Enhancing Hate Speech Annotations with Background Semantics

***Content warning**. This research aims to tackle hate speech and contains examples of triggering and harmful language.*

Paula Reyero Lobo

Knowledge Media Institute
The Open University, UK

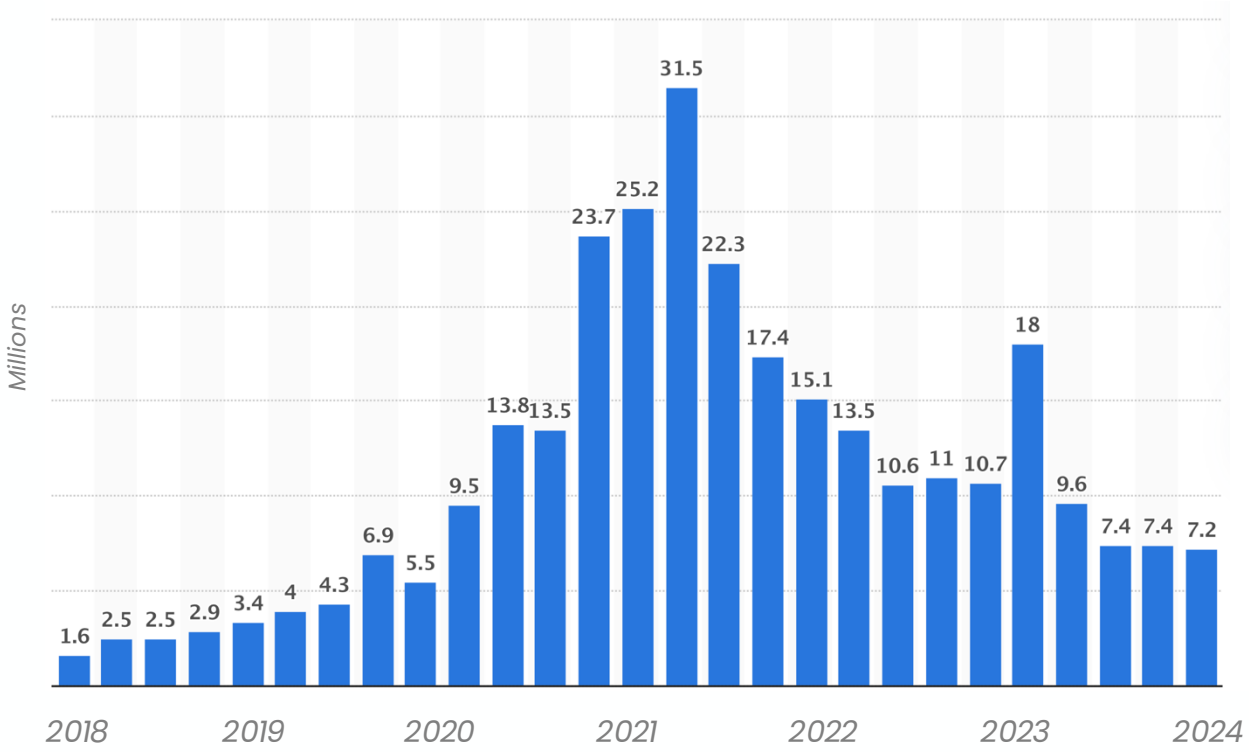
23 October 2024

Advised by:
Enrico Daga, Harith Alani, Miriam Fernandez

Funded by the European Union

Introduction

Content in millions removed on Facebook



Statista research Department, 12 September 2024

The Open University

The high incidence of hate speech on online platforms requires automated detection methods

Introduction

Does this comment contain **hate speech**,

defined as "*bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics*"?

The kebabs are a bunch of homosexual rapist deviants.

Does the message mention or is about...

gender

sexuality

race...?

Suck farts out of obese sheboons

Human annotations are crucial for training hate speech detection models

Introduction

Does this comment contain **hate speech**,

defined as "*bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics*"?

