From Legal to Technical Concept: Towards an Automated Classification of German Political Twitter Postings as Criminal Offenses

Frederike Zufall[1†], Tobias Horsmann[2†], Torsten Zesch[2]
[1] Waseda Institute for Advanced Study, Waseda University, Tokyo, Japan
[2] Language Technology Lab, University of Duisburg-Essen, Germany
f.zufall@kurenai.waseda.jp
{tobias.horsmann|torsten.zesch}@uni-due.de

Abstract

Advances in the automated detection of offensive Internet postings make this mechanism very attractive to social media companies, who are increasingly under pressure to monitor and action activity on their sites. However, these advances also have important implications as a threat to the fundamental right of free expression. In this article, we analyze which Twitter posts could actually be deemed offenses under German criminal law. German law follows the deductive method of the Roman law tradition based on abstract rules as opposed to the inductive reasoning in Anglo-American common law systems. This allows us to show how legal conclusions can be reached and implemented without relying on existing court decisions. We present a data annotation schema, consisting of a series of binary decisions, for determining whether a specific post would constitute a criminal offense. This schema serves as a step towards an inexpensive creation of a sufficient amount of data for automated classification. We find that the majority of posts deemed morally offensive actually do not constitute a criminal offense and still contribute to public discourse. Furthermore, laymen can provide sufficiently reliable data to an expert reference but are, for instance, more lenient in the interpretation of what constitutes a disparaging statement.

1 Introduction

The Internet is frequently used for discussing a variety of topics and an important medium for the exchange of opinions, considered crucial for healthy democratic societies. However, the rough tone in the Internet frequently leads to defamatory or abusive comments in these discussions. The EU has tried to tackle the problem by defining the term 'illegal hate speech'.[1] Additionally, in 2017, the European Commission published a communication entitled 'Tackling Illegal Content Online' aiming for enhanced responsibility of online platforms.[2] Independently from these recent developments on the EU level, Germany adopted the 'Network Enforcement Act'[3] in 2017. The Act provides for a regulatory framework for 'illegal content'[4] on social network platforms like Twitter or Facebook. It imposes the obligation on these providers to delete illegal content upon notification within seven days; in case of evidently illegal content within 24 hours.[5] From a practical point of view, given the number of statements on social media along with their possible notification, feasibility and accuracy of the required legal assessment becomes an important issue. Natural Language Processing might thus provide the necessary means to assist the legal assessment.

In this work, we investigate at which point morally offensive statements in social media constitute defamatory offenses under the German Criminal Code (StGB)[6], thus representing 'illegal content' according to the Network Enforcement Act and thereby triggering a deletion obligation for platform providers.[7] We analyze the legal decision-making process to determine defam-

[†]Equal contribution.

[1]Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law and national laws transposing it.

[2]COM(2017) 555 final.

[3]Netzwerkdurchsetzungsgesetz v. 1.9.2017 (BGBl. I S. 3352).

[4]See § 1(3) 'rechtswidrige Inhalte'.

[5]See § 3(2)(2),(3) of the Act. It is however doubtful whether these strict procedural requirements violate EU law, namely Art. 3, Art. 14 e-Commerce Directive (2000/31/EC) i.e. require acting 'expeditiously' after obtaining knowledge.

[6]Strafgesetzbuch v. 13.11.1998 (BGBl. I S. 3322).

[7]It is not guaranteed that a judge would necessarily arrive at the same conclusion, but a lawyer's expertise serves as a strong indicator for potentially punishable conduct.

Proceedings of NAACL-HLT 2019, pages 1337–1347
Minneapolis, Minnesota, June 2 - June 7, 2019.

atory offenses (§ 185 to § 187 StGB), which also clarifies the tension between the right to honor and the freedom of expression. Due to its additional complexity, we leave out incitement to hatred against a national, racial, religious or ethnic group or segments of the population (§ 130 StGB) as an offense against public peace in this paper. Furthermore, we investigate automated detection of postings protected by the freedom of expression in order to assist social media moderators. We focus in particular on the process of inexpensive and scalable data annotation, as access to legal expertise is a major bottleneck for providing a sufficient amount of data for classifier training.

2 Related Work

An automated detection of Internet discourse in which individuals or groups are verbally attacked has been intensively investigated under a variety of names, for instance: abusive language (Waseem et al., 2017), 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) or socially unacceptable discourse (Fišer et al., 2017).

