Knowledge Graphs: Semantics, Machine Learning, and Languages
M. Acosta et al. (Eds.)

doi:10.3233/SSW230002

Knowledge-Grounded Target Group Language Recognition in Hate Speech

Paula REYERO LOBO [a,1], Enrico DAGA [a] Harith ALANI [a] and Miriam FERNANDEZ [a]
[a] *Knowledge Media Institute, The Open University, United Kingdom*
ORCiD ID: Paula Reyero Lobo https://orcid.org/0000-0001-5238-4550, Enrico Daga https://orcid.org/0000-0002-3184-5407, Harith Alani https://orcid.org/0000-0003-2784-349X, Miriam Fernandez https://orcid.org/0000-0001-5939-4321

Abstract. Hate speech comes in different forms depending on the communities targeted, often based on factors like gender, sexuality, race, or religion. Detecting it online is challenging because existing systems are not accounting for the diversity of hate based on the identity of the target and may be biased towards certain groups, leading to inaccurate results. Current language models perform well in identifying target communities, but only provide a probability that a hate speech text contains references to a particular group. This lack of transparency is problematic because these models learn biases from data annotated by individuals who may not be familiar with the target group. To improve hate speech detection, particularly target group identification, we propose a new hybrid approach that incorporates explicit knowledge about the language used by specific identity groups. We leverage a Knowledge Graph (KG) and adapt it, considering an appropriate level of abstraction, to recognise hate speech-language related to gender and sexual orientation. A thorough quantitative and qualitative evaluation demonstrates that our approach is as effective as state-of-the-art language models while adjusting better to domain and data changes. By grounding the task in explicit knowledge, we can better contextualise the results generated by our proposed approach with the language of the groups most frequently impacted by these technologies. Semantic enrichment helps us examine model outcomes and the training data used for hate speech detection systems, and handle ambiguous cases in human annotations more effectively. Overall, infusing semantic knowledge in hate speech detection is crucial for enhancing understanding of model behaviors and addressing biases derived from training data.

Keywords. hate speech, semantic enrichment, knowledge graphs, language models

1. Introduction

One of the challenges when addressing online hate speech is the extensive use of specialized language that is specific to the communities that are most frequently targeted. A motivating example is illustrated by **Figure 1**. We show two posts from a well-known

[1]Corresponding Author: Paula Reyero Lobo, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom; E-mail: paula.reyero-lobo@open.ac.uk.

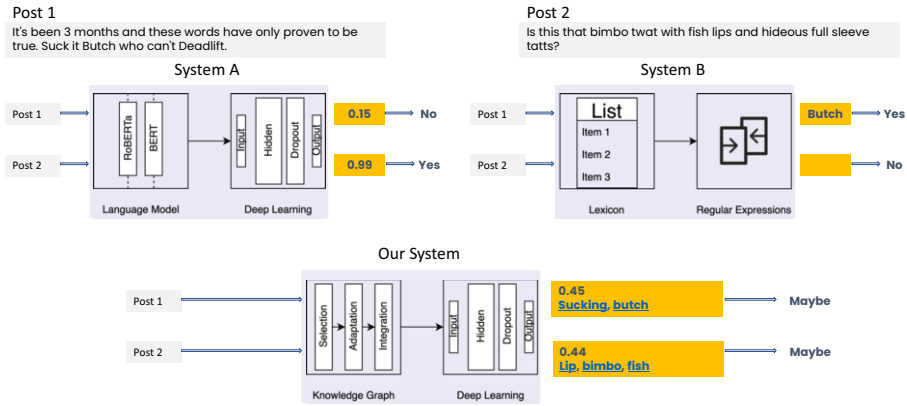

Figure 1. Hate speech recognisers based on language sensitive to gender and sexual orientation. Our approach (**Our System**) embeds a knowledge graph in a deep learning model to give a probability estimate that is competitive with state-of-the-art language models (**System A**), while providing more semantic information supporting the prediction than the existing linguistic approaches of lower accuracy (**System B**).

hate speech corpus [1]. To make appropriate decisions about the hateful nature of these posts, it is crucial to be familiar with the language being used.

Looking at it from the perspective of gender and sexual orientation, certain terms like "butch" and "bimbo" carry connotations related to a woman's masculinity or physical appearance. Depending on the context, these terms can be used to insult or reinforce social stereotypes. These subtle differences in language make the task of recognising hate speech highly dependent on the specific identity groups involved and the context in which the language is used [2].

There are two main approaches to recognise target group language in hate speech. The first, which we name **System A**, are supervised learning approaches [3,4,5] and the state-of-the-art (SOTA) are based on language models [6]. While they can achieve high performance, they only give probabilities of the posts containing references to particular identity groups. Only the second post in our example would contain sensitive language with high probability, but it leaves no further information as to why. This lack of supporting information is concerning due to the subjective interpretations and biases of human annotators. Judgement of hate speech varies significantly according to demographic characteristics [7,8], as it is to be expected that any human annotator will lack familiarity with the language of a particular target. Thus, integrating grounding knowledge may help to better understand predictions, but crucially to make the model more robust to biases in the training datasets. The second, which we name **System B**, are linguistic approaches displaying higher transparency [9,10,11]. They provide references using a list of terms (lexicon) or regular expression patterns and would identify, for example, relevant terminology in the first post [12]. However, they are less accurate as they only capture a sparse representation of the language sensitive to an identity group.

In this work, we aim to integrate a Knowledge Graph (KG) to enrich state-of-the art language model predictions with the entities supporting a decision, while preserving an optimal predictive performance (**Our System**). Following our motivating example, even if the model gives borderline probabilities for the particular posts, the additional semantic information helps to understand the prediction better. Representing terms as en-

tities provides useful semantic relations and properties such as definitions or synonyms, which we can exploit when auditing the model and data, as well as developing the hybrid approach.

Our contributions can be summarised as follows. We propose a conceptual framework to combine semantic knowledge in the form of a KG with an existing deep learning architecture (§3). Specifically, we propose a novel entity weighting scheme to effectively adapt a KG to text classification. We conduct a thorough quantitative and qualitative evaluation of our proposed hybrid learning framework (§4). Particularly, by comparing it against SOTA approaches on recognising language references to gender and sexual orientation in a variety of hate speech datasets. Our proposed semantically enriched approach displays equivalent performance to the use of language models, with transparency and higher generalisability to external datasets. The rest of the paper is structured as follows. (§2) summarises the related work, and (§5) our proposed approach, its strengths and limitations, and concludes the work. The instructions for accessing the data, code for training hybrid models in new domains, or applying them to new data, are published in a public repository[2].

