A Legal Approach to Hate Speech –
Operationalizing the EU’s Legal Framework against the Expression of
Hatred as an NLP Task

Anonymous ACL submission

Abstract
We propose a ‘legal approach’ to hate speech detection by operationalization of the decision as to whether a post is subject to criminal law into an NLP task. Comparing existing regulatory regimes for hate speech, we base our investigation on the European Union’s framework as it provides a widely applicable legal minimum standard. Accurately judging whether a post is punishable or not usually requires legal training. We show that, by breaking the legal assessment down into a series of simpler sub-decisions, even laypersons can annotate consistently. Based on a newly annotated dataset, our experiments show that directly learning an automated model of punishable content is challenging. However, learning the two sub-tasks of ‘target group’ and ‘targeting conduct’ instead of an end-to-end approach to punishability yields better results. Overall, our method also provides decisions that are more transparent than those of end-to-end models, which is a crucial point in legal decision-making.

1 Introduction
Social media provides the platform for the expression of opinions along with their widespread dissemination. Unrestricted freedom of expression, however, bears the risk of harming certain groups of people - rendering the regulation of hate speech a potential instrument against discrimination. To do so at scale, automated detection systems are required to aid the moderation process. While research on hate speech detection is well-established, defining ‘hate speech’ remains challenging. Datasets encode all kinds of (partly incompatible) notions of hatefulness or offensiveness (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Poletto et al., 2020; Fortuna et al., 2021) that make it difficult to decide which postings would justify restricting freedom of speech through deletion. Ultimately, a subset of especially hateful content can be considered punishable by law and thus would not fall under the legal right to freedom of expression. As there exist competing legal standards for the regulation of hateful expressions, the selection requires discussion.

Competing Legal Standards
On the international level, Article 4 of the ‘International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)’¹ binds the signatory states to punish incitement to racial discrimination against any race or group of persons of another colour or ethnic origin by their respective national law. However, the convention does not cover discrimination based on religion and is limited in its legal effect, as various states have made reservations. This is especially the case for the U.S., where the expression of hatred toward any group is constitutionally widely protected by the Free Speech Clause of the First Amendment (Fisch, 2002). Consequently, as US law does not provide for any legal provision prohibiting hate speech as an act of speech, it cannot serve as a base for a detection system.

In Europe, however, the prevention of discrimination against and segregation of a target group (thereby ensuring the members’ acceptance as equal in a society) is considered such an important prerequisite for democracy that it may justify the restriction of free speech. The Council of Europe has set up an additional protocol to the ‘Convention on Cybercrime’, concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems.² However, the Protocol has not been ratified or even signed by all Member States of the Council of Europe and is subject to several reservations.³

¹General Assembly resolution 2106 (XX) of 21 Dec 1965.
²ETS No. 189, 28.01.2003.
³Bulgaria, Hungary, Ireland, the Russian Federation and the U.K., for instance, did not sign the Protocol. Countries like Austria, Belgium, Italy, Sweden, Switzerland and Turkey signed, but did not (yet) ratify it.
Figure 1: Scope of the EU Framework’s legal standard. It defines a common core of punishable offenses.

Legally and practically more relevant is the following instrument: the European Union (EU) has, after long debate, set up a common regime with a Framework Decision that fully binds all of its Member States to make incitement to hatred or violence a punishable criminal offense. The framework also affects U.S. social-media platforms as long as the offender or the material hosted is located within the EU. Its importance has also been emphasized by the ‘EU Code of conduct on countering illegal hate speech online’ that the EU Commission agreed with IT companies like Facebook, Twitter, and Youtube. Furthermore, the EU’s new proposal of a Digital Services Act aims to create new obligations for large online platforms regarding illegal content. The regulation would not only be directly applicable in all EU Member States, but also apply to providers established outside the EU if they provide their services to recipients in the Union. Hence, the EU Framework Decision not only provides a minimum standard for handling hate speech by criminal law, 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.

As Figure 1 shows, each Member State may still go beyond the framework’s minimum requirements and define higher standards. Germany, for instance, provides for a broader definition of the possible protected target group by including ‘sections of the population’, e.g. refugees otherwise not being covered as they cannot clearly be distinguished by race, ethnic, or national origin. However, the Framework Decision allows member states to make the incrimination depend on additional requirements.

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 of our further studies.