The majority of the work focuses on the English language with few exceptions for instance for German (Ross et al., 2016), Dutch (Oostdijk and van Halteren, 2013), Italian (Del Vigna et al., 2017) or Slovene (Fišer et al., 2017). The dataset annotated in Fišer et al. (2017) is the only one that includes a coarse-grained binary annotation category indicating if an utterance violates Slovene law. To the best of our knowledge, automatic determination as to whether the (textual) content of a posting constitutes a criminal offense has never been previously attempted. Previous work focused on detecting postings with socially unacceptable content but without considering actual legal implications for freedom of expression.

Approaches that bring together Natural Language Processing with the legal perspective are in contrast significantly fewer, especially considering the fact that the legal evaluation depends on the applicable legal regime. Previous work focused on predicting the outcome of court trials, which all have in common that they derive their data from a rather large set of court-provided information. Bruninghaus and Ashley (2003) works on a combination of U.S. case law and normative rules: they experiment with clustering and regression models for predicting the outcome of U.S. cases. Katz et al. (2017) predicts U.S. supreme court rulings by using a random forest classifier; Kastellec (2010) investigates mappings from case facts to court decisions as outcomes. Waltl et al. (2017) predicts the outcome of decisions in German tax law. Aletras et al. (2016) predicts decisions of the European Court of Human Rights. Deriving data from court decisions might be an approach that is practical if relevant case law exists for the respective legal problem, which particularly makes sense from the perspective of the Anglo-American common law system.[8]

3 Operationalising Legal Assessment

Unlike under Anglo-American common law, for legal systems based on Roman law ('civil law' systems), the dogmatic perception of the respective legal disposition lies at the heart of legal decision-making. Our approach thus differs from the above-cited works by placing the focus on the abstract concept of an existing legal norm. The advantage of our approach is therefore that we pursue a solution to address legal problems by creating new data out of abstract legal rules, independently of whether they have been decided by a court. We rely solely on the Internet posting for this consideration, which is the same information available to moderators of social media platforms. To build the bridge from legal thinking to a technical implementation, we start by analyzing the legal requirements for social media content. We find that the decision-making process to determine criminal offenses can be formulated as a sequence of binary decisions when applying the legal dependencies between German criminal law and the fundamental rights of the individual as shown in Figure 1. The derived schema of binary decisions is shown in Figure 2, which we will use in the following section. We now turn to a discussion and analysis of the legal decision process to clarify how we derived this sequence of binary decisions.

[8] 'Common law' refers to the Anglo-American legal system that derives the law from judicial decisions, in contrast to the 'civil law' system of continental Europe that focuses on the abstraction of legal concepts in codified statutory law. See: B.A. Garner (2001) A Dictionary of Modern Legal Usage (2nd, revised ed.) New York: OUP.