2. Related Work

Content moderation systems generally focus on defining policies to protect any identity group or individual targeted [13]. Nevertheless, the specific sociolinguistic aspects of harmful expressions [14,15] make this phenomenon different for each target. A system focused on recognising hate directed to a specific group would not generalise to a different identity [2]. Similarly, linguistic nuances across group identities significantly impact the annotation of training datasets and lead to inconsistent labelling. Humans who labelled the data have different cultural backgrounds or beliefs [7], are exposed to language sensitive to groups with whom they may not self-identify [16,17], and their subjective interpretations of hate speech differ [18,19], especially when recognising identity groups that are frequent hate targets [20]. Lexical biases, where algorithms associate hate with any language that refers to a particular minority group [10,21], make it critical to analyse hate in terms of the identities targeted.

Due to these issues, one stream of work (**System B**) has been based on the manual selection of terms or expressions to recognise language references to identity groups in hate speech data. To a greater extent, these consist of direct references to members of an identity group (so-called group identifiers, identity terms or identity mentions) [10,12,9], or expressions that comprise potentially offensive language depending on the context, including slurs and objectifying outdated terms, as well as reclaimed slurs [11]. However, these approaches have mainly been used in the development of techniques to mitigate lexical bias in hate speech [22,23], or to measure the effectiveness of such mitigation techniques [24,25,26,27]. Existing approaches based on structured knowledge can only partially cover the prejudice towards identity groups in hate speech training data.

Another prominent line of work (**System A**) relies on supervised learning to more effectively recognise language references to identity groups [4,6,3,5]. These references may refer to broad groups based on sensitive attributes such as gender, race, sexual ori-

[2] https://github.com/preyero/hate-speech-identities

entation, disability or religion; or to the specific affected communities within an identity (e.g. women, transgender, male or other gender subgroups). To the best of our knowledge, these systems may display high performance, but only output a probability to indicate how much language sensitive to a group the post contains. However, in addition to the problems already discussed, language models also acquire biases from the large corpora used in pretraining [28]. Using additional sources of knowledge can produce models that are less flexible and more robust to specific annotation schemes, domain and contexts than deep learning models alone [29].

Semantic knowledge integration has helped to address bias related to the data annotation [30], to generalise better to unseen data [31], and to explain model predictions [32]. These examples imply, on the one hand, that it should be plausible to better overcome the discussed training data specificities given an adequate source of knowledge representation of the language from these communities. On the other, additional knowledge could enrich probabilistic model predictions to better understand them. A common challenge for hybrid approaches remains in finding the correct level of abstraction when applying semantic knowledge to a downstream task [29]. In this work, we propose a novel hybrid approach for grounding deep learning predictions in relevant knowledge for the task (**Our System**). Prior adaptation to the language distribution of the downstream task enables to integrate a KG simply and effectively into the model, without sacrificing predictive performance.

Particularly in hate speech, our approach allows focusing the detection from the target's perspective, as it highlights the specific entities representing language references that influence the prediction. Attending only to the signals learned with standard supervised learning (System A) has shown to acquire annotation and lexical biases from the training data and, in the worst cases, has lead marginalization and censorship of communities at risk [33,34]. Linguistic approaches (System B) to probe these systems for fairness fall short in addressing the language that refers to these communities [25,24]. With the domain specific constraints set by a knowledge-grounded approach (Our System), we intend to bring more focus on the language sensitive to frequent target communities to better understand and prevent online hate.

3. Hybrid Approach

In this section, we present a conceptual framework for supporting text classification with semantic knowledge [3]. First, we present the rationale for selecting semantic knowledge and describe the information leveraged in our approach (§3.1). Second, we present an adaptation phase. The goal is to assign weights to the KG entities based on their relevance in a pretraining corpus from the task domain (§3.2). Finally, we describe the integration of the adapted KG with a deep learning architecture (§3.3). The resulting six hybrid model versions are described in **Table 1**.

3.1. Semantic Knowledge Selection

A Knowledge Graph is a structured representation of knowledge that captures relationships between entities in a particular domain. It is a type of knowledge representation

[3]Note that, while we have selected semantic knowledge to cover specifically language from gender and sexual orientation, this is not fixed in our approach, and the KG can be exchanged.

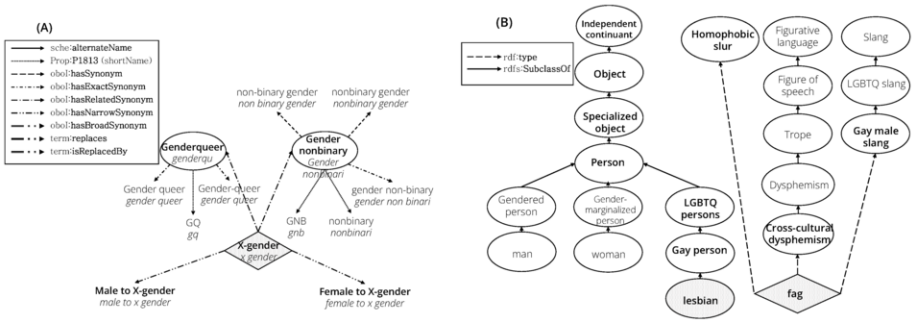

Figure 2. Knowledge Graph class (◯) and instance (◇) entities (*Gender, Sex, and Sexual Orientation* [35]). **(A)** Matching "X-gender" to text. **(B)** Hierarchical entity expansion of "lesbian" and "fag". Entities in bold are used to match or assign weight to the entities, respectively.

often used in combination with Machine Learning (ML) techniques, as they can help to improve model performance and interpretability in a variety of tasks including search, question answering and natural language understanding [29,36].

To select the particular KG for the task, we explore a wide range of existing KGs that could have comprehensive information about language sensitive to a group, including well-known KGs such as wikidata, DBPedia, and YAGO [37,38,39]. While some of these KGs contain information related to many identities, we selected the Gender, Sex and Sexual Orientation Ontology [35] as a more specific source of knowledge to base hate predictions on the language of a target group. This KG aims to be an integrated vocabulary system to address the lack of standardised gender and sexual orientation terminology in healthcare and includes, to this date, over 16,000 entities and 292 properties. To the best of our knowledge, it is the most comprehensive and up to date to describe these two common hat targets that we can consider for assessing our approach.

