Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection

Sylvia Jaki
University of Hildesheim
Institut für Übersetzungs- und Fachkommunikation
Universitätsplatz 1, 31141 Hildesheim, Germany
jakisy@uni-hildesheim.de

Tom De Smedt
University of Antwerp
Computational Linguistics Research Group
Lange Winkelstraat 40-42, 2000 Antwerpen, Belgium
St Lucas School of Arts Antwerp
Experimental Media Research Group

May 2018

Abstract
Discussion about the social network Twitter often concerns its role in political discourse, involving the question of when an expression of opinion becomes offensive, immoral, and/or unethical hate speech, and how to deal with it. This paper analyzes over 55,000 right-wing German hate tweets from the period between August 2017 and April 2018 to give an insight into what disparaging verbal behavior from extremist right-wing users looks like, who is targeted, and how. Since reported hate speech on social media now has to be deleted within 24 hours as per the recent NetzDG law, there is increasing need for technology that allows for an automated detection of hate speech. We intend to show that such technology is feasible today by proposing a method that automatically detects right-wing hate speech with 84% accuracy.

Keywords: hate speech, social media, natural language processing, machine learning
1 Introduction

During the 2015 European refugee crisis, nearly half a million refugees arrived in Germany (Eurostat 2017), or more than double of the year before, with Germany’s Chancellor Angela Merkel famously stating “Wir schaffen das” (‘we can do this’). Most of the refugees were young men from Muslim countries, reportedly including some Islamic State militants disguised as asylum seekers (Reuters 2016). During this time, the country also witnessed a number of violent incidents, such as the 2015 New Year’s Eve sexual assaults, where groups of young male refugees sexually assaulted women in Cologne, or the December 2016 Berlin attack, where an Islamic State militant drove a hijacked truck into a Christmas market, killing 12 and injuring 56. This sharply polarized the German sentiment towards refugees (YouGov 2016). In wake of these events, the 2017 German federal elections experienced a considerable rise in right-wing populism, with the political party Alternative für Deutschland (AfD) achieving a striking success with 12.6% of the votes.

Social media such as Twitter and Facebook are believed to have played an important role in the electoral debate, mainly in propelling the increasingly polarizing rhetoric. To illustrate this, in prior work Conover, Ratkiewicz, Francisco et. al. (2011) used a cluster analysis to examine 250,000 Twitter messages (tweets) posted in the lead-up to the US congressional midterm elections, and found a highly segregated partisan structure with few retweets between left and right-wing Twitter users. Davey & Ebner (2017:23) analyzed online right-wing partisan structure for the 2017 German federal elections, and report an increased connectivity between far-right activists and their followers. Such segregated “echo chambers” of like-minded users lend themselves to more extreme sentiments than would be the case in face-to-face interactions (Colleoni, Rozza, & Arvidsson 2014). Ebner claims that only a small minority of German Facebook users (5%) are responsible for most of the online hate speech.¹ These users can mostly be traced to AfD and the Austrian Identitäre Bewegung (IB), and attempt to stage the impression that topics such as the relation between immigration and crime preoccupy a much larger majority.

¹http://faktenfinder.tagesschau.de/inland/hasskommentare-analyse-101.html
Another recent study by Müller & Schwarz (2017) shows a significant correlation between increased German hate speech on social media and physical violence towards refugees in Germany. The NetzDG law (Netzwerkdurchsetzungsgesetz), effective in Germany as of October 1, 2017, now forces social media networks to remove reported offensive content within 24 hours, with remarkable consequences such as one AfD politician being temporarily suspended from Twitter. For the most part, the response by IT companies has been cautious. Twitter has argued that ‘no magic algorithm’ exists for detecting problematic content (Twitter 2016), and Facebook has stated that it will take another ‘5 to 10 years’ to develop the proper AI detection systems (Reuters 2018).

While far from perfect, we were able to automatically detect hateful German content with about 84% accuracy. In this paper, we discuss the steps taken to collect our data (section 2), we provide a qualitative and quantitative analysis of what constitutes right-wing German hate speech from a linguistic perspective (sections 3 and 4), and we outline the technical details of our machine learning approach (section 5). Finally, we discuss some of the societal issues with implementing such technology (section 6).

2 Data Collection

Between August 2017 and April 2018, we have collected a corpus of over 55,000 hateful tweets from 100+ right-wing German Twitter users, using the Pattern toolkit (De Smedt & Daelemans 2012) and the Twitter API. A subset of over 20,000 anonymised tweets and 2,000 images has been made publicly available for academic study in the POLLY corpus (De Smedt & Jaki 2018). To collect the data, we manually identified 112 subversive Twitter profiles and then automatically collected their tweets. Subversive profiles are those that (1) post tweets containing racial slurs, profanity, and violent rhetoric, (2) do so repeatedly and consistently, and (3) indicate far-right ideology, either directly (i.e., in the profile description) or indirectly (e.g., through coded language or visual metaphors). For each profile, we started by searching Twitter for derogatory keywords, such as Neger (‘nigger’), and examined the search results. If a profile posted vigorously about refugees, and had three or more tweets that could be considered offensive or illegal, it entered the pool. Many of these profiles retweet each other extensively, which supports Conover et al. (2011) and Davey & Ebner (2017).
Over 25% of the profiles in our dataset contain neo-Nazi or extremist right-wing cues in their username. For example, some usernames contain neo-Nazi ciphers and abbreviations. Five usernames end in -88, a substitution cipher for HH or Heil Hitler (Turner-Graham 2015). Two end in -18 (AH, Adolf Hitler) and another two in -59 (EI, Eil Hitler). One username contains two lightning emoticons for SS, which in combination with the historic pictures posted by the user is an obvious reference to Third Reich SS. Two usernames contain an IB-prefix, referring to the Austrian Identitäre Bewegung (Hentges, Kökgiran, & Nottbohm 2015). Three usernames contain references to obscure Nazi occultism, e.g., mentioning Thule Society or the Holy Grail (Goodrick-Clarke 1993). Four usernames can be associated with Teutonic mythology, mentioning Norse gods of war such as Thor and Freya, or Norse mythological creatures such as frost giants. Three usernames mention the Black Front, the Aryan race, or nerve gas, and two profiles support neo-Nazi music bands and hooliganism. Other cues of right-wing extremism are racial slurs, used in three usernames, and grandiose proclamations such as being a ‘Guardian of Germany’ or a ‘Defender of the West’ (another three).