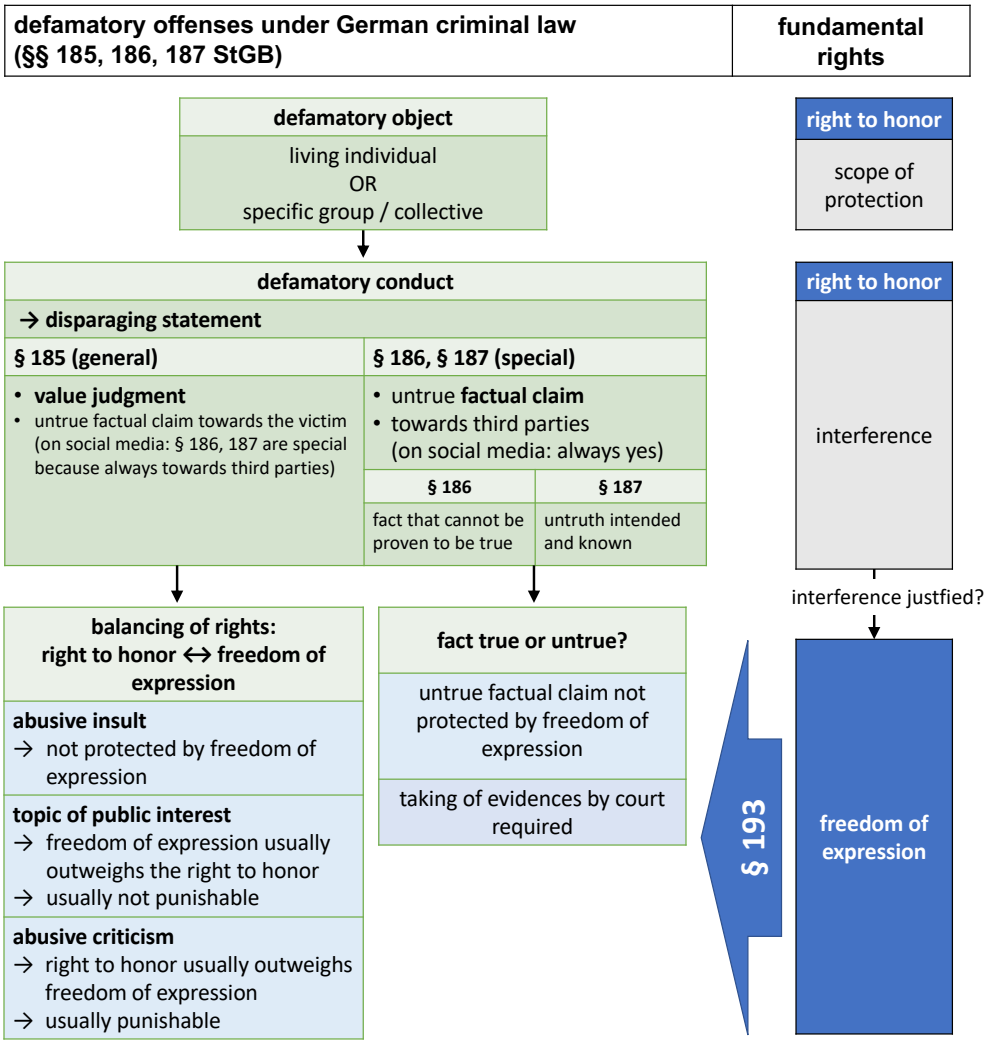

Figure 1: Conditions for defamatory offenses and fundamental rights' background under German law

Scope So what constitutes 'illegal content' that the Network Enforcement Act is targeting? The legal definition of the term 'illegal content'[9] is referring to offenses stipulated in the German Criminal Code. These references include, inter alia, defamatory offenses in § 185 to § 187 StGB[10] that cover the criminal punishment of insulting or defamatory statements. Accordingly, if a statement posted on social media fulfills the required elements of these offenses, the provider has the above-described obligation based on the Network Enforcement Act to delete said statement upon notification.[11] For this paper, we exclude § 130 StGB[12], that covers incitement to hatred against a national, racial, religious group or a group defined by their ethnic origins, due to an additional complexity of the assessment.

3.1 The Relevance of Fundamental Rights

To understand their elements in detail, it is crucial to refer to the more general legal concept behind these criminal offenses: as illustrated in Figure 1 the intention behind § 185 to § 187 StGB is leading to the protection of the victim's personality right, namely their right to honor under the German Constitution.[13] It is this right that is potentially at stake when social media users are disseminating statements with third parties as potential victims.

[9] See § 1(3) 'rechtswidrige Inhalte'.

[10] § 185: 'insult'(*Beleidigung*), § 186: 'defamation' *Üble Nachrede*) and § 187: 'intentional defamation' (*Verleumdung*). The reference to these defamatory offenses has however been criticized in literature, see: Erbs/Kohlhaas, Strafrechtliche Nebengesetze, 220. EL Juli 2018, § 1 NetzDG, Rn. 16-18.

[11] See § 3(2)(2),(3) of the Act.

[12] § 130, 'incitement to hatred' (*Volksverhetzung*).

[13] Derived from Art. 2(1) and Art. 1(1) of the German Constitution (Grundgesetz); BVerfGE 35, 202; E 54, 148, 155.

