DH-FBK at SemEval-2022

Leveraging annotators' disagreement and multiple data views for patronizing language detection

Alan **Ramponi** & Elisa **Leonardelli**
Fondazione Bruno Kessler, Trento, Italy

16th International Workshop on Semantic Evaluation (July 14-15, 2022 – Seattle, USA)

Background: Patronizing and condescending language (PCL)

What is PCL? Language use denoting superior attitude towards others, who are talked down or depicted in a compassionate way [Pérez-Almendros et al., 2020]

- **Subtle**: often unconscious, good-natured
- **Undesirably conveys harm**: promotes stereotypes & superiority mindset

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

For definitions of PCL categories refer to [Pérez-Almendros et al., 2020]

Background: Subjectivity of PCL detection

Challenges PCL is a linguistic phenomenon that human annotators often perceive differently due to background and sensibility, and thus annotate in different ways

a_1 a_2

Data and task: SemEval-2022 Task 4 overview

Data *"Don't Patronize Me!"* annotated dataset (v1.4) [Pérez-Almendros et al., 2020]

- 10,469 *en* paragraphs from the news of 20 English-speaking countries (2010–18) from "News on Web" corpus [Davies, 2013]
- Each paragraph mentions one of ten selected vulnerable communities
    - E.g., disabled, homeless, immigrant, migrant, poor families, refugee, women, amongst others

Task setup Given an input paragraph *P*:

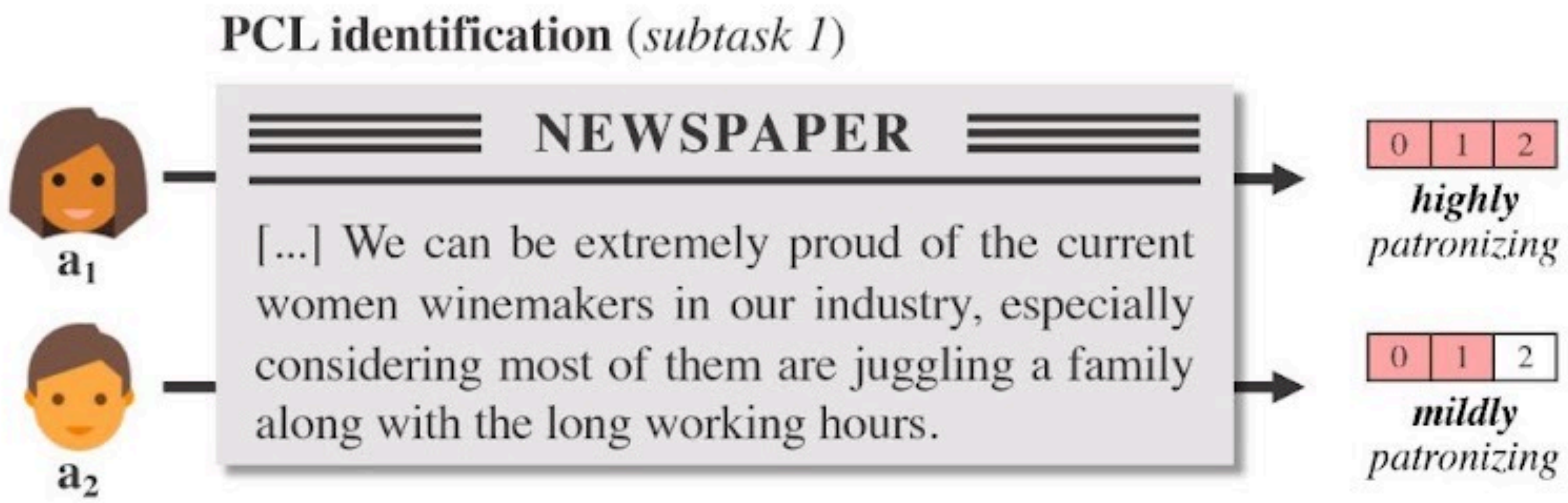
- **PCL identification** (Subtask 1): identify whether *P* entails any form of PCL
- **PCL classification** (Subtask 2): determine PCL forms expressed by *P* (if any)

Don't Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities (Perez Almendros et al., COLING 2020)
Corpus of News on the WEB (NEW): 3+ Billion Words from 20 Countries, Updated Every Day (Davies, 2013). Available online at: https://corpus.byu.edu/now/

Data and task: A closer look at the annotation

Subtask 1 PCL identification

- **Annotators a_1, a_2 labeled all *Ps*:** 0 (*no PCL*), 1 (*borderline*), 2 (*highly PCL*)



