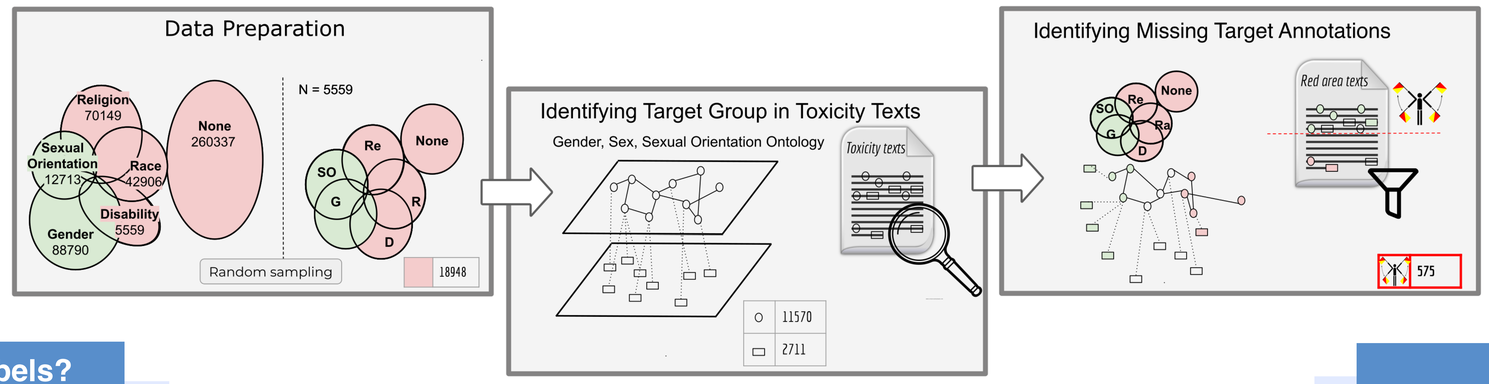
Supporting Online Toxicity Detection with Knowledge Graphs

Paula Reyero Lobo; Enrico Daga, PhD; Harith Alani, Prof.
Knowledge Media Institute, The Open University, UK

*Annotators often struggle to agree on which group is a **toxicity target***

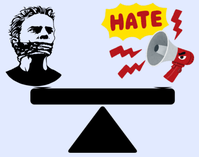
Semantic Knowledge for validating toxicity target labels



3% of~19k toxicity texts were potentially mislabeled!

Why using toxicity target labels?

Since toxicity detection systems have shown **bias towards the content** of the groups they aim to protect, identifying their information is essential.



Our approach

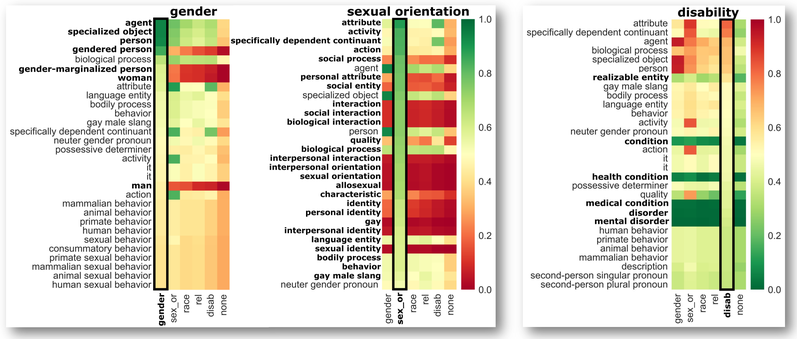
Use **semantic knowledge** about two specific toxicity targets to assess missing target information in toxic speech annotations.

Data

Gender, Sex, Sexual Orientation *(GSSO) with over 14k entities (slang, culturally specific gender identities) [1].*

*Balanced samples of texts referring to demographic groups from **Jigsaw's Toxicity 445k** corpus [2].*

Knowledge graph entities representative of different toxicity targets



We were able to automatically identify mentions of target group in toxicity texts

The higher the frequency, the more representative of the interest (**green**) or the other groups (red).

Gender and sexual orientation mentions in toxicity texts

"The only snowflakes I see are the insecure white supremacists and the Nazi **pansies**."

"Good Lord the **gay** brings out the worst in some 'TradCaths' understanding of Christianity"

"The progressive values are the ones the elitist left are slowly making norm, **LGBTQ**, **Gender neutrality**, break down of families and Government control of kids, open borders and the influx of Islam into every nation."

Sexual Orientation

Gender

"Most liberals are white **hon**."

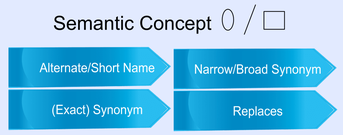
"**Abortion** and Planned Parenthood is quite the misnomer as **abortion** should never be used as **birth** control or for **gender selection**, however, I do agree that **abortion** for **pregnancy** due to **rape** is necessary though."

"The problem with your thesis is that you are treating this "disorder" like any other mental disorder where the cure is to rid yourself of the symptoms, feelings, urges, etc. Whereas the only reason the DSM and WHO list **gender dysphoria** as an illness is that it causes distress and dysfunction; however, the recommended cure by both organizations is to acknowledge the disconnect and support the person in **transitioning**."

We found target group mentions missed by the annotators

Methods

1. Search for **semantic concepts frequent** only in each group sample.
   ↳ Mapping their semantic properties in toxicity texts.



2. Use knowledge graph entities to identify mentions of toxicity targets **in texts not labelled as gender or sexual orientation.**

Findings

We find a **close relationship** between semantic concepts and group samples in toxicity texts:

Percentage of representative entities			
Sexual Orientation	Gender	Disability	Other
86.67% (26/30)	23,33% (7/30)	20% (6/30)	0%

We identified mentions of target groups that were **too complex** to be labelled correctly.
↳ 575 texts **not identified**.

Conclusion

Semantic knowledge offers a solution to **validate and sanitise** large datasets that are difficult to annotate!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 860630.

Contact

Paula Reyero Lobo
Knowledge Media Institute, The Open University
Milton Keynes, UK
paula.reyero-lobo@open.ac.uk

References

1. Kronk, C., and Dexheimer, J. W. "Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and Workflow." *Journal of the American Medical Informatics Association*, 27(7): 1110–1115 (2020).
2. Borkan, D., et al. "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification." *In Companion Proceedings of the 2019 World Wide Web Conference,* 491-500 (2019).

Check our code for more details!