Contributions In this paper, we translate the legal framework as defined in the EU Framework Decision 2008/913/JHA into a series of binary decisions. We show that the resulting annotation scheme can be used by laypeople to reliably produce a legal evaluation of posts that is comparable to those of legal experts, making dataset generation for this task feasible. Based on the resulting dataset, we experiment with directly learning an automated model of punishable content. The discouraging results of the end-to-end approach and ethical considerations lead us to proposing two sub-tasks instead: ‘target group’ and ‘targeting conduct’ detection. We show that the sub-tasks can be more reliably learned and also provide for better explainability and higher transparency, which is a crucial point in legal decision-making. We make our dataset and models publicly available to foster future research in that direction.

2 Operationalizing Legal Assessment

We begin our investigation by operationalizing the relevant part of the Framework Decision (FD) into a sequence of binary decisions that can be reliably annotated (see Figure 2 for the final decision tree). In a way, we are translating the plain text of the legal definition into an actionable algorithm.

Article 1(1) FD states that the following intentional conduct is punishable:

(a) publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin;
(b) the commission of an act referred to in point (a) by public dissemination or distribution of tracts, pictures or other material;

The punishable conduct addressed in paragraph (a) refers to the oral expression of hatred, while paragraph (b) broadens the scope to public dissemination or distribution of tracts, pictures or other
As shown in Figure 2, Art.1(1)(a) refers to the following potential targets: a group of persons or a member of such a group defined by reference to race, colour, religion, descent, or national or ethnic origin (see Example 1).

2.1 Target Group

As shown in Figure 2, Art.1(1)(a) refers to the following potential targets: a group of persons or a member of such a group defined by reference to race, colour, religion, descent, or national or ethnic origin (see Example 1).

Example 1: Distinguishable groups.

The scope also covers individuals in case they are targeted as a member of an aforementioned group, as illustrated in Example 2.

Example 2: Individuals as members of a group.

‘Race and ‘colour’ are discriminating grounds that can be understood interchangeable. ‘Religion’ refers broadly to persons defined by reference to their religious convictions or beliefs (Recital (8)).

Recital (7) clarifies that ‘descent’ points to persons or groups of persons who descend from persons who could be identified by characteristics like race or colour. It is not necessary that all these characteristics still be present in the respective persons. Hence, the descendants would be protected regardless, e.g., descendants of people of Jewish faith even in cases where they do not share this faith anymore. ‘National origin’ or ‘ethnic origin’ are both distinguishing grounds that require reference to a specific nationality or ethnic group. Statements that refer to ‘foreigners’ or ‘refugees’ without further specification are not covered, as these references are considered too general.

2.2 Targeting Conduct

With respect to the target group as a victim, Art.1(1)(a) requires at least one of the following acts to be committed by the potential offender: (i) inciting hatred, or (ii) inciting violence.

Regarding the definition and understanding of these acts, freedom of expression needs to be taken into consideration through Art.7(1), which ultimately refers to Article 11(1) of the EU Charter of Human Rights. By preventing segregation, the intent is to protect minorities from being deprived of their human dignity as equal members of society. Punishing expressions is only justified in the respective cases if the legal interest in preventing discrimination outweighs the right to free speech – which is likewise a precondition for democracy.

Within these limits, the Framework Decision itself does not provide for a more detailed definition of ‘inciting hatred’ and ‘inciting violence’, but entrusts the Member States with elaborating the interpretation in national case law. For our annotation guidelines, we draw here from German case law, which provides for long-standing settled decision-making practice for these terms.

Inciting ‘Inciting’ has been defined as ‘conduct influencing emotions and intellect of others’. A key element of the definition is the clear intent to influence others. To outweigh freedom of expression, the conduct has to go beyond mere rejection or contempt and means more than merely endorsing.

---

Example 1:

- French people are frog eaters. (nationality)
- Black people = slaves! (race)
- Muslims are all terrorists! (religion)
- Sinti and Roma = awful parasites! (ethnic origin)

Example 2:

- you fucking muslim should leave our country!
- This dirty american bitch, typical american, lying son-of-a-bitch, out of our country!
Hatred The Framework Decision limits, in Recital (8), the notion of ‘hatred’ as such based on race, colour, religion, descent, or national or ethnic origin. In other words, ‘hatred’ expressed against a specific group, but which is unrelated to the belonging to this group, is not covered. We draw here again on German case law, where the act of incitement to hatred needs to be ‘objectively capable and subjectively intended to create or intensify an emotionally enhanced, hostile attitude (towards the respective group)’. Example 3 illustrates comments that fit these criteria.

