK team to SemEval-2022 Task 4, aiming at detecting PCL towards vulnerable communities in English media texts. Motivated by the subjectivity of human interpretation, we propose to leverage annotators' uncertainty and disagreement to better capture the shades of PCL in a multi-task, multi-view learning framework. Our approach achieves competitive results, largely outperforming baselines and ranking on the top-left side of the leaderboard on both PCL identification and classification. Noticeably, our approach does not rely on any external data or model ensemble, making it a viable and attractive solution for real-world use.

1 Introduction

Detecting patronizing and condescending language (PCL) is an open, challenging, and underexplored research area in natural language processing (Pérez-Almendros et al., 2020; Wang and Potts, 2019). A patronizing and condescending attitude is expressed as a good-natured and beneficial attitude from a person of authority towards others, who are typically depicted in a subtly compassionate way (Pérez-Almendros et al., 2020).

PCL is a mildly perceived phenomenon. It is often unconscious, driven by good intentions, and expressed through *flowery wordings* (Wong et al., 2014; Huckin, 2002). This makes PCL identification and classification difficult both for NLP systems and human annotators (cf. Figure 1), as it cannot be linked to specific words. Nonetheless, it undesirably conveys harmful messages in many ways, as it promotes stereotypes, and a superiority and discriminatory mindset (Fiske, 1993). This is particularly damaging when used by large-audience media outlets, since it drives greater exclusion of already vulnerable communities (Pérez-Almendros et al., 2020). Automatically detecting PCL has the potential to enable a range of applications and research directions, such as suggestion tools for news editors to mitigate condescension in writing before publication, and studies on the interplay between condescension and sociodemographic factors.



[...] We can be extremely proud of the current women winemakers in our industry, especially considering most of them are juggling a family along with the long working hours.

[...] We can be extremely proud of the current women winemakers in our industry, especially considering most of them are juggling a family along with the long working hours.

Figure 1: Example showing that patronizing and condescending language is a subtle linguistic phenomenon that human annotators (a_1, a_2) often perceive differently, and thus annotate in different ways.

To encourage research on patronizing and condescending language detection, the SemEval-2022 Task 4 (Pérez-Almendros et al., 2022) has recently been proposed. The shared task aims at investigating methods for the identification of PCL (subtask 1; Figure 1, top) and the categorization of the linguistic techniques which are used to express it (subtask 2; Figure 1, bottom) on English news stories mentioning vulnerable communities (Section 2).

In this paper, we present the **DH-FBK** entry for the SemEval-2022 Task 4 (Pérez-Almendros et al., 2022). Motivated by the subtle nature of

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 324 - 334
July 14-15, 2022 ©2022 Association for Computational Linguistics

Example	Category
"We can be extremely proud of the current women winemakers"	Unbalanced power relations
"The inclusion of a refugee team"	Shallow solution
"An immigrant to a developed country lives in two worlds"	Presupposition
"women must wake up"	Authority voice
"trapped in the prison of poverty"	Metaphor
"more than 400 suspected asylum seekers are awaiting their fate"	Compassion
"how talented disabled people can be"	The poorer, the merrier

Table 1: Examples of text excerpts expressing patronizing and condescending language, along with their category.

patronizing language and the subjectivity of human interpretation, we propose a multi-task, multi-view learning approach which leverages annotators' uncertainty and disagreement as auxiliary tasks (Section 3) in judging for PCL presence (subtask 1) or category (subtask 2). Further, we investigate the effectiveness of sequentially fine-tuning on subtasks of increasing complexity (subtask 1 $\mapsto$ subtask 2), as well as the use of additional information such as the geographical provenance of news outlets.

Our systems achieve competitive results on the SemEval-2022 Task 4, outperforming the organizers' RoBERTa (Liu et al., 2019) baseline by a large margin (subtask 1: $+8.8$ F_1; subtask 2: $+26.9$ F_1) and consistently ranking on the top-left side of the leaderboard (subtask 1: 18th out of 78 teams; subtask 2: 13th out of 49 teams) without using any external data or ensemble strategy, making it a viable solution for real-world use. We make our code publicly available to the research community to encourage future work on this direction.[1]

2 Data and task description

In this section, we present relevant details on data and the associated shared task. We firstly summarize the dataset (Section 2.1) and describe the task setup (Section 2.2). Next, we focus on the data annotation process as it is central for understanding our methods (Section 2.3).

