Human Label Variation in Linguistic Annotation

Sara Tonelli
satonelli@fbk.eu

Linguistic Annotation

Why linguistic annotation?

- Corpus studies
- Training data for supervised learning (NLP models)
- Benchmarks
- Performance comparisons

Linguistic Annotation

Ingredients of Linguistic Annotation:

1. A set of **instances** to annotate
2. A target **phenomenon** described by guidelines
3. **Annotation scheme** described by guidelines
4. Annotation by **multiple subjects**
5. Measure of **inter-annotator agreement**
6. Aggregation by *majority voting*

Assumptions on Linguistic Annotation

Assumption: Natural language expressions have a single and clearly identifiable interpretation
Reality: Evidence of genuine disagreement from large-scale annotation projects.

> “ *What does disagreement tell us?* ”

Possible sources of disagreement:

1. Disagreement due to noise (e.g., spammers)
2. Disagreement due to inaccurate guidelines
3. Disagreement due to ambiguity
4. Disagreement due to item difficulty
5. Disagreement due to subjectivity

Sources of Disagreement

Disagreement due to inaccurate guidelines

> “ *2nd wave about to be a bitch* ”

Task: Offensive language detection

Sources of Disagreement

Disagreement due to ambiguity

> “ *Dude this guy is serious?*
> *And trump retweeted this??????* ”

Task: Offensive language detection

Sources of Disagreement

Disagreement due to item difficulty

> “*Trump is a walking petri dish. His goal is to spread the virus to as many people as possible*”

Task: Offensive language detection

Sources of Disagreement

Disagreement due to subjectivity

“ ***Premise***: *The important thing is to realise that it’s way past time to move it*

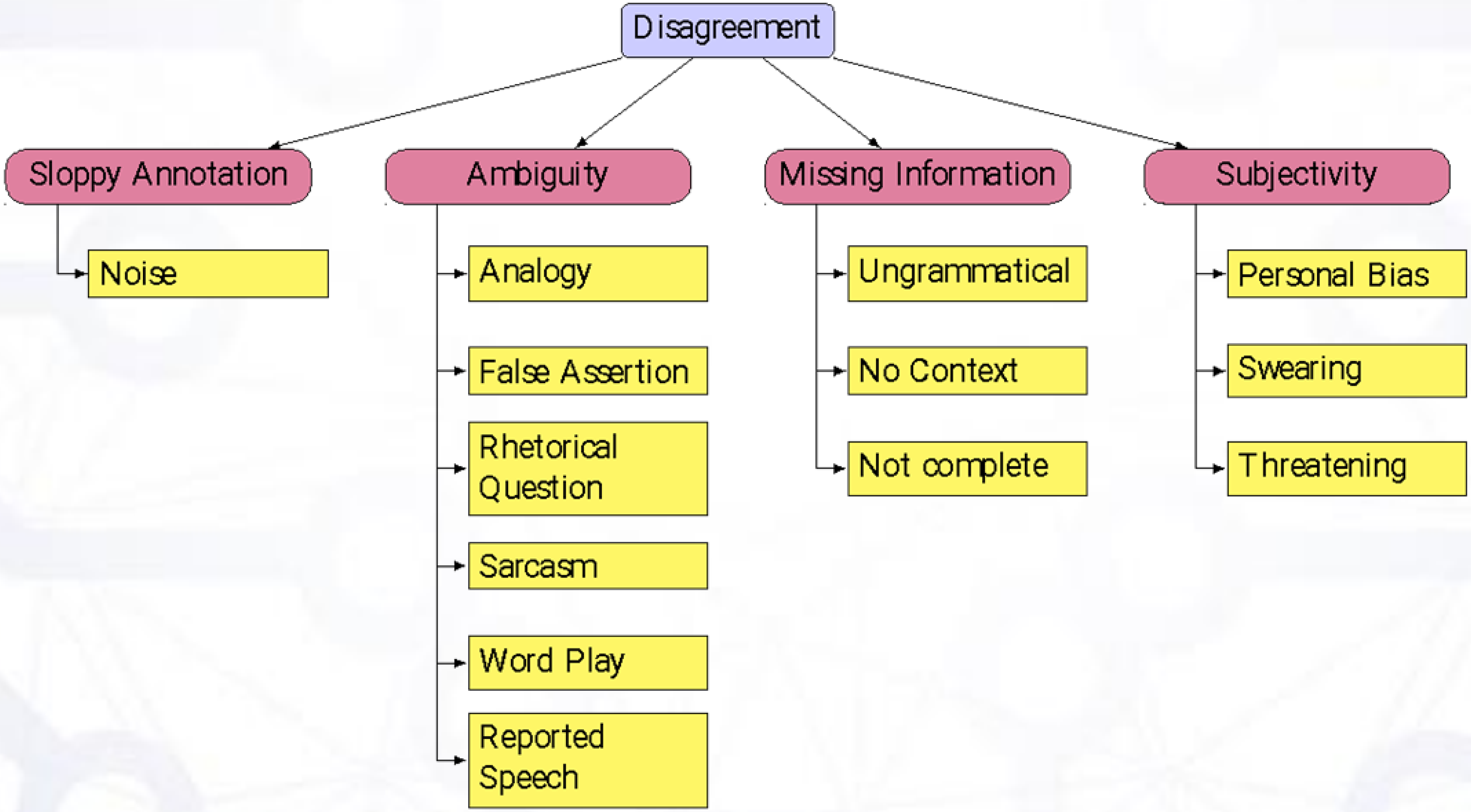
Hypothesis: *It cannot be moved, now or ever* ”

Task: Textual entailment

From Baan, Aziz, Plank, Fernandez, 2022 (EMNLP)

