Topic Refinement in Multi-level Hate Speech Detection

Tom Bourgeade[1(✉)], Patricia Chiril[3], Farah Benamara[1,2], and Véronique Moriceau[1]

[1] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
{tom.bourgeade,farah.benamara,veronique.moriceau}@irit.fr
[2] IPAL, CNRS-NUS-ASTAR, Singapore, Singapore
[3] University of Chicago, Chicago, IL, USA
pchiril@uchicago.edu

Abstract. Hate speech detection is quite a hot topic in NLP and various annotated datasets have been proposed, most of them using binary generic (hateful vs. non-hateful) or finer-grained specific (sexism/racism/etc.) annotations, to account for particular manifestations of hate. We explore in this paper how to transfer knowledge across both different manifestations, and different granularity or levels of hate speech annotations from existing datasets, relying for the first time on a multilevel learning approach which we can use to refine generically labelled instances with specific hate speech labels. We experiment with an easily extensible Text-to-Text approach, based on the T5 architecture, as well as a combination of transfer and multitask learning. Our results are encouraging and constitute a first step towards automatic annotation of hate speech datasets, for which only some or no fine-grained annotations are available.

1 Motivation

Hate Speech (HS hereafter) has become a widespread phenomenon on social media platforms like Twitter, and automated detection systems are thus required to deal with it. In spite of no universally accepted definition of HS, these messages may express threats, harassment, intimidation or *"disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic"* [26]. HS may have different topical focuses: misogyny, sexism, racism, xenophobia, etc. Which can be referred to as *hate speech topics*. For each HS topic, hateful content is directed towards specific *targets* that represent the community (individuals or groups) receiving the hatred.[1] HS is thus, by definition, *target-oriented*, and it involves different ways of linguistically expressing hateful content such as references to racial or sexist stereotypes, the use of negative and positive emotions, swearing terms, etc., all of which have to be considered if one is to train effective automated HS detection systems.

[1] For example, black people and white people represent possible targets when the topical focus is *racism* [31], while women are the targets when the topical focus is *misogyny* or *sexism* [22].
Warning: *This paper includes tweets that may contain instances of vulgarity, degrading terms and/or hate speech.*

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 367–376, 2023.
https://doi.org/10.1007/978-3-031-28238-6_26

Indeed, such systems would be invaluable for a variety of applications, from automated content classification and moderation, to (potentially malicious) community detection and analysis on social media [9].

To that end, various datasets of human-annotated tweets have been proposed, most often using binary *generic* (e.g., HS/not HS), or multi-label *specific* schemas (e.g., racism/sexism/neither). Unfortunately, due (in great parts) to the lack of clear consensus on these HS annotation schemas [21], gathering enough data to train models that generalize these concepts effectively is difficult. Various approaches have been proposed to palliate these issues: for example, transfer learning has been successfully used in a variety of NLP settings, in particular thanks to the Transformer architecture [33], which allows to leverage large quantities of unannotated text, by fine-tuning pre-trained models such as BERT [7] on tasks for which annotated data is more sparse, such as HS detection [1,17,24,25].

A complementary type of approach is Multi-Task Learning (MTL) [5,18,23], in which one can leverage different tasks and datasets by jointly training a single architecture on multiple objectives at once, sharing all (or parts) of its parameters between them. [32] were the first to showcase how MTL might be used to generalize HS detection models across a variety of datasets, and later on, [16].

Recently, [4] experimented with transferring specific manifestations of hate across HS topics on a varied set of such datasets, showing that MTL could be used to jointly predict both the hatefulness and the topical focus of specific HS instances.

These studies, however, usually consider generic and specific HS datasets as independent (train on one set and test on another) without accounting for common properties shared between both different manifestations of hate, as well as different levels or granularity of annotation. We take here a different perspective and investigate, to our knowledge for the first time, HS detection in a Multi-Level scenario, by answering the following question: *Could instances of generic HS be refined with specific labels, using a model jointly trained on these two levels of annotations?* To this end, we propose:

1. **An easily extensible multitask and multilevel setup designed for HS topic refinement of generic HS instances**, based on the T5 architecture [29], which can be used to generate new specific HS labels (see Fig. 1).
2. **A qualitative and error analyses of the refined labels produced by this approach**, applied to two popular generic HS datasets from the literature.