We show how we leverage this information following the examples in **Figure 2**. KG entities can capture semantic concepts (e.g., "gay male slang") or their concrete examples (e.g., "fag"). Additionally, there are object properties to make connections between entities, and data properties to describe the specific values or attributes of entities. For example, the wikidata property `Prop:P1813` indicates that the literal "GQ" can be a short name for the entity "Genderqueer". We use this richer representation of terms as entities to facilitate the matching of the KG to the texts **(A)**. We use properties like `rdf:type` (to link instances to their classes, such as "fag" to "homophobic slur") and `rdfs:SubClassOf` (to make hierarchical connections between the classes, e.g., "lesbian" to "gay person"), to exploit the KG structure in our hybrid approach **(B)**.

3.2. Knowledge Graph Adaptation

One major challenge when applying semantic knowledge in combination with ML to address particular tasks is the level of abstraction of its information [40]. The existing entities can sometimes encapsulate information that is too abstract or too fine-grained for the task at hand. In this work, we propose an adaptation phase that allows us to weigh the KG entities based on pretraining data. The aim is to give more relevance to those entities that better encapsulate the group language and information, which adds an additional dimension to the factual and structural information of previous hybrid approaches [41].

3.2.1. Search for Pretraining Data

To learn the weights of the KG we selected a balanced subset of the Jigsaw Toxicity dataset. For a full description of the datasets used in this work for training and validation, including references, descriptions, and statistics, see (§4.1) and **Table 2**. To create the *Jigsaw Sample* we selected all the texts annotated with sexual orientation, a total of 12,713 texts, and a random same size set that includes 16,850 texts annotated as related to gender. We then built a stratified sample of texts from all the remaining identities (i.e., religion, race, disability, and none) as the negative class. This provided us a balanced dataset of 50,852 texts, with 50% of them related to the class, and 50% related to any other identities.

3.2.2. Entity recognition

To determine whether an entity appears or not in a given text (*entity matching*) we take into account, not just the entity's label, but also its alternative names and existing synonyms. For example, in **Figure 2 (A)** we observe that "X-gender" is defined by "Male to X-gender","Female to X-gender", "Genderqueer" and "Gender nonbinary". In addition to these terms, we also consider stemming variations, such as "genderqu" and "gender nonbinari". Any text that contains any of these expressions is considered a text where the entity appears. Specifically, the KG properties shown in **Figure 2 (A)** are used to derive the synonyms for the entity matching. We develop a search index based on the Whoosh 2.7.4 library (https://whoosh.readthedocs.io) to speed up the entity matching, and obtain the stemming variations with its Porter stemmer native function.

3.2.3. Entity weighting

Finally, we consider two types of entity weighting schemes: (i) based on the frequencies of entities in the pretraining data and (ii) based on the learned coefficients of a ML model used for the domain task, in this case, the binary classification.

Entity weighting based on frequency (**DocF**) The weight provided to the entities is based on the ratio of appearance of that entity within the positive sample (i.e., all the texts related to gender and sexual orientation) vs. the negative sample (i.e., all the text related to any other identity). Lets D_p be the set of all texts related to the class, and D_n be the set of all text related to any other identity. Given an entity e_i, we consider the occurrences e_i in D_p (D'_p), and the occurrences e_i in D_n (D'_n). The weight of the entity $w(e_i)$ is then computed as $w(e_i) = D'_p/D_p - D'_n/D_n$, such that $w \in$ [-1, 1].

Entity weighting based on ML coefficients (**LR** and **MultiNB**) This approach provides weights to the different entities based on the coefficients defined by a machine learning model. The coefficients reflect how discriminative the entities are when predicting whether a particular text refers (or not) to the class. We use two different ML models for the task: Logistic Regression (LR) and Multinomial Naive Bayes (MultiNB). As input to train the ML models, we provide for each pretraining sample: (i) a class label (i.e., whether the text contains any language references to gender and sexual orientation) and (ii) the one-hot-encoding of the entities resulting from the *entity matching* explained above. That is, the ML models use entities as features for the classification. The resulting coefficients reflect the feature importance and how much each entity contributes to the prediction.

Additional modification of the weighing schemes based on hierarchical entity expansion To test whether the KG structure could serve us to better refine the adaptation, we propose a modification affecting the entities to be included in the weighting process. **Figure 2 (B)** shows in bold the process of expanding both an entity that is class (e.g., "lesbian") or an instance (e.g., "fag"). Every class expands up to its top-level using the `rdfs:SubclassOf` property to gather, e.g., that "lesbian" is a Gay Person, LGBTQ person, Person, and so on. For every instance, we would expand based on the `rdf:type` property and also include that "fag" is a Homophobic slur, Cross-cultural dysphemism, and Gay male slang, for example. Thus, in this modified version of the weighting scheme, an entity e_i is considered mentioned in a text if the entity e_i itself, any of its subclasses, or any of its instances appears in the text.

3.3. Knowledge Graph Integration

This phase describes how the adapted KG can be embedded with a deep learning architecture. In the following, we present the two main components of the proposed hybrid learning framework.

Semantic component Our hybrid approach considers an adapted KG (KG with pretraining weights) in the feature extraction. The weights of the entities found in the training samples constitute the feature vectors that are used as inputs. That is, the input for the deep learning component is a sparse vector representation, where the non-zero components are the weights of the entities in the training samples. Compared to contextualised word embeddings, the KG-based feature extraction provides a lower dimensional numerical representation of the input texts. We compare our hybrid approach to pretrained transformer architectures used in the SOTA, where RoBERTa [42] is the best-performing as compared to BERT [43] and the Universal Sentence Encoder [44].

Deep learning component The deep learning architecture used in the SOTA for recognising target group language in hate speech consists of a Feed-foward Multilayer Neural Network with a dropout layer and M binary layers for classification, one for each group identity (gender, sexual orientation, religion, race, disability, national origin, and age) [6]. That is, for a given input text, the model provides M probabilities indicating whether it contains any language related to each group. Because our work is focused on gender and sexual orientation, we only consider the probability of belonging to any of these two classes.