FONDAZIONE
BRUNO KESSLER

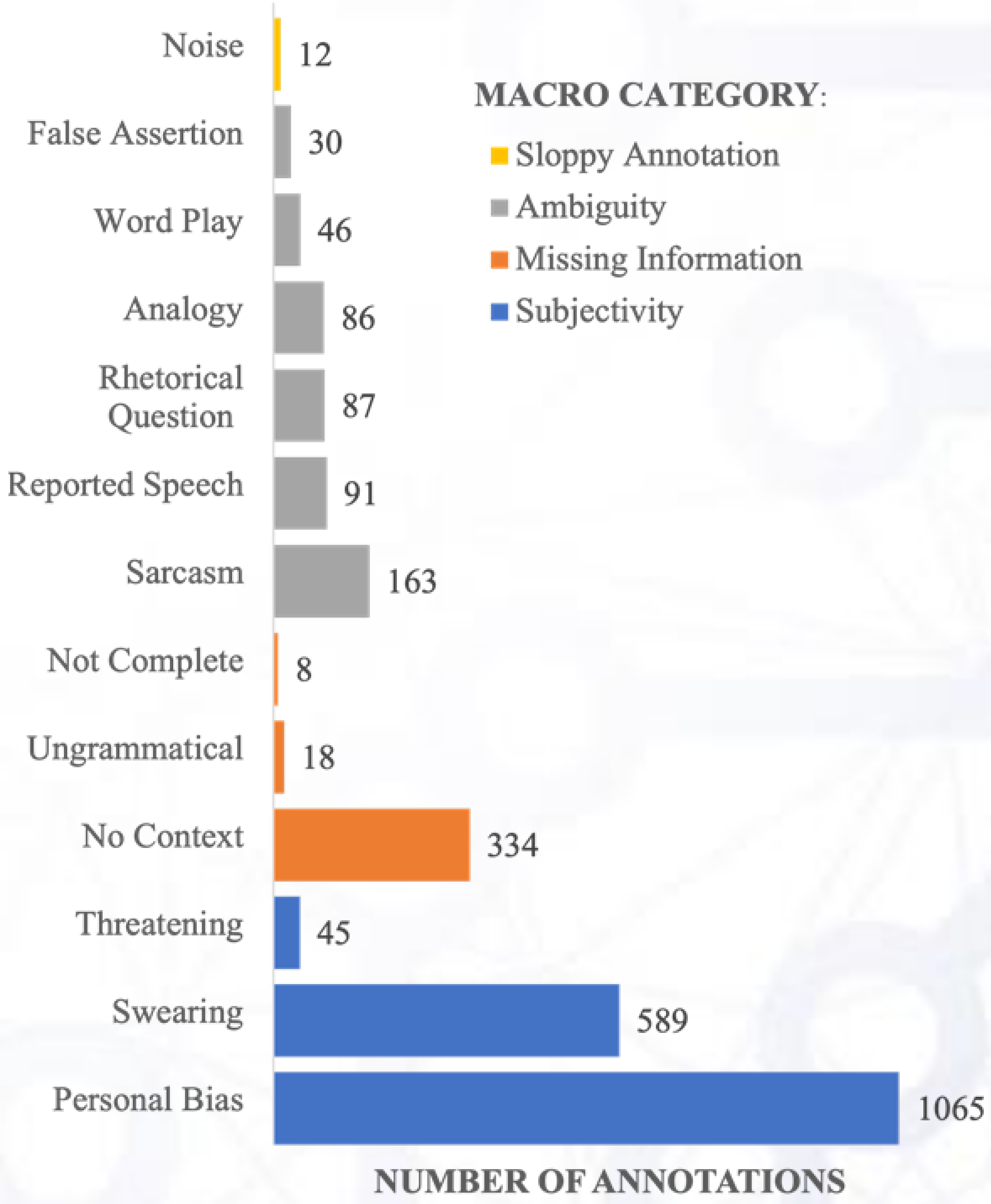
A Taxonomy of Disagreement



Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek (2023) Why Don’t You Do It Right? Analysing Annotators’ Disagreement in Subjective Tasks. Proceedings of EACL