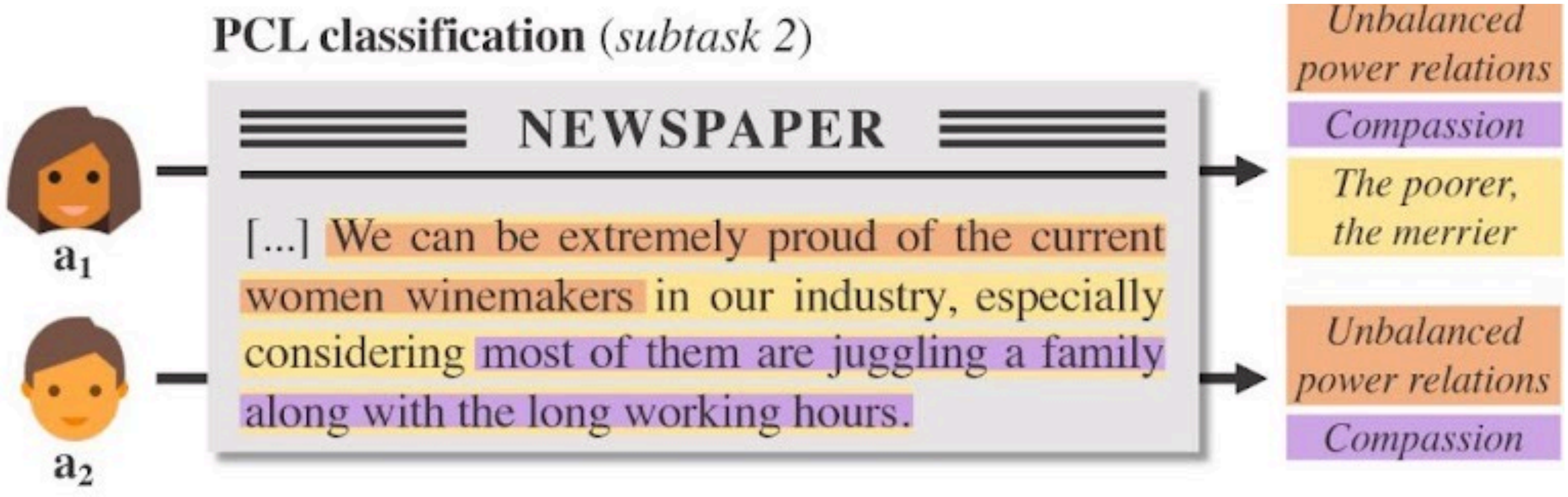
Gold labels Sum of decisions mapped to binary – {0, 1}→**NO-PCL**, {2, 3, 4}→**PCL**

> 💡 **Idea**
> *The raw 5-point scale score can be viewed as a joint notion of **uncertainty** and **agreement** between annotators*

Data and task: A closer look at the annotation

Subtask 2 Characterization of PCL-containing *Ps* with PCL categories

- **Annotators a_1, a_2 identified & categorized PCL-expressing spans** within *P*



Each span exhibits 1+ labels, depending on agreement of annotators on PCL presence/type

