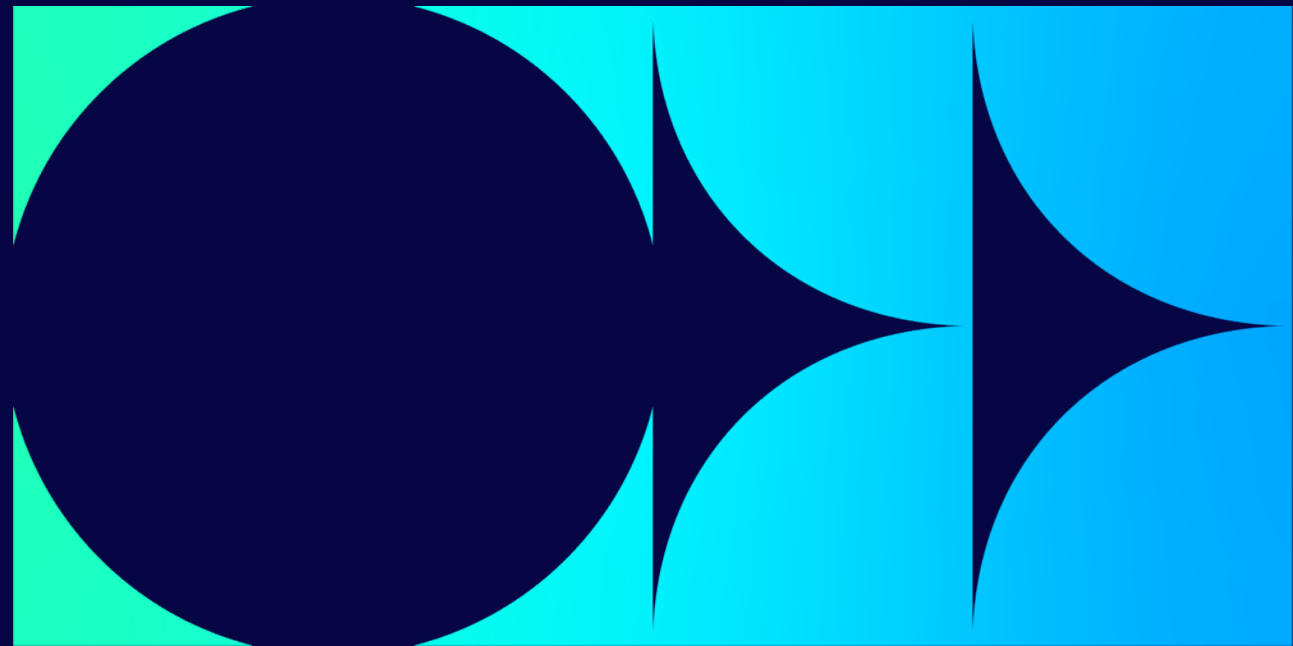
SEMANTiCS
Leipzig 2023

Knowledge-Grounded Target Group Language Recognition in Hate Speech

Paula Reyero Lobo
Knowledge Media Institute
20-22 September 2023

Advised by:
Enrico Daga, Harith Alani, Miriam Fernandez



The Open University

Funded by
the European Union

Introduction

Does this post mention or is about...

gender

sexuality

race...?

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

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

The Open University

Introduction

Hate speech contains **nuanced**, **specialized language** to attack frequently target groups.

Does this post mention or is about...

gender

sexuality

race...?

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

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

The Open University

Introduction

Hate speech contains **nuanced**, **specialized language** to attack frequently target groups.

Does this post mention or is about...

gender

sexuality

race...?

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

A lesbian who appears "masculine"

Slang term for conventionally, sexualized, and unintelligent woman

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

The Open University

This language makes target groups in hate speech especially difficult to understand and perceive.

Problem statement

Does this post mention or is about **gender or sexuality?**

Post 1

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

Label

Sexuality

None

Post 2

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

Label

Gender

The label bias in training datasets calls for the need of **interpretability** when building automated detection systems.

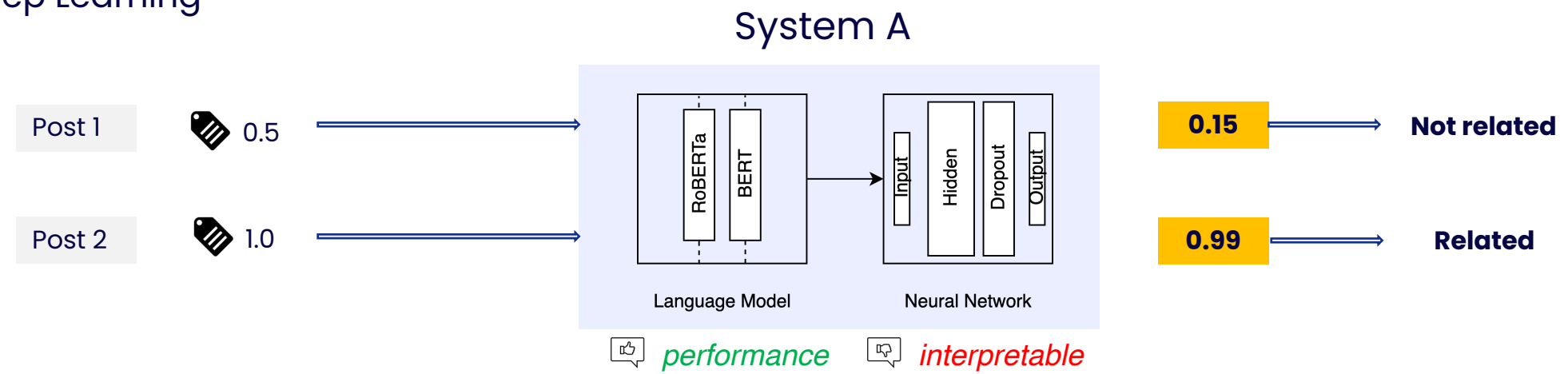
Kennedy, Brendan, et al. "Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale." Language Resources and Evaluation (2022): 1-30.