- Muslims are deceitful parasites enjoying life thanks to hard working german citizens!!
- Bring back the slaves! #niggerarenohumans

Example 3: Comments inciting hatred.

Violence While ‘hatred’ refers to the creation of a hostile attitude, inciting ‘violence’ shall ‘give rise to the determination of others to commit violence’. Violent measures do not just comprise assault, but also violent expulsion or pogroms. Example 4 illustrates comments inciting violence.

- U.S. citizens should be hunt down and deported!
- Burn all Muslims in their mosques!

Example 4: Comments inciting violence.

2.3 Optional Qualifiers

Art.1(2), however, grants one exception to the minimum standard, as seen in Figure 1. Member States may predicate the offense on the additional requirements of the disturbance of public order or threatening, abusive or insulting conduct. In other words, a Member State may stipulate that the conduct is only punishable if it also leads to a disturbance of public order, or if the conduct is also threatening, abusive, or insulting. As these additional requirements are only required by a few Member States, we do not operationalize them.

3 Feasibility Study

To test our decision tree annotation scheme, we first perform a feasibility study, where we assess the quality of annotations produced by our annotation scheme against direct annotation. We also assess the reliability of an assessment by legal experts to establish an upper bound for this task.

Setup We asked public prosecutors from one of the two cybercrime prosecution centers in Germany to provide the ground truth for punishability based on §130 of the German Criminal Code – which implements the EU Framework. As prosecutors would be obliged to open an investigation for each punishable post, we provided a set of 156 ‘made-up’ hate speech posts in German. These were never openly published and are thus not punishable. The prosecutors did not use our decision tree, but decided based on their legal training and expertise. As a control condition, we asked layperson annotators to perform a direct annotation. Annotators were provided with the legal text of §130 and decided whether a post was punishable using their understanding of the legal code. Finally, we asked layperson annotators to follow our multi-label annotation scheme, from which we can automatically derive whether a post is punishable or not, depending on the combination of our labels.

Results Figure 3 shows the inter-annotator agreement (IAA) per setup in the feasibility study. 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). However, the high kappa between expert prosecutors shows that sufficient legal expertise enables consistent judgements.

Using our annotation scheme increases consistency between annotators and agreement with experts. Thus, based on the success of the feasibility study, we adapt our annotation scheme to fit the EU framework and produced the full dataset, described in the next section.

\[^9\text{BGHSt 21, 371 (372); BGHSt 46, 212 (217)}\]
\[^8\text{BGH 3.4.2008 – 3 StR 394/07}\]

Figure 3: Cohen’s Kappa for the Punishable label for different annotation schemes in the feasibility study.

\[^{10}\text{As §130 of the German Criminal Code is a transposition of the minimum standard set by the EU Framework Decision (see Section 2), the results obtained in this way should be generalizable to EU law.}\]

\[^{11}\text{The made-up posts are comparable in nature to realistic posts. See next Section 4 for a more detailed description.}\]
### 4 Punishable Hate Speech Dataset

In this section, we describe how the full dataset was created. All posts in the dataset are in German.

#### 4.1 Data Sources

Social-media posts were sampled and requested from a multitude of sources with the primary goal of obtaining sufficient examples of punishable hate speech. Table 1 provides an overview of the final composition of the dataset.

<table>
<thead>
<tr>
<th>Source</th>
<th>#</th>
<th>% Punishable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Made-up</td>
<td>157</td>
<td>13.5</td>
</tr>
<tr>
<td>Web search</td>
<td>80</td>
<td>6.2</td>
</tr>
<tr>
<td>Anti hate speech initiatives</td>
<td>88</td>
<td>10.2</td>
</tr>
<tr>
<td>GermEval2019 (abuse, insult)</td>
<td>425</td>
<td>0.9</td>
</tr>
<tr>
<td>GermEval2019 (other)</td>
<td>250</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1: Composition of the dataset by source.

#### 4.2 Annotation Scheme & Process

The full dataset was annotated by two paid laypersons, who were provided with an annotation manual based on the operationalisation described in Section 2 with further explanations, instructions, and examples. To measure annotation quality, a subset of 101 posts was annotated by a fully-qualified lawyer using the same annotation scheme.