Gold labels Paragraph level

💡 **Idea**

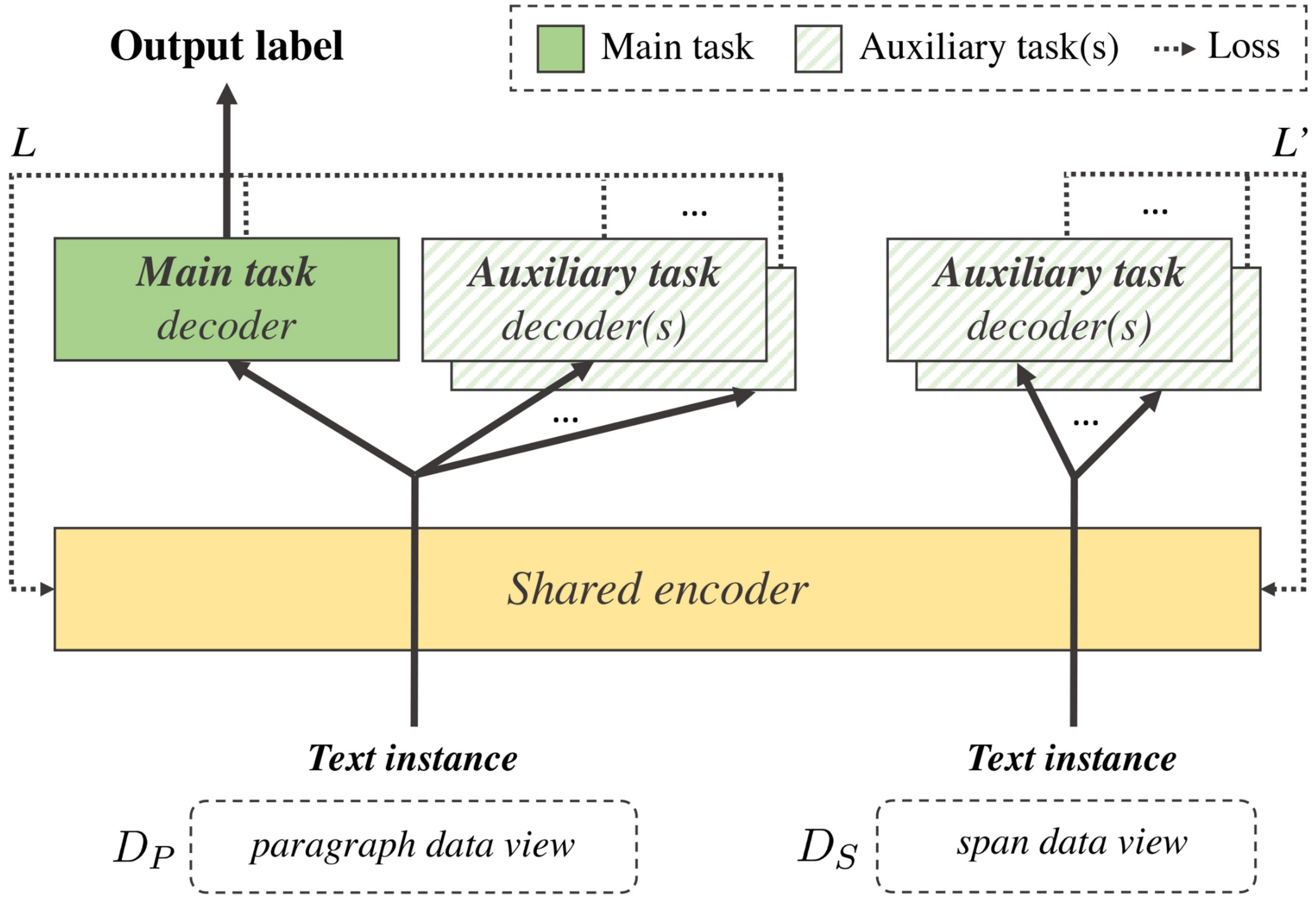
*Per-span, per-type agreement information on a 2-point scale reflects **annotators' interpretation and sensibility***

➢ *leveraged to model different shades of PCL*

Methods: General framework

Multi-task learning framework

- Shared encoder: common representation
- Main task decoder: for the end task
  - i.e., PCL detection or PCL classification
- Auxiliary task decoder(s): for providing useful signals to improve the main task



Leveraging multiple views Different forms (or *views*) of the dataset

- **Paragraph data view** (D_P): dataset in its standard form (i.e., paragraphs)
- **Span data view** (D_S): dataset consisting of all PCL-expressing spans from D_P