2.1 "Don't Patronize Me!" data

The organizers of SemEval-2022 Task 4 provide participant teams with the "Don't Patronize Me!" annotated dataset, originally introduced in Pérez-Almendros et al. (2020) and further updated for the purpose of the shared task (v1.4). The dataset comprises a selection of 10,469 paragraphs published in the news of 20 English-speaking countries[2] from all over the world between 2010 and 2018, and sampled from the "News on Web" corpus (NoW; Davies, 2013). Each paragraph mentions one of ten selected vulnerable communities (i.e., *disabled*, *homeless*, *hopeless*, *immigrant*, *in need*, *migrant*, *poor families*, *refugee*, *vulnerable*, and *women*). These communities have been chosen because they are often target of PCL. Notably, attention has been paid to balance paragraphs across communities and news outlets' countries. For further details, we refer to Pérez-Almendros et al. (2022, 2020).

2.2 Task setup

The SemEval-2022 Task 4 challenge is divided into the following two subtasks:

1. **PCL identification**: given an input paragraph, identify whether it entails any form of PCL. Formally, this is a binary classification task;

2. **PCL classification**: given an input paragraph, decide what linguistic techniques are used to express the condescension (if any). This is a multi-label classification task, with 7 possible labels – i.e., *unbalanced power relations* (UNB), *shallow solution* (SHA), *presupposition* (PRE), *authority voice* (AUT), *metaphor* (MET), *compassion* (COM), and *the poorer, the merrier* (THE). These categories follow a validated PCL taxonomy (Pérez-Almendros et al., 2020) that we summarize in Appendix A. Examples for each category are in Table 1.

As PCL is a subtle and mild phenomenon, the annotation process was not straightforward (Pérez-Almendros et al., 2020). In the following section, the annotation scheme followed by dataset creators

[1] The source code is available at https://github.com/dhfbk/pcl-detection-disagreement.

[2] Covered English-speaking country codes: au, bd, ca, gb, gh, hk, ie, in, jm, ke, lk, my, ng, nz, ph, pk, sg, tz, us, and za.

Annotation task	Individual decisions (a_1,a_2)	Score	Instances	Gold label
Subtask 1: *"Does the paragraph contain any form of PCL?"* Values: $0, 1, 2$	(0,0)	0	8,529	No
	(0,1), (1,0), *	1	947	
	(1,1), *	2	144	Yes
	(2,1), (1,2), *	3	458	
	(2,2)	4	391	
Subtask 2: *"Which PCL category does the span express (if any)?"* Values: $c_i, c_j \in C$, NONE	(c_i, NONE), (NONE, c_i) $(c_i, c_j)_{c_i \neq c_j}$, $(c_j, c_i)_{c_j \neq c_i}$	1	1,359	c_i c_i, c_j
	(c_i, c_i)	2	1,401	c_i

Table 2: The annotation process from individual annotators' decisions to gold labels for both subtasks. For subtask 1, two annotators (a_1, a_2) assigned value 0 (no PCL), 1 (mild PCL) or 2 (high PCL) to each instance of the dataset. The "Score" column indicates the sum of their decisions. In subtask 2, the annotators further characterized the PCL instances of subtask 1, by identifying exact text spans and determining categories of the PCL (if any). We generalize with $c_i, c_j \in C$ two of the $|C| = 7$ possible PCL categories which annotators could have chosen, to show the process in case of disagreement. The "Score" column indicates the number of annotators which agreed on the category. *Includes cases of total disagreement – i.e., (0,2) and (2,0) – resolved by a third annotator a_3.

is presented, as it is especially relevant for understanding the data itself and our methods.

2.3 Annotation process

The dataset has been manually labeled by expert annotators[3] following a two-step process as described below. The resulting annotations are the gold-standard reference for the subtasks (Section 2.2).

Subtask 1 The annotation was performed by three annotators. Two annotators labeled the whole dataset (a_1 and a_2), while a third one (a_3) intervened in case of clear disagreement between a_1 and a_2. Authors reported that "*this annotation step proved more difficult than expected, stemming from the often subtle and subjective nature of PCL*" (Pérez-Almendros et al., 2020). Because of this difficulty, annotators were given the possibility to assign each paragraph a value 0 (no PCL), 1 (borderline), or 2 (highly PCL). Information about the annotation is available as a 5-point scale, which reflects a joint notion of uncertainty and agreement between annotators. For subtask 1, organizers map values into a binary form (i.e., $\{0, 1\} \mapsto$ NO-PCL, $\{2, 3, 4\} \mapsto$ PCL), evaluating systems accordingly. As anticipated, cases of total disagreement (i.e., a_1: 0 and a_2: 2, and viceversa) received a third independent annotation by a_3.[4] If a_3 considered the paragraph not to contain PCL, a borderline case, or an otherwise clear PCL case, the paragraph was assigned a value of 1, 2 or 3, respectively. This has the effect to leave extreme values (0 and 4) reserved for clear-cut cases. The number of PCL-expressing paragraphs is 993 (9.5%). A summary of the annotation process for subtask 1 is in Table 2, top.