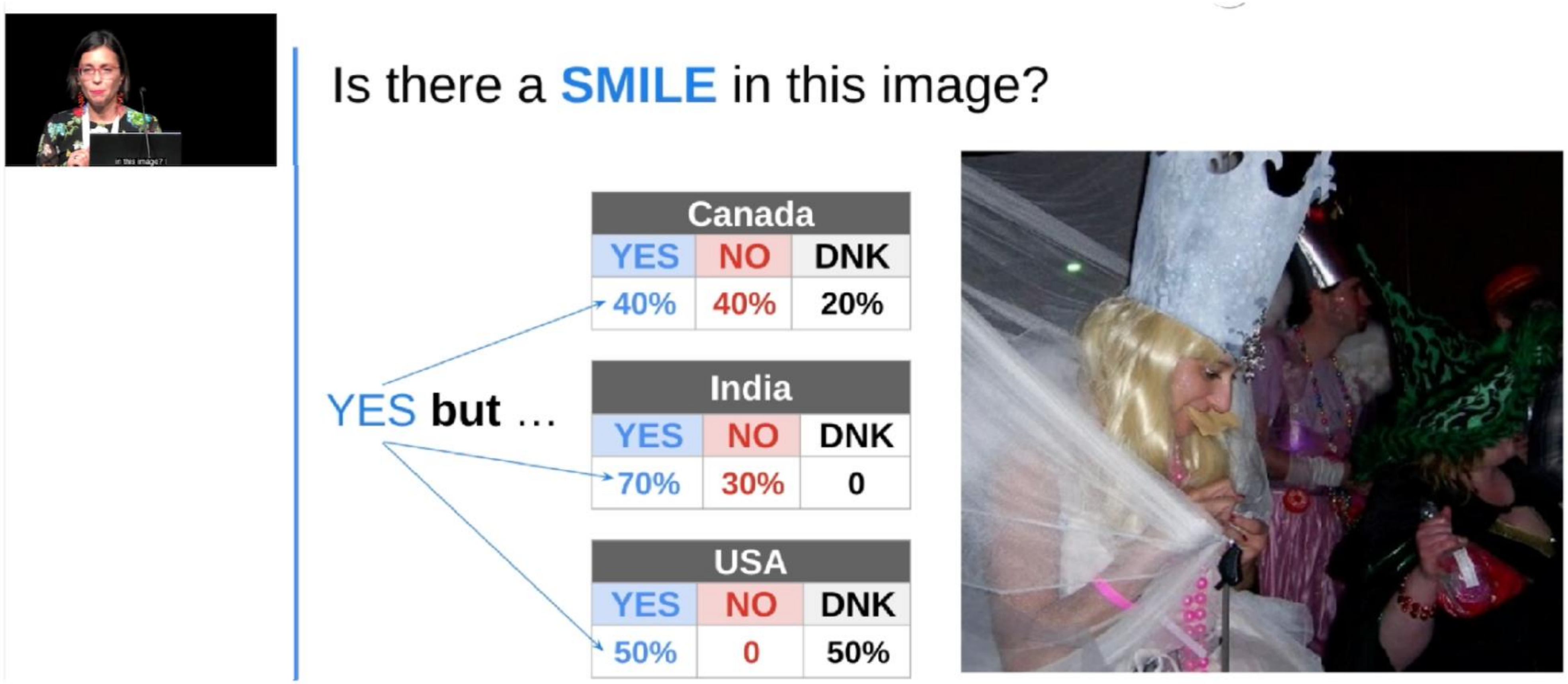
Disagreement in Hate Speech

2,574 tweets with disagreement manually assigned to one of these categories by a trained linguist



Disagreement not only in Text

Laura Aroyo's NeurIPS Keynote in 2023



Is there a **SMILE** in this image?

YES **but** …

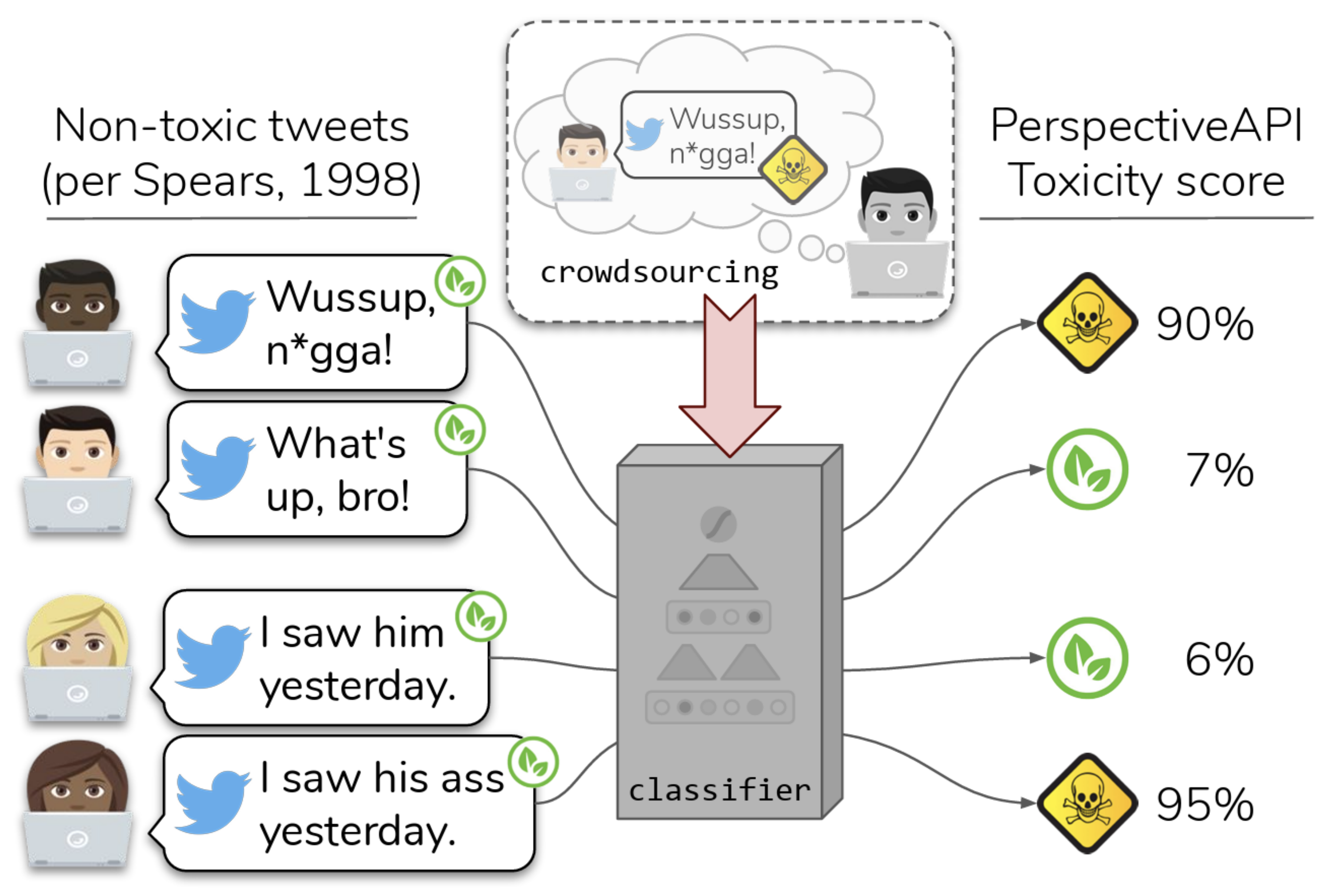
Canada		
YES	NO	DNK
40%	40%	20%

India		
YES	NO	DNK
70%	30%	0

USA		
YES	NO	DNK
50%	0	50%

https://slideslive.com/39015341/the-many-faces-of-responsible-ai?ref=speaker-55217