Related Work

Does this post mention or is about **gender or sexuality?**

Post	Label
Post 1: 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?	👍

- Deep Learning



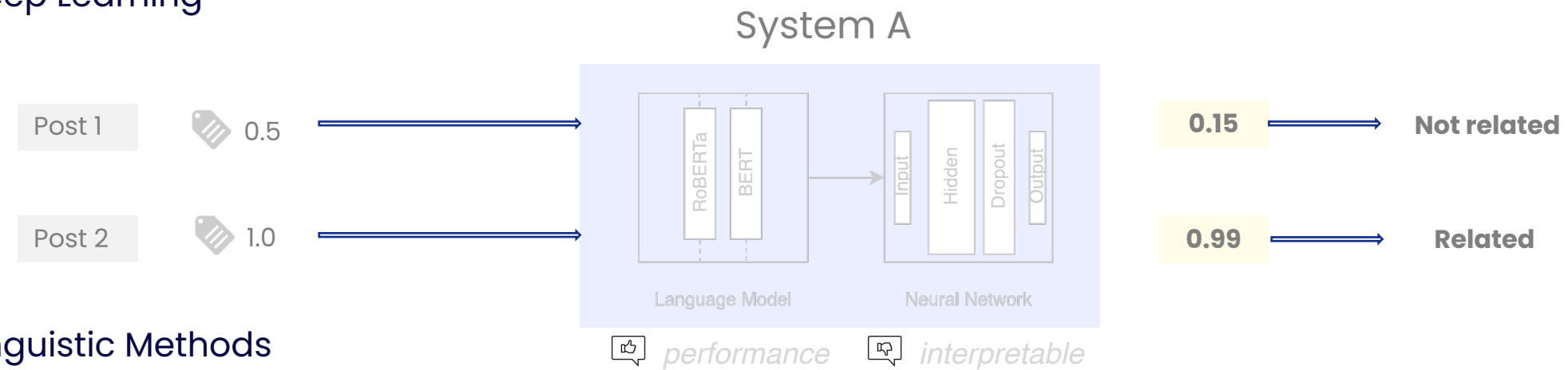
Sachdeva, Pratik, et al. "Targeted Identity Group Prediction in Hate Speech Corpora." Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). 2022.

Related Work

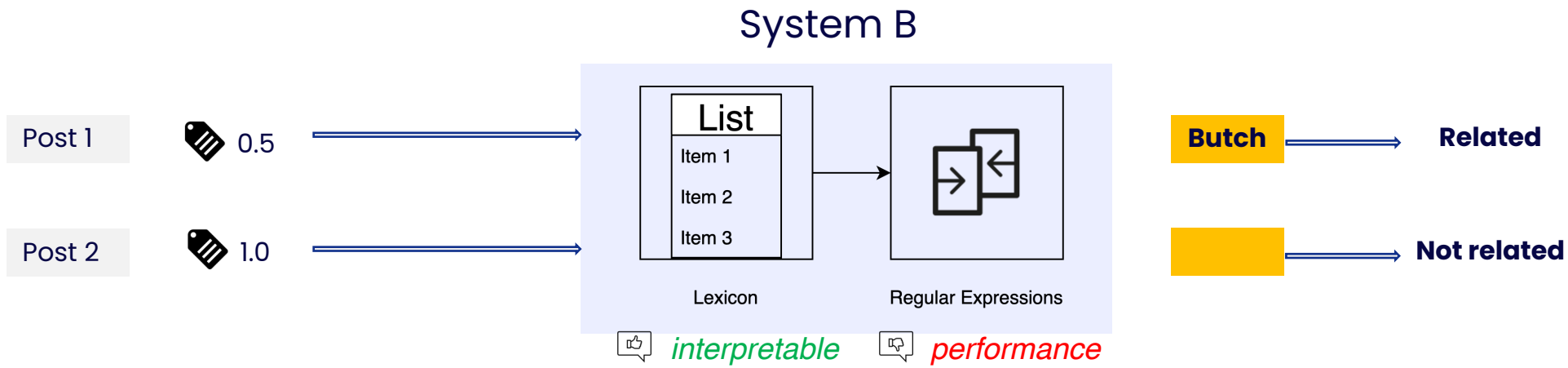
Does this post mention or is about **gender or sexuality?**

Post 1	Label
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?	👍

- Deep Learning



- Linguistic Methods



Smith, Eric Michael, et al. ""I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.