Subtask 2 In the second round of annotation, paragraphs previously labeled as containing PCL were further characterized by a_1 and a_2. The aim is two-fold: (i) identify the paragraph segments (or spans) that express PCL, and (ii) categorize each of them into one or more PCL categories (cf. Section 2.2). As a consequence, each identified span exhibits one or multiple labels, depending on whether one or both annotators identified it, and on their agreement on the type(s) of condescension expressed by the text segment. This results in a per-span per-type agreement information on a 2-point scale (1 or 2). Organizers frame subtask 2 as a paragraph-level classification problem, and thus each paragraph can express zero or more condescension types based on the resulting 2,760 span annotations (2.8 annotations per paragraph, on average). An overview of the annotation process for subtask 2 is presented in Table 2, bottom.

3 Methods

Models proposed for PCL identification and classification are all based on multi-task learning (Caruana, 1997) and use multiple views of input data, inspired by Clark et al. (2018). In this section, we firstly introduce the general framework on which all our models are based on (Section 3.1).

[3] Dataset authors reported the annotators' background is on the fields of communication, media, and data science.

[4] According to the first dataset release, these account for 5.5% of the annotations (Pérez-Almendros et al., 2020).

Then, we provide details on data- and task-specific components that we use in our systems, namely dataset views (Section 3.2) and auxiliary tasks (Section 3.3). Lastly, we present the composition of our final models (Section 3.4).

3.1 General framework

Our approach is based on multi-task learning, a learning paradigm that aims to leverage training signals of related tasks at the same time by exploiting a shared representation in the model (Caruana, 1997). In all our models, we employ a *main* task, namely PCL identification or classification, which is a task of direct interest. Additionally, we employ *auxiliary* tasks (see Section 3.3), namely tasks which can provide useful signals to potentially improve the performance on the main task.

All our models use RoBERTa-base (Liu et al., 2019) as shared encoder, and a separate decoder for each task. This way, all tasks benefit from mutual signals encoded by a shared contextualized representation that is jointly fine-tuned during training.

The input is a text instance that is encoded using byte-pair encoding (BPE; Sennrich et al., 2016), whereas the output label is given by task-specific decoders which consist of a linear classification layer and operate on the contextual embedding of the special `[CLS]` classification token.

In our models, each auxiliary task makes use of a specific form (or view) of the original dataset, which we introduce in the following (Section 3.2). When training with multiple views of data, each input batch to the model consists of examples from a single data view, and the loss function is only activated for tasks associated to that data view. For further details on multi-view (or multi-dataset) training, refer to van der Goot et al. (2021). An overview of the framework is presented in Figure 2.

3.2 Data views

Our models employ different forms (or views) of the original "Don't Patronize Me!" dataset provided by organizers. Specifically, we use (i) paragraph and (ii) span views as detailed as follows.

Paragraph data view (D_P) This corresponds to the dataset in its standard form – i.e., whole paragraphs – as provided by organizers (Section 2.1).

Span data view (D_S) A dataset consisting of all text excerpts – i.e., paragraph substrings – that have been marked as expressing patronizing and condescending language. As a result, this dataset represents a different view of D_P data, where only snippets of PCL are included. Examples of text instances in D_S are reported in Table 1.

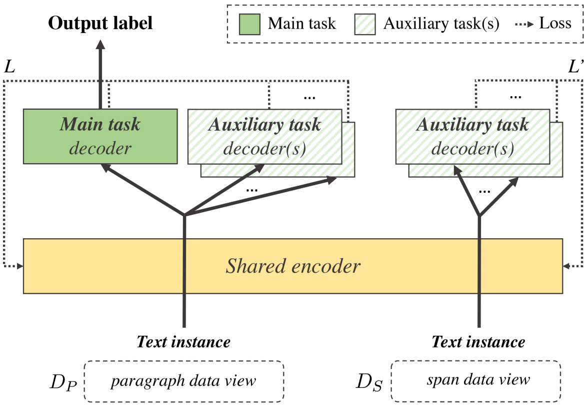


Figure 2: A high-level overview of our multi-task, multi-view learning framework.