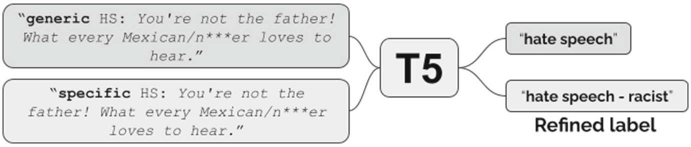
"generic HS: You're not the father! What every Mexican/n***er loves to hear."
"specific HS: You're not the father! What every Mexican/n***er loves to hear."
T5
"hate speech"
"hate speech - racist"
Refined label

Fig. 1. Illustration of our topic refinement approach based on the T5 architecture

2 Datasets

As our main objective is investigating the problem of *transferring knowledge from different datasets, with different annotation granularity and different topical focuses*, we leverage six manually annotated HS corpora from previous studies. We selected these datasets as they are freely available to the research community. Among them, two are generic (`Davidson` [6] and `Founta` [14][2]), and four are specific about four different HS topics: *misogyny* (the Automatic Misogyny Identification (AMI) dataset collection from both `IberEval` [11] and `Evalita` [10]), *misogyny and xenophobia* (the `HatEval` dataset [2]), and *racism* and *sexism* (the `Waseem` dataset [34]). Each of these HS topics targets either gender (sexism and misogyny) and/or ethnicity, religion or race (xenophobia and racism). In Table 1 we summarize the corpora used in this study.

For the purpose of our experiments, we performed some simplifying split and merge operations on their classes, and their associated labels. For all datasets, we considered the respective'negative' (i.e., not HS) classes to be equivalent, and used the unified negative-class label `"nothing"`. In addition, as we are using both generic and specific HS datasets, we merged positives instances from generic datasets in a unified generic class labelled `"HS"`. The Offensive and Abusive instances were removed from these datasets, as these concepts often co-exist with HS, but without a clear distinction [21, 27].

For the specific HS corpora, we made the simplification of merging the classes related to sexism and misogyny into the single unified label `"HS-sexist"`. Similarly, we merged racism and xenophobia into the unified label `"HS-racist"`. These labels are designed with T5's *text-to-text* nature in mind (cf. next section): the generic HS label overlaps part of the specific ones, thus a "misprediction" (or more accurately, a partial prediction in this multi-level scenario setup) at training time should only incur a partial error signal (e.g. predicting only "hate speech" in the *specific HS* task, the correct label being "hate speech - racist", incurs less error than predicting "nothing") (see Table 1).

As noted by a number of previous works [12, 13, 20, 21], these types of merging of classes/labels may not be desirable, as each dataset has its own annotation schema. However, as the goal of this work is to explore the viability of HS topic refinement with currently available datasets, we chose to use this simplified annotation schema, and thus consider this added source of label noise to be part of the experimental setting. Addressing these issues, by expanding or reworking this set of labels will likely be explored in future work.

3 Experiments and Cross-Dataset Evaluation

3.1 Models

We rely primarily on a T5 (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer) architecture [29]. We also experiment with a RoBERTa [19] model, which we use here in an MTL architecture, as a point of comparison for evaluating the performances of these two models across datasets, outside of label refinement (see Sect. 4).

[2] At the moment of collecting the data, from the original dataset (http://ow.ly/BqCf30jqffN) we were able to retrieve only 44,898 tweets. See [20] for more details.

Table 1. General overview of the datasets used in this study.

Dataset	Original classes and sizes (with our T5 labels in bold)	T5 task prefix
Davidson	**HS**: Hate (1,430); **nothing**: Neither (4,160)	generic HS
Founta	**HS**: Hate (1,996); **nothing**: Normal (37,889)	generic HS
Waseem	**HS-racist**: Racism (1,957); **HS-sexist**: Sexism (3,216); **nothing**: None (11,315)	specific HS
HatEval	**HS-racist**: Immigrant (2,617); **HS-sexist**: Women (2,845); **nothing**: Not HS (7,509)	specific HS
Evalita	**HS-sexist**: Misogyny (2,245); **nothing**: Not Misogyny (2,755)	specific HS
IberEval	**HS-sexist**: Misogyny (1,851); **nothing:** Not Misogyny (2,126)	specific HS

T5 proposes a way to unify text generation and classification tasks in NLP, by reframing all of them as *text-to-text* problems. This allows the model to both better leverage its pre-training on large quantities of unsupervised text data, but also greatly simplifies MTL setups. Indeed, instead of requiring additional per-task label-space projection layers, the same fine-tuned weights can be used to perform each desired task, which can be indicated to the model by prepending input instances with some task-specific prefix text. MTL with RoBERTa, on the other hand, is traditionally performed by constructing some kind of projection layer (or layers) for each task in the training set, each with their separate target label-space.