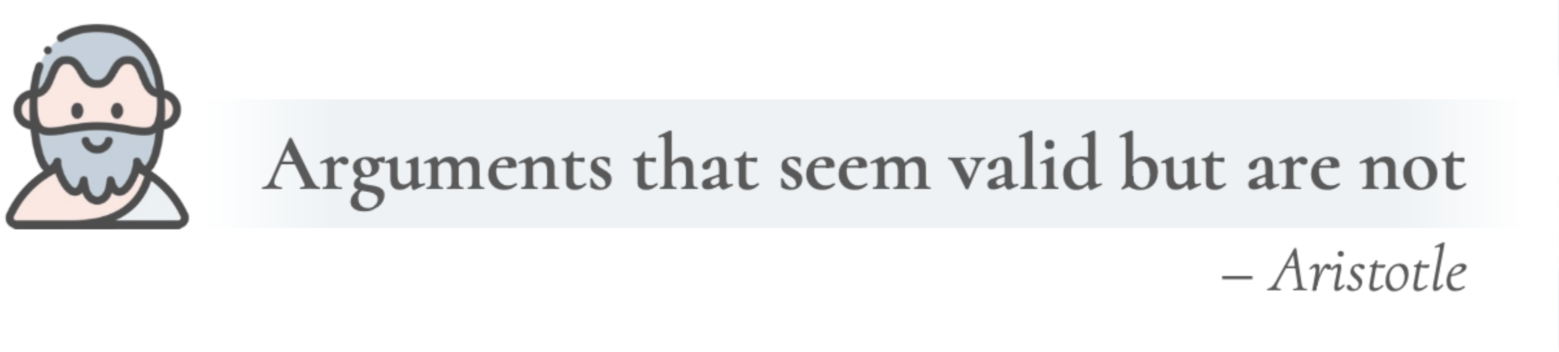
Agreement = Discrimination?



Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith (2019) The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

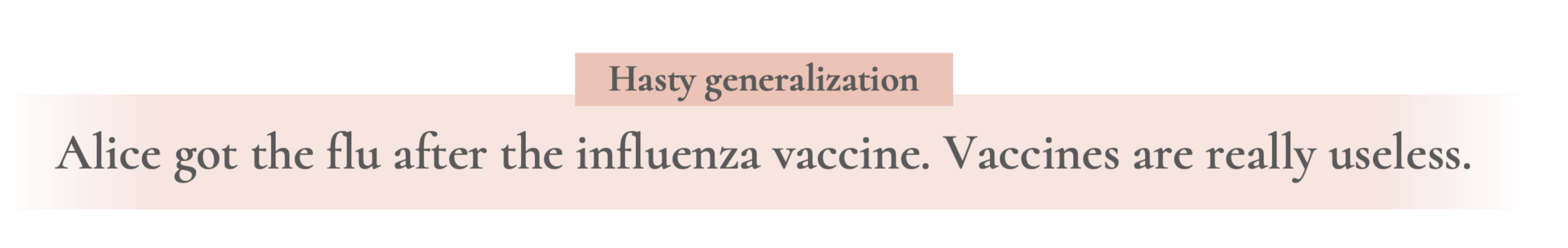
How should we deal with disagreement?

Task: Fallacy detection



Arguments that seem valid but are not

– *Aristotle*



Hasty generalization

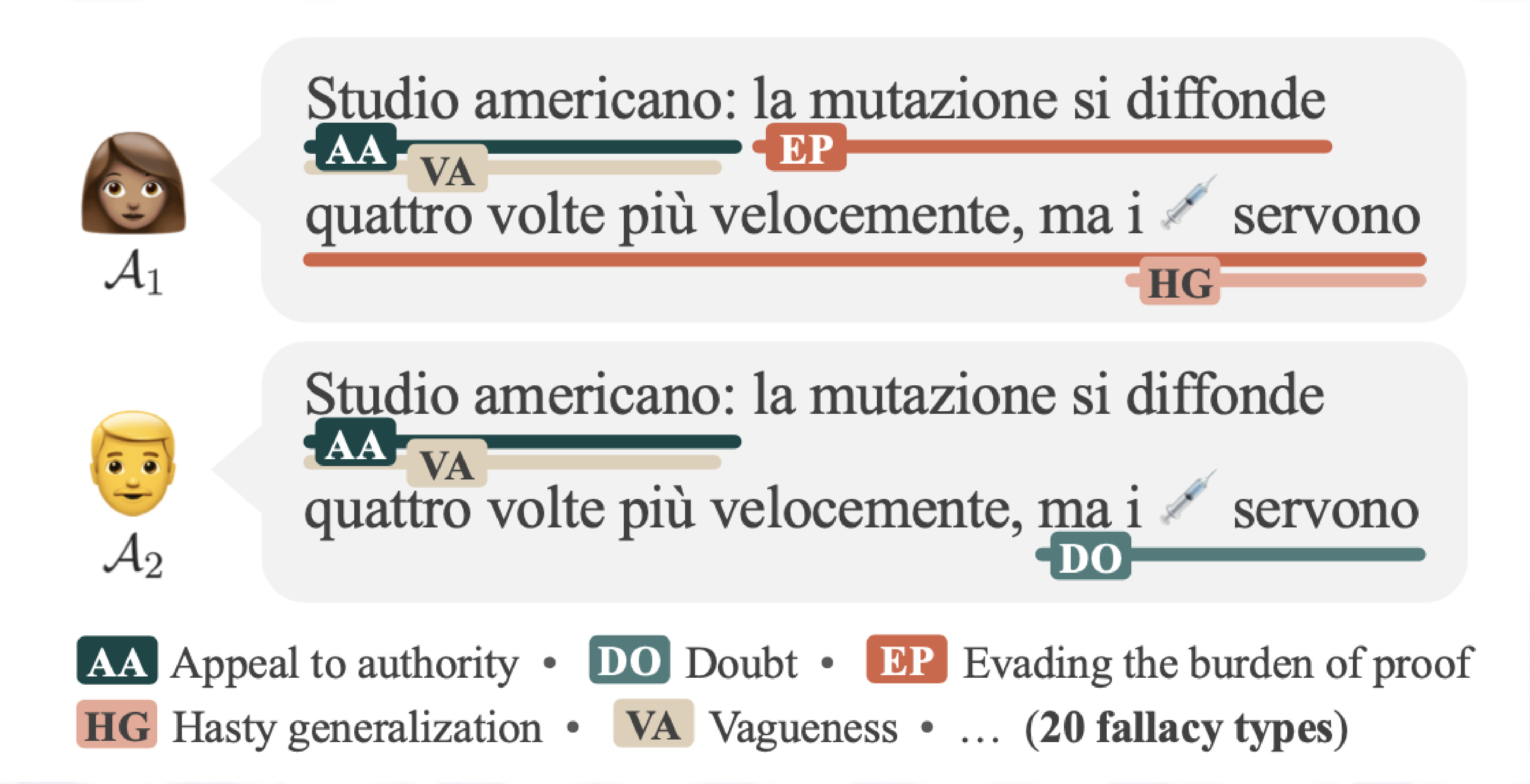
Alice got the flu after the influenza vaccine. Vaccines are really useless.