We annotate whether a group of persons or a group member was mentioned in a post and, if so, whether the group is distinguishable by any reference to race, descent, or national or ethnic origin. In case a group is explicitly mentioned, we also annotate the surface form used in the comment. We created a short list of frequently attacked groups and asked annotators to choose one of these or ‘other’ (‘Group Category’ annotation). We include groups not covered by the EU framework like women or refugees, as they might be relevant for future detection tasks regarding other legal regimes.

We also annotate possible targeting conduct described in Section 2.2, i.e. inciting hatred and inciting violence. If a relevant group is targeted and any targeting conduct is present, a post is considered punishable. Table 2 provides some examples of the resulting annotation.

#### 4.3 Analysis

We analyze the IAA among laypersons and the expert annotator in terms of Cohen’s Kappa as shown in Table 3. Aggregated results on target group and targeting conduct are quite reliable (kappa between .52 and .70), while kappa for the punishable label is rather low (.33 to .43). People agree on the facts (group, conduct), but disagree on the interpretation.

Table 1 displays the distribution of punishable posts. It is noteworthy that in the GermEval2019 data a surprisingly low proportion (under 1%) of abusive or insulting comments are actually punishable under EU law. This highlights that hate speech detection and detecting illegal content are fundamentally different tasks.

---

12 Volunteers did not participate in subsequent annotation efforts.
Table 2: Example annotations from the resulting dataset. Surface form referring to groups is underlined.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Group</th>
<th>Target Group</th>
<th>Targeting Conduct</th>
</tr>
</thead>
<tbody>
<tr>
<td>So dummi wie Gutmenschen hält sind! Stupid as do-gooders are!</td>
<td></td>
<td>Other</td>
<td>✓ ✓</td>
</tr>
<tr>
<td>Die linken Hünersöhne! Ballert die alle ab!</td>
<td></td>
<td>Left-Wing</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Es brennen noch zu wenige Synagogen. Still too few synagogues on fire.</td>
<td></td>
<td>Jews</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>Muslims sind alles Vergewaltiger! Schützt deutsche Frauen! Muslims are all rapists! Protect our German women!</td>
<td></td>
<td>Muslims</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Table 3: Inter-annotator agreement (Cohen’s Kappa) between laypersons and domain expert.

Figure 4 shows the confusion matrix between the two layperson annotators regarding the group annotation from our short list (subset of 392 posts having a group mention). The largest target group is foreigners/migrants, which is not explicitly protected under EU law. Differences between annotators mainly arise due to the ‘None’ and ‘Other’ categories, while the largest disagreement is within closely related categories like ‘left-wing/green party’ and ‘other politicians’.

Each group is referred to by a wide variety of different surface forms. Table 5 lists selected examples of surface forms in the dataset. The median number of surface forms per group is 20 (min=3, max=135), showing that automatic detection will have to deal with a high variance. The ‘other’ category contains a wide range of different types of groups like law enforcement, vegans, jobless, football clubs, or media outlets that we might consider as distinct groups in a revised annotation scheme.

Figure 4: Confusion matrix of non-expert annotators.

5 Automated Detection

To study the extent to which our annotated data can serve as a basis for automated detection, we train a baseline classifier that takes a post as input and estimates whether the post is punishable. For model training, differences between annotators were adjudicated by a legal expert. The agreement numbers reported in Section 4 are thus not applicable for the following experiment.

Setup Fine-tuned BERT (Devlin et al., 2018) models have proven to be strong baselines for various NLP tasks, so we follow this practice using GBERT-base (Chan et al., 2020). The model is trained for 20 epochs using a batch size of 16 and NLL loss. For optimization, we choose bias-corrected Adam (Kingma and Ba, 2015), with a learning rate of $2e^{-5}$. The learning rate is linearly increased up to its peak during the first 10% of training.

13For the implementation, we use HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019).
ing and then linearly decreased. These choices follow the recommendations of Mosbach et al. (2020) for increasing training stability when fine-tuning BERT. For evaluation, we perform a stratified 10-fold CV. We also include a random classification baseline for the punishable label. We set \( p = 0.5 \) and thus yield a recall of 0.5 and a precision that corresponds to the overall ratio of punishable posts in the dataset, i.e. 0.04.