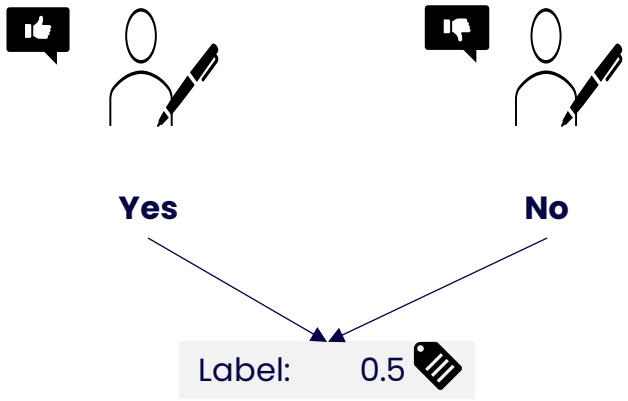
Does the message mention or is about...

gender

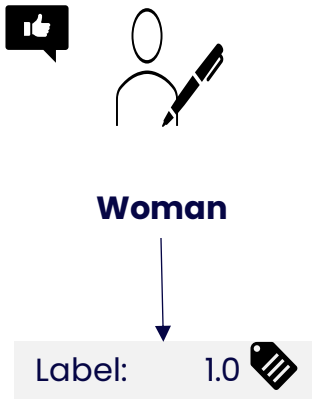
sexuality

race...?

The kebabs are a bunch of homosexual rapist deviants.



Suck farts out of obese sheboons



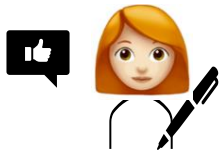
The disagreement between annotators challenges the use of majority votes and the accuracy of AI models

Related work

Does this comment contain **hate speech**,

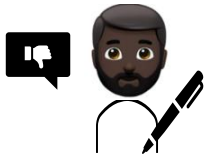
defined as "*bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics*"?

The kebabs are a bunch of homosexual rapist deviants.



No

Label: 0.0



Yes

Label: 1.0

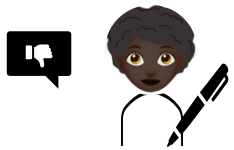
Does the message mention or is about...

gender

sexuality

race...?

Suck farts out of obese sheboons



Woman

Label: 1.0

Problem statement

Slang, derogatory, offensive term targeting **Black women**

The **kebabs** are a bunch of **homosexual rapist** deviants.

Suck farts out of obese **sheboons**

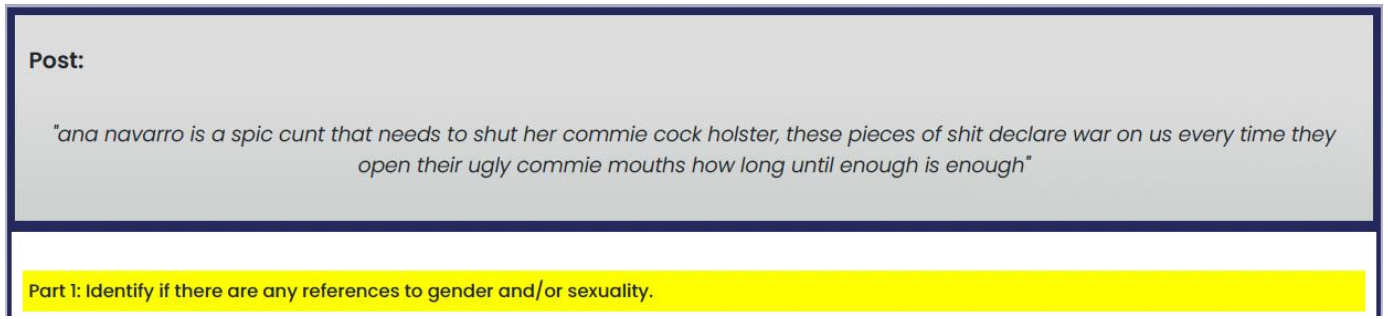
Ethnic slur to refer to a **Muslim**

The use of slang, non-standard, and domain specific language makes groups targeted by hate speech especially difficult to understand and perceive.