As a result, we obtain the six different hybrid model versions described in **Table 1**. The hyperparameters are the same for training all models, using 8 as the size of the training batches, the number of hidden layers set to 256, and a 0.05 dropout rate in the Feed-forward neural layer.

Table 1. Hybrid models (in bold) resulting from the different adaptation schemes (§3.2) used for hybrid feature extraction. H.E indicates the model variation when including hierarchical entity expansion.

Version	Description	H.E
HybridDocF	Features based on the ratio of entity occurrences	**HybridDocF_h**
HybridLR	Features based on coefficients of entities in a linear regression	**HybridLR_h**
HybridMultiNB	Features based on the coefficients of entities in a multinomial Naive Bayes model	**HybridMultiNB_h**

Table 2. Number and (%) of texts that are related to each identity group in training and validation datasets. A text may relate to none, one or more identity groups.

Identity	Jigsaw	Jigsaw$_{Sample}$	Measuring Hate Speech	Gab Hate Corpus*	HateXplain	XtremeSpeech$_{English}$
Gender	88790(19.82%)	16850(33.14%)	14825(37.47%)	568(7.27%)	1375(11.15%)	145(5.49%)
S. Orientation	12713(2.84%)	12713(25.00%)	7719(19.51%)	355(4.54%)	1643(13.32%)	39(1.48%)
Religion	70149(15.66%)	12683(24.94%)	6578(16.63%)	1347(17.24%)	3781(30.66%)	79(2.99%)
Race	42906(9.58%)	9674(19.02%)	12635(31.93%)	1711(21.90%)	4597(37.27%)	34(1.29%)
Disability	5559(1.24%)	4918(9.67%)	1120(2.83%)	241(3.08%)	54(0.44%)	
Origin			7744(19.57%)	1202(15.38%)	642(5.21%)	
Economic					9(0.07%)	23(0.87%)
Age			1051(2.66%)			
Politics				3063(39.2%)		
Any other					712(5.77%)	701(26.56%)
Total	448000	50852	39565	7813	12334	2639

4. Evaluation

In this section, we present our evaluation setup (§4.1) as well as the quantitative results against SOTA approaches for recognising language sensitive to gender and sexual orientation in hate speech (§4.2). (§4.3) provides the results from our qualitative evaluation. Specifically, an error analysis of the best-performing hybrid model (§4.3.1) and a data and model prediction analysis guided by the KG (§4.3.2).

4.1. Experimental Setup

This section describes the datasets used for training and testing our proposed hybrid models, and the baselines and metrics used for evaluation.

4.1.1. Data

We consider five datasets for training and testing our models. See **Table 2** for specific statistics and data descriptions.

Jigsaw [45]: To the best of our knowledge, this is the largest public toxicity corpus containing annotations of identity groups, with 448k annotated posts from the Civil Comments platform. These texts are annotated with a binary indicator of toxicity (toxic/non-toxic) and with the identity groups mentioned in them. Group annotations are based on the following identities: gender, sexual orientation, race, religion, disability or no mention of an identity group.

Measuring Hate Speech [20]: This dataset constitutes the largest hate speech training corpus and was used in the SOTA [6]. It contains 39,565 texts collected from Reddit, Youtube, and Twitter, and annotated with gender, sexual orientation, race, religion, age, disability and national origin identities. The gender and sexual orientation categories constitute 56.98% of the dataset.

Gab Hate Corpus [1]: This commonly used dataset contains 7813 texts collected from Gab, which were deemed hateful by the annotators and provide additional annotations for gender, sex, race, religion, disability, and political ideology.

XtremeSpeech English [46]: The complete dataset contains 5,180 texts collected from Facebook, Twitter and WhatsApp. The dataset is not yet public, but the authors have kindly shared with us a subset of 2,639 texts written in English that focuses on Kenya as a geographic location. These texts contain dangerous, derogatory and exclusionary

speech and are annotated considering the following identities: gender, sexual orientation, religion, race, and economic status.

HateXplain [47]: One of the first datasets that included annotations for identity groups. The corpus provides 12,334 texts collected from Twitter and Gab, and those deemed hateful provide annotations for gender, sexual orientation, race, religion, and national origin identity groups.

We use a subset the Jigsaw dataset as pretraining data for the KG adaptation (see §3.2). The hybrid models (§3.3) and the RoBERTa_base baseline are trained using the *Measuring Hate Speech* corpus with the same data preparation used by [6], and evaluated using the *HateXplain*, *XtremeSpeech English* and *Gab Hate Corpus*. While soft labels are used for training the models (i.e., the proportion of annotations for each text), majority voting is considered in the validation datasets for consistency with the baseline evaluation.

4.1.2. Baselines

We select the most representative System A (supervised learning) and System B (linguistic) approaches as baselines. As System A, a **RoBERTa_base** [6] model sets the upper bound in terms of performance. However, this model does not provide any insights on why texts are associated with a particular identity group and only learns from the training data. As System B, **Toxic Debias** [11] is the list of terms and regular expressions most commonly used for the identification of texts containing sensitive language towards minoritized groups in hate speech. From the 53 potentially offensive and 26 non-offensive mentions to these groups, 47 expressions refer to gender and sexual orientation. We highlight the 14 non-offensive and 33 possibly offensive mentions in our publicly available repository.

4.1.3. Evaluation Metrics

For comparability issues, we adopt the same evaluation as in the supervised learning baseline and consider Accuracy and F1 scores according to a 0.5 threshold, and two threshold-agnostic metrics: the Area under the ROC Curve (ROC AUC) and Area under the Precision-Recall Curve (PR AUC).

4.2. Evaluation Results

This section reports on the effectiveness of our hybrid approach for recognising language sensitive to gender and sexual orientation identities in hate speech datasets. First, we compare hybrid models against the best-performing baseline (System A) with a 5-fold cross-validation (**Figure 3**) for comparability with the original paper [6]. Second, we test the robustness of the linguistic, supervised, and hybrid learning approaches to different data contexts with a thorough evaluation on datasets external to training (**Table 3**). We include, to the best of our knowledge, all published datasets on hate speech that have consider identity groups in their annotation. We note that the linguistic baseline (System B) does not require training. For simplicity, we only include in the table the hybrid models with hierarchical entity expansion as they are the best-performing ones.