**Results** The fine tuned model achieves an average \( F_1 \) of .39 (P .69; R .28), which shows that legal hate speech classification is complex and not easily solvable by baseline models. The mere presence of language inciting hate or violence is not a sufficient signal, but the model needs to learn in addition whether (i) the hate is directed against an object, (ii) the object is a group, (iii) the group is protected under the given law.

This is exemplified looking at some misclassifications. In Example 6.1, no group is targeted; yet, the model classified the post as punishable. In Example 6.2, the model erroneously judged the post to be punishable, even though the group is not protected under the EU framework. Implicit or metaphorical references to a group were not detected by the model (e.g. Example 6.3).

**Example 6**: Cases misclassified by detection model

We also trained separate models for the prediction of target groups and targeting conduct. The last row in Table 4 shows results of applying those two models to derive punishability. In terms of \( F_1 \) score, the subtask approach is comparable to the direct approach of estimating punishability (.42 vs .39). Looking at the performance of the subtask models, we see mixed results. While the **Group of persons**, **Distinguishable by nationality, race, religion, ethnicity and Inciting violence** models produce good results (.71 – .83), the models for **Individual as member of group** and **Inciting hatred** failed making accurate predictions (.00 – .11). Both are rare in the dataset (14 positive cases each), making them difficult labels to learn.

### 6 Generalizing beyond EU Law

So far, we have presented a case study of operationalizing a specific legal standard (i.e. the EU Framework Decision). The underlying methodology can be generalized in a straightforward way. Instead of directly predicting whether a post is punishable or not, we divide the problem into two subtasks, (i) group detection and (ii) conduct detection, each of which can be tackled separately, depending on the applicable legal regime. This approach offers higher explainability of model decisions, an aspect that is crucial for legal decision-making.

#### 6.1 Group Detection

If we were able reliably to detect all groups referred to in a comment, we could take the list of protected groups and only consider those relevant under a certain legal standard. In this way, our approach would also generalize beyond EU law.

However, groups are often referenced by a variety of different surface forms, some of which are only metaphorically related to the group (e.g. ‘Goldstücke’; Engl. ‘gold pieces’ for people of color, see Table 5). Consequently, we cannot use Named Entity Recognition (Ritter et al., 2011) for group detection, as, e.g. ‘women’ are a common target group, but not a named entity. A better fit seems Entity Linking (Derczynski et al., 2015), which would (depending on the underlying knowledge base) find explicitly mentioned groups. However, groups can also be implicitly mentioned (7.1) or as part of a co-reference chain (7.2).

### Table 4: Overview of prediction results. See Table 6 in the Appendix for label distributions.

<table>
<thead>
<tr>
<th>Category</th>
<th>Surface Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>People of Color</td>
<td>#negersindkainemenschen, afle, himbo, dunkler teint, nafirs, neger, negroude goldstücke, schwarze, sklaven</td>
</tr>
<tr>
<td>Jews</td>
<td>dreckiges judenpack, judenschwein, zentralrat der juden, jüdischer zombie, rattenvolk, zionisten</td>
</tr>
<tr>
<td>Muslims</td>
<td>#islamisierung, #muslime, islamlobbys, bärtigen kinder-schänder, dtib imams, dreckige kopftuchmadchen, gotteskrieger, isis-schlampen, muslim-ungeziefer, scharia</td>
</tr>
<tr>
<td>Nationality/Origin</td>
<td>pro-erdogan türener, abschaum afrikas, araber, schützlingsige, deutsche kartoffel, deniz, nafirs, polnische hurensohne</td>
</tr>
</tbody>
</table>

### Table 5: Examples of surface forms of target groups

<table>
<thead>
<tr>
<th>Category</th>
<th>Surface Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>People of Color</td>
<td>#negersindkainemenschen, afle, himbo, dunkler teint, nafirs, neger, negroude goldstücke, schwarze, sklaven</td>
</tr>
<tr>
<td>Jews</td>
<td>dreckiges judenpack, judenschwein, zentralrat der juden, jüdischer zombie, rattenvolk, zionisten</td>
</tr>
<tr>
<td>Muslims</td>
<td>#islamisierung, #muslime, islamlobbys, bärtigen kinder-schänder, dtib imams, dreckige kopftuchmadchen, gotteskrieger, isis-schlampen, muslim-ungeziefer, scharia</td>
</tr>
<tr>
<td>Nationality/Origin</td>
<td>pro-erdogan türener, abschaum afrikas, araber, schützlingsige, deutsche kartoffel, deniz, nafirs, polnische hurensohne</td>
</tr>
</tbody>
</table>

\[\text{P} \quad \text{R} \quad \text{F}_1\]

