General Response

As some reviews have questioned whether a legal approach to hate speech is valid, we would like to stress that:

- We are grounding our approach on a legal definition of hate speech. This definition will be narrower than previously used definitions of hate speech (offensiveness etc.), but to the advantage of a binding, democratically discussed and legitimated and ultimately actionable concept.
- We introduce and investigate these points with a case study on an operationalisation of the EU Framework Decision. We also discuss a potential operationalisation beyond this framework and welcome future research on different legal frameworks.

The Meta-Review suggested the following revisions:

The dataset is relevant small and the majority of the data points are labeled as hate speech. You could perhaps extend it to make it more balanced. You could optionally include other languages other than German as suggested by Reviewer 1s3q.

- The dataset is already more "balanced" than other hate speech datasets wrt. to how many illegal posts are contained (see Table 1 which shows that e.g. the GermEval dataset contains less than 1% of such posts). It could be made bigger by “diluting” it with non-hateful comments, but that would rather make it less balanced.
- Including other languages is challenging, as we lack the cultural background to even judge the meaning of many posts. We also do not have access to legal experts for other languages.
- The dataset's intended purpose was the validation of our operationalisation, which we see as the primary contribution of the paper. The creation of a larger, more varied and representative dataset (also covering more languages) is a subject for subsequent work.

Explain better which groups of people/minorities are protected or not under the EU framework (see comment by Reviewer 2NeX).

- We understand that Reviewer 2NeX points out that due to the fact that the EU framework does only protect certain groups, "frequently targeted groups" might not be covered by our operationalisation. This is correct, but based on a democratic decision on the respective legal framework. We show that our approach could be generalised and adapted by adding other groups (Sec. 6). But we also have to emphasise here that the purpose of our operationalisation is not hate speech in general or detecting any desirable target group, but an operationalisation of a legal definition of hate speech. Such a legal definition will always be narrower, but has benefits in terms of binding force and actionability.
- As we stated in our response to Reviewer 2NeX, we agree on the importance of empirical social science research that investigates protection and discrimination beyond legal obligation. On the other hand, there is the aspect of balancing this interest against a potential overblocking of protected free speech. This is why we limit our discussion here to the legal and NLP level.
- Otherwise, we clearly indicate under Section 2.1 that only groups "defined by reference to race, colour, religion, descent, or national or ethnic origin" are subject to the EU Framework. We also illustrate these groups in Example 1 (nationality: French people; race: black people; religion: muslims; ethnic origin: Sinti and Roma) and explain that groups like “refugees” do not fall under this definition as they are not clearly distinguishable by one of these criteria (e.g. consisting of several nationalities etc.).

Include simple baselines such as random and majority class.
We have included a random baseline (F=0.07). The majority class is “legal” in our case, so P=R=F=0. We thus decided against including the majority baseline.

Choice of legal framework:
Another issue that was questioned by most reviewers is the choice of legal framework. We are well aware that other legal standards exist and use the study of the EU Framework Decision as a starting point to discuss challenges of a legal operationalisation. It is important to note here that a legal definition of “hate speech” will always be narrow and cannot cover all kinds of offensive or hateful posts. This is due to the necessity to define clear-cut and precise obligations as they are binding and actionable. The NLP tasks needed here are thus different from those previously discussed for “hate speech detection”. An adaptation to different legal systems will be part of that task.

In the following, we address the individual reviews.

Review URwF

“widely applicable legal minimum standard” Widely applicable where? And why?
This is a summarising sentence in our abstract that we further specify in Section 1 as the application depends on the location of the platform provider or the hosting server. As we elaborate, the new Digital Services Act would also lead to a direct application to US platform providers, for instance.

“Overall, our method also provides for better explainability” This is a very strong claim.
In comparison to a holistic end-to-end approach, our approach is better explainable at the label-level. We clarified this in the paper.

“harming certain groups of people” And their democratic participation disproportionately
Exactly. We will add this in the final draft.

“rendering the regulation of hate speech an instrument against discrimination” Another big claim, back this up with a reference.
This is an introductory sentence indicating that regulating hate speech against a group may be an instrument against discrimination. For instance, already the title of “The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)” that targets inter alia hate speech illustrates this intention.

“(Fortuna and Nunes, 2018; Poletto et al., 2020; Schmidt and Wiegand, 2017)” Fortuna and Nunes 2021 is a great reference here.
Thank you, citation added.

“Ultimately, a subset of especially hateful content can be considered punishable by 041 law and thus would not fall under freedom of expression” Do the authors mean in a legal sense?
Yes.