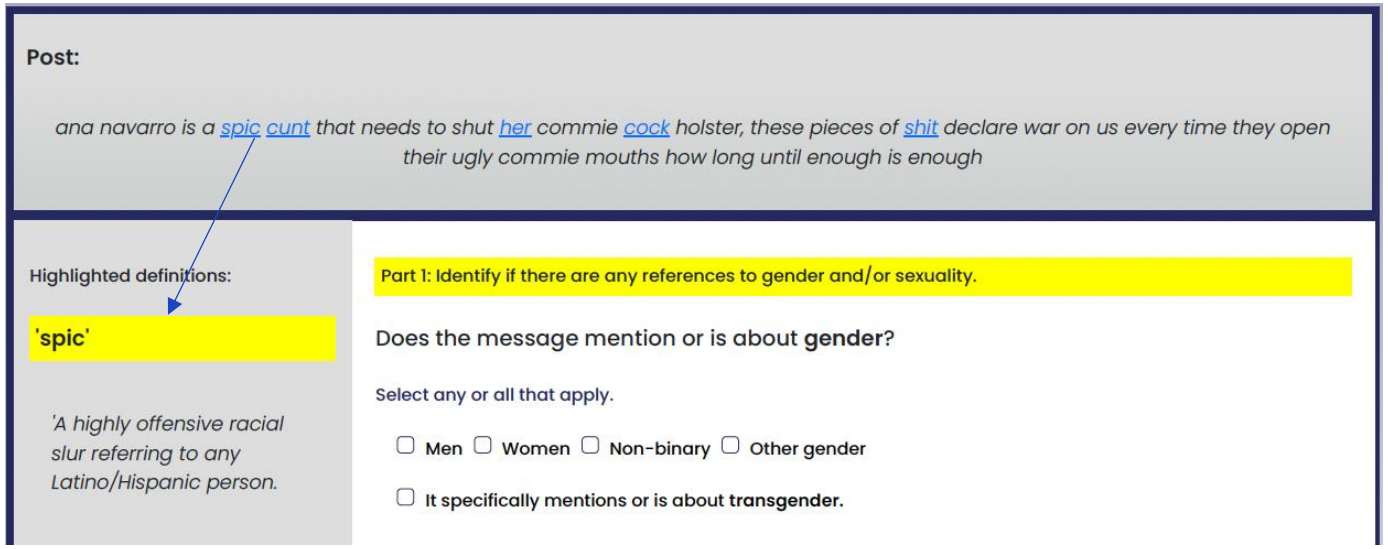
Our approach

To compensate for the lack of familiarity with domain language, we introduce semantics to enhance comprehension in hate speech annotation tasks.

A post:



