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 for better explainability and higher transparency, 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 an 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 (Fortuna and Nunes, 2018; Poletto et al., 2020; Schmidt and Wiegand, 2017) 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 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]

Legally and practically more relevant is the following instrument: the European Union (EU) has,

[1] General Assembly resolution 2106 (XX) of 21 Dec 1965.

[2] ETS No. 189, 28.01.2003.

[3] Bulgaria, 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.

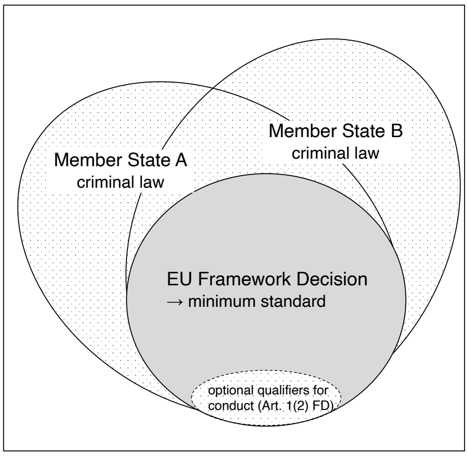

Figure 1: Scope of the EU Framework's legal standard. It defines a common core of punishable offenses.

after long debate, set up a common regime with a *Framework Decision*[4] 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.[5] Furthermore, the EU's new proposal of a Digital Services Act aims to create new obligations for large online platforms regarding illegal content.[6] 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 subtasks 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 material. For the detection of social-media posts,

[4] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. In the remainder of this paper, we shall refer to this as 'EU law' or 'EU Framework Decision' for simplification.

[5] `https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985`

[6] Proposal of 15.12.2020, COM(2020) 825 final.

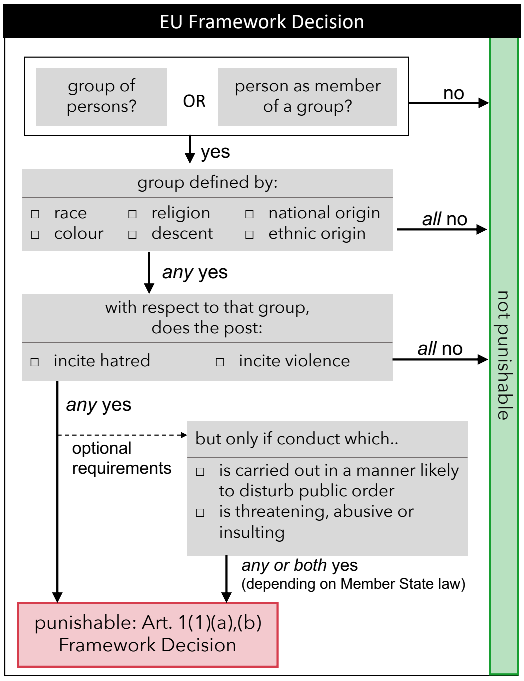

Figure 2: Decision tree derived from legal framework

there is no added value in implementing these actions separately, as they are always met in case of public social-media posting on the Internet.

In a simplified way, two main questions have to be answered: (1) does a statement address a protected group? and (2) does it target that group by inciting hatred or violence? We address these as (1) *target group* and (2) *targeting conduct*.

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).

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

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.

- you fucking muslim should leave our country!
- This dirty american bitch, typical american, lying son-of-a-bitch, out of our country!

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 Art.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'.[7] 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.

[7] BGHSt 21, 371 (372); BGHSt 46, 212 (217)

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)'.[8] 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'.[9] 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.

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

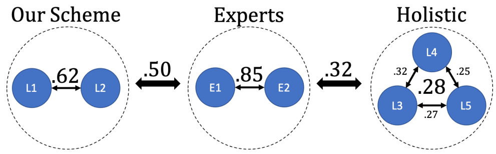

Figure 3: Cohen's Kappa for different annotation schemes in the feasibility study.

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.[10] 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.[11] 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.

[10] 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] The made-up posts are comparable in nature to realistic posts. See next Section 4 for a more detailed description.

Source	#	% Punishable
Made-up	157	13.5
Web search	80	6.2
Anti hate speech initiatives	88	10.2
GermEval2019 (abuse, insult)	425	0.9
GermEval2019 (other)	250	0.0

Table 1: Composition of the dataset by source.

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.

Made-up We include the 'made-up' examples from the feasibility study, re-annotated according to the EU framework. The examples were produced by volunteers, who were instructed to write short texts presumably constituting 'incitement to hatred' against the list of target groups mentioned in Figure 4. Participants also received instances of real hate speech as examples for their artificial posts. 9 participants created a total of 157 short texts. The resulting statements are nearly indistinguishable in form from real examples, but we have no way of controlling for topic biases that might have been introduced via this process.