“Article 4 of the ‘International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)” Which body?
This is a UN convention adopted by the General Assembly as indicated in the footnote.

“the offender or the material hosted is located within the EU” Or the target right or did i misunderstand when i read it?
No, the location of the target is not relevant here. It is just as we wrote: it is only applicable if "the offender or the material hosted is located within the EU" (Art. 9(2) of the Framework Decision).

"Hence, the EU Framework Decision not only provides a minimum standard ... but it also ... would trigger the broadest regulatory obligations" Se it's not a "minimum" if it's the broadest globally. Considering China, Turkey and Iran would also be relevant here and why you're not selecting to be based in their content moderation legislation

It is just as we write it: it is a minimum standard in terms of material scope ("What is considered illegal hate speech?"). In terms of territorial scope ("Where does the standard apply?") we write explicitly (the relevant section was omitted from the quote): “but it is also the regime that -- in connection with the new Digital Services Act -- would trigger the broadest regulatory obligations for large platform providers inside and outside the EU.” The new Digital Services Act, just as we write in the paragraph above this sentence, would apply to US platform providers even if they are not situated in the EU, but provide their services to EU citizens. This is not the case for China, Turkey, Iran where Google, Youtube etc. are all blocked.

"Based on all these considerations, the Framework Decision's minimum standard may stand in for a general legal approach to hate speech and serve as the basis" This is a subjective decision that the authors make, so this ought to be written as an explicit choice.

This is the concluding sentence of a full page of discussion under the heading "competing legal standards". We stand by our analysis.

"Target Group" Section lacks a discussion of power. The examples show that there is no consideration of social hegemonies in law, but for hate speech we consider power imbalances b/c we're interested protecting marginalised groups.

As we described above, the intention of the EU Framework Decision is to protect marginalized groups.

"Agreement in the control condition (holistic annotation) is very low, which is in line with previous findings of low IAA for hate speech annotations (Ross et al., 2016).” Yes but much higher than this.

Ross et al. reported alpha values between .18 and .29. The annotations using only the legal text in our study produced alpha .28.

"Source" Why not add the Ross et al. Dataset?

"Composition of the dataset by source" If the authors argue for the ease for lay people to annotate then why only create such a small dataset?

The primary goal of our work is to evaluate the feasibility of the presented operationalisation on a variety of different sources, the resulting dataset is a byproduct of this evaluation and not the main contribution.

The dataset by Ross et al. (2016) has a large topical overlap with our data as it also contains tweets addressing the refugee crises. While adding it may give a more complete picture, we don’t expect it to be dissimilar enough from the data we have to drastically alter the results.

"Punishable Hate Speech Dataset" It would be helpful to have a table of data split sizes and class distributions

We added class distributions in the appendix and reference them in the paper.

"volunteers" Did these volunteers also label data?

No, the volunteers only provided examples from the “Made Up” portion of the dataset. We adjusted the paper to clarify.

"but we have no way of controlling for topic biases that might have been introduced via this process.” Why not? 150 should be possible to manually examine
Upon inspecting the Made Up portion of the dataset, we noticed that it contains mostly generalised attributions of character and behaviour, but contains only few references to recent events related to the groups.

“We performed a manual search of Twitter, comment sections of online newsrooms, law forums, court databases as well as news articles 316 resulting in 80 instances.” This is surprisingly low. What were your search terms?

We searched for cases in which hateful social media posts were brought to court, hence the limited number of entries in this category. We clarify this in the paper.

“fully-qualified lawyer” An expert in the area?

Yes, the “fully-qualified lawyer” has the qualification to work as a prosecutor or criminal judge.

“law enforcement” Don’t include cops.

We are not sure what is meant with that comment.

“The model 401 is trained for 20 epochs using a batch size of 16 402 and NLL loss.” Why so many epochs? Most commonly i see about 4-5 epochs of fine-Tuning with bert-based modelling.

We followed recommendations by Mosbach et al. (2020). However, on average models stopped improving after ~6-7 epochs per fold. As this is not the main concern of the paper, we omitted model behaviour during training from the discussion of our results.

“bias- 403 corrected Adam” Citation needed.

Thank you, citation added.

“F1 of .39 411 (P .69; R .28),” For a binary task this is exceedingly low. Was this really the best model? How does a majority class or random choice baseline compare?