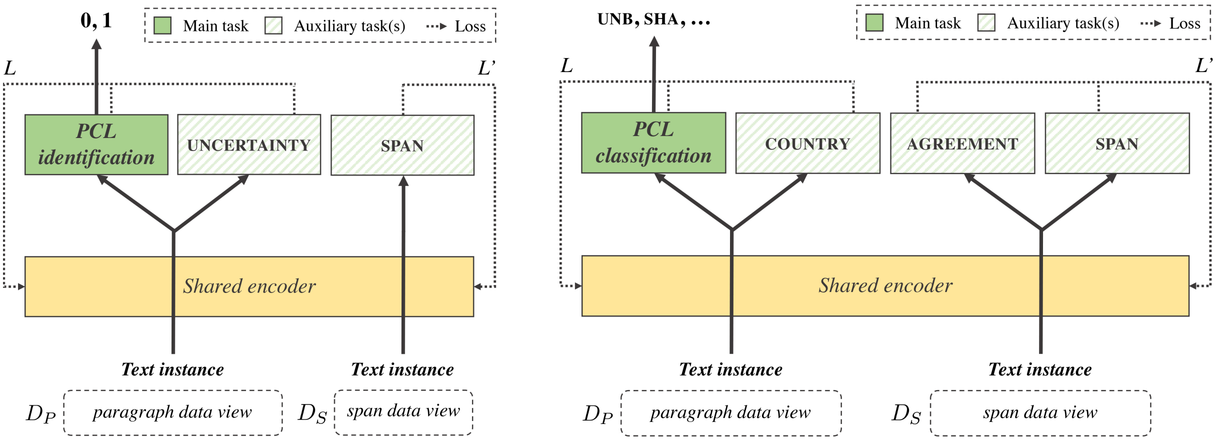
These different data views are used by specialized task decoders, as presented in Section 3.3.

3.3 Auxiliary tasks

In this section, we describe the auxiliary tasks we used in one or more of our final models (Section 3.4), along with the data view each task uses. For details on the interplay between main and auxiliary tasks, we refer the reader to Section 4.1.

Paragraph uncertainty level (UNCERTAINTY) This task is used for subtask 1 to consider different annotators' point of view in identifying PCL. We use the aggregated 5-point scale score (cf. Section 2.3, subtask 1) assigned to each paragraph as auxiliary task. Although disaggregated annotators' decisions have not been made available for the shared task, we argue that the combined annotation provided by organizers can be viewed as a joint notion of uncertainty and agreement between annotators in identifying PCL, and is thus valuable information that can inform PCL identification. Since this is a paragraph-level information, this auxiliary task uses the D_P data view. For each paragraph, a label $l \in \{0, 1, 2, 3, 4\}$ must be predicted.

Span agreement level (AGREEMENT) This task is used for subtask 2 as it potentially drives useful signals for PCL classification. We hypothesize that the number of annotators which agree on a particular PCL category for a span (cf. Section 2.3, subtask 2) is a crucial information as it can provide the main task with different shades of PCL based on annotators' interpretation and sensibility. The



(a) MTMW(UNC+SPAN) model for subtask 1.

(b) MTMW(AGR+COU+SPAN) model for subtask 2.

Figure 3: Our selected models for PCL identification (subtask 1) and classification (subtask 2).

agreement level is a span-level information, and thus we use the D_S view for the associated task. For each span, a label $l \in \{1, 2\}$ must be predicted.

Span categorization (SPAN) This auxiliary task is used for both subtask 1 and 2, as we argue that making use of small and focused units of information such as condescending text excerpts would be useful to inject knowledge in the encoder about which paragraph segments are important for recognizing PCL. The task uses the D_S data view, and for each span a label $l \in \{$UNB, SHA, PRE, AUT, MET, COM, THE$\}$ has to be predicted (i.e., one among the condescending types in Section 2.2).