Web search We performed a manual search of Twitter, comment sections of online newsrooms, law forums, court databases as well as news articles resulting in 80 instances.

Anti hate speech initiatives We include 88 hate speech comments collected by the initiative 'respect!' of the Demokratiezentrum Baden-Württemberg.

GermEval2019 Data samples from the subtask two corpus of GermEval 2019, a shared task on the identification of offensive language (Struß et al., 2019), were also included. We add 425 tweets of 'abuse' and 'insult' category. We also add 250 of the 'other' category in order to have some tweets with the same topic that are non-punishable.

4.2 Annotation Scheme & Process

The full dataset was annotated by two paid layperson annotators. We provided them with an annotation manual based on the legal requirements described in Section 2 with further explanations, instructions, and examples. To measure annotation quality, a subset (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. The full list of target groups used in our study can be seen in Figure 4.

- Euthanasia is the only way to deal with all the **disabled people**, they should be buried.
- You should gas all the **Jews**. All they want is your money.
- The **oil eyes** should set themselves on fire with their gasoline.

Example 5: Explicitly mentioned groups.

Beyond groups, we 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 as well as between 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.

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

Statement	Group Category	Target Group: Group of persons	Target Group: Distinguishable by nationality, race, religion, ethnicity	Target Group: Explicit	Targeting Conduct: Incite hatred	Targeting Conduct: Incite violence	**Punishable**
So dumm wie Gutmenschen halt sind! Stupid as do-gooders are!	Other	✓		✓			
Die linken Hurensöhne! Ballert die alle ab! Left-wing sons of bitches! Shoot 'em all!	Left-Wing	✓		✓		✓	
Es brennen noch zu wenige Synagogen. Still too few synagogues on fire.	Jews	✓	✓			✓	✓
Muslime sind alles Vergewaltiger! Schützt deutsche Frauen! Muslims are all rapists! Protect our German women!	Muslims	✓	✓	✓	✓		✓

Table 2: Example annotations from the resulting dataset. Surface form referring to groups is underlined.

		L1/ L2	L1/ Exp	L2/ Exp
	Group Category	**.77**	**.70**	**.67**
Group	Group of persons	.49	.82	.55
	Individual as group member	.14	.24	.48
	Nationality, race, religion, ...	.52	.42	.67
	Any target group	**.53**	**.52**	**.70**
Conduct	Inciting hatred	.11	.39	.00
	Inciting violence	.56	.64	.74
	Any targeting conduct	**.56**	**.69**	**.52**
	Punishable	**.33**	**.43**	**.37**

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

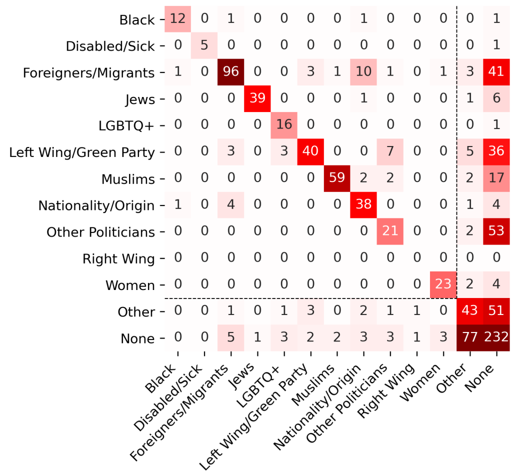

Figure 4: Confusion matrix of non-expert annotators.

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.

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.[12]

Setup Fine-tuned `BERT` (Devlin et al., 2018) models have proven to be strong baselines for various NLP tasks, so we follow this practice[13], 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, with a learning rate of $2e^{-5}$. The learning rate is linearly increased up to its peak during the first 10% of training 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.

Results The model achieves an average F_1 of .39 (P .69; R .28), which shows that the task is complex

[12]For model training, differences between annotators were adjudicated by a legal expert. The IAA reported in Section 4 are thus not applicable.

[13]For the implementation, we use *HuggingFace Transformers* (Wolf et al., 2020) and *PyTorch* (Paszke et al., 2019).