Methods: Auxiliary tasks and associated data views

Paragraph uncertainty level (*uncertainty*): 5-point scale score assigned to *P*

- Label space: {0, 1, 2, 3, 4}, data view: D_P, suitable for: subtask 1

Span agreement level (*agreement*): 2-point scale score assigned to spans in *P*

- Label space: {1, 2}, data view: D_S, suitable for: subtask 2

Span categorization (*span*): classification of PCL-expressing text excerpts

- Label space: {UNB, SHA, PRE, …}, data view: D_S, suitable for: subtask 1, 2

News outlet country (*country*): classification of provenance country

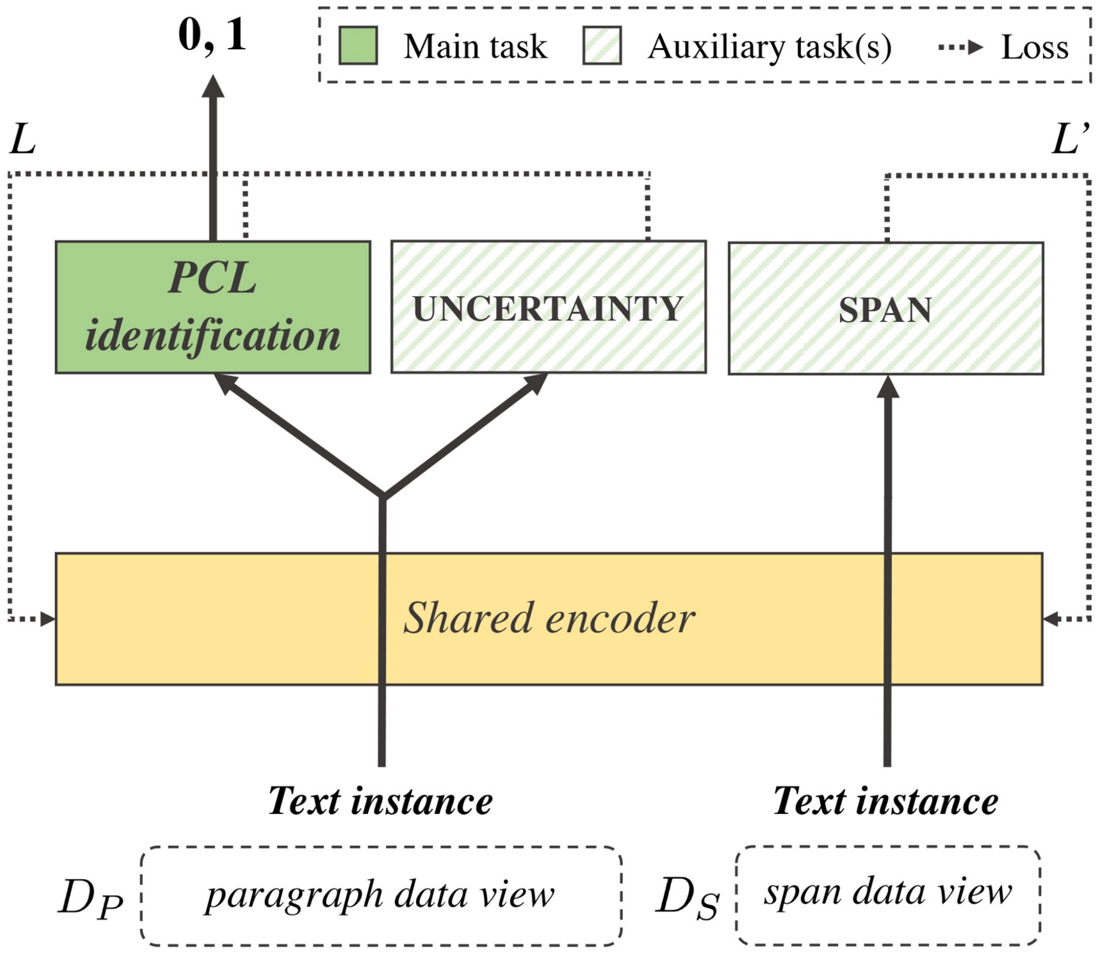
- Label space: {au, bd, ca, gb, gh, hk, …}, data view: D_P, suitable for: subtask 1, 2