These are the numbers for the fine-tuned model, trained on the punishable label. As written later in the same section, using the submodels with the decision tree slightly improves results. We added a random classifier. We do not add a majority class baseline, as this would yield precision = recall = f1 = 0 for the “punishable” label.

“the task” Which task?

The task to classify a post holistically as “punishable” (end-to-end). We clarified this in the paper.

“In contrast to regular hate speech detection ... the model needs to learn ... whether (i) the hate is directed (ii) the object is a group, (iii) the group is protected” The last is different from hate speech & maybe 2) by the list the law cares about is much smaller than what hate speech scholars care about.

This is true. Since different stakeholders are interested in different sets of groups, we propose to create a general group identifier capable of detecting all kinds of groups, independent of their legal status. Such a module could be used in legal hate speech detection as well as classical hate speech detection, by defining a subset of relevant groups for each task.

“We also trained a separate model to predict the target group and targeting conduct subtasks and then derive the punishability accordingly. Table 4 shows the results.” Rework these into a single sentence.

We made the phrasing clearer.

“predict the target group” How is this different from NER and only mapping to a small set of entities?
As we explain in section 6.1, the two tasks are somewhat similar but not all group identifiers are named entities, e.g. "women".

"challenging task" And perhaps impossible?

We believe that with the right operationalisation, even tasks that require high precision and a nuanced understanding of human language can be solved. By sharing the task with the academic community, we hope to spark interest in the legal approach to hate speech, resulting in improved models that eventually may achieve performance comparable to human experts.

“Generalizing beyond EU Law” Drop this section maybe, as any given legislation that is operationalised must be subject to the conditions of the legal text

Yes, but the point of our paper is to investigate a possible approach for operationalisation that could be adapted to different legal texts. We see the question of “generalizing” as an open research question.

“In this way, our approach 472 would also generalize beyond EU law” How would this be beyond eu law? This would assume a similar legal construction of illegal speech.

Yes, but as long as “target group” and “targeting conduct” are key elements, it would generalise. As this is also the approach taken by the UN convention ICERD (see Sect. 1), this is very likely.

"Slovene (Fišer et al, 536 2017)" This uses a legally informed framework.

Thanks for pointing this out. We corrected this for the final draft and contrasted their approach with ours.

"Predicting the legal status of a comment might infringe on the fundamental right of ‘free speech’...." Why not just refer to that what the manuscript tries to identify is not legal in the eu?

We do not understand that comment.

Review 1s3q

Summary Of Weaknesses:

The study is limited to the German language and the German legislation on hate speech. While there is a discussion section on the generalizability of the approach, it is just a theoretical discussion with no experimental basis.

We acknowledge that there is potential in further generalisation, however, as stated in the paper the EU Framework Decision has is already large in regulatory scope and relevance. The study is conducted in German, but our scheme is not specific to the German language and repeating annotation efforts in different languages is an extension for future work.

Other details are missing too, e.g., what is the web search that yielded 80 posts?

We searched for cases in which hateful social media posts were brought to court, hence the limited number of entries in this category. We clarified this in the paper.

I found particularly surprising that the only established dataset employed in this study is GermEval, which is not a hate speech dataset. Indeed, only a very small percentage of hate speech is found there, while there exist several available datasets annotated with hate speech in several languages (notably, HatEval from SemEval 2019, but see https://hatespeechdata.com/).
You are right that the GermEval 2019 dataset does not predominantly contain hate speech. We sampled from it to include adversarial examples containing similar language as actually punishable statements.

HateEval 2019 did not include German data.

Re-annotating an existing dataset of hate speech may be time-consuming, but it would help benchmarking the benefits of the proposed methods, and generate an incredibly useful resource along the way.

We have already done that (partially) with re-annotating the GermEval dataset. In this paper, we focus on analysing the legal side, translating that into NLP tasks, and developing the annotation scheme. Extending that to other datasets is future work.

Review 2NeX

Summary Of Weaknesses:

I fully appreciate the reasons (explained in Section 1) for using the EU legal framework in the work of the paper, but I'm not entirely convinced that using this framework (and with German-specific refinements) "may stand in for a general legal approach to hate speech and serve as the basis of further studies" (lines 112-114). In other words, the legal provisions that the paper uses to establish the ground truth seem too EU-specific (and at times even too German-specific, lines 211-213) and it is not entirely clear how they could be used in other parts of the world (Section 6 is rather speculative and not particularly informative), even if EU requirements may "trigger the broadest regulatory obligations for large platform providers inside and outside the EU" (lines 98-100). I would recommend re-branding the paper as a case study of how EU and German legal provisions can be used to establish the ground truth for punishable hate speech detection in Germany, with experiments that investigate the extent to which this detection can be automated in that particular country. This case study may then help others carry out similar work in other countries with very different legal frameworks, but I would avoid implying that the particular EU framework (with German-specific refinements) should be the common denominator for the entire world.