A post with **background semantics**: highlighted entities with definitions

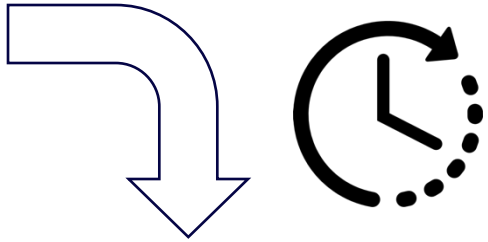


Data Repository

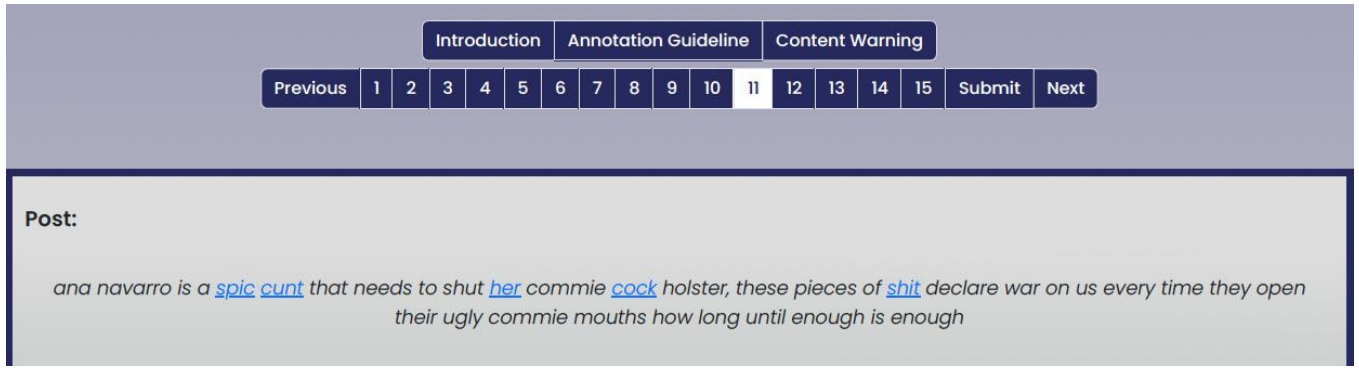
Our approach

Phase 1

One week later



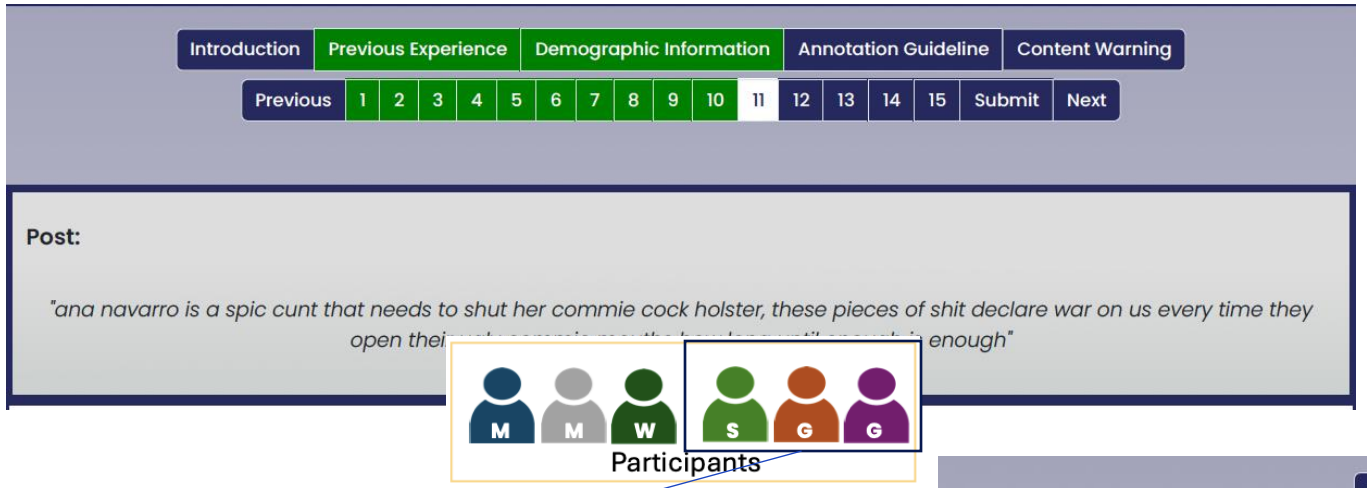
Phase 2



Data Repository

Our approach

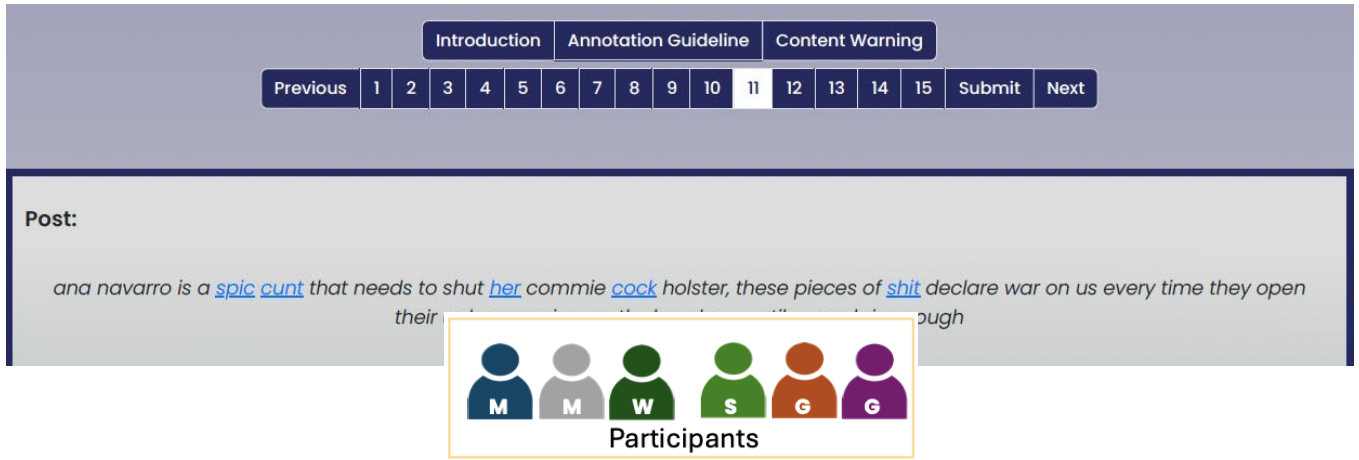
Phase 1



Each text has annotations from at least three non-cisgender and/or non-heterosexual annotators

One week later

Phase 2



Data Repository

We collect **2,880** annotations from Prolific participants with different genders and sexual orientations, obtaining an average of 6 annotations per text.

Evaluation

1. Agreement

How supplementing text with semantics affects hate speech annotation agreement?

2. Correlation

How do semantics impact convergence between annotators from hate speech target and non-target groups?

3. Change

How does the identification of hate target groups change after introducing semantics?

The Open University

Results

1. How supplementing text with semantics affects hate speech annotation agreement?

Krippendorff's Alpha Scores

Gender Labels	Phase 1	Phase 2	Δ
other gender	0.242	0.087	-0.155
non-binary	0.151	0.095	-0.056
gender unclear	0.069	0.035	-0.035
transgender	0.386	0.381	-0.006