Our approach

Does this post mention or is about **gender or sexuality?**

- Hybrid Learning

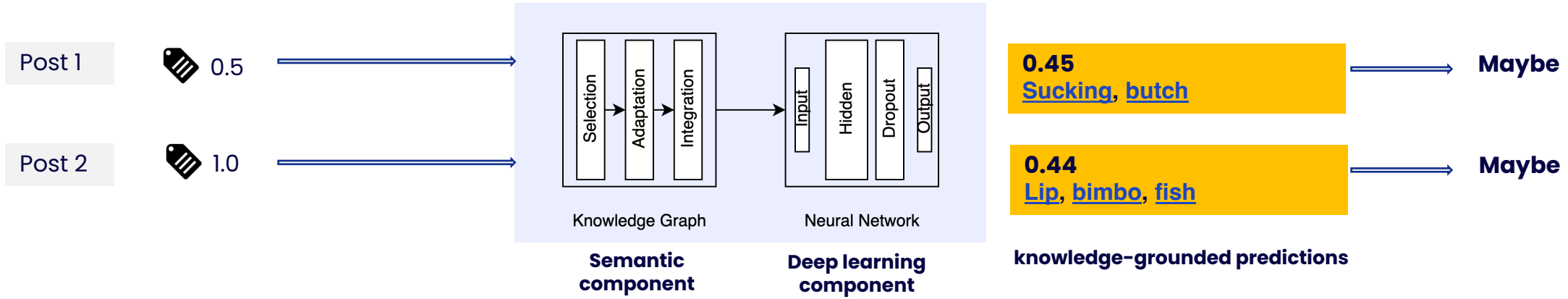
Post 1

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?

Our System



Our approach

Does this post mention or is about **gender or sexuality?**

- Hybrid Learning

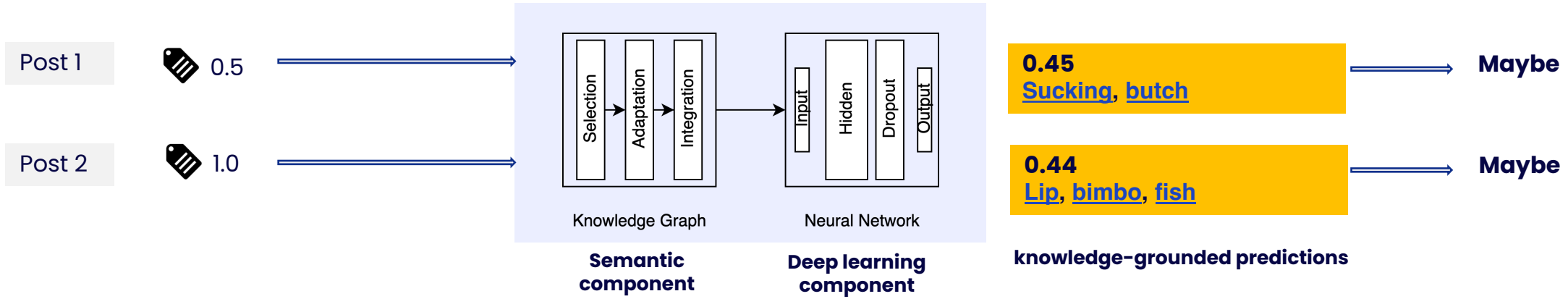
Post 1

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?

Our System



Instance: butch

Term IRI: http://purl.obolibrary.org/obo/GSSO_000338

Definition: 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".

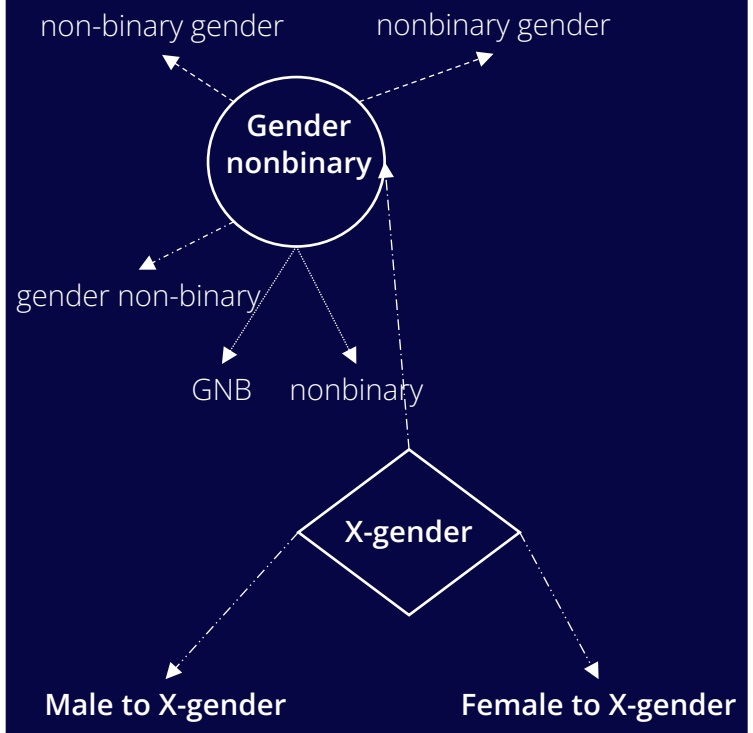
Annotations

- **comment:** In the first half of the 20th century, "butch" was used as a gay male slang term referring to a masculine gay man.

Kronk, Clair, et al. "Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow." Journal of the American Medical Informatics Association. 2020.