News outlet country (COUNTRY) We employ this auxiliary task for our subtask 2 model.[5] We hypothesize that the diverse background culture of countries results in variation in language use, and thus could have an impact on the expression of patronizing and condescending language. The news outlet provenance is provided by organizers along with the original dataset. The auxiliary task uses the D_P view, and the label to be predicted is a country code, i.e., $l \in \{$au, bd, ca, gb, gh, hk, ie, in, jm, ke, lk, my, ng, nz, ph, pk, sg, tz, us, za$\}$.

3.4 Models

Our approaches to PCL identification and classification are all centered on the hypothesis that leveraging annotators' uncertainty and disagreement during training is beneficial for capturing the subtle language which characterizes PCL. This is in line with recent work emphasizing the importance of modeling annotators' disagreement in subjective tasks (Davani et al., 2022; Leonardelli et al., 2021; Uma et al., 2021) and initiatives supporting the release of disaggregated annotations in NLP (Abercrombie et al., 2022).

We participated in the SemEval-2022 Task 4 challenge with two submissions for both subtasks. To this end, we built three models. Two models are our best systems – as evaluated on the development set (cf. Section 4.3) – on subtask 1 (Section 3.4.1) and subtask 2 (Section 3.4.2). We submit them for the corresponding subtasks. The third model was instead built aiming at a generic and unified solution, and represents our second submission for both subtasks (Section 3.4.3).

3.4.1 MTMW(UNC+SPAN) model for PCL identification (subtask 1)

In this multi-task, multi-view model (MTMW), we consider PCL identification as the main (binary classification) task, and employ UNCERTAINTY and SPAN as auxiliary tasks, as described in Section 3.3. We refer to this model to as MTMW(UNC+SPAN). An overview of the resulting approach for PCL identification is presented in Figure 3a.

3.4.2 MTMW(AGR+COU+SPAN) model for PCL classification (subtask 2)

In this multi-task, multi-view model (MTMW), the main task is PCL classification, whereas AGREEMENT, COUNTRY and SPAN are auxiliary tasks. For PCL classification, each label is modeled through a dedicated binary classification decoder. We use MTMW(AGR+COU+SPAN) to refer to this approach in the remainder of this paper. We graphically

[5] We have experimented with this auxiliary task on subtask 1 too; however, we noticed a substantial performance degradation compared to using uncertainty only (cf. Section 4.3).

present the model in Figure 3b, and refer the reader to Section 3.3 for details on auxiliary components.

3.4.3 Sequential PCL identification and classification (subtasks 1 & 2)

Given that PCL classification is a fine-grained version of the PCL identification task, we argue that sequentially fine-tuning encoder weights on tasks of increased complexity should be beneficial to improve the performance on both subtasks. This approach borrows the idea from modern data-centric adaptation methods in NLP (Ramponi and Plank, 2020) such as continued pretraining of language models (Gururangan et al., 2020), however employing it at the fine-tuning stage, similarly to intermediate-task transfer (Phang et al., 2018).

We firstly run the PCL identification model described in Section 3.4.1. Then, we use the resulting fine-tuned encoder weights as initialization for the encoder of the PCL classification model (Section 3.4.2) with SPAN auxiliary only.[6] Finally, we fine-tune it on subtask 2. This results in a single model that has incrementally learnt the complexity of PCL detection as a whole. Prediction of labels for subtask 1 were done simply considering a paragraph as containing PCL if it exhibits at least a PCL category label in subtask 2. We refer to this approach to as SEQ. FINE-TUNING model.

4 Experiments

In this section, we first outline the experimental setup (Section 4.1). Then, we present the results of our models (Section 4.2), as well as additional analyses and discussion (Section 4.3).

4.1 Experimental Setup

We implemented our models using the MaChAmp v0.2 toolkit (van der Goot et al., 2021) and employ RoBERTa-base (Liu et al., 2019) as shared encoder since it has been shown to outperform other commonly used pretrained language models on PCL detection tasks (Pérez-Almendros et al., 2020).

For training, we use default hyperparameters (Appendix B) and fine-tune each model – roughly 110M trainable parameters – for 10 epochs on a single GPU.[7] We use a cross-entropy loss with balanced class weights to give equal importance to all classes during fine-tuning, and thus emphasizing underrepresented classes in training data. The multi-task learning loss is computed as $L = \sum_t \lambda_t L_t$, where L_t is the loss for task t, and λ_t the corresponding weighting parameter. In our experiments, we empirically set $\lambda_t = 1$ for the main task, and $\lambda_t = 0.25$ for auxiliary tasks.[8]

	P	R	F$_1$
Organizers' baseline	39.35	65.30	49.11
MTMW(UNC+SPAN)	64.23	52.68	**57.89**
SEQ. FINE-TUNING	53.99	55.52	54.74

Table 3: Official test set results of our models compared to the organizers' RoBERTa baseline on PCL identification (subtask 1). P: Precision; R: Recall; F$_1$: F$_1$ score over the positive class. Best results are in bold.