We also experimented with BERT-like models which are domain-adapted for HS and toxic language detection, such as fBERT [30], HateBERT [3], or ToxDectRoBERTa [36], but they yielded similar cross-dataset performances, and so to conserve space, we do not present these results.

3.2 Experiments and Results

For the T5 model, we initially experimented with different prefixes and task labels configurations, but settled on "generic HS:" and "specific HS:", for the generic and specific HS datasets, respectively. In this setup, the model is fine-tuned without task or dataset specific information added, but rather, only the level of HS classification available and/or requested (*is HS present or not?* vs. *which specific topic of HS?*). We refer to this particular configuration using unified prefixes as T5-Refine.

To ascertain how well this configuration is able to learn both of these tasks, we perform a comparative evaluation of performance across datasets alongside other configurations, similar but not intended for topic refinement. As such, we also trained our models with MTL architectures as follows.

RoBERTa-MTL: This is a RoBERTa-base classifier, in the "classic" MTL configuration with one dedicated classification layer per task/dataset (a simple linear projection of the [CLS] token; see [7] or [19] for more details), on the same set of multi-level datasets. (output labels: HS/nothing for Davidson & Founta; HS-sexist/nothing for Evalita & IberEval; HS-racist/HS-sexist/nothing for Waseem & HatEval);

T5-MTL: This is a fine-tuned T5-base model with task-specific prefixes (the names of the corresponding datasets) (output labels: HS/HS-racist/HS-sexist/nothing for all datasets), used here as an intermediate point of comparison between the previous two models (i.e., RoBERTa-MTL and T5-Refine).

Table 2. Comparative evaluation of our models across generic vs. specific HS datasets.

Test sets	**Generic**			**Specific**			**All**		
Model	P	R	$F1$	P	R	$F1$	P	R	$F1$
RoBERTa-MTL	65.14	71.09	67.23	80.84	81.15	77.68	73.49	76.44	72.78
T5-MTL	65.91	64.07	64.83	78.56	75.95	75.79	72.64	70.39	70.66
T5-Refine	63.00	65.06	63.92	79.32	73.59	73.62	71.68	69.60	69.08

We trained T5-Refine on all the training datasets combined (with generic/specific HS task prefixes) while RoBERTa-MTL and T5-MTL models were trained in a multi-task fashion (one head/task prefix per dataset) on the train set of each dataset. Experiments were performed with the AllenNLP [15] and Huggingface Transformers library [35]. Models were trained for a maximum of 12 epochs, with early stopping (patience 4 on validation loss), a batch size of 6, and gradient accumulation of 12. For T5 (RoBERTa) we use the AdaFactor (AdamW) optimizer with a learning rate =1e-3 (1e-5), determined by manual hyperparameter fine-tuning.

Table 2 presents the aggregated averaged results in terms of F-score ($F1$), precision (P), and recall (R) for the three models when tested on: all generic HS test sets (Davidson and Founta), all specific HS (Waseem, HatEval, Evalita, and IberEval) test sets, and all 6 combined test sets.

Table 3 present a more detailed view of these results, in terms of macro F1-scores only (for conciseness): for clarity, the multi-topic datasets (HatEval and Waseem) have been split into single-topic subsets (HatEval sexist/Waseem sexist and HatEval racist/Waseem racist). Then, for each dataset, “HS” and “not HS” correspond to each respective (sub)set’s relevant binarized HS positive and negative classes (HS[-sexist/-racist]/nothing), alongside the Macro Averaged F1-scores. As can be observed, our HS topic refinement model, T5-Refine, despite training under the most difficult configuration (unified label-space and topic-level merged task prefixes), does not showcase significantly degraded cross-dataset performance, compared to the more task dedicated models.

4 Hate Speech Topic Refinement

Using the trained T5-Refine model, we can thus request it to produce specific HS labels for instances of generic HS datasets, here, Davidson and Founta, by simply switching to the specific task prefix at inference time. Table 4 presents a few illustrative examples, of what we consider to be successfully refined labels (examples #1–4), as well as errors (examples #5–9).