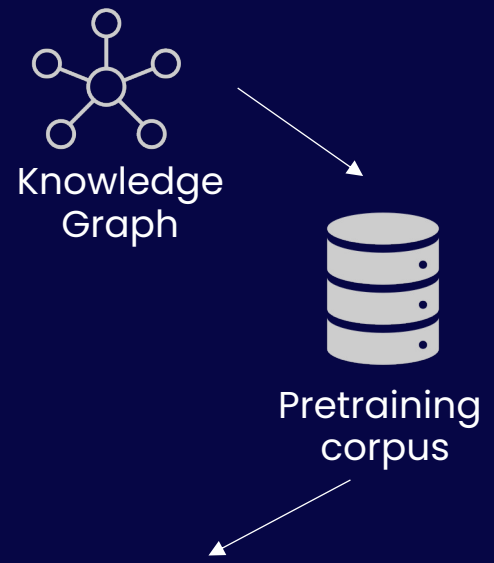
Hybrid Approach

1. Selection



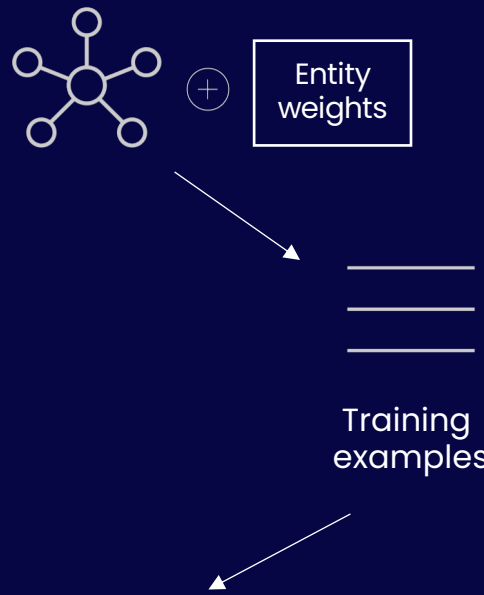
Knowledge Graph

2. Adaptation



Feature extractor model

3. Integration

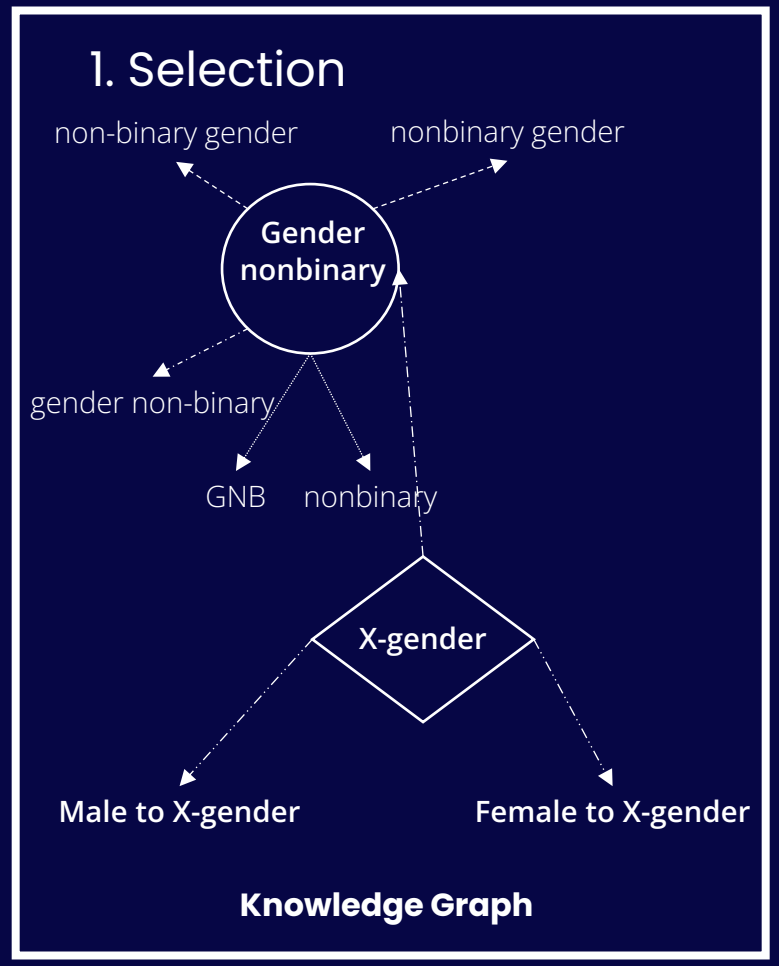


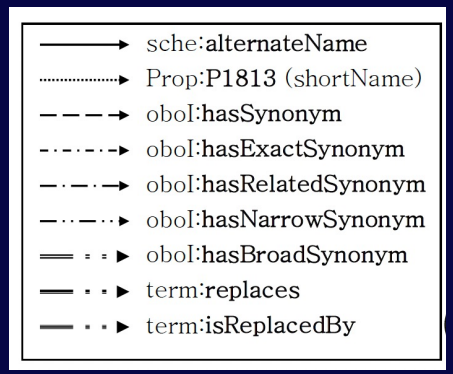
Deep Learning Model input

The Open University

Recognition of Target Group Language in Hate Speech Corpora