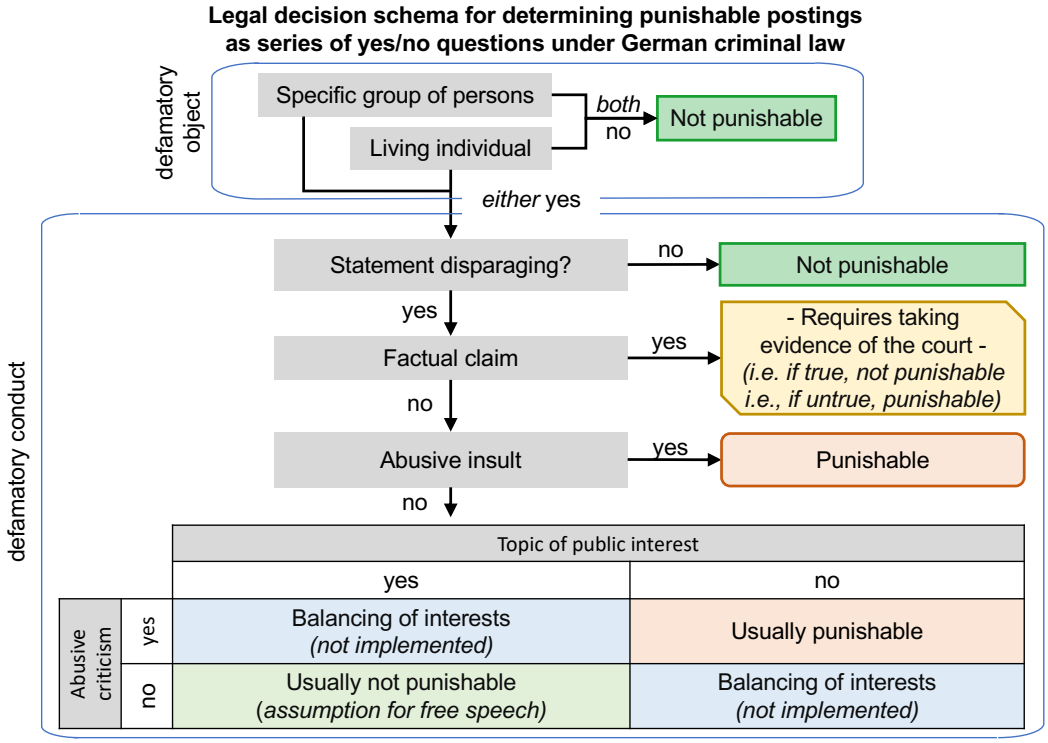

Figure 2: Series of binary decisions for determining criminal offenses under German criminal law

3.2 Defamatory Object

Consequently, the scope of protection of § 185 to § 187 StGB follows the respective interpretation of the right to honor. Thus, all three offenses share the approach to the possible victim as a holder of the right to honor: a **living individual** that might be addressed by a name, personal pronoun or user-mention as shown in Example 1.

(a) Are *you* kidding?
(b) *John* is this true?
(c) @*user* I don't believe you.

Example 1: Addressing living individuals

A **group of persons** can be considered as a potential victim if that group is distinguishable from the general public such that every member of that group could feel their honor is infringed as shown in Example 2.[14]

(a) *My school's language teachers* are all idiots.
(b) *The female students of this year's grad class* are all dump.

Example 2: Addressing distinguishable group of persons (highlighted in italic)

[14]BGHSt 19, 235, 238.

Consequently, only certain groups do qualify as potential defamatory object. Example 3 illustrates groups that would be too broad to be distinguishable from the general public.[15]

(a) All international conflicts are caused by *men*.
(b) *Refugees* out!!

Example 3: Counterexamples for addressing too unspecific or large groups

Collective entities such as governments or press companies with a recognized social role and who act with a collective, single will are included in the right to honor as shown in Example 4.[16]

(a) *The federal government* is lying.
(b) I don't like *the Christian Democratic Party*
(c) *The New York Times* got it all wrong.

Example 4: Addressing collective entities (highlighted in italic)

We translate these conditions of § 185 to § 187 StGB into an either/or-question, respectively whether *either* a living individual *or* a specific group is an object of the respective statement.

[15]These groups may, however, still qualify as a potential victim of 'incitement to hatred' (§ 130).
[16]See § 194 StGB; BGHSt 6, 186, 191, 192.

3.3 Defamatory Conduct

Disparaging Statement The next step in the legal assessment is then the existence of insulting or defamatory conduct with respect to the above-mentioned object, in the form of an expressed disparaging statement. This requirement is again shared by § 185 to § 187 StGB. It is already fulfilled by expressing contempt or disrespect through the allegation of shortcomings that could reduce the victim's social standing as shown in Example 5.[17]

(a) John is *an idiot.*
(b) The government is *lying*.
(c) Minister M *slept during the discussion*.