Thanks a lot for your suggestions. We also believe that the paper may serve as a starting point for similar research. Our key point here is to suggest and introduce this as a new task. While it is true that the specific requirements vary across legal systems, the NLP tasks of "target group" and "targeting conduct" detection also correspond to the ICERD approach, thus making it very probable that our approach would indeed be a good fit internationally. Beyond this general point of suggesting future NLP tasks, we indeed tried to investigate more detailedly an operationalisation for the EU framework respectively. And we are of course open and interested in further research criticizing our work or suggesting a different approach.

Although grounding hate speech detection decisions on a legal framework makes the ground truth less ad hoc and less subjective, it also substantially reduces the types of hate speech to be blocked. As the paper points out, the EU framework does not protect, for example, targeting "foreigners" or "refugees" in general (lines 187-188), or targeting "women" in general (line 347), or people in terms of sexual orientation. In that sense, the legal approach to hate speech detection advocated by this paper will probably not protect several groups frequently targeted by hate speech. The paper includes two additional target groups not protected by the EU framework ("women", "LGBTQ+") in its experiments, and points out (in the Ethical Considerations section) that the presented approach allows adding other groups. But still, the legal approach of the paper may be a good excuse for social networks and other on-line fora to do less to protect frequently targeted
groups they are not legally required to protect. I would recommend at least including some more
discussion on this point, and making it clearer (right from the introduction) that the paper is about
detecting legally punishable hate speech, not hate speech in general.

We also encourage and see the importance of social science research that investigates
protection and discrimination beyond legal obligation. On the other hand, there is the aspect
of balancing this interest against a potential overblocking of protected free speech. This is
why we limit our discussion here to the legal and NLP level.

Regarding the difference between detecting punishable posts and hate speech, we believe
that title, abstract and introduction (as well as the body of the paper) make that distinction
very clear.

The majority (750 out of 1,000) of the posts of the dataset are hate speech (legally punishable, or
not legally punishable, but still hate speech). Experimenting on this dataset cannot really tell us
how well systems would be able to detect punishable hate speech (the goal of the paper) in a real
life social network or similar platform, where most posts would probably not be hate speech. In
other words, the ratio of non-hate speech posts is not realistic.

Based on the German Network Enforcement Act, social platform providers are obliged to
report potentially punishable postings to the Federal Criminal Police. These reported
postings are - to a large extent - somehow offensive. This is a very practical and direct
application of our research and we are, in fact, involved in developing a detection system
specifically for this use case. Our paper also implies that the task is performed on a dataset
with already somehow offensive postings, not on a whole social media corpus.

The dataset in the supplementary material files contains only 163 posts, whereas Table 1 indicates
that the dataset contains 1,000 posts.

This is correct. As we write in the paper in the “Ethical
considerations” section under “Release
of the Data”: due to restrictions by the General Data Protection Regulation we were not
allowed to (re-)publish the whole dataset, only the 163 made-up posts.

Please clarify what “with the same topic” means in lines 326-327.

With the same topics as the other posts in the dataset. We have clarified that in the paper.

Please clarify how the “Punishable” agreement scores were obtained in Table 3. Were the annotators
also asked to directly label posts as punishable or not? Or is the “Punishable” decision obtained
automatically from the target group and targeting conduct decisions (lines 274-275 and 353-355)? If
the latter, the poor agreement on "Punishable" seems to be implying that the proposed guidelines
(decision tree) is not producing sufficient agreement on the end decision.

The punishable decision was automatically derived using the decision tree. If the final
decision should be derived automatically at all is currently discussed within the research
community. Our model (identifying group and conduct) opens the possibility to leave the
final decision to a human, which is legally necessary e.g. in Germany. We also strongly
believe that this has advantages over deploying a black-box AI.

Line 539: It is not true that “Interdisciplinary work combining NLP with a legal perspective has
mostly focused on predicting the outcome of legal decisions” (lines 539-541). See, for example, the
proceedings of the NLLP workshop series (https://sites.google.com/view/nllp/home).
Thank you for pointing this out, it was phrased in an ambiguous way. Of course there is work on legal NLP in general. We have rephrased this.