gender not-referring	0.187	0.269	**0.081**
men	0.267	0.396	**0.129**
women	0.370	0.529	**0.159**
Sexuality Labels	Phase 1	Phase 2	Δ
asexual	0.229	0.206	-0.023
sexuality unclear	0.086	0.065	-0.021
heterosexual	0.147	0.151	0.004
sexuality not-referring	0.305	0.329	**0.024**
other sexuality	0.202	0.254	**0.053**
homosexual	0.597	0.654	**0.056**
bisexual	0.213	0.295	**0.082**
General Questions	Phase 1	Phase 2	Δ
Hate speech?	0.318	0.321	0.003
Hate speech targeting?	0.255	0.260	0.005
About sexuality?	0.370	0.409	**0.039**
About gender?	0.211	0.396	**0.186**
average	0.256	0.285	0.113

Adding background semantics increased agreement by **11.3% on average** for gender and sexuality groups.

Results

2. How do semantics impact convergence between annotators from hate speech target and non-target groups?

Pearson's Correlation Scores

	Phase 1				Phase 2			
	M	W	S	G	M	W	S	G
other	**0.46**	0.38	0.4	nan	0.0	**0.12**	0.06	nan
non-binary	0.28	0.11	**0.44**	nan	0.08	0.1	**0.3**	nan
unclear	0.26	0.07	**0.31**	nan	0.03	-0.04	**0.05**	nan
transgender	0.4	0.3	**0.63**	nan	0.5	0.39	**0.55**	nan
not-referring	0.24	**0.28**	0.27	nan	**0.35**	0.26	0.3	nan
men	0.24	**0.39**	0.3	nan	0.39	0.44	**0.56**	nan
women	0.44	**0.46**	0.41	nan	0.6	0.55	**0.62**	nan