Example 5: Disparaging statements

From the perspective of the underlying fundamental rights, it is this disparaging statement which constitutes the interference with the potential victim's right to honor. The existence of a *disparaging statement* is implemented by a *yes/no-question*.[18]

Value Judgment or Factual Claim? As simplified in Figure 1, the legal assessment then varies between § 185 StGB as general disposition and § 186 and § 187 StGB with special rules and an increased penalty range.

For the different scope of these dispositions, the difference between the legal terms *'value judgment'* and *'factual claim'* (i.e. the assertion of facts, may they be true or untrue) is crucial. Value judgments constitute expressions of personal opinions as shown in Example 6:[19]

(a) *Merkels decisions are bullshit.*
(b) *@user I don't like you.*

Example 6: Value judgments

A factual claim can be clearly classified as true or untrue and is accordingly capable of proof as shown in Example 7.[20]

(a) *I saw Mr. A buying drugs yesterday evening.*
(b) *Minister M slept during the discussion.*

Example 7: Factual claim

§ 185 StGB, stipulating the insult ('Beleidigung'), comprises value judgments and untrue factual claims, irrespective of their dissemination towards third parties. § 186 and § 187 StGB on the other side provide for special rules for the assertion or dissemination of untrue facts, i.e. towards third parties. As the publication of statements on social media constitutes an 'assertion' or 'dissemination', untrue facts - for our study - are only treated by § 186, § 187 StGB. This reduces the scope of § 185 StGB to value judgments only.

From the perspective of the right to honor, only untrue factual claims may constitute a violation, while the assertion of true facts is always covered by the freedom of expression.[21] The distinction has consequences on the procedural level: because only the assertion of untrue facts violates the right to honor, during criminal proceedings, the court has to assess the truth by taking evidence. A technical implementation of this assessment would therefore require access to unlimited knowledge that goes beyond the textual information on which we work. Accordingly, we stop our examination in case of a factual claim.[22]

3.4 Value Judgments: Balancing of Rights

As the distinction between value judgment and factual claim is an alternative decision,[23] we continue our implementation for value judgments. In criminal proceedings, the court would have to consider at this point once more fundamental rights: value judgments - being not classifiable as untrue or true - generally fall under the scope of the freedom of expression of the potential offender.[24]

In the German Criminal Code, this is reflected by § 193 StGB: even if a statement falls under the scope of said criminal offenses, it might still be

[17] BGHSt 36, 145, 148.

[18] Regarding the mere *expression* of a statement, we assume a *yes-answer* for statements published on social media that are subject to this study, and therefore do not implement this condition.

[19] BVerfGE 61, 1; for Twitter postings: OLG Karlsruhe, 24.10.2018, 6 U 65/18.

[20] RG 41, 193; 55, 131; BVerfGE 94, 8.

[21] BVerfGE 99, 185, 197; E 97, 381, 403.

[22] Consequently, we do not implement subsequent conditions of § 186, § 187 StGB, as shown in Figure 1, respectively whether facts cannot be proven to be true (§ 186 StGB) or whether the untruth was intended and known (§ 187 StGB).

[23] Ambiguous statements that are based on facts, but are overall characterized by a valuation of these facts, fall under the category of 'value judgments'.

[24] Art. 5(1)1 of the German Constitution (Grundgesetz). According to Art. 5(2) the freedom of expression then again is limited by the right to honor.

justified based on § 193 StGB as exercise of legitimate interests. The most prominent example of one of these conflicting interests is the offender's freedom of expression. On the constitutional level, then, the decision of whether a social media posting constitutes a punishable criminal offense and leads to the platform provider's deletion obligation can thus ultimately be perceived as a balancing between freedom of expression and the right to honor.

Consequently, the court would have to balance these concurrent rights depending on the case at hand. But how could that balancing, usually comprising an evaluation of various factors, be carried over to a technical implementation? Over the years, German case law from the Federal Constitutional Court has developed guidelines for this balancing to be considered by the judge, which take the step of implying the typical outcome of the balancing. We implement these guidelines in three yes/no-questions:[25]

Abusive Insult Statements that constitute breaking a taboo by themselves and intend only the defamation of the victim without any substantiated contribution are classified as *'abusive insult' (Formalbeleidigung)*. According to settled case law, these statements are already excluded from the scope of freedom of expression.[26] Consequently, a justification based on § 193 StGB is, in this regard, denied and the elements of § 185 StGB are fulfilled along with a violation of the right to honor. Given these severe consequences for free speech, the German Constitutional Court has so far only once approved a statement as constituting an 'abusive insult' as shown in Example 8:[27]