Methods: Models

We design **3 models** which leverage annotators' uncertainty & disagreement

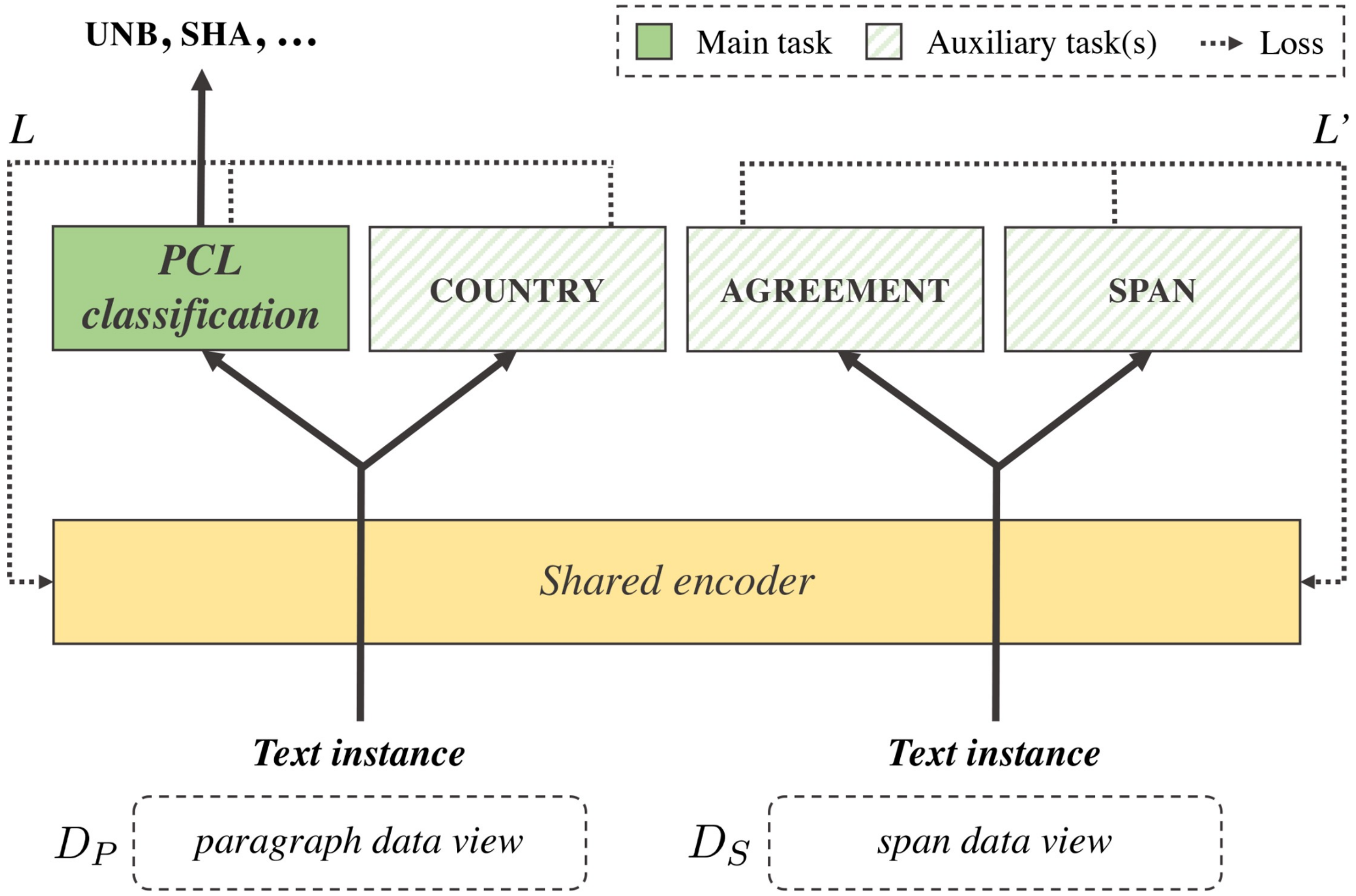
(1) MTMW(UNC+SPAN)

model for subtask 1



(2) MTMW(AGR+COU+SPAN)

model for subtask 2



(3) SEQ. FINE-TUNING

model for subtask 1 and 2

sequential fine-tuning approach inspired by [Gururangan et al., 2020]

1. Finetune on *subtask 1*
2. Use (1)'s weights to finetune on *subtask 2*
3. Use model to predict both *subtask 1* and *2*

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks (Gururangan et al., ACL 2020)

Experiments: Setup

All our models are based on **MaChAmp** v0.2 toolkit [van der Goot et al., 2021]

- **Encoder**: RoBERTa-base, with default hyperparameters and 10 epochs
- **Training loss**: cross-entropy with balanced class weights
- **Auxiliary tasks' weights**: empirically, λ=0.25 (λ=1 for main task)
- **Model selection**: stratified 5-fold cross-validation, shared task metrics
  - Subtask 1: F_1 score over positives, Subtask 2: macro F_1 score

No additional external data for training

No model ensembles – focus on *environmental impact* and *real-world usage*

Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP (van der Goot et al., EACL 2021)

Experiments: Results on test set

Comparison of test set scores to organizers' (RoBERTa-base) baseline

PCL identification

subtask 1

	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

🏆 **18th** / 78 teams

🏆 **13th / 49 teams**

PCL classification

subtask 2

	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

Analysis: Auxiliary tasks and role of disagreement

Contribution of auxiliary tasks (*main insights*)

- Subtask 1: overall, *uncertainty* as auxiliary consistently improves performance over the baseline
- Subtask 2: *agreement* as auxiliary provides signals orthogonal to *country* (i.e., they help each other)

	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	$\mathbf{58.38}_{\pm 3.7}$
	+ 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	$\mathbf{38.81}_{\pm 2.9}$

Role of uncertainty and disagreement

- Subtask 1: F_1 score across uncertainty/agreement levels suggests a prominent role of uncertainty in worsening performance, rather than disagreement
- Subtask 2: similar analysis confirms that instances exhibiting disagreement are more difficult to classify

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

Conclusion

- PCL feeds stereotypes, strengthens power-knowledge relationships, and perpetuates discrimination towards vulnerable communities

- PCL detection depends on **annotators’ interpretation and sensibility**
  - Future efforts should start considering annotators-centric NLP for subjective tasks

- Leveraging **annotators’ uncertainty and disagreement is beneficial**
  - A multi-task, multi-view learning allows to consider different perspectives

- Our approach achieves **competitive results on PCL detection**
  - No need for external data sources or model ensembles