Gender Labels: non-cisgender annotators (G) as the group targeted

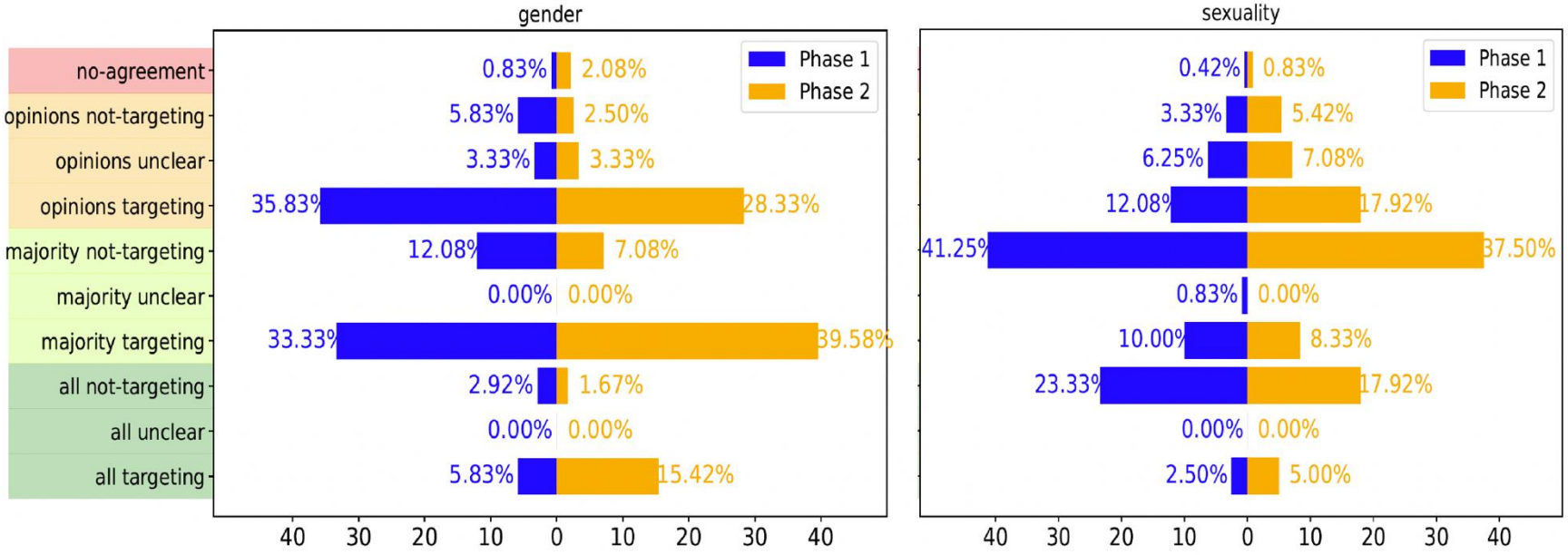
	Phase 1				Phase 2			
	M	W	S	G	M	W	S	G
asexual	0.31	0.1	nan	**0.68**	**0.6**	0.41	nan	0.37
unclear	**0.23**	0.15	nan	0.17	**0.14**	-0.01	nan	0.07
heterosexual	**0.24**	0.14	nan	0.23	**0.45**	0.11	nan	0.39
not-referring	0.33	**0.45**	nan	0.41	0.52	0.35	nan	**0.54**
other	**0.49**	0.19	nan	0.43	0.29	0.26	nan	**0.54**
homosexual	0.72	**0.83**	nan	0.66	0.79	**0.89**	nan	0.67
bisexual	**0.49**	0.45	nan	0.4	0.46	**0.64**	nan	0.33

Sexuality Labels: non-heterosexual annotators (S) as the group targeted

When semantics increased inter-annotator agreement, it was because annotations from non-target groups **aligned more closely** with those from the target group.

Results

3. How does the identification of hate target groups change after introducing semantics?



Each text may be targeting, not targeting, or unclear as decided by all, the majority, or at least two annotators (opinions).