Finding 1. *Our proposed hybrid approach is simple and effective, displaying a comparable performance to the SOTA supervised learning (System A) approaches based on language models.*

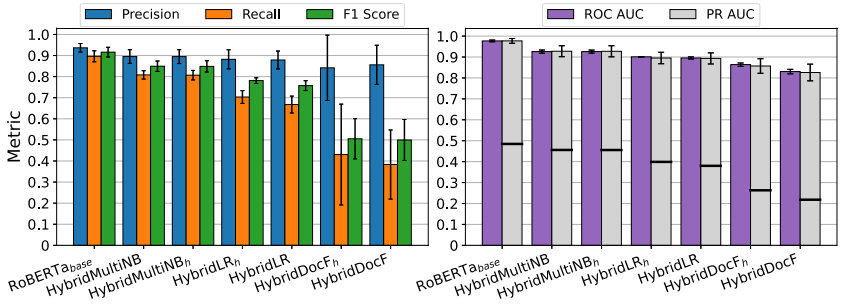

Figure 3. Supervised and hybrid learning model cross-validation results over the training corpus (Measuring Hate Speech). *ML-based hybrid models can be as effective as language models in recognising language references to gender and sexual orientation in hate speech.*

Table 3. Results of the linguistic, supervised learning and hybrid models when testing out of training domain (Gab Hate Corpus, XtremeSpeech$_{English}$, and HateXplain). *Semantic knowledge makes the model more robust to changes in domain and context.*

	Gab Hate Corpus				XtremeSpeech$_{English}$				HateXplain			
Model	Accuracy	$F1$	ROC AUC	PR AUC	Accuracy	$F1$	ROC AUC	PR AUC	Accuracy	$F1$	ROC AUC	PR AUC
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
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
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**

As seen in **Figure 3**, the hybrid models based on ML coefficients (LR and MultiNB) obtain competitive results with a RoBERTa_base model. They outperform the frequency-based models (DocF), particularly in terms of Recall and F1 Score, with lower standard variation across folds. The differences in incidence rates (horizontal black lines of the PR AUC bar to show the proportion of positive predictions across folds) indicate that the HybridDocF predictions are less aligned with the transformer and other hybrid-based models. The figure also shows that hierarchical entity expansion outperforms their counterparts for LR and DocF models, and remains the same in the MultiNB setting.

Finding 2. *The hybrid models display higher generalisability when applied to external datasets than baseline approaches.*

Table 3 shows the generalisability to external validation datasets. As expected, performance drops when evaluating these models with data of a different nature to that used during training (see §4.1.1 and **Table 2** for details on the platform sources, data characteristics and distribution in annotations). This is true, especially in the XtremeSpeech corpus, which captures data from Kenya and English is used in combination with Swahili for some texts. We show, however, that the generalisability of our hybrid models is higher than the baseline, since the gap with the upper bound set by the language models drops with respect to the in-domain evaluation. Aside from enhancing transparency, the introduction of semantic knowledge is key to making these models more robust to context, data and domain changes.

Finding 3. *Our hybrid models display higher performance than the linguistic (System B) baseline while also providing higher levels of interpretability.*

All our proposed hybrid methods outperform Toxic Debias in all metrics except Accuracy in *XtremeSpeech English* and *Gab Hate Corpus*. This is due to imbalanced dataset

Table 4. Error analysis in a False Positives (FP) and False Negatives (FN) sample. A.E indicates the categories that are associated with possible Annotation Errors. N indicates the number of errors found in our sample. *Semantic knowledge provides a better understanding of training data and model outcomes.*

Category (FP)	Definition	A.E	N
Demographic descriptor	Direct explicit reference to a member of the identity group.	X	**117**
Targeted language	Insults, sexually explicit or topics related to the group.	X	20
Implicit reference	Referes 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 domain.		19
Missed by method	Mention not correctly found or given importance by model.		**85**

conditions. As shown in **Table 2**, these datasets have a lower number of texts from the positive class. System B Accuracy drops below both supervised and hybrid approaches when the proportion of true positives is higher (*HateXplain*), where a model predicting only one class would have a lower chance of obtaining high scores. We observe how performance is significantly higher for the hybrid models in all other evaluation metrics.

In addition to outperforming the linguistic baseline, our hybrid approach provides a higher level of interpretability. While the lexicon only provides terms recognised in the text to categorise it as being associated with the group (e.g., the term "fag"), our hybrid methods provide entities, and with them, their semantic structure. In **Figure 2** we see that, in addition to the label, the KG structure informs about the fact that it is a Gay male slang, a Homophobic slur, and holds different meanings across cultures (i.e., a cross-cultural dysphemism). Similarly, the properties in the KG would also inform that "faggot" is a related synonym, and that it can be replaced by "gay man". KG properties and relations provide a much richer level of knowledge representation than simple terms. This richer source of semantic knowledge has helped to achieve a competitive hybrid baseline with the one based on language models.

4.3. Exploiting Semantic Enrichment

This section highlights that the KG is also instrumental to enhance the model's transparency and robustness to problems in hate speech training datasets. We begin our qualitative evaluation with an in-depth error analysis (§4.3.1), and extend it to audit how the training datasets capture language related to these groups (§4.3.2).

4.3.1. Error Analysis

Using a thematic analysis approach [48], we identify emerging typologies of errors and group them into distinct categories (**Table 4**). We focus on LR-based hybrid model with hierarchical entity expansion (HybridLR_h) because it outperforms the MultiNB-based models in two of the three validation datasets (**Table 3**). We translate errors into distinct categories considering: (i) each text, (ii) its group identity annotation, (iii) the model predicted probability for the gender and sexual orientation identities and, (iv) the list of entities provided by the model ranked by their weight. Our analysis consists of a 100-quartile random error sample in the validation datasets, to cover equally errors in all ranges of predicted probabilities. Sampling in each of the 3 validation datasets results in 280 texts.