To judge the quality of these newly produced labels, we sample 600 instances (200 from each of: [gold = HS | predicted = nothing]; [gold = <any> | predicted = HS - sexist]; [gold = <any> | predicted = HS - racist], where <any> stands for all the possible gold labels) for each of the two generic HS datasets, and compare the predicted labels with the dataset’s gold labels, but also with our own human re-annotation[3]

[3] Performed by a computational scientist and two of the authors of this paper.

Table 3. Detailed evaluation results per-dataset (F1-scores).

Label	**Generic**		**Specifc (Gender)**				**Specifc (Race)**	
	Davidson	Founta	Evalita	IberEval	HatEval sexist	Waseem sexist	HatEval racist	Waseem racist
				RoBERTa-MTL				
HS	87.53	30.64	83.11	88.01	67.74	78.09	62.64	80.53
not HS	96.51	96.85	86.18	92.07	53.24	93.31	34.80	96.36
Macro	**92.02**	**63.75**	**84.65**	**90.04**	**60.49**	**85.70**	48.72	**88.44**
				T5-MTL				
HS	93.82	25.64	64.97	91.41	63.73	68.25	59.90	96.42
not HS	80.45	97.75	71.22	84.97	51.66	92.62	63.96	79.08
Macro	87.13	61.70	68.09	88.19	57.70	80.44	**61.93**	87.75
				T5-Refine				
HS	79.47	24.31	73.65	91.49	42.38	72.07	38.51	74.42
not HS	93.68	97.16	79.65	85.22	63.49	92.25	58.39	94.38
Macro	86.58	60.74	76.65	88.35	52.93	82.16	48.45	84.40

of those same instances. For both datasets, after manually re-annotating with specific HS labels, the final label was assigned according to a majority vote (at least two annotators always ended up agreeing, so no adjudication was necessary).[4] For Founta, the re-annotations process shows that in ~19% of the cases the instances gold-labelled as "HS" belong to a type of abusive language different from the ones investigated in this paper (e.g., offensive language, reporting/denunciation of hate speech, homophobia, islamohobia, etc.), which were re-annotated as out-of-scope. We obtain similar findings for Davidson, though at a larger scale (~57%). After discarding the instances re-annotated as out-of-scope, we obtained a "soft" agreement (coercing HS - racist and HS - sexist labels as equivalent to the generic HS gold label) with the gold labels of 25% for Founta, and 70% for Davidson. In contrast, the refined HS labels exactly match the human re-annotations in 52% of the in-scope instances for Davidson, and in 44% for Founta. While not perfect, overall, the annotators agree almost twice as often with the model-refined labels than with the gold labels for Founta. For Davidson this agreement instead decreases by 18%.

Qualitatively, we believe the main cause of mis-refinement stems from the significant number of merely offensive or abusive instances having been misannotated as hateful in model's training data, when they should be distinct according to datasets' annotation schemes (see last example of Table 4), which is a known problem in HS detection [12,28]. For example, in Davidson, all the instances containing the substring "b*tch" are gold-labelled as HS, regardless of context of use. After re-annotating, 19% were found to be actually HS - sexist, and 78% out-of-scope (more than 70% offensive). Similarly, the substring "f*g" was gold-labelled as HS, with 88% re-annotated as out-of-scope (mostly offensive, with less than 18% found to be homophobic). This is likely the cause of a number of false positive refined labels, which we

[4] Fleiss' kappas for the three-way re-annotation: 0.59 for Davidson and 0.62 Founta.

Table 4. Examples of refined labels obtained from our approach.

#	Dataset	Instance	Gold Label	Refined Label
1	Davidson	*Our people. Now is the time for the Aryan race 2 stand up and say“no more”. Before the mongerls turn the world into a ghetto slum.*	HS	HS-racist
2	Davidson	*RT @USER: It’s unattractive when girls act ghetto*	nothing	HS-sexist
3	Founta	*US attack/siege caused “1/3 #Yemeni #children acutely malnourished”- Says @USER #EndYemenSiege [URL]*	HS	nothing
4	Founta	*@USER @USER Don’t think the world is as ignorant as you.Just because you think a certain law doesn’t exist,doesn’t make it true,you look foolish.*	HS	nothing
5	Founta	*Islamic State says U.S. ’being run by an idiot’ [URL]*	HS	HS-racist
6	Founta	*I just watched a video with a crowd of white ppl shouting n**ga & goin crazy to songs about black men killing each other & it made me so sad*	HS	HS-racist
7	Davidson	*@USER: Lowkey called that faggot a faggot.*	HS	HS-sexist
8	Davidson	*Happppppy Birthdayyyy lol. Niggahs is really 21 in this bitch . [URL]*	HS	HS-sexist
9	Davidson	*#SomethingIGetAlot Are you... asian? black? Hawaiian? gay? retarded? drunk?*	HS	HS-sexist