A disabled person is called *"cripple"*

Example 8: Abusive insult

Topic of Public Interest For statements that contain a contribution to the public discourse with respect to a particular relevant topic of public interest, the settled case law of the German Federal Constitutional Court mandates a presumption in favor of free speech.[28]

Merkel prostitutes herself for the German car industry costing tax payers

Example 9: Topic of public interest

Example 9 comments on the right to stay of refugees, by this participating to the public debate in Germany about refugees from Syria. Accordingly, such statements usually outweigh the right to honor. They thus usually can be made, justified as having a legitimate interest based on § 193 StGB, therefore *usually not punishable*.

Abusive Criticism Finally, as *'abusive criticism' (Schmähkritik)* settled case law has defined statements that go beyond plausible criticism by primarily intending to abusively offend the victim, hereby neglecting a substantiated contribution.[29] In Example 10, the statement:

Minister M, that asshole, is lying to all of us!! noone has money to pay for this...

Example 10: No abusive criticism

despite the word 'asshole', still contributes to the public discourse, which is why its primary purpose is not (only) to abusively offend. Abusive criticism thus usually leads to favoring the right to honor over freedom of expression. Without justification pursuant to § 193 StGB, such statements are therefore *usually punishable*.

4 Proof of Concept

In this section, we now use the schema in Figure 2 to annotate data and learn more about the reliability of an automated classification.

4.1 Dataset

In order to legally assess social media postings, we first need to annotate a corpus as a starting point for an analysis. Randomly sampling postings from the Internet is a possible strategy to collect data for an annotation, but we would not have any certainty that enough offending postings occur. Therefore, we decide to use an existing corpus that has already been annotated for moral offensiveness. We use the corpus provided by the GermEval shared

[25] As illustrated in Figure 2, the judge would perform the balancing freely based on all circumstances (which we do not implement) if there is no *'abusive insult'* and if *'topic of public interest'* and *'abusive criticism'* are both yes or both no.

[26] Maunz/Dürig, Grundgesetz-Kommentar, 84. EL August 2018, Art. 5 Abs. 1, Rn. 62.

[27] BVerfGE 86, 1, 45 (*"Krüppel"*).

[28] BVerfGE 7, 198, 212.

[29] BVerfGE 61, 1, 12; E 82, 272 (284); for Twitter postings: LG Berlin, 13.10.2012, 33 O 434/11.

Decision	Agreement Acc	Cohen's κ
Living individual	.985	.961
Specific group	.940	.809
Disparaging	.925	.867
Factual claim	.925	.821
Abusive insult	.925	.820
Of public interest	.940	.855
Abusive criticism	.866	.678
Joint-decision	.821	.720

Table 1: Agreement between two legal experts on two-hundred randomly selected postings

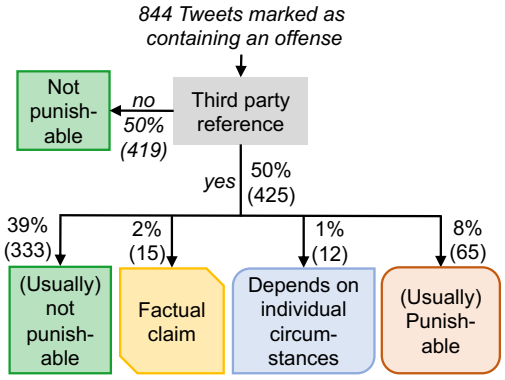

Figure 3: Legal categorization of annotated Tweets that were marked as containing an offense

task for *detecting offensive language usage* (Ruppenhofer et al., 2018). This dataset contains a mixture of German Twitter postings with a focus on German politics that are marked if the tweet is considered morally offensive from the subjective perception of the annotator. We work with a subset of 1,100 postings from this corpus, two-thirds of the postings (844) are marked as morally offensive. This enables us to investigate which statements commonly found in political debates are protected by the freedom of expression and which are not.

Annotation The reference annotation of these postings is provided by a fully-qualified lawyer of German law applying the schema in Figure 2. We additionally received 200 postings from a second fully-qualified lawyer in order to compute an agreement score between the two legal experts, which is shown in Table 1. We report accuracy and Cohen's κ (Cohen, 1960) for each decision and show the agreement for a *joint-decision* where we treat all decisions for a posting as a single decision. The legal experts disagree slightly on the assumption of abusive criticism. This is not surprising as the evaluation of courts might differ in different instances, especially regarding the balancing of interests in the case at hand.