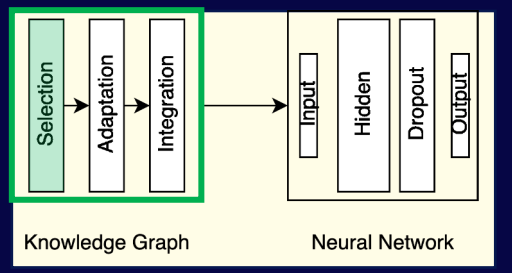
Hybrid Approach





- Knowledge graph

Gender, Sex, and Sexual Orientation Ontology



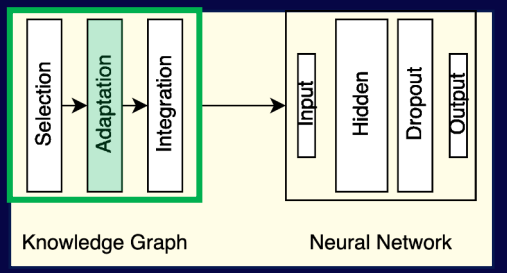
Semantic component

Deep learning component

The Open University

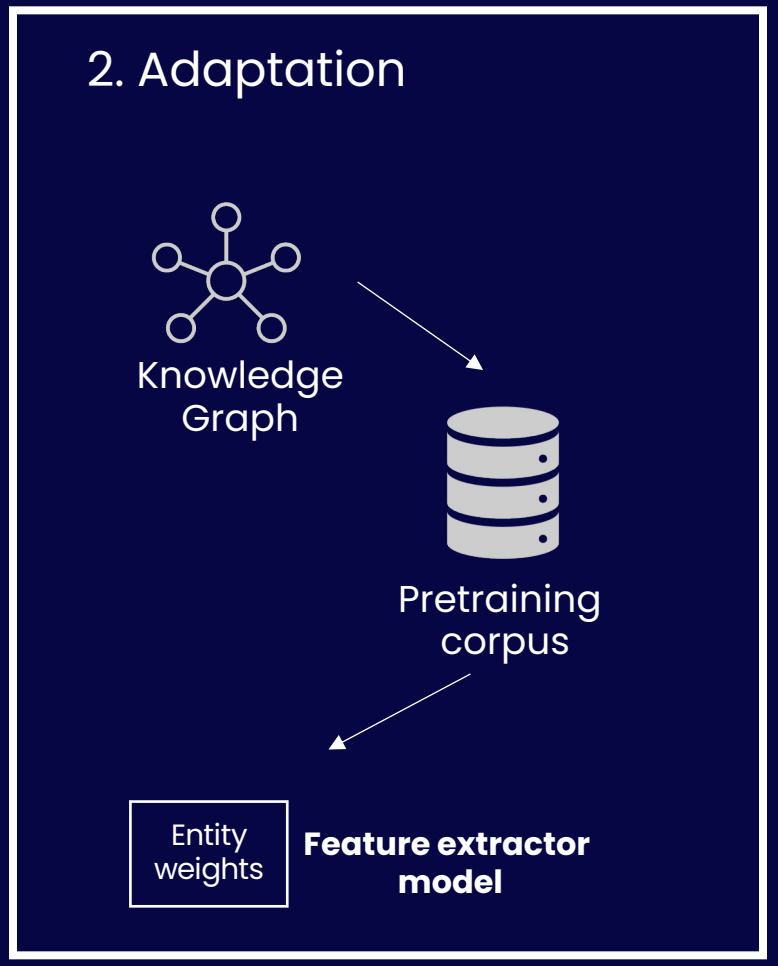
Kronk, Clair, et al. "Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow." Journal of the American Medical Informatics Association. 2020.