FAINA dataset: dataset for fine-grained fallacy detection with human label variation, embraces multiple plausible answers and natural disagreement

Alan Ramponi, Agnese Daffara, and Sara Tonelli (2025) Fine-grained Fallacy Detection with Human Label Variation. In *Proceedings of the NAACL*.

How should we deal with disagreement?

- Large inventory of 20 fallacy types
- Fine-grained annotation at the span level with potential overlaps



$\mathcal{A}_1$

Studio americano: la mutazione si diffonde quattro volte più velocemente, ma i 💉 servono

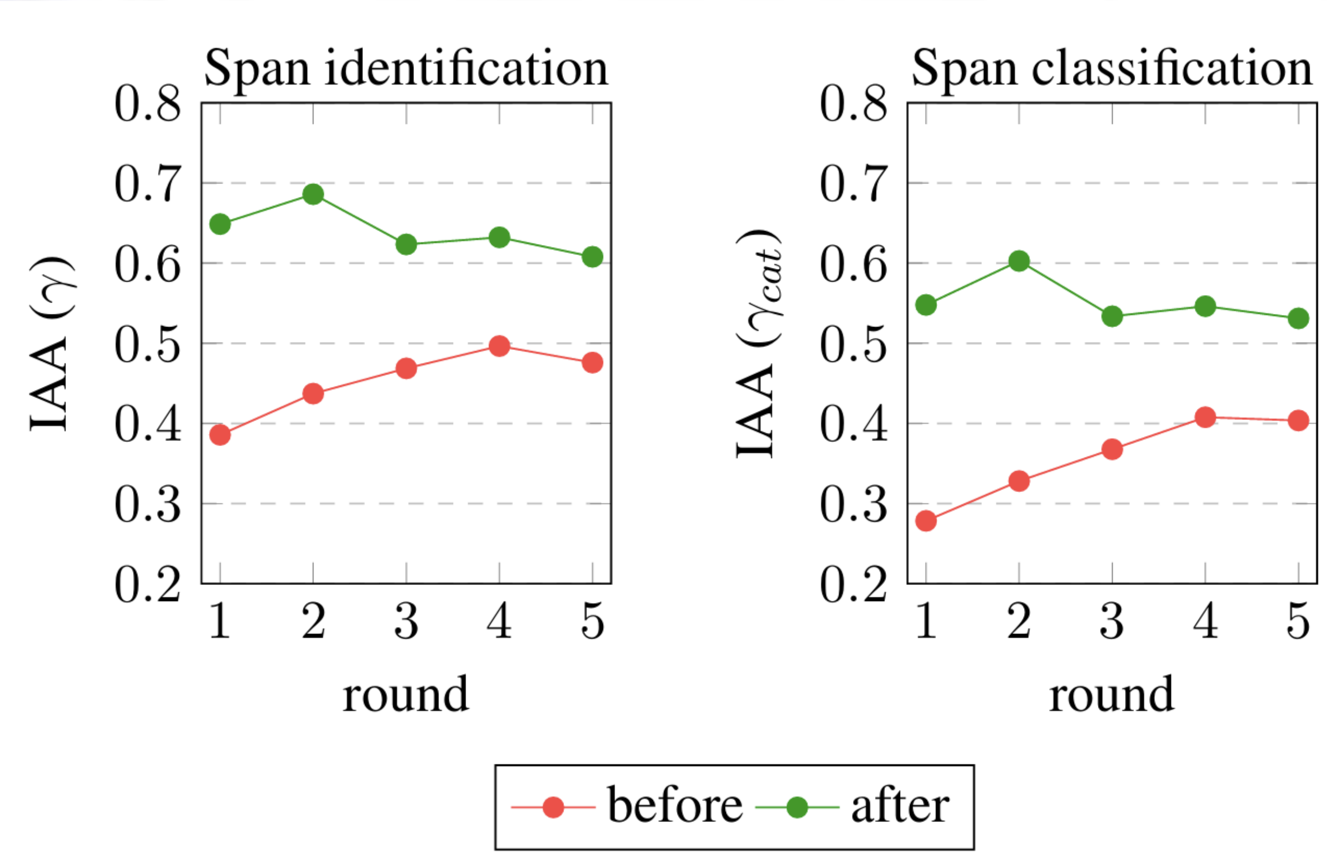
$\mathcal{A}_2$

Studio americano: la mutazione si diffonde quattro volte più velocemente, ma i 💉 servono

AA Appeal to authority • DO Doubt • EP Evading the burden of proof
HG Hasty generalization • VA Vagueness • … (**20 fallacy types**)

How should we deal with disagreement?

- Minimize annotation errors while keeping signals of human label variation
- Intrinsically difficult task: fallacy nuances, inventory, granularity, overlaps
- Reduce doubts and annotation errors through discussion rounds



Discussions are necessary but disagreement can be resolved only partially.

Classification and disagreement?

How can we use data with disagreement for training and evaluation?