We solely rely on data provided by organizers, and use the provided 80% train and 20% development split as one fold, additionally creating the remaining 80%/20% splits in a stratified fashion. To avoid confounding our results, we ensure training splits for the span data view do not contain any text excerpt appearing in development data of the paragraph data view. This results in 5 folds that we use for selecting models for our submissions. For official test set evaluation, we then submit the selected models trained on the provided data split.

For the purpose of the shared task, models for PCL identification (subtask 1) are evaluated using F_1 score over the positive class, whereas models for PCL classification (subtask 2) are evaluated based on macro-average F_1 score. We employ the same metrics at the model selection stage.

4.2 Results

We present the official results for our proposed models for subtask 1 and 2 in Table 3 and Table 4, respectively. We also include scores of the organizers' RoBERTa baseline for informed comparison based on the shared task metrics.

Subtask 1 As shown in Table 3, our submitted MTMW(UNC+SPAN) model outperforms the RoBERTa baseline by a large margin, with most of the benefit coming from the precision metric. This indicates that the uncertainty and agreement

[6] This choice is motivated by the need for a simpler and unified solution for both subtasks. We decided to leave country and agreement auxiliaries out for this submission due to negligible differences in performance on subtask 2 (Section 4.3).

[7] NVIDIA Tesla V100-SXM2.

[8] Except for AGREEMENT, for which we empirically set $\lambda_t = 0.125$ since it is actually *auxiliary of an auxiliary task*. A thorough investigation on the impact of the weighting parameter on individual auxiliary tasks is left for future work.

	UNB	SHA	PRE	AUT	MET	COM	THE	F_1
Organizers' baseline	35.35	0.00	16.67	0.00	0.00	20.87	0.00	10.41
MTMW(AGR+COU+SPAN)	52.46	36.22	26.95	**37.71**	**31.86**	**45.95**	**30.30**	**37.35**
SEQ. FINE-TUNING	**54.00**	**46.73**	**28.07**	22.22	29.73	44.28	20.69	35.10

Table 4: Official test set results of our models compared to the organizers' RoBERTa baseline on PCL classification (subtask 2). UNB: Unbalanced power relations; SHA: Shallow solution; PRE: Presupposition; AUT: Authority voice; MET: Metaphor; COM: Compassion; THE: The poorer, the merrier; F_1: macro-average F_1. Best results are in bold.

level as well as the use of focused text excerpts do play a positive role in PCL identification. On the other hand, while also the SEQ. FINE-TUNING approach provides better results compared to the baseline, it scores lower than MTMW(UNC+SPAN). A reason for this behavior can be attributed to the way we infer labels for this subtask (i.e., based on predictions for subtask 2, as anticipated in Section 3.4.3), or due to *catastrophic forgetting* (McCloskey and Cohen, 1989), a phenomenon in which prior knowledge is largely forgotten when learning a new task. On the shared task leaderboard, MTMW(UNC+SPAN) ranked 18th out of 78 teams.

Subtask 2 Similarly to results for subtask 1, both our submitted systems largely outperform the RoBERTa baseline, as shown in Table 4. We notice that the SEQ. FINE-TUNING approach still scores lower than a tailored approach for subtask 2, i.e., MTMW(AGR+COU+SPAN). Specifically, MTMW(AGR+COU+SPAN) shows an absolute improvement of +26.9 points in macro-average F_1 score over the RoBERTa baseline, with a clear advantage over the SEQ. FINE-TUNING model on underrepresented classes (i.e., AUT, MET and THE). Our MTMW(AGR+COU+SPAN) model ranked 13th out of 49 participating teams on the official leaderboard. Overall, our models do not require any external data or model ensemble, and we think this makes them viable approaches for real-world use.

4.3 Analysis and discussion

In order to provide insights for future work on PCL detection, we conduct analyses on the contribution of auxiliary tasks to performance of our models (Section 4.3.1), and an in-depth study on the role of uncertainty and disagreement (Section 4.3.2).