argue should not be annotated/refined as HS: for example, reporting of HS, either correctly (#3) or incorrectly refined (#5–6), or offensive language (#8).

Due to our limited unified specific HS labels, the model also struggles with instances containing neither sexist or racist HS (example #7), or those containing multiple simultaneous HS topics (#9): in both cases, a potential solution could be to add training datasets which are annotated for more varied and/or multiple targets per instance, such as [8] for example. Despite those issues, the model was still successful at producing a number of coherent refined labels (examples #1–2), or even “corrected” negative labels for some instances (examples #3–4).

5 Conclusion and Perspectives

In this paper, we show that multilevel and multitask learning for the purpose of topic refinement in HS appears to be a viable way to palliate the relative lack of specific HS annotated data. We experimented with a T5 architecture which presents a number of advantages for future improvements: namely, it is significantly easier to extend after-the-fact, as new tasks and datasets may be further fine-tuned on, without having to modify the model’s architecture to accommodate new labels or levels of annotation. This may enable taking into account other topics of HS, such as homophobia, ableism, etc., which may be present in smaller quantities in generic HS datasets, through the use of Few-Shot learning, for example.

Acknowledgments. We would like to thank Walid Younes for helping in the re-annotation process. This work has been carried out in the framework of the STERHEOTYPES project funded by the Compagnia San Paolo ’Challenges for Europe’. The research of Farah Benamara is also partially supported by DesCartes: The National Research Foundation, Prime Minister’s Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

References

1. Alonso, P., Saini, R., Kovács, G.: Hate speech detection using transformer ensembles on the HASOC dataset. In: Karpov, A., Potapova, R. (eds.) SPECOM 2020. LNCS (LNAI), vol. 12335, pp. 13–21. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60276-5_2
2. Basile, V., et al.: SemEval-2019 Task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA, pp. 54–63. Association for Computational Linguistics (Jun 2019). https://doi.org/10.18653/v1/S19-2007, https://aclanthology.org/S19-2007
3. Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: HateBERT: retraining BERT for abusive language detection in english. In: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pp. 17–25. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.woah-1.3, https://aclanthology.org/2021.woah-1.3
4. Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V., Patti, V.: Emotionally informed hate speech detection: a multi-target perspective. Cogn. Comput. **14**(1), 322–352 (2021). https://doi.org/10.1007/s12559-021-09862-5
5. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167. ICML 2008, Association for Computing Machinery, New York, NY, USA (Jul 2008). https://doi.org/10.1145/1390156.1390177, https://doi.org/10.1145/1390156.1390177
6. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11(1), pp. 512–515 (May 2017), https://ojs.aaai.org/index.php/ICWSM/article/view/14955
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
8. ElSherief, M., et al.: Latent hatred: a benchmark for understanding implicit hate speech. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 345–363. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.29, https://aclanthology.org/2021.emnlp-main.29
9. Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., Novak, P.K.: Retweet communities reveal the main sources of hate speech. PLOS ONE **17**(3), e0265602 (2022). https://doi.org/10.1371/journal.pone.0265602, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265602
10. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (AMI). In: Caselli, T., Novielli, N., Patti, V., Rosso, P. (eds.) Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, 12–13 Dec 2018. CEUR Workshop Proceedings, vol. 2263. CEUR-WS.org (2018). http://ceur-ws.org/Vol-2263/paper009.pdf

11. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at IberEval 2018. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., de Albornoz, J.C. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, 18 Sep 2018. CEUR Workshop Proceedings, vol. 2150, pp. 214–228. CEUR-WS.org (2018). http://ceur-ws.org/Vol-2150/overview-AMI.pdf
12. Fortuna, P., Soler, J., Wanner, L.: Toxic, Hateful, Offensive or Abusive? What are we really classifying? An empirical analysis of hate speech datasets. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France, pp. 6786–6794. European Language Resources Association (May 2020). https://aclanthology.org/2020.lrec-1.838
13. Fortuna, P., Soler-Company, J., Wanner, L.: How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Information Processing & Management **58**(3), 102524 (2021). https://doi.org/10.1016/j.ipm.2021.102524, https://www.sciencedirect.com/science/article/pii/S0306457321000339
14. Founta, A.M., et al.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Twelfth International AAAI Conference on Web and Social Media (Jun 2018). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909
15. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform. In: Proceedings of Workshop for NLP Open Source Software (NLP-OSS). Melbourne, Australia, pp. 1–6. Association for Computational Linguistics (Jul 2018). https://doi.org/10.18653/v1/W18-2501, https://aclanthology.org/W18-2501
16. Kapil, P., Ekbal, A.: A deep neural network based multi-task learning approach to hate speech detection. Knowl. Based Syst. **210**, 106458 (2020). https://doi.org/10.1016/j.knosys.2020.106458, https://www.sciencedirect.com/science/article/pii/S0950705120305876
17. Kovács, G., Alonso, P., Saini, R.: Challenges of hate speech detection in social media. SN Comput. Sci. **2**(2), 1–15 (2021). https://doi.org/10.1007/s42979-021-00457-3
18. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, pp. 4487–4496. Association for Computational Linguistics (Jul 2019). https://doi.org/10.18653/v1/P19-1441, https://www.aclweb.org/anthology/P19-1441
19. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs] (Jul 2019)
20. Madukwe, K., Gao, X., Xue, B.. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. pp. 150–161. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.alw-1.18, https://aclanthology.org/2020.alw-1.18
21. Malmasi, S., Zampieri, M.: Challenges in discriminating profanity from hate speech. J. Exp. Theor. Artif. Intell. **30**(2), 187–202 (2018). https://doi.org/10.1080/0952813X.2017.1409284
22. Manne, K.: Down Girl: The Logic of Misogyny. Oxford University Press (2017)
23. Martínez Alonso, H., Plank, B.: When is multitask learning effective? Semantic sequence prediction under varying data conditions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain. vol. 1, Long Papers, pp. 44–53. Association for Computational Linguistics (Apr 2017), https://aclanthology.org/E17-1005

24. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35(17), pp. 14867–14875 (May 2021). https://ojs.aaai.org/index.php/AAAI/article/view/17745
25. Mutanga, R.T., Naicker, N., Olugbara, O.O.: Hate speech detection in twitter using transformer methods. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **11**(9) (2020). https://doi.org/10.14569/IJACSA.2020.0110972, https://thesai.org/Publications/ViewPaper?Volume=11&Issue=9&Code=IJACSA&SerialNo=72
26. Nockleby, J.T.: Hate speech. In: L.W. Levy., K.L. Karst. (eds.), Encyclopedia of the American Constitution, 2nd edn. pp. 1277–1279 (2000)
27. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2021)
28. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Resour. Eval. **55**(2), 477–523 (2020). https://doi.org/10.1007/s10579-020-09502-8
29. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html
30. Sarkar, D., Zampieri, M., Ranasinghe, T., Ororbia, A.: fBERT: a neural transformer for identifying offensive content. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic, pp. 1792–1798. Association for Computational Linguistics (Nov 2021). https://doi.org/10.18653/v1/2021.findings-emnlp.154, https://aclanthology.org/2021.findings-emnlp.154
31. Silva, L., Mondal, M., Correa, D., Benevenuto, F., Weber, I.: Analyzing the targets of hate in online social media. In: Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016, pp. 687–690. AAAI Press (2016). 10th International Conference on Web and Social Media, ICWSM 2016; Conference date: 17–05-2016 Through 20–05-2016
32. Talat, Z., Thorne, J., Bingel, J.: Bridging the Gaps: multi task learning for domain transfer of hate speech detection. In: Golbeck, J. (ed.) Online Harassment. HIS, pp. 29–55. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78583-7_3
33. Vaswani, A., et al.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017). https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
34. 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, pp. 88–93. Association for Computational Linguistics (Jun 2016). https://doi.org/10.18653/v1/N16-2013, https://aclanthology.org/N16-2013
35. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, https://aclanthology.org/2020.emnlp-demos.6
36. Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., Smith, N.: Challenges in automated debiasing for toxic language detection. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 3143–3155. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.274, https://aclanthology.org/2021.eacl-main.274