Knowledge-grounded Target Group Language Recognition in Hate Speech

Paula Reyero Lobo, Enrico Daga, Harith Alani, Miriam Fernandez
Knowledge Media Institute, The Open University, UK

PROBLEM

Post 1

#AltRight #NRx #tcot #Anarchists #Libertarian #GOP #Republicans #Conservative #GamerGate #ProudBoys #NewRight #Q #TheAwakening #MAGA #Pegida #Afd #Orban #Putin #Brexit #Qanon #QAnon #qanon #TheGreatAwakening #WalkAway #Qanuck #TheStorm

It's been 3 months and these words have only proven to be true. Suck it **Butch** who can't Deadlift.

Post 2

Is this that **bimbo** twat with fish lips and hideous full sleeve tatts?

butch

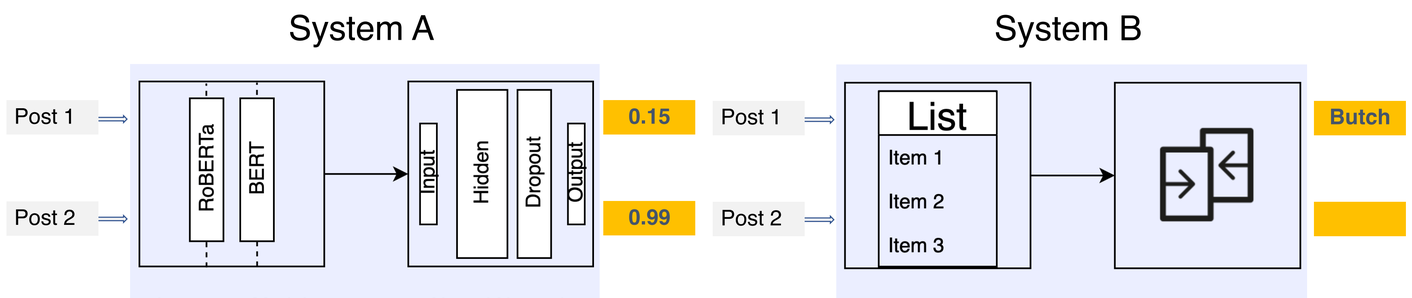
Traditionally, a lesbian who appears "masculine" or acts in a "masculine" manner. However, the term is often used to describe a wide variety of gender expressions and identities adjacent to masculinity or "butchilinity".

bimbo

A slang term for a conventionally attractive, sexualized, naïve, and unintelligent woman. It is often used to describe women who are blond, have curvaceous figures, heavy makeup, and revealing clothing. It is commonly associated with "the dumb blonde" stereotype.

Hate speech is expressed through language particularly sensitive to the target communities, which makes it harder to understand and perceive.

EXISTING APPROACHES



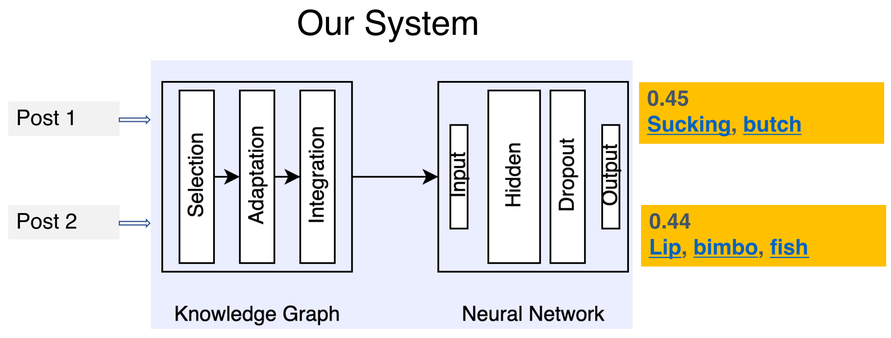
Deep Learning Based

performance *interpretable*

Linguistic Based

interpretable *performance*

HYBRID MACHINE LEARNING



Our framework integrates knowledge about the target groups to ground predictions to the entities supporting a decision, while preserving optimal predictive quality.