4.3.1 Contribution of auxiliary tasks

Our submitted models for PCL identification and classification leverage training signals coming from selected auxiliary tasks (cf. Section 3.4). These model variants have been chosen after a thorough performance evaluation on the development splits. In Table 5 we report the mean and standard deviation of F_1 scores for each model with different auxiliary task configurations. We denote with "Our single task baseline" our baseline RoBERTa model with no multi-task nor multi-view learning. Note that this is different from organizers' RoBERTa baseline, since we employ a different hyperparameter setup with the addition of class weights (Section 4.1). We do not include the organizers' baseline here due to incomparability –

	Model	F_1 score
subtask 1	Our single task baseline	$56.73_{\pm 3.2}$
	Multi-task setup	
	+ COUNTRY	$55.99_{\pm 2.7}$
	+ UNCERTAINTY	$56.92_{\pm 3.2}$
	+ COUNTRY, UNCERTAINTY	$57.74_{\pm 3.5}$
	Multi-task, multi-view setup	$55.69_{\pm 2.0}$
	+ COUNTRY	$57.35_{\pm 1.9}$
	+ UNCERTAINTY	**58.38**$_{\pm 3.7}$
	+ COUNTRY, UNCERTAINTY	$57.53_{\pm 4.6}$
subtask 2	Our single task baseline	$37.02_{\pm 2.8}$
	Multi-task setup	
	+ COUNTRY	$36.26_{\pm 2.3}$
	Multi-task, multi-view setup	$38.25_{\pm 3.6}$
	+ COUNTRY	$37.16_{\pm 2.3}$
	+ AGREEMENT	$37.53_{\pm 0.8}$
	+ COUNTRY, AGREEMENT	**38.81**$_{\pm 2.9}$

Table 5: Contribution of auxiliary tasks to performance of models for subtask 1 (top) and subtask 2 (bottom). We report mean and standard deviation of F_1 scores on development splits. The multi-task, multi-view setups use SPAN by default. Results for AGREEMENT and its combinations in the multi-task setup of subtask 2 are omitted, since they would require the D_S view, and thus only refer to the multi-task, multi-view configuration.

we only have access to results on a single development split. For reference, organizers reported 48.29 F_1 on subtask 1, and 13.40 F_1 on subtask 2 for their baseline. Multi-task setups refer to configurations where D_P-based auxiliaries are used, whereas multi-task, multi-view setups also exploit the D_S view, thus using SPAN as default auxiliary task on all experiments.

Subtask 1 Table 5 (top) shows the results of models with different auxiliary task combinations on subtask 1 development data. All multi-task, and multi-task multi-view setups outperform the baseline, with the only exception of the multi-task configuration solely using COUNTRY as auxiliary task. Overall, the use of UNCERTAINTY as auxiliary task consistently improves the performance over the baseline, even when coupled with other auxiliaries. The best results are obtained when employing a multi-task, multi-view setup with UNCERTAINTY only (58.38 F_1), suggesting that information coming from the COUNTRY auxiliary is not as useful as in the multi-task scenario. We hypothesize this behaviour could be attributed to the use of SPAN in the multi-task, multi-view setup, which indirectly provides a more useful inductive bias for PCL identification. In future work we aim to further dig into this aspect, exploring various loss weight configurations to assess the strength of this finding.

Subtask 2 We present in Table 5 (bottom) the contribution of auxiliary tasks on subtask 2 development data. Similarly to subtask 1, COUNTRY in the multi-task setup is the only auxiliary task that does not improve the performance over our single task RoBERTa baseline. However, when coupled with AGREEMENT in the multi-task, multi-view configuration, it provides the best overall performance over all model variants (38.81 F_1). This suggests that the AGREEMENT auxiliary provides signals orthogonal to COUNTRY, as shown by the multi-task, multi-view alternatives employing these auxiliaries in isolation. By a closer look, the performance of the multi-task, multi-view setup alone (i.e., SPAN only) are highly competitive, confirming our hypothesis that using PCL-expressing text excerpts is beneficial for PCL detection as a whole.

4.3.2 Role of uncertainty and disagreement

To delve into the role of agreement and uncertainty, we further study performance of our best systems on subtask 1 and 2 as a function of the agreement/uncertainty level. To the goal, we use the predictions of our models on development splits.

level	0	1	2	3	4
F_1	49.27	44.67	27.32	33.39	41.95

Table 6: Performance of our model for subtask 1 as a function of different levels of uncertainty/disagreement. F_1 scores are averages over the five development splits.