Analysis Figure 3 shows the annotation results of the postings marked as *morally offensive*. We find that about half of the postings have to be categorized early on as *not punishable* for not containing a defamatory object, i.e. no living individual addressed or the addressed group is too unspecific. The remaining half is still to a large extent *usually not punishable*, mostly because the posts still contribute to a topic of public interest, despite of being disparaging. A small number of cases are either factual claims that would require taking evidence by the court or value judgments that do not concern topics of public interest. Thus, despite containing statements that may be deemed morally offensive, the vast majority of statements are legally acceptable, i.e. protected by the freedom of expression. The punishable cases often contain insulting buzzwords such as *slut, fat-ass* or *scumbag* when directed at a private individual, not at a person of public interest. Furthermore, punishable statements addressing a specific group use more frequently offending comparisons or descriptions but no typical single or two-word insults. However, it is important to recall that the dataset has a focus on political debates. Accordingly, most statements tackle a topic of public interest, and are thus considered *usually not punishable* granting a high degree of protection under the freedom of expression.

This analysis also shows that shared tasks such as *OffensEval* (Zampieri et al., 2019) tackle essentially only one step in the legal assessment, namely whether a statement is *disparaging*. Thus, they fall short of valuing the freedom of expression, which is in particular a problem for public discourse such as political debates, where opinions are often accompanied by 'bad' language.

4.2 Automated Detection

For an automated detection, it would seem straight forward to distinguish between *punishable* and *not punishable* postings. This approach requires an extremely large amount of data for each of the two classes, which we do not have. The data distribution is skewed with the *punishable* class being extremely small, which makes this direct ap-

	Decision	Acc	Class	F1
Defam. Object	Living individual	.794	Yes	.685
			No	.847
	Specific group	.835	Yes	.431
			No	.903
	either of the above	.744	Yes	.775
			No	.702
Defamatory Conduct	Disparaging	.727	Yes	.656
			No	.774
	Factual claim	.977	Yes	.000
			No	.989
	Abusive insult	.984	Yes	.000
			No	.992
	Is of public interest	.715	Yes	.514
			No	.798
	Abusive criticism	.952	Yes	.036
			No	.975

Table 2: Averaged 10-fold CV results for each decision on 1,100 Tweets, using an LSTM

proach infeasible. Instead, we investigate how well each of the binary decisions shown in Figure 2 can be learned independently, which has a less skewed distribution. We use a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) for classification.[30] We use the 300-dimensional German pre-trained word embeddings provided by Grave et al. (2018), which are trained on the German common crawl.

Table 2 shows averaged 10-fold CV results for each decision point. We observe that the accuracy is close to the underlying distribution of the two classes. The classification of the *defamatory object* has a mediocre performance. In particular, an insufficient coverage of group names and names of individuals in the dataset seem to be the main cause as the *no* classes usually perform considerably better than the *yes* classes. The classification of the decisions under *defamatory conduct* follows a similar trend. The few positive instances: *factual claim*, *abusive insult* and *abusive criticism* prevent a reliable distinction of these cases.

The next step would be to investigate how well the classification works in sequence, i.e. continuing the classification with the positively categorized instances of the previous step. However, the independent classification shows already that the amount of data is insufficient. Therefore, we turn next to the more pressing question of how to generate more data in a scalable way, especially without relying on expensive legal experts as annotators.

[30] We train 30 epochs, with 0.2 dropout and initialize using Glorot, and ReLU as activation function.

5 Data Annotation by Laymen

A scalable annotation of more data requires that laymen can be instructed in a way that enables them to solve the task at hand. Laymen are readily available, for instance via crowdsourcing but also as student assistants who can be more cheaply employed than legal experts for annotating data.

Setup We compare the annotation performance of both random crowd workers and student assistants. The crowd workers and the student assistants were required to speak German. We have no information on the educational background of the crowd workers, but we ensured that the student assistants were not students of law-related subjects. We prepared a simplified manual[31] based on Figure 2, which is supplemented with text examples for each decision to guide the layman through the annotation of each decision. We use the crowdsourcing platform figure-eight.com to let crowd workers and student assistants re-annotate the 1,100 postings for which we have a reference annotation by a legal expert. Each posting is annotated by three annotators.