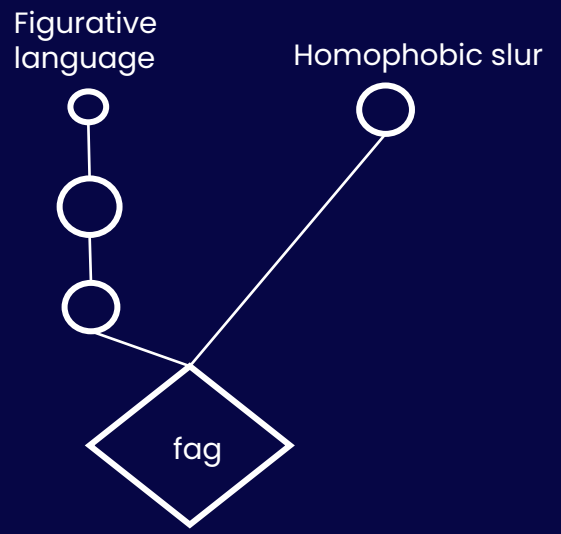
Hybrid Approach



- Pretraining data

Jigsaw Toxicity Sample

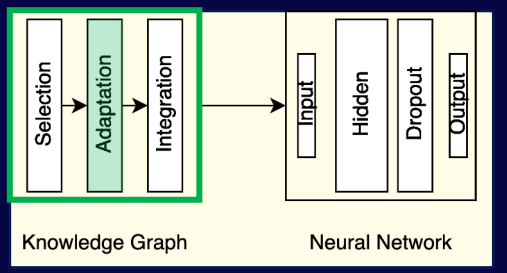


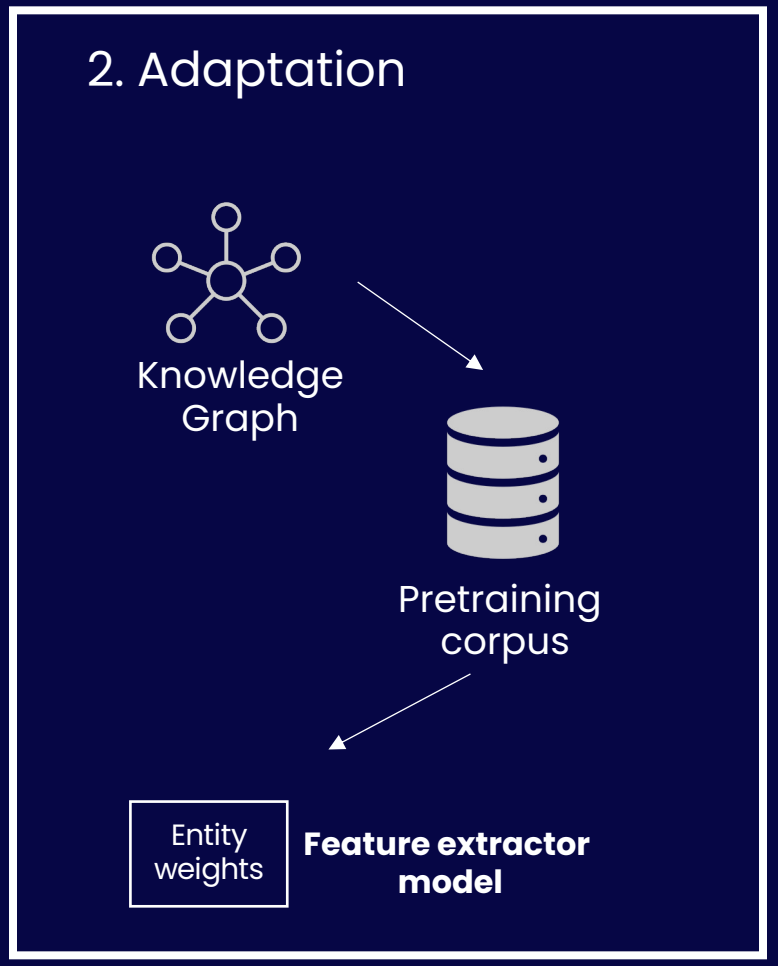


The Open University

Borkan, Daniel, et al. "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. Companion Proceedings of the World Wide Web Conference. 2019.

Hybrid Approach





- Entity weighting*

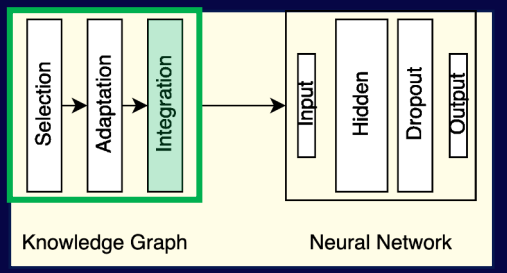
Frequency-based
DocF

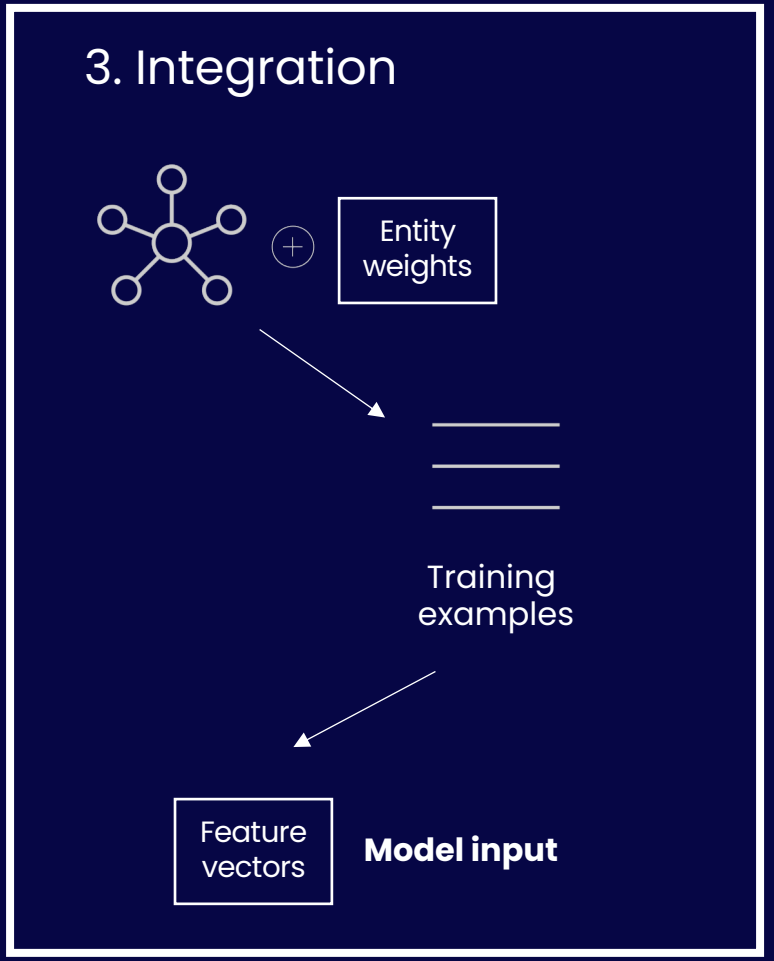
Machine Learning coefficients
Logistic Regression
Multinomial Naïve Bayes

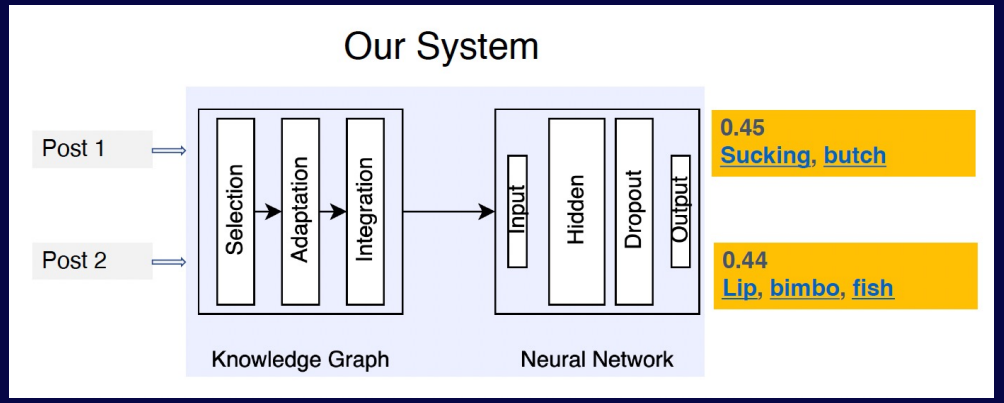
* Hierarchical entity expansion: an entity (e_i) appears in a text if e_i, any of its subclasses, or any of its instances appears in the text