Subtask 1 To investigate the impact of uncertainty/agreement on the performance of our model for subtask 1, we first divided each development split by uncertainty/agreement level (cf. “Score”, Table 2, top). Then, we calculated the per-level F_1 score on each split. Finally, we averaged the per-level performance on all folds. We report the experimental results in Table 6. It is interesting to observe how in this task the uncertainty/agreement levels 0, 2 and 4 reflect agreement between annotators (0+0, 1+1, or 2+2), but performance for score 2 – where both annotators agree in being uncertain – is much worse. This suggests a prominent role of uncertainty in worsening classifier’s performance, rather than disagreement. On the other hand, uncertainty and disagreement represent two sides of the same coin, as the less certain and clear a decision is, the greater probability is to have disagreement between annotators.

Subtask 2 We perform a similar analysis for subtask 2. The agreement level (cf. “Score”, Table 2, bottom) has a clear effect on model’s performance: out of the 2,760 PCL-expressing spans, only 44% of 1,359 spans with an agreement level of 1 are correctly labeled, compared to 56% of 1,401 spans for which both annotators signaled a form of PCL. Considering paragraphs with a single PCL-expressing span only, this results to 35% of 801 examples for agreement level 1, and 52% of 798 examples for agreement 2, further confirming that instances exhibiting disagreement are more difficult to classify.

5 Conclusion

In this paper, we presented our submitted systems to SemEval-2022 Task 4. We showed that leveraging annotators’ uncertainty and disagreement during training in a multi-task, multi-view framework is beneficial for the identification and classification of patronizing and condescending language, and achieves competitive results on the official leaderboard without relying on any external data or model ensemble. We also showed that sequential fine-tuning is a viable alternative to tackle PCL identification and classification jointly, with the goal of reducing the use of computational resources during inference, although it obtains lower performance compared to our tailored solutions. A thorough analysis on the impact of diverse auxiliary tasks on the performance of our models for PCL detection, and an investigation on the role of uncertainty and disagreement further confirmed the importance of considering annotators' point of view in PCL detection. As future work, we aim to test the presence and assess the impact of spurious lexical biases in the dataset (Ramponi and Tonelli, 2022) and extend our models to other genres, such as social media (Wang and Potts, 2019). We hope this work will encourage future efforts towards annotators-centric NLP, on PCL detection and other subjective tasks more broadly.

Acknowledgements

Part of this work was funded by the PROTECTOR European project (ISFP-2020-AG-PROTECT-101034216-PROTECTOR). This research was also supported by the PERCEPTIONS project under European Union's H2020 Research & Innovation Action, Grant Agreement No. 833870.

References

Gavin Abercrombie, Valerio Basile, Verena Rieser, Sara Tonelli, and Alexandra Uma, editors. 2022. *Proceedings of the 1st International Workshop on Perspectivist Approaches to NLP*. European Language Resources Association, Marseille, France.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Mark Davies. 2013. Corpus of news on the web (now): 3+ billion words from 20 countries, updated every day. Available online at `https://corpus.byu.edu/now/`.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

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

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

Carla Pérez-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)*. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, Washington, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multitask learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. 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 (EMNLP-IJCNLP)*, pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

Gloria Wong, Annie O Derthick, EJR David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and social problems*, 6(2):181–200.

Appendix

A PCL categories

In the following, we provide the precise definitions of the PCL categories according to Pérez-Almendros et al. (2020).

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.*”

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.*”

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 cliches are also considered to be examples of presupposition.*”

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.*”

Metaphor “*[Metaphors] 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. [...] Euphemisms are considered as an example of metaphors.*”

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.*”

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’.*”

Hyperparameter	Value
Optimizer	AdamW
β_1, β_2	0.9, 0.99
Dropout	0.3
Epochs	10
Batch size	32
Learning rate (LR)	0.0001
LR scheduler	Slanted triangular
Decay factor	0.38
Cut fraction	0.2
Main task loss weight	1
Aux task loss weight	0.25
Aux’s aux task loss weight	0.125

Table 7: Hyperparameter values used for all our experiments.

B Hyperparameters

The hyperparameter setting for all our models is presented in Table 7. This reflects the default MaChAmp’s hyperparameter values (van der Goot et al., 2021), with the addition of loss weights, as introduced in Section 4.1, and 10 epochs of training as suggested in the original RoBERTa publication (Liu et al., 2019).

C Credits

People icons included in Figure 1 are by `https://icons8.com/`.