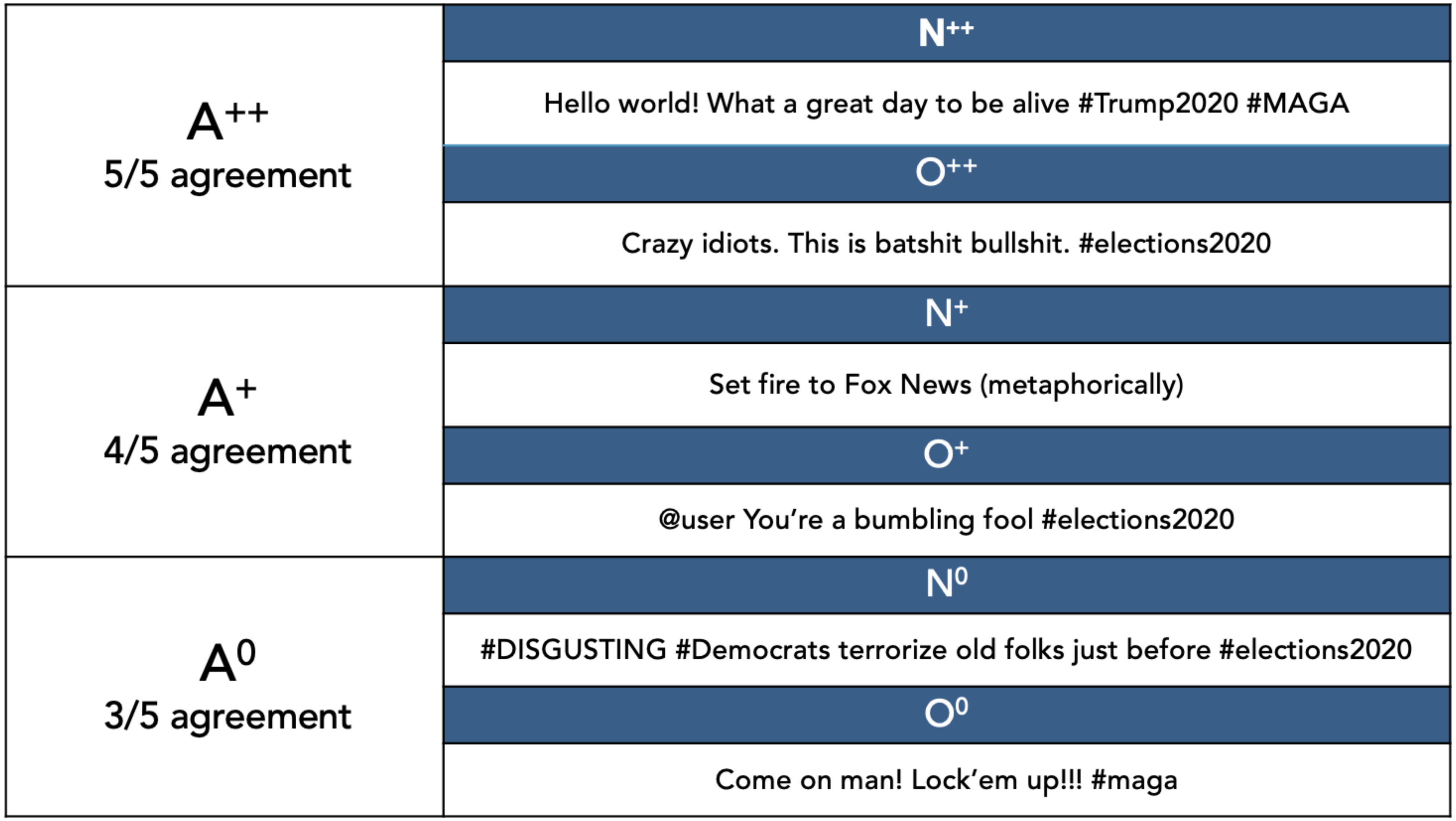
Train as many classifiers as annotators, perform multi-view classification, one test set for each annotator and then average

When several annotators are available, first cluster by preference and then train different classifiers and ensemble (Akhtar et al., HCOMP 2020)

FONDAZIONE BRUNO KESSLER

Should we eliminate disagreement?

Create a dataset for offensive language detection with 2,700 tweets on Covid, Trump and Black Lives Matter, each annotated by 5 annotators using crowdsourcing platform