The Open University

Hybrid Approach





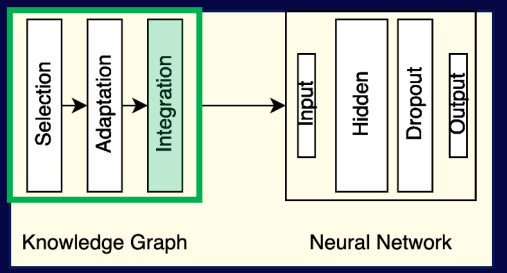


Semantic component

Deep learning component

The Open University

Hybrid Approach



- Training data

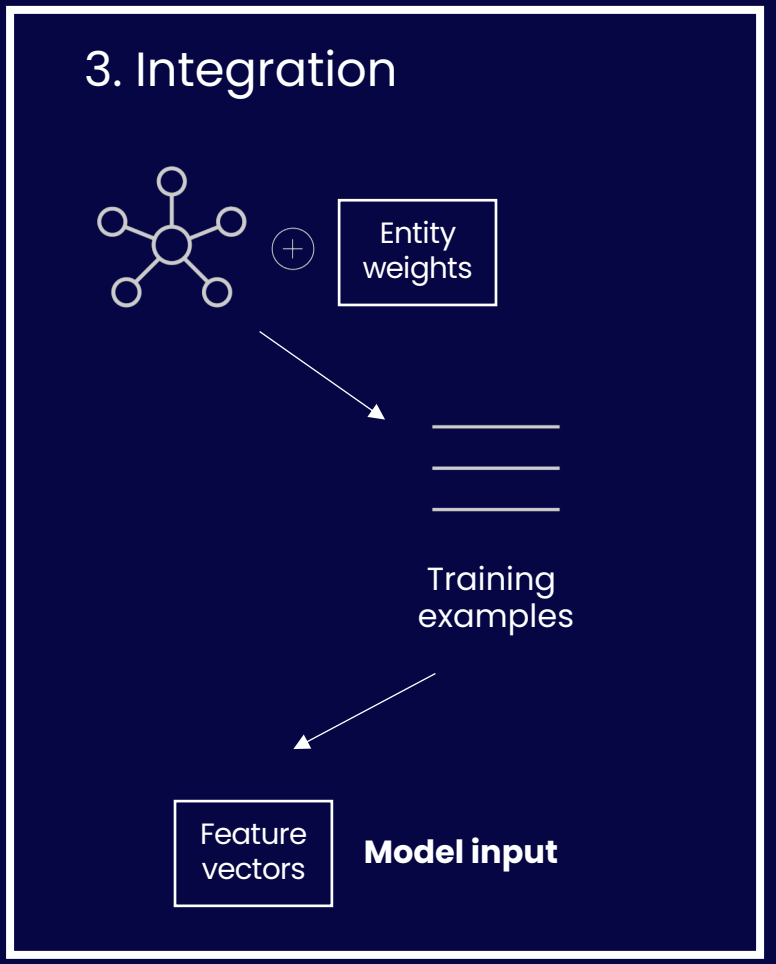
  Measuring Hate Speech

- Evaluation data

  Gab Hate Corpus
  HateXplain
  Xtremespeech

- Baselines

  RoBERTa base (System A)
  Toxic Debias (System B)



Sachdeva, Patrik, et al. "The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism." In Proceedings of the 1st Workshop on Perspectivist Approaches to NLP. 2022.

Sachdeva, Patrik, et al. "Targeted Identity Group Prediction in Hate Speech Corpora.." In: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). 2022.

Zhou, Xuhui, et al. "Challenges in Automated Debiasing for Toxic Language Detection." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021.

Results

Quantitative analysis

	Model	Gab Hate Corpus				XtremeSpeech$_{English}$				HateXplain			
		Accuracy	$F1$	ROC AUC	PR AUC	Accuracy	$F1$	ROC AUC	PR AUC	Accuracy	$F1$	ROC AUC	PR AUC
Lexicon [1]	Toxic Debias	**91.81**	58.82	74.82	40.20	**94.01**	52.41	72.96	31.12	84.43	67.36	79.01	52.66
Hybrid	HybridDocF$_h$	91.30	51.15	84.52	54.55	93.97	53.91	87.02	47.05	79.45	43.55	78.37	55.96
	HybridLR$_h$	90.64	**62.42**	89.30	64.38	90.79	49.27	88.36	50.79	83.48	67.72	88.15	68.35
	HybridMultiNB$_h$	89.36	61.11	90.13	68.24	90.38	47.74	87.26	51.80	85.63	73.57	91.38	78.37
PLM [2]	RoBERTa$_{base}$	88.85	61.55	**93.06**	**70.32**	92.99	**57.67**	**93.67**	**57.38**	**89.91**	**80.22**	**95.60**	**86.46**

Our hybrid approach outperforms existing interpretable approaches and is, to some extent, competitive with the use of pre-trained language models (PLMs).