<table>
<thead>
<tr>
<th>Category</th>
<th>Surface Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group of persons</td>
<td>.81 .85 .83</td>
</tr>
<tr>
<td>Individual as member of group</td>
<td>.00 .00 .00</td>
</tr>
<tr>
<td>Distinguishable by nationality, etc.</td>
<td>.79 .71 .75</td>
</tr>
<tr>
<td>Inciting hatred</td>
<td>.25 .07 .11</td>
</tr>
<tr>
<td>Inciting violence</td>
<td>.70 .73 .72</td>
</tr>
<tr>
<td>Punishable (random)</td>
<td>.04 .50 .07</td>
</tr>
<tr>
<td>Punishable (direct)</td>
<td>.69 .28 .39</td>
</tr>
<tr>
<td>Punishable (submodels + decision tree)</td>
<td>.41 .43 .42</td>
</tr>
</tbody>
</table>
Thus, we argue that annotating data for groups referenced in the text (even implicitly) is a prerequisite for 'group detection' as a stand-alone NLP task. Once this is established, it can be used to find the best methods for group detection. A possible way to find surface variants might be to compile a list of common surface forms and compare the closest synonyms for a group as computed over a more general corpus.

6.2 Conduct Detection

For specific targeting conduct like inciting violence, detecting the most common actions patterns like 'kill GROUP' or 'burn GROUP' might be a promising approach. This would also limit the number of false positives, e.g. when someone 'threatens' to burn a candle instead. For this task, semantic role labeling (Gildea and Jurafsky, 2002) or using frames (Baker, 2014) could be useful, but existing resources like FrameNet seem not specific enough, as they put 'threat' under the COMMITMENT frame (in the sense of 'committing to harm someone').

In general, there is a high level of metaphor, irony and sarcasm in the comments, which poses serious challenges to all conduct detection methods. Even though irony and sarcasm are not legal terms as such, they might have an influence on the assessment as to whether a targeting conduct like inciting hatred is given. Accordingly, these cases can be captured at the annotation level as in dubio pro reo, i.e. not punishable.

7 Related Work

Automated detection of offensive Internet discourse has been intensively studied under a variety of names, for instance: abusive language (Waseem et al., 2017) or content (Kiritchenko et al., 2020), ad hominem arguments (Habernal et al., 2018), aggression (Kumar et al., 2018), cyberbullying (Xu et al., 2012; Macbeth et al., 2013), hate speech (Warner and Hirschberg, 2012; Ross et al., 2016; Del Vigna et al., 2017), offensive language usage (Razavi et al., 2010), profanity (Schmidt and Wiegand, 2017), threats (Oostdijk and van Halteren, 2013) and socially unacceptable discourse (Fišer et al., 2017). While most early work focused on English, now there is also a growing body of work in other languages, e.g., German (Ross et al., 2016), Italian (Del Vigna et al., 2017), and Dutch (Oostdijk and van Halteren, 2013). All of those works use a non-legally informed definition of the construct to be detected. A notable exception is work on Slovene by Fišer et al. (2017) who relied on annotators interpretation of the legal definition without further breaking down those decisions.

There is a body of work in NLP with a legal perspective focused on predicting the outcome of court decisions (Aletras et al., 2016; Katz et al., 2017; Brunninghaus and Ashley, 2003; Kastellec, 2010; Waltl et al., 2017), to the best of our knowledge our work is the only effort to operationalise a legal framework for hate speech. However, the dependence on existing court decisions makes it difficult to work with legal problems where relevant case law is not available as a data source. To overcome this problem, Zufall et al. (2019) translated statutory rules for defamatory offenses into a series of annotatable binary decisions.

The importance of finding groups for hate speech analysis has also been stressed by Kiritchenko et al. (2020). As offenses against groups are often implicitly framed, Sap et al. (2020) introduce Social Bias Frames that make the attacked group explicit. As group detection can work with any set of group categories, it can also be adapted to cover non-Western groups (Sambasivan et al., 2021).