Finding 3. *The semantically enriched predictions provided by the adapted KG enhance the transparency of the model, which helps to better understand model errors and to detect possible annotation errors.*

As shown in **Table 4**, we identified seven distinct categories of errors. We first analysed the false positives (FP) errors (i.e., where the model indicated that the texts mention gender and sexual orientation identities but annotators indicated the opposite). Our analysis reveals that, while most of these texts were not annotated as related, they contain relevant entities, including (i) *demographic descriptors* such as woman, man, girls, male, females, trans, girlfriend, homosexual, gay or lesbian, (ii) *targeted language*, such as insults and sexually explicit references (e.g., sexual assault), and/or (iii) thematically related entities, like birth, feminist, or lgbt. Less distinctive cases include the use of *implicit references* such as pronouns to refer to members of the group. It is important to highlight that all of these instances could be interpreted as annotation errors rather than model errors since the annotators may have missed relevant information in the text. While more experiments are needed to provide robust conclusions, these results seem to indicate that the information provided by the KG could be key to further investigating annotation disagreements. We find only 3 examples that are incorrectly categorised by the model due to polysemy (e.g., "straight" not meaning a sexual orientation), which were clear model errors.

We then analysed the false negative (FN) errors (i.e., instances where our model said that the text did not belong to the categories of gender and sexual orientation and the annotators indicated the opposite). The first category identified is *No Reference*. These are texts that do not display any term associated with sex and sexual orientation. These can also constitute annotation errors, where the annotator wrongly associated the text with these identities. The second category identified is *Missed at content*. These are errors where KG had the relevant entities, but they were not recognised within the text due to spelling mistakes (e.g., "feminisium", "gayfagsex"), or because those entities did not appear in the training corpus used during the KG adaptation phase (e.g., "sexism"). The third category identified is *Missed by method*. This reflects errors where either the KG did not contain the relevant information due to lack of coverage (e.g., "gayzors, "lezbos", "fellatiate", "madam", or "negress") or the relevant entities had a low weight assigned during the KG adaptation phase (e.g., "transphobe", "prostitutes", or "polygamy"). These issues constitute 65% of the errors. In some cases, entities related to the group receive a low weight during KG adaptation due to having noisy synonyms (e.g., "t word" as related synonym of "tranny" and found in texts with "t* word"). These observations could help to improve the specificity of the KG by revisiting which properties to use as synonyms for the entity recognition. Similarly, lack of coverage can also be due to an insufficient level of granularity with the KG (e.g., "daughter" and "son" as synonyms for "child", which is not expressing gender). These insights provide relevant information for improving both the KG and the proposed hybrid solutions.

Overall, we are able to identify these error categories guided by the additional semantic information provided by the hybrid approach. The issues identified along the model pipeline will help us in our future work to refine our hybrid models and enhance its performance. We also draw attention to the finding that grounding predictions on knowledge can help us to better understanding not only model errors, but the ambiguous cases that exist within the data that may be harder to classify by human annotators.

Table 5. KG entities sorted by feature importance that represent the language related to Gender and Sexual orientation in a sample of true predictions and errors. *Semantic knowledge displays hard-to-classify cases for the model and the human annotators.*

Sample	Target Group Language
True Positives	woman, man, LGBTQ, LGBT, .lgbt, man who has sex with men, r/lgbt, male gender identity, lesbian woman, female gender identity, gay man, .gay, transgender person, Black man, gay, homo, gay person, gender, heterosexual, homosexuality, feminist, lesbian, asexual and homoromantic person, gai, A-Gay, heterosexual person, gay identity, human homosexuality, sak veng (long hair), queer sexual orientation, transgender, same-gender marriage, marriage, transgenderism, heterosexuality, womanism, pederasty, lesbian identity, sexuality, heterosexual identity, lesbianism, homosexualism, personal identity, homophobia, queer identity, person who menstruates, mixed-orientation marriage, single person, *sex, feminism, partner, marital partner, sex worker, fag, faggot, masculism, pussy, hers, thot, rape, menstruation, bitch*
Missing in annotation	woman, man, LGBT, woman of color, .lgbt, man who has sex with men, r/lgbt, male gender identity, female gender identity, .gay, gay, gay person, heterosexual, homosexuality, feminist, lesbian, asexual and homoromantic person, gai, A-Gay, gay identity, human homosexuality, queer sexual orientation, transgender, same-gender marriage, marriage, interpersonal orientation, womanism, lesbian identity, sexuality, homosexualism, personal identity, lesbianism, homophobia, queer identity, single person, abusive person, sex, interpersonal attraction, *partner, faggot, semen, pussy, hers, bitch*
Missing in prediction	man who has sex with men, feminist, marriage, homophobia, person who menstruates, sex, **interracial marriage**, **sex work client**, *marital partner,* ***parent****, sex worker, fag, faggot,* ***rapist****,* ***female gender role****, pussy,* ***abortion****,* ***morphological enlargement****, hers,* ***vagina****, thot,* ***penis****, rape, domestic violence,* ***she****,* ***bimbo****,* ***sexual abstinence****, cunt, bitch,* ***he****,* ***whore****,* ***slut****,* ***fuck****,* ***Mrs.****,* ***rainbow flag***

4.3.2. Auditing Training Datasets

Motivated by our error analysis, we exploit our semantically enriched method to assess how hate speech training dataset captures identity group language. The result of this analysis can be seen in **Table 5**. We follow the same approach in (§4.3.1) and draw a 100-quartile sample of true predictions, which includes 286 texts. We then use the elbow method [49] to filter those entities that are more relevant considering the weights provided by the HybridLR_h model. Within this category (*True Positives*) we show those relevant entities extracted from texts where both, the model and the human annotators, agreed that the text was related to these identities. We note that these lists are not intended to provide an exhaustive list of all the language related to gender and sexual orientation. Nevertheless, they provide the minimum set of KG entities required to identify language references to these identities, and gives valuable insights to better understand how common hate targets are captured by the hate speech training datasets.

We conduct the same analysis in the samples with mismatches of annotation and model predictions. Within the category *Missing in annotation*, we display the relevant entities in texts indicated as related by the model, but not by the human annotators. Using the same data from the error analysis, the list includes language from 147 texts, and highlights relevant entities that the annotators may have missed when assessing the texts.

Within the category *Missing in prediction*, we display the relevant entities in texts indicated as related by the human annotators, but not by the model. The list highlights relevant entities that, while available within the KG, were not given enough relevance during the hybrid approach. Entities in italics correspond to those not included by the elbow point due to having a lower weight, but required to identify related texts. Entities in bold are unique to the texts missed in the prediction. This important entities highlight the complexity of learning language, as some of these entities may only be related to the gender and sexual orientation identities in specific contexts (e.g, f*ck as a swear word, or being sexually explicit). The same is true for entities that appear in texts that are only sometimes annotated as related (e.g. woman, LGBT, gay).