The Open University

[1] *Zhou, Xuhui, et al. "Challenges in Automated Debiasing for Toxic Language Detection ." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021.*

[2] *Sachdeva, Pratik, et al. "Targeted Identity Group Prediction in Hate Speech Corpora." Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). 2022.*

Results

Qualitative analysis

We identify gender and sexuality related language in samples of human and model (dis)agreement.

Our hybrid approach is interpretable in that it identifies features that contribute to predict the class.

In our case, the language related to gender and sexuality.

Does this post mention or is about **gender or sexuality?**

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 .

Not related

0.83
gay, homosexuality, insult

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

Related to gender

0.49
beard, vagina, she

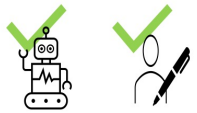
The Open University

Results

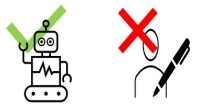
Qualitative analysis

Our error analysis finds that 97% of the false positives were **annotation errors**.

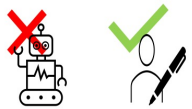
The entities in red were mentioned in texts that annotators deemed as “not related”.



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*

Does this post mention or is about **gender or sexuality?**

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 .

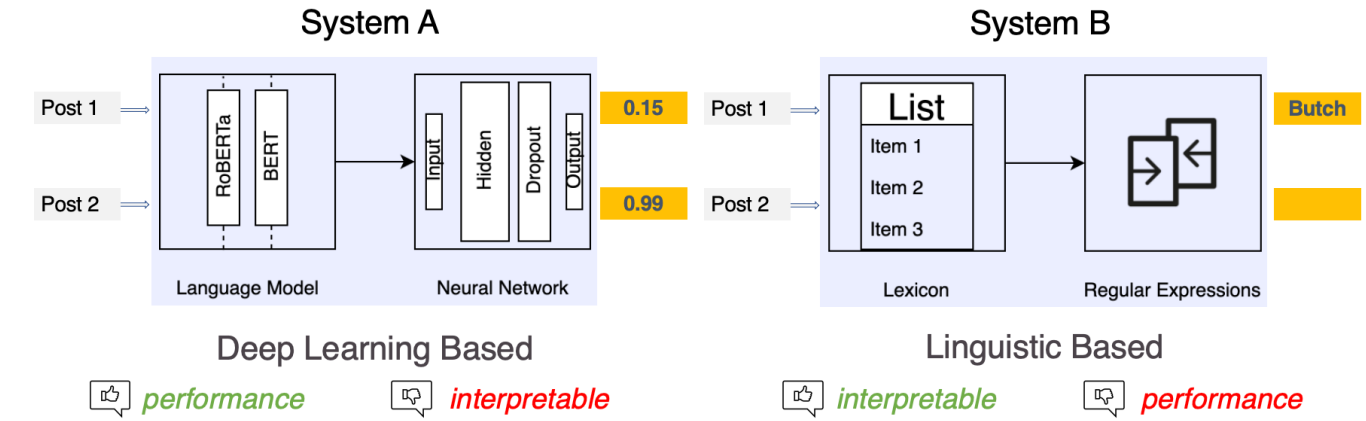
Not related

0.83
gay, homosexuality, insult

Conclusion

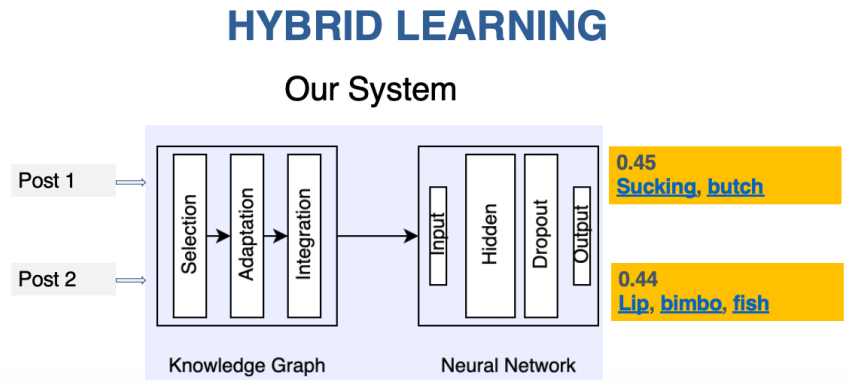
Hybrid learning is key for enhancing

- Interpretability
- Robustness
- Handling annotation errors



Limitations

- KG are human-intensive
- Moving from use case to a general case scenario



The Open University

Open questions

To what extent a hybrid architecture can…

- ***be an extrinsic interpretation of deep learning/opaque models?***
- ***prevent or correct data annotation errors?***
- ***audit bias in hate speech detection systems?***

Combining a KG with deep learning provides interpretability in the detection of target groups in hate speech

To what extent a hybrid architecture can...

- *interpret deep learning/opaque models?*
- *prevent or correct data annotation errors?*
- *audit bias in hate speech detection systems?*

Knowledge-Grounded Target Group Language Recognition in Hate Speech
Reyero Lobo, P., Daga, E., Alani, H., & Fernandez, M.

Contact

paula.reyero-lobo@open.ac.uk

www.linkedin.com/in/paula-reyero-lobo-116449170

https://orcid.org/0000-0001-5238-4550

The Open University

Funded by the European Union