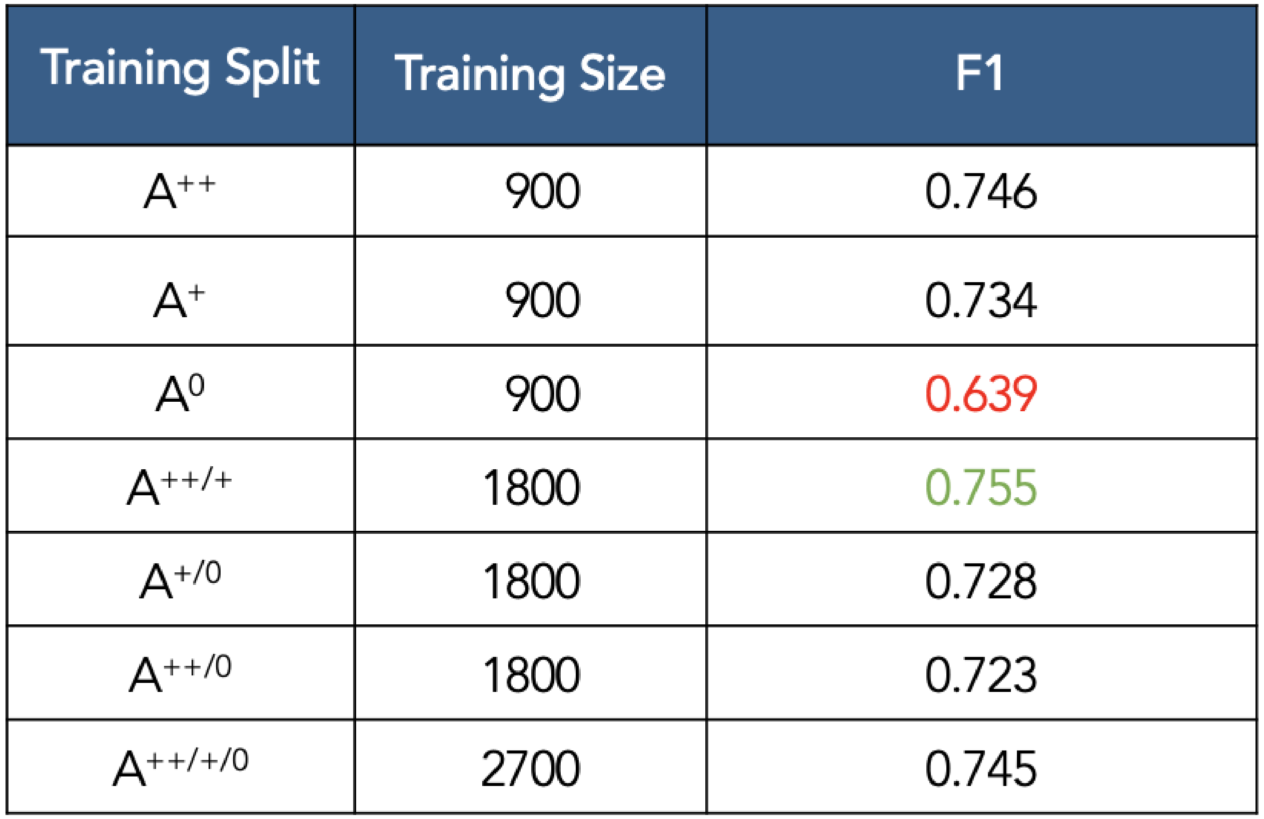


A^{++} 5/5 agreement	N^{++}
	Hello world! What a great day to be alive #Trump2020 #MAGA
	O^{++}
	Crazy idiots. This is batshit bullshit. #elections2020
A^{+} 4/5 agreement	N^{+}
	Set fire to Fox News (metaphorically)
	O^{+}
	@user You're a bumbling fool #elections2020
A^{0} 3/5 agreement	N^{0}
	#DISGUSTING #Democrats terrorize old folks just before #elections2020
	O^{0}
	Come on man! Lock'em up!!! #maga

Leonardelli et al. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. Proceedings of EMNLP 2021

Training on data with different agreement

Removing training data with high agreement yields to better results



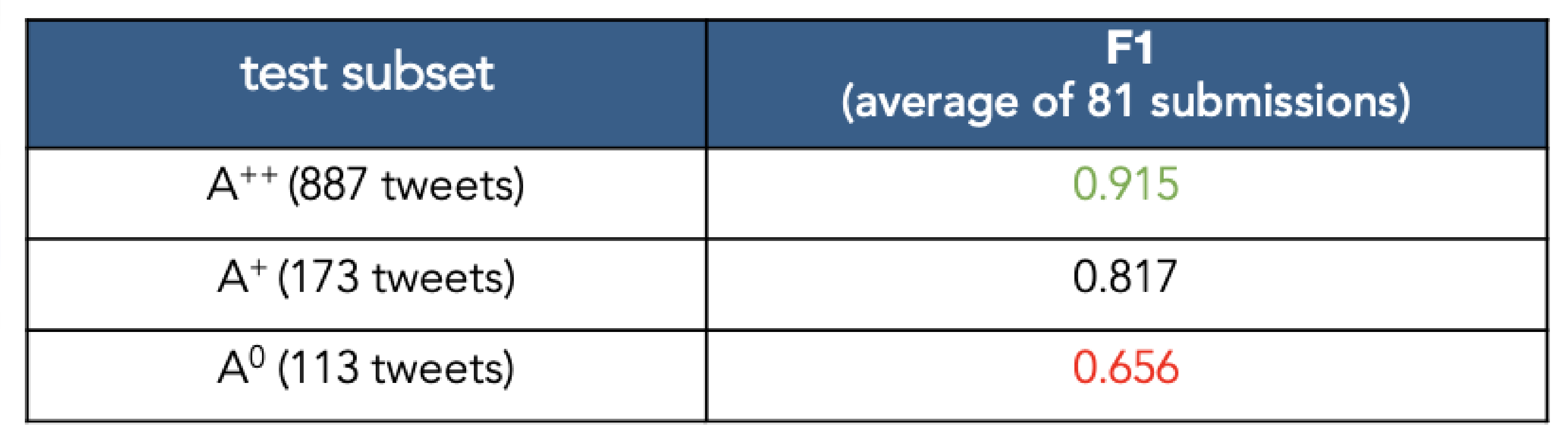
Training Split	Training Size	F1
A^{++}	900	0.746
A^{+}	900	0.734
A^{0}	900	0.639
$A^{++/+}$	1800	0.755
$A^{+/0}$	1800	0.728
$A^{++/0}$	1800	0.723
$A^{++/+/0}$	2700	0.745

FONDAZIONE
BRUNO KESSLER

Is there disagreement in Shared task data?

Offenseval 2021: Majority of submissions > .90 F1

We randomly sample and reannotate around 1,000 tweets from the test set: 90% are A^{+} or A^{++}



test subset	F1 (average of 81 submissions)
A^{++} (887 tweets)	0.915
A^{+} (173 tweets)	0.817
A^{0} (113 tweets)	0.656

Performance drastically decreases when agreement is low

Lessons Learnt

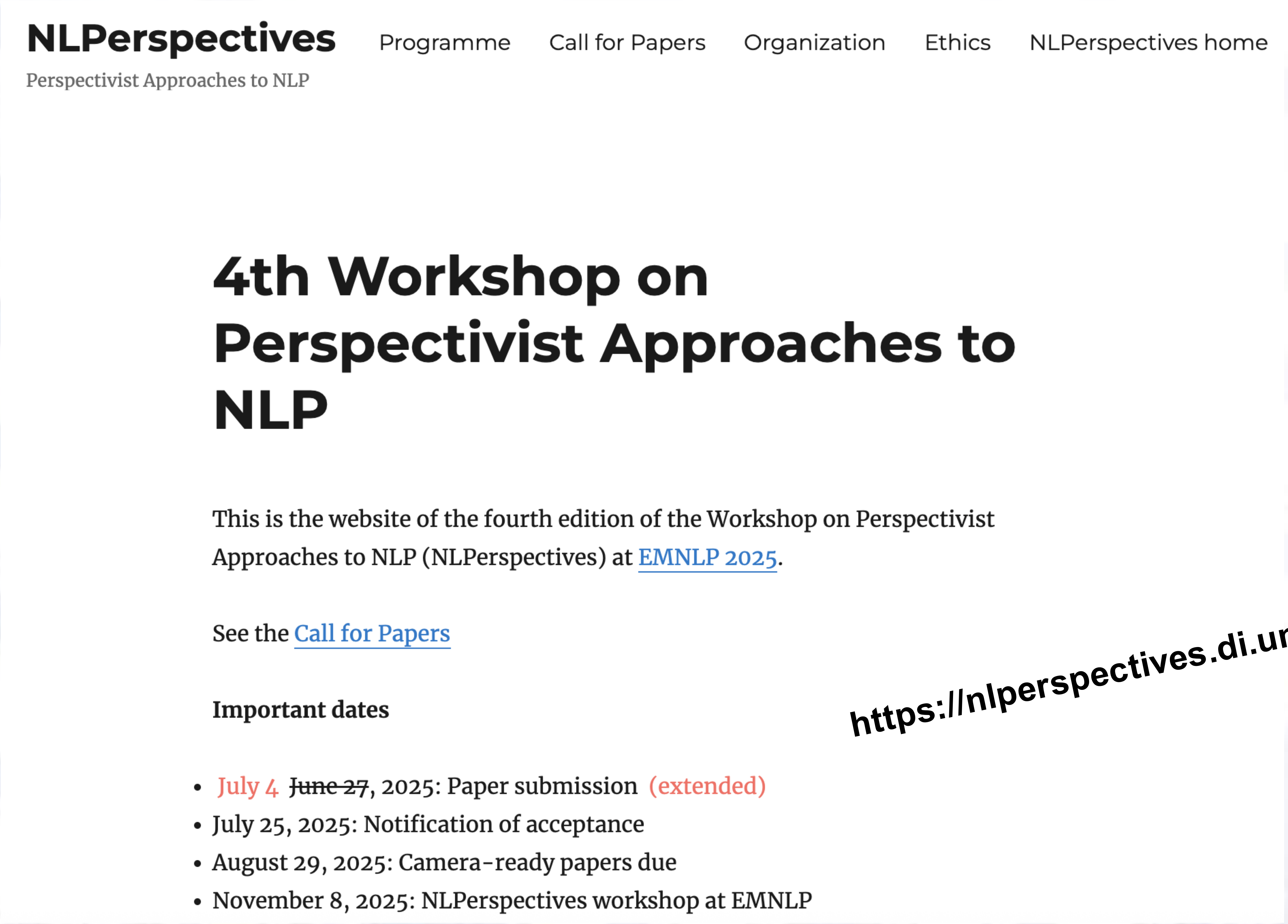
When creating linguistic datasets, control for the presence of (dis)agreement in training and test set