Finding 4. *A knowledge-grounded approach for understanding hate from the perspective of gender and sexual orientation identities helps to identify language relevant for their recognition in hate speech, as well as the terminology that may be associated with either model or annotation errors.*

5. Conclusion

We present a novel hybrid approach for grounding deep learning predictions in semantic knowledge relevant to the recognition of language references to gender and sexual orientation in hate speech. First, selecting a KG as semantic knowledge is a richer form of structured knowledge than existing linguistic approaches, providing novel semantically enriched predictions that are as effective to the use of black-box language models. Second, an adaptation phase based on machine learning allows finding an optimal representation level, which is a major challenge for applying semantic knowledge to downstream tasks. Finally, we propose a simple and effective feature-based approach to integrate the adapted KG to a neural network. Our evaluation on gender and sexual orientation demonstrates that a knowledge-grounded approach is key to enhance model transparency, robustness, and handling of annotation errors. Particularly, as it can highlight vocabulary for better understanding how training data captures identity group language, what are the type of errors in the model and, more interestingly, the ambiguities in human annotations.

We acknowledge we only evaluate our approach on two particular groups and one KG. Further research on other target groups would underline the value of knowledge-based approaches to hate speech detection. Similarly, considering a variety of KG domains and sizes would provide valuable insights on how to integrate them more effectively. KGs are generally costly to generate and maintain, and sometimes their coverage may not be sufficient for the task [50]. Our work however shows that, when this knowledge is available, it can positive complement and enhance a standard deep learning approach. We acknowledge the limitations of hate speech evaluation using standard performance metrics and leave as future work settings specific to the task [51,52] tailored to these identities. In terms of annotation findings, our semantically enriched models uncover references in 97% false positive errors. A more exhaustive analysis is needed to investigate the reasons behind these disagreements and the extent to which these cases constitute difficult to classify training examples that could improve hate speech recognisers [53]. Nevertheless, analysing hate in terms of the groups targeted is critical due to the subtlety of this language, which makes the recognition of hate speech even more difficult for annotators to understand and perceive [20].

To conclude, we particularly emphasise that this work does not aim to infer an individual's sensitive attributes [54]. This work rather aims to attend to the sociolinguistic aspects in hate speech in the hope of better contextualising automatic recognition systems with the language use of the social realities they imply.

References

[1] Kennedy B, Atari M, Davani AM, Yeh L, Omrani A, Kim Y, 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.

[2] Yoder M, Ng L, Brown DW, Carley K. How Hate Speech Varies by Target Identity: A Computational Analysis. In: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics; 2022. p. 27-39. Available from: https://aclanthology.org/2022.conll-1.3.

[3] Silva L, Mondal M, Correa D, Benevenuto F, Weber I. Analyzing the targets of hate in online social media. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 10; 2016. p. 687-90.

[4] Mossie Z, Wang JH. Vulnerable community identification using hate speech detection on social media. Information Processing & Management. 2020;57(3):102087. Available from: https://www.sciencedirect.com/science/article/pii/S0306457318310902.

[5] Waseem Z, Hovy D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: Proceedings of the NAACL Student Research Workshop. San Diego, California: Association for Computational Linguistics; 2016. p. 88-93. Available from: https://aclanthology.org/N16-2013.

[6] Sachdeva P, Barreto R, Von Vacano C, Kennedy C. Targeted Identity Group Prediction in Hate Speech Corpora. In: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Seattle, Washington (Hybrid): Association for Computational Linguistics; 2022. p. 231-44. Available from: https://aclanthology.org/2022.woah-1.22.

[7] Sap M, Swayamdipta S, Vianna L, Zhou X, Choi Y, Smith NA. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 5884-906. Available from: https://aclanthology.org/2022.naacl-main.431.

[8] Gubitz SR. Race, Gender, and the Politics of Incivility: How Identity Moderates Perceptions of Uncivil Discourse. Journal of Race, Ethnicity, and Politics. 2022;7(3):526–543.

[9] Nozza D, Bianchi F, Lauscher A, Hovy D. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 26-34. Available from: https://aclanthology.org/2022.ltedi-1.4.

[10] Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and Mitigating Unintended Bias in Text Classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 67–73. Available from: https://doi.org/10.1145/3278721.3278729.

[11] Zhou X, Sap M, Swayamdipta S, Choi Y, Smith N. 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. Available from: https://par.nsf.gov/biblio/10308662.

[12] Smith EM, Hall M, Kambadur M, Presani E, Williams A. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 9180-211. Available from: https://aclanthology.org/2022.emnlp-main.625.

[13] Calabrese A, Ross B, Lapata M. Explainable Abuse Detection as Intent Classification and Slot Filling. Transactions of the Association for Computational Linguistics. 2022 12;10:1440-54. Available from: https://doi.org/10.1162/tacl_a_00527.

[14] Saha K, Kim SC, Reddy MD, Carter AJ, Sharma E, Haimson OL, et al. The Language of LGBTQ+ Minority Stress Experiences on Social Media. Proc ACM Hum-Comput Interact. 2019 nov;3(CSCW). Available from: https://doi-org.libezproxy.open.ac.uk/10.1145/3361108.

[15] Kwarteng J, Perfumi SC, Farrell T, Third A, Fernandez M. Misogynoir: challenges in detecting intersectional hate. Social Network Analysis and Mining. 2022;12(1):166.

[16] Goyal N, Kivlichan ID, Rosen R, Vasserman L. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. Proc ACM Hum-Comput Interact. 2022 nov;6(CSCW2). Available from: https://doi-org.libezproxy.open.ac.uk/10.1145/3555088.

[17] Mastromattei M, Basile V, Zanzotto FM. Change My Mind: How Syntax-based Hate Speech Recognizer Can Uncover Hidden Motivations Based on Different Viewpoints. In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022. Marseille, France: European Lan-

guage Resources Association; 2022. p. 117-25. Available from: https://aclanthology.org/2022.nlperspectives-1.15.

[18] Rottger P, Vidgen B, Hovy D, Pierrehumbert J. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 175-90. Available from: https://aclanthology.org/2022.naacl-main.13.

[19] Kazienko P, Bielaniewicz J, Gruza M, Kanclerz K, Karanowski K, Miłkowski P, et al. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. Information Fusion. 2023;94:43-65. Available from: https://www.sciencedirect.com/science/article/pii/S1566253523000167.