EVALUATION

Benchmark Results

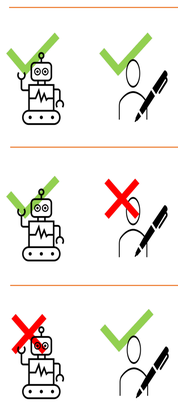
Model	Accuracy	$F1$	ROC AUC	PR AUC
Gab Hate Corpus				
Toxic Debias	91.81	58.82	74.82	40.20
HybridDocF$_h$	**91.30**	51.15	84.52	54.55
HybridLR$_h$	90.64	**62.42**	89.30	64.38
HybridMultiNB$_h$	89.36	61.11	**90.13**	**68.24**
RoBERTa$_{base}$	88.85	61.55	93.06	70.32
XtremeSpeech$_{English}$				
Toxic Debias	94.01	52.41	72.96	31.12
HybridDocF$_h$	**93.97**	**53.91**	87.02	47.05
HybridLR$_h$	90.79	49.27	**88.36**	50.79
HybridMultiNB$_h$	90.38	47.74	87.26	**51.80**
RoBERTa$_{base}$	92.99	57.67	93.67	57.38
HateXplain				
Toxic Debias	84.43	67.36	79.01	52.66
HybridDocF$_h$	79.45	43.55	78.37	55.96
HybridLR$_h$	83.48	67.72	88.15	68.35
HybridMultiNB$_h$	**85.63**	**73.57**	**91.38**	**78.37**
RoBERTa$_{base}$	89.91	80.22	95.60	86.46

Error Analysis

Category (FP)	Definition	A.E	N
Demographic descriptor	Direct explicit reference to a member of the group.	X	117
Targeted language	Insults, sexually explicit, or topics related to the group.	X	20
Implicit reference	Refers to a group member using pronouns.	X	10
False match	Incorrectly flagged due to polysemy.		3

Category (FN)	Definition	A.E	N
No reference	No language related to the group.	X	26
Missed at content	Not identified at validation, due to misspellings or being out-of-training.		19
Missed by method	Mention not correctly found or given importance by model.		85

Knowledge-grounded Predictions Analysis



woman, man, LGBTQ, transgender person, Black man, gay, homo, heterosexual, homosexuality, feminist, lesbian, queer sexual orientation, same-gender marriage, pederasty, sexuality, homophobia, queer identity, *sex, feminism, marital partner, sex worker, fag, faggot, masculism, pussy, thot, rape, menstruation, bitch*

woman, man, LGBT, woman of color, gay, heterosexual, homo-sexuality, feminist, lesbian, queer sexual orientation, transgender, same-gender marriage, womanism, sexuality, homophobia, queer identity, abusive person, sex, *partner, faggot, semen, pussy, bitch*

feminist, homophobia, sex, sex work client, *marital partner, parent, sex worker, fag, faggot,* ***rapist, female gender role****, pussy,* ***abortion****, vagina, thot,* ***penis****, rape, domestic violence,* ***she, bimbo, sexual abstinence****, cunt, bitch,* ***whore, slut, fuck, Mrs., rainbow flag***

I don't think I am being gullible, […]. Not to mention **homosexuals** and the mentally ill. If it is an international racist insult, it falls flat.

No

0.83
gay, homosexuality, insult

It's History you know. Mary Beard told me so and she has a **vagina** from Oxford.

Yes

0.49
vagina, beard, she

LESSONS LEARNED

- ❑ Our hybrid models are effective as the state-of-the-art models (System A), while adjusting better to domain and data changes.
- ❑ Entities in a Knowledge Graph (KG) are a richer representation than simple terms (System B), resulting in more transparency and competitive performance with language models.
- ❑ Grounding predictions in explicit knowledge is key to understand model outcomes, the data used to train hate speech detection systems, and the ambiguous cases in human annotations.

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