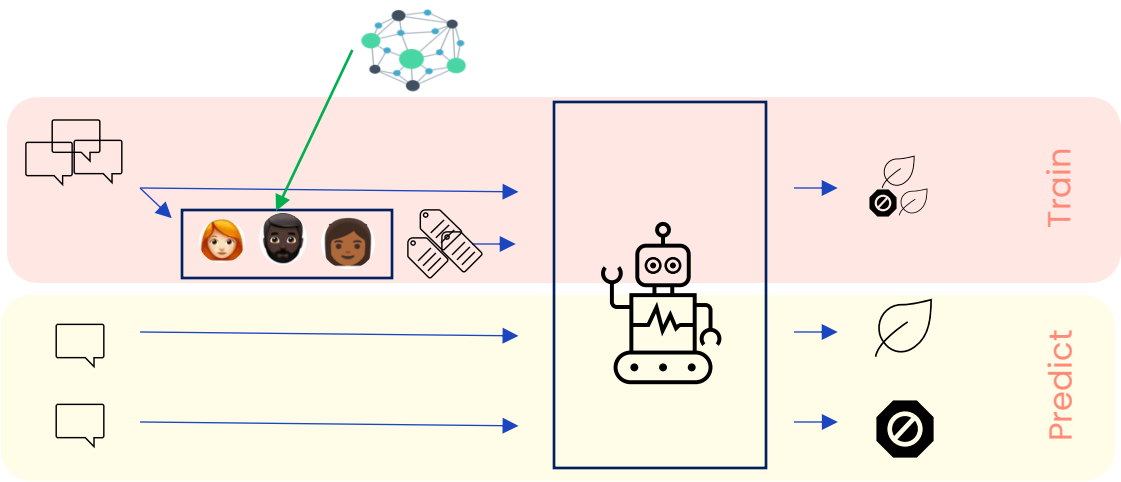
Semantics helps to **resolve hard-to-annotate text** (i.e., posts with a lower agreement or initially not deemed as pointing to hate-speech target groups).

Conclusion

Background semantics is key for enhancing

- Inter-annotator agreement
- Alignment with annotators from the targeted group
- Comprehension and knowledge acquisition in all groups of annotators

Incorporating Semantics in Hate Speech Annotation

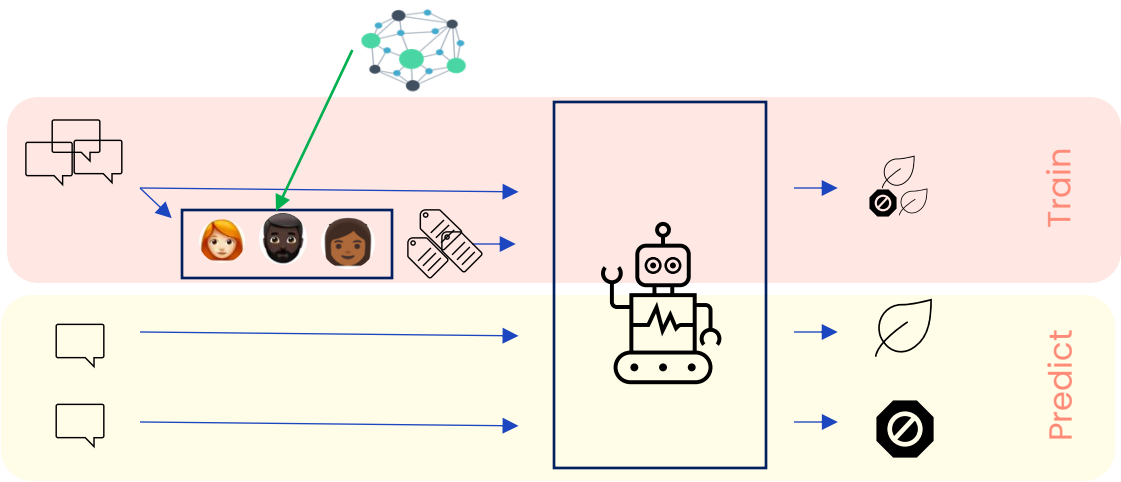


Conclusion

Background semantics is key for enhancing

- Inter-annotator agreement
- Alignment with annotators from the targeted group
- Comprehension and knowledge acquisition in all groups of annotators

Incorporating Semantics in Hate Speech Annotation



Limitations

- Semantics being less effective when hate speech is implicit
- Testing impact on ML performance

The Open University

Semantics helps annotators to process hateful terminology often used to target individuals or groups in social media

Enhancing Hate Speech Annotations with Background Semantics

Reyero Lobo, P., Daga, E., Alani, H., & Fernandez, M.

Data Repository

Code Repository

Contact

paula.reyero-lobo@open.ac.uk

The Open University

Funded by the European Union