The annotation results are shown in Table 3. It is to be expected that some annotators will perform better than others, but distinguishing the 'good' from the 'bad' is an additional challenge, which we will not deal with here. Instead, we aggregate the annotations of all participants in a voting-like fashion, taking in each case the majority vote for each decision.[32] This provides us with an approximation of the average layman performance on this task, which is the key information that we are interested in.

Analysis The results show that student assistants solve this task considerably better than crowd workers. In particular, the recognition of references to *specific group* poses the biggest challenges for crowd workers, which also explains why this group performs much more poorly than the student assistants. As shown in Figure 2, the evaluation for a post ends if neither a living individual nor specific group is addressed. If either of the first two decisions is incorrect, an annotator automatically makes up to five additional follow-up errors. The student assistants applied the manual considerably more consistently than the crowd

[31] `github.com/Horsmann/NAACL-2019-legal`

[32] We restrict the comparisons to postings for which we have three votes of the respective sub-group.

Decision	All users Acc	κ	Crowd-workers Acc	κ	Student assistants Acc	κ
Living individual	.822	.628	.800	.555	.826	.655
Specific group	.745	.357	.600	.192	.817	.502
Disparaging	.649	.401	.475	.158	.765	.574
Factual claim	.654	.381	.492	.131	.774	.577
Abusive insult	.590	.285	.400	.005	.669	.436
Is of public interest	.672	.263	.592	-.043	.691	.388
Abusive criticism	.589	.161	.575	-.049	.530	.224
Joint-decision	.357	.201	.175	.050	.383	.250

Table 3: Agreement between the reference annotation by a legal expert and the aggregated laymen annotations of: *all users* (on 1,000 posts), only *crowd-workers* (on 402 posts) and only *student assistants* (on 390 posts). Results for *crowd-workers* and *student-assistants* are limited to postings where all three votes per posting were provided by users from the respective group.

workers, leading to fewer follow-up errors. Determining the referenced individual is also frequently challenging when several Twitter users are referenced by an at-mention, which introduces an uncertainty that the statement might refer to one of the linked users. We also find that the laymen tend to apply a more lenient interpretation of what is *disparaging* and consider many statements as non-disparaging, i.e. already an allegation of shortcomings[33], which could reduce the victim's social standing is disparaging in the legal sense.

The annotation results of the student assistants are encouraging for obtaining sufficient training data for a larger study on automated classification, i.e. a correct automated classification of the first two decisions would already be able to exclude many cases that do not have to be deleted based on the Network Enforcement Act.

6 Conclusion

We investigated which offenses found in German political Tweets constitute defamatory offenses under German criminal law, that social media operators are obliged to delete under the Network Enforcement Act. Following the dogmatic approach of civil law systems, we started with an analysis of the legal framework for defamatory offenses in the German Criminal Code along with its foundations in the balancing between the potential offender's freedom of expression and the potential victim's right to honor. We derived from this consideration a schema suited for data annotation consisting of a sequence of binary decisions to determine if a statement constituted a defamatory offense, which we used for annotating data. We find that the majority of the morally offensive postings in our dataset still contribute to the public discourse and are, hence, protected by the freedom of expression. We also investigated if laymen can be instructed to use this annotation schema to facilitate an inexpensive annotation of more data for classifier training. We find that laymen suited to the task can be found, but in particular the legal notions of a specific *group of persons* and the scope of what is considered *disparaging* are challenging for them.

In future work, we will investigate the usefulness of layman-annotated data for an automated classification. Furthermore, we will expand our work by investigating additionally the criminal offense of *incitement to hatred* (§ 130 StGB) and its implication on the freedom of expression.

Acknowledgements

We would like to thank Tatiana Günster for taking the time to provide us with a second legal opinion. Furthermore, we would like to thank Emilie Mathieu for helpful corrections and Michael Wojatzki for the helpful discussions about the user-study design.

[33] e.g. *I am not sure whether John knows what he's doing.*

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*.

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.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

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.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, pages 1735–1780.

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.

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.

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.

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.

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.

Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand. 2018. GermEval 2018: Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval 2018: Shared Task on the Identification of Offensive Language*, Vienna, Austria.

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.

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.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The International Workshop on Semantic Evaluation (SemEval)*.