Release disaggregated data

Disagreement cases are not “wrong annotations” but often reflect natural disagreement, they should be accounted for also in Shared tasks

FONDAZIONE BRUNO KESSLER

Shared Tasks and Other initiatives



4th Workshop on Perspectivist Approaches to NLP

This is the website of the fourth edition of the Workshop on Perspectivist Approaches to NLP (NLPerspectives) at EMNLP 2025.

See the Call for Papers

Important dates

- July 4 ~~June 27~~, 2025: Paper submission (extended)
- July 25, 2025: Notification of acceptance
- August 29, 2025: Camera-ready papers due
- November 8, 2025: NLPerspectives workshop at EMNLP

https://nlperspectives.di.unito.it/

Shared Tasks and Other initiatives

Learning With Disagreement evaluation initiative

AIM: provide a unified testing framework for learning from disagreements, using datasets containing information about disagreements

E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, and M. Poesio (2023) SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023).

Le-Wi-Di task 2023

Le-Wi-Di
Learning With Disagreements
Task 11 @SemEval2023

Focus on subjective tasks

Organisers provide 4 datasets with disagreement:

- HS-Brexit dataset (hate speech)
- ArMIS dataset (misogyny and sexism)
- ConvAbuse dataset (abusive language between agents)
- MD-Agreement dataset (offensive language)

Participants are encouraged to develop methods able to capture agreements/disagreements, rather than focus on developing the best model

FONDAZIONE
BRUNO KESSLER

Le-Wi-Di Evaluation

Focusing on subjective disagreement, the existence of a ‘truth’ cannot be assumed.

→ two metrics: one “standard”, but also one that captures the multiplicity of annotation

1. **Soft evaluation**: evaluates how well the model's probabilities reflect the level of agreement among annotators. It is measured using cross-entropy, as a model that correctly predicts the distribution of labels produced by the crowd for each item will have low cross-entropy.

2. **Hard evaluation**: evaluates how well the results submitted align with the preferred (gold) interpretation. Each item is considered correct if it assigns the maximum probability to the preferred interpretation (if available). It is measured using micro-F1

Le-Wi-Di Evaluation

> 130teams subscribed to the Codalab competition page, **30 teams** participated in the evaluation phase of our task submitting their predictions for at least a dataset.

The majority of the teams (21) submitted predictions for all datasets, while one team submitted for three datasets and two teams for two datasets. A few teams (6) submitted their predictions for only one dataset

Le-Wi-Di task 2025 (in progress)

Le-Wi-Di
Learning With Disagreements
Task 11 @SemEval2023

Organisers provide 4 datasets with disagreement:

- CSC dataset (sarcasm)
- MP dataset (irony)
- Par dataset (paraphrasing)
- VEN dataset (natural language inference)

Tasks: provide probability distribution and predict annotation of single annotators

15 teams participated in the competition, results still to be published

FONDAZIONE
BRUNO KESSLER

Conclusions (1)

We should embrace disagreement to make linguistic annotation (and AI systems) more inclusive and «realistic», capturing the different perspectives existing in real life

Future directions include **Participatory NLP**, where stakeholders are empowered and involved in the creation of linguistic resources from the beginning

Be aware of whose perspectives shape decisions and how consensus is navigated

Su Lin Blodgett. Prospects for Participatory Approaches in NLP. Invited talk at NLPerspectives workshop (2022)

Core Principles of Participatory Design (2)

“enable those who will be impacted by a technology to have a voice in its design, without needing to speak the language of professional technology design”

involve the voices of marginalized users and communities in “decision processes that will affect them” in democratic ways

J. Gregory. Scandinavian Approaches to Participatory Design. International Journal of Engineering Education. 2003.
T. Robertson and J. Simonsen. Challenges and Opportunities in Contemporary Participatory Design. Design Issues. 2012.

Conclusions (3)

Use participatory NLP to unpack assumptions, goals and methods within linguistic annotation

However, it is impossible to boil linguistic perspectives down to identity characteristics, complexity of bounding speaker communities and language varieties

...but things in the NLP community are already changing!

Thank you !

satonelli@fbk.eu
https://dh.fbk.eu

Stefano Menini

Elisa Leonardelli

Alessio Palmero Aprosio

Alan Ramponi

Agnese Daffara