8 Conclusion

We operationalize a 'legal approach to hate speech' by translating the requirements of the EU Framework Decision into a series of annotation steps that can be reliably performed by laypersons. However, we show that learning a model of whether a post is punishable remains challenging. We thus propose to tackle two subtasks instead: group detection and conduct detection. Depending on the applicable legal framework, a final decision on the legal status of a comment can then be derived from the combination of detected group and conduct. Relying on subtasks comes with the added benefit of increased transparency and explainability compared to black-box models. This is crucial for systems that potentially interfere with human rights, such as the balance between freedom of expression and the prevention of discrimination. Hence, we recommend this modular approach as the preferred way of composing systems for legal decision-making.
Ethical Considerations

Predicting the legal status of a comment might infringe on the fundamental right of ‘free speech’. On the other hand, we are targeting the worst tail-end of the distribution – the kind of hate speech that is putting democracy in danger by inciting hatred and violence in a society. Not addressing hate speech and its foregoing automated detection methods would give further rise to possible discrimination, making it a problem for equal participation in a democracy. As our approach introduces a layer of algorithmic transparency not found in traditional methods, we believe that the importance of this research outweighs its dangers.

Annotation Process  Regarding our made-up examples, we conducted a survey with nine students, asking them to create short texts that presumably constitute ‘incitement to hatred’ (see Section 4). This survey was approved by the ethics committee of ANONYMIZED. The final annotation of the dataset was carried out by two paid annotators, who were compensated above the local minimum wage. Annotators were warned about the offensive nature of the data and instructed only to annotate 50 comments a day to mitigate the effect of fatigue.

Race and Gender  The EU Framework Decision explicitly requires the conduct to be directed against a “group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin” (Art.1(1)(a) Framework Decision). It is thus a necessary legal requirement which is meant to protect the aforementioned groups and to prevent discrimination. We also use the groups ‘women’ and ‘LGBTQ+’, as these are often the targets of hate speech. Our model explicitly allows for adding other groups in order to adapt to differing legal standards.

Deploying Systems for Legal Decision-making

Systems used in the context of legal decision-making or, more generally, systems that filter specific content should be used with great care and in view of the potential interference with human rights, namely the right to free speech. We explicitly do not recommend using any legal decision-making system without human supervision. We consider the improved transparency of our model to be an important step in allowing prosecutors to understand the reasons behind flagging a certain comment as potentially punishable.

Release of the Data  As our dataset consists of postings that could be traced back to individuals, it contains personal data in the sense of the EU General Data Protection Regulation (GDPR). To comply with this legal standard, and given the sensitive nature of the data, we do not make any of the real postings publicly available. We do, however, publish the made-up examples generated during the feasibility study.
A Data Distribution in Automated Experiments

Note that the number of total annotations is larger than 1000, since some posts contained multiple groups.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>false</th>
<th>true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group of persons</td>
<td>541</td>
<td>465</td>
</tr>
<tr>
<td>Individual as member of group</td>
<td>992</td>
<td>14</td>
</tr>
<tr>
<td>Distinguishable by nationality, etc.</td>
<td>744</td>
<td>262</td>
</tr>
<tr>
<td>Inciting hatred</td>
<td>992</td>
<td>14</td>
</tr>
<tr>
<td>Inciting violence</td>
<td>866</td>
<td>120</td>
</tr>
<tr>
<td>Punishable</td>
<td>966</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 6: Label distribution for automated detection experiments.

<table>
<thead>
<tr>
<th>Group Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>341</td>
</tr>
<tr>
<td>Foreigners/Migrants</td>
<td>155</td>
</tr>
<tr>
<td>Other</td>
<td>103</td>
</tr>
<tr>
<td>Left Wing/Green Party</td>
<td>93</td>
</tr>
<tr>
<td>Muslims</td>
<td>81</td>
</tr>
<tr>
<td>Other Politicians</td>
<td>69</td>
</tr>
<tr>
<td>Nationality/Origin</td>
<td>49</td>
</tr>
<tr>
<td>Jews</td>
<td>46</td>
</tr>
<tr>
<td>Women</td>
<td>29</td>
</tr>
<tr>
<td>LGBTQ+</td>
<td>17</td>
</tr>
<tr>
<td>People of Color</td>
<td>15</td>
</tr>
<tr>
<td>Disabled/Sick</td>
<td>6</td>
</tr>
<tr>
<td>Right Wing</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7: Distribution of adjudicated group categories in the dataset.

References


System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.