	P	R	F_1
Group of persons	.81	.85	.83
Individual as member of group	.00	.00	.00
Distinguishable by nationality, etc.	.79	.71	.75
Inciting hatred	.25	.07	.11
Inciting violence	.70	.73	.72
Punishable (direct)	.69	.28	.39
Punishable (submodels + decision tree)	.41	.43	.42

Table 4: Overview of prediction results

Category	**Surface Form**
People of Color	#negersindkeinemenschen, affe, bimbo, dunkler teint, nafris, neger, negroide goldstücke, schwarze, sklaven
Jews	dreckiges judenpack, judenschwein, zentralrat der juden, jüdischer zombie, rattenvolk, zionisten
Muslims	#islamisierung, #muslime, islamlobbys, bärtigen kinderschänder, ditib imams, dreckige kopftuchmädchen, gotteskrieger, isis-schlampen, muslim-ungeziefer, scharia
Nationality/ Origin	pro-erdogan türken, abschaum afrikas, araber, schlitzäugige, deutsche kartoffel, deniz, nafris, polnische hurensöhne

Table 5: Examples of surface forms of target groups

and not already solved by the baseline model. In contrast to regular hate speech detection, 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).

1) DEPORT DEPORT [...] DEPORT
2) Faggots should be castrated and locked up!
3) A metro we build, a metro we build, a metro we build from Jerusalem to Auschwitz, a metro we build!

Example 6: Cases misclassified by detection model

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. 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 make accurate predictions ($.00 - .11$). Both are rare in the dataset (14 positive cases each), making it difficult to learn these from the data. Having more positive cases should bring performance up to levels comparable to the other annotations.

Conclusion The performance of our baseline systems is rather low, indicating that an automatic prediction of legal decisions is a challenging task. Experiments on subtasks suggest that more data may improve performance, but obtaining examples of these rare phenomena poses a challenge in itself. In the future, we recommend using a modular subtask approach, as doing so lends explainability to model decisions - a crucial property in a system interacting with fundamental rights.

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). However, we argue that the underlying methodology can be generalized in a straightforward way. Instead of directly predicting whether a post is punishable or not, we should 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).

1) [...] For them the sport [football] is like. I put a goat on the field, 22 holy warriors and whoever knocks it up first, wins.
2) No mercy for **terrorists**. We have declared war on **Islam**. **They** had 800 years to reform. Time is up!

Example 7: 1) Implicit targeting of Muslims. 2) Muslims target group only identifiable by coreference chains.

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, as our dataset indicates that calls to some actions are quite common. 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), Dutch (Oostdijk and van Halteren, 2013) and Slovene (Fišer et al., 2017). All of those works use a non-legally informed definition of the construct to be detected.

Interdisciplinary work combining NLP with a legal perspective has mostly focused on predicting the outcome of court decisions (Aletras et al., 2016; Katz et al., 2017; Bruninghaus and Ashley, 2003; Kastellec, 2010; Waltl et al., 2017). 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 or not remains challenging. We thus propose to tackle independently the subtasks of *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 task, we do not make any of the real postings publicly available. We do, however, publish the made-up examples generated during the feasibility study.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *PeerJ Computer Science*.

Collin F. Baker. 2014. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, MD, USA. Association for Computational Linguistics.

Stefanie Bruninghaus and Kevin D. Ashley. 2003. Predicting Outcomes of Case Based Legal Arguments. In *Proceedings of the International Conference on Artificial Intelligence and Law*, pages 233–242, New York, NY, USA. ACM.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

William B. Fisch. 2002. Hate speech in the constitutional law of the united states. *The American Journal of Comparative Law*, (50):463–492.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. 51(4).

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Comput. Linguistics*, 28(3):245–288.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396. Association for Computational Linguistics.

Jonathan Kastellec. 2010. The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees. *Journal of Empirical Legal Studies*, 7(2):202–230.

Daniel Martin Katz, Michael J. Bommarito, II, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4):1–18.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2020. Confronting abusive language online: A survey from the ethical and human rights perspective.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11. Association for Computational Linguistics.

Jamie Macbeth, Hanna Adeyema, Henry Lieberman, and Christopher Fry. 2013. Script-based story matching for cyberbullying prevention. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 901–906.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv*.

Nelleke Oostdijk and Hans van Halteren. 2013. N-Gram-Based Recognition of Threatening Tweets. In *Computational Linguistics and Intelligent Text Processing*, pages 183–196, Berlin, Heidelberg. Springer Berlin Heidelberg.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer-Verlag.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.

Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Julia Maria Struß, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the Outcome of Appeal Decisions in Germany's Tax Law. In *Electronic Participation*, pages 89–99, Cham. Springer International Publishing.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frederike Zufall, Tobias Horsmann, and Torsten Zesch. 2019. From Legal to Technical Concept: Towards an Automated Classification of German Political Twitter Postings as Criminal Offenses . In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, NAACL HLT '19, pages 1337–1347. Association for Computational Linguistics.