[20] Sachdeva P, Barreto R, Bacon G, Sahn A, von Vacano C, Kennedy C. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022. Marseille, France: European Language Resources Association; 2022. p. 83-94. Available from: https://aclanthology.org/2022.nlperspectives-1.11.

[21] Xu A, Pathak E, Wallace E, Gururangan S, Sap M, Klein D. Detoxifying Language Models Risks Marginalizing Minority Voices. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. p. 2390-7. Available from: https://aclanthology.org/2021.naacl-main.190.

[22] Kennedy B, Jin X, Mostafazadeh Davani A, Dehghani M, Ren X. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 5435-42. Available from: https://aclanthology.org/2020.acl-main.483.

[23] Zhang G, Bai B, Zhang J, Bai K, Zhu C, Zhao T. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4134-45. Available from: https://aclanthology.org/2020.acl-main.380.

[24] Attanasio G, Nozza D, Hovy D, Baralis E. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1105-19. Available from: https://aclanthology.org/2022.findings-acl.88.

[25] Cai Y, Zimek A, Wunder G, Ntoutsi E. Power of Explanations: Towards automatic debiasing in hate speech detection. arXiv e-prints. 2022 Sep:arXiv:2209.09975.

[26] Sen I, Samory M, Wagner C, Augenstein I. Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 4716-26. Available from: https://aclanthology.org/2022.naacl-main.347.

[27] Chuang YS, Gao M, Luo H, Glass J, Lee Hy, Chen YN, et al. Mitigating Biases in Toxic Language Detection through Invariant Rationalization. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Online: Association for Computational Linguistics; 2021. p. 114-20. Available from: https://aclanthology.org/2021.woah-1.12.

[28] Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence. 2022;4(3):258-68.

[29] Futia G, Vetrò A. On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI—Three Challenges for Future Research. Information. 2020 Feb;11(2):122. Available from: http://dx.doi.org/10.3390/info11020122.

[30] Reyero Lobo P, Daga E, Alani H, Fernandez M. Semantic Web technologies and bias in artificial intelligence: A systematic literature review. Semantic Web. 2023;14(4):745-70. Publisher: IOS Press.

[31] Cui L, Wu Y, Liu S, Zhang Y. Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational

Linguistics; 2021. p. 2328-37. Available from: https://aclanthology.org/2021.emnlp-main.179.

[32] Sridhar R, Yang D. Explaining Toxic Text via Knowledge Enhanced Text Generation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics; 2022. p. 811-26. Available from: https://aclanthology.org/2022.naacl-main.59.

[33] Haimson OL, Delmonaco D, Nie P, Wegner A. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. Proc ACM Hum-Comput Interact. 2021 oct;5(CSCW2). Available from: https://doi-org.libezproxy.open.ac.uk/10.1145/3479610.

[34] Thiago DO, Marcelo AD, Gomes A. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. Sexuality & culture. 2021;25(2):700-32.

[35] Kronk CA, Dexheimer JW. Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. Journal of the American Medical Informatics Association. 2020 06;27(7):1110-5. Available from: https://doi.org/10.1093/jamia/ocaa061.

[36] Breit A, Waltersdorfer L, Ekaputra FJ, Sabou M, Ekelhart A, Iana A, et al. Combining Machine Learning and Semantic Web: A Systematic Mapping Study. ACM Comput Surv. 2023 mar. Available from: https://doi.org/10.1145/3586163.

[37] Vrandečić D, Krötzsch M. Wikidata: A Free Collaborative Knowledgebase. Commun ACM. 2014 sep;57(10):78–85. Available from: https://doi.org/10.1145/2629489.

[38] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag; 2007. p. 722–735.

[39] Suchanek FM, Kasneci G, Weikum G. Yago: A Core of Semantic Knowledge. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07. New York, NY, USA: Association for Computing Machinery; 2007. p. 697–706. Available from: https://doi.org/10.1145/1242572.1242667.

[40] d'Avila Garcez A, Lamb LC. Neurosymbolic AI: The 3rd Wave. arXiv e-prints. 2020 Dec:arXiv:2012.05876.

[41] Hamilton K, Nayak A, Božić B, Longo L. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. Semantic Web. 2022 nov:1-42. Available from: https://doi.org/10.3233%2Fsw-223228.

[42] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.

[43] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[44] Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, et al. Universal sentence encoder. arXiv preprint arXiv:180311175. 2018.

[45] Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. CoRR. 2019;abs/1903.04561. Available from: http://arxiv.org/abs/1903.04561.

[46] Maronikolakis A, Wisiorek A, Nann L, Jabbar H, Udupa S, Schuetze H. Listening to Affected Communities to Define Extreme Speech: Dataset and Experiments. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1089-104. Available from: https://aclanthology.org/2022.findings-acl.87.

[47] Mathew B, Saha P, Yimam SM, Biemann C, Goyal P, Mukherjee A. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence. 2021 May;35(17):14867-75. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/17745.

[48] Braun V, Clarke V. Using thematic analysis in psychology. Qualitative research in psychology. 2006;3(2):77-101.

[49] Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st international conference on distributed computing systems workshops. IEEE; 2011. p. 166-71.

[50] Janowicz K, Yan B, Regalia B, Zhu R, Mai G. Debiasing Knowledge Graphs: Why Female Presidents

are not like Female Popes. In: ISWC (P&D/Industry/BlueSky); 2018. .

[51] Röttger P, Vidgen B, Nguyen D, Waseem Z, Margetts H, Pierrehumbert J. HateCheck: Functional Tests for Hate Speech Detection Models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 41-58. Available from: https://aclanthology.org/2021.acl-long.4.

[52] Calabrese A, Bevilacqua M, Ross B, Tripodi R, Navigli R. AAA: Fair Evaluation for Abuse Detection Systems Wanted. WebSci '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 243–252. Available from: https://doi-org.libezproxy.open.ac.uk/10.1145/3447535.3462484.

[53] Leonardelli E, Menini S, Aprosio AP, Guerini M, Tonelli S. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. arXiv preprint arXiv:210913563. 2021.

[54] Keyes O. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proc ACM Hum-Comput Interact. 2018 nov;2(CSCW). Available from: https://doi-org.libezproxy.open.ac.uk/10.1145/3274357.