Some of these cues could be coincidental, especially the ciphers. For example, -88 might be a reference to a birth date. But this is ruled out by examining the user’s profile picture and profile content. Several of these users also have a profile picture with a völkisch rune (Mees 2008), crusader heraldry (Takács 2015), Hitler parodies, portraits of or quotes by Nazi frontmen, or various other explicit imagery, such as German landmarks burning in flames, rioting, skulls, wolf packs, or medieval knights. Additionally, most profiles have a description with anti-refugee statements (e.g., Islamkritiker, Gegen Islamismus), pro-national socialism (National Sozialistisch, NS Jetzt), or pro-white power statements (White Power Worldwide).

We also collected a companion corpus of 50,000 ‘safe’ tweets for comparison. These include about 20,000 tweets by more than 35 elected German politicians, who we expected to talk about relevant topics without posting offensive content, and 30,000 German tweets by as many Twitter users, talking about family, work, holidays, and so on.

Despite assertions by the EU that IT companies now remove over 70% of reported hateful content (EU Memo 18-262, January 2018),\(^2\) over 90% of the profiles in our hate speech corpus were still online in April 2018. In the 9-month period from August to April, we observed less than 10 suspensions out of 100+ subversive profiles.

---

3 Qualitative Analysis

To establish a more comprehensive picture of what online hate speech looks like and how it is used in context, a quantitative analysis of political communication on Twitter should go hand with a qualitative approach (Pal & Gonawela 2017). To this end, we have examined a random subsample of 2000 tweets from our corpus to address the following questions: (1) When is a tweet hate speech? (2) Who is targeted in those tweets? (3) What kind of speech acts are used to disparage groups or individuals? (4) How is figurative language used? (5) What does the rhetoric reveal about the users? The examples we present are either parts of tweets or tweets as a whole, and will be noted as they are, i.e., including grammatical or spelling mistakes.

Not all the tweets in the sample are hateful. There are also non-disparaging politically motivated messages, such as “Heute wird es sich entscheiden, Freunde!” (‘today it will be decided, my friends’), tweeted as a lead-up to the German federal elections. Likewise, users may simply post holiday photos or extend good wishes, as in “Schönen Urlaub dir/euch!” (‘have a nice holiday’). Non-political posts are less likely to be offensive, with the exception of football tweets like “Herrlich raus du bist ein schieß trainer” (‘Herrlich, get lost you crappy coach’), which is due to a connection between right-wing extremism and football ultras (Pilz n.d.). But non-political tweets are rare in these profiles (<10%), suggesting that Twitter is used purposefully for propaganda. The majority of tweets relate to political parties and political ideology, or address immigration and purported crimes by refugees.

3.1 Hate Speech Definition

According to the EU-memo 18-262, IT companies have made considerable progress in removing online hate speech. But a EU-wide and legally binding definition of hate speech is still in the offing, placing the responsibility in the hands of private companies. The absence of a legal framework also presents problems for automatic hate speech detection systems: What exactly are we detecting? In part, defining hate speech is so difficult because we are dealing with a highly heterogeneous phenomenon that does not even necessarily always involve attitudes of hatred (Meibauer 2013:3; Brown 2017:432). Many instances of online group-focused enmity are currently in limbo between legally acceptable and illegal language use (Bundeszentrale für politische Bildung 2017).
The Encyclopedia of the American Constitution defines hate speech as “any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic” (Nockleby 2000). As Sponholz (2018:48) sums up, hate speech can be seen as the communicative production of human inferiority. Under German criminal law, illegal forms of hate speech include incitement to criminal behavior (Öffentliche Aufforderung zu Straftaten, §111 StGB), incitement of the masses (Volksverhetzung, §130 StGB), insult (Beleidigung, §185 StGB), defamation (Üble Nachrede, §186 StGB), slander (Verleumdung, §187 StGB), coercion (Nötigung, §240 StGB), and intimidation (Bedrohung, §241 StGB) (Puneßen 2016). A form that is not usually mentioned in this context, but also relevant, is defamation or slander of a politician (Üble Nachrede und Verleumdung gegen Personen des politischen Lebens, §188 StGB) (Griffen 2017). There is a conflict between the new NetzDG law, which demands that illegal hate speech online be censored, and §5 of the German Grundgesetz, which protects freedom of speech and prohibits censorship. However, as per §5(2), the latter principle is only valid as long as it does not infringe on other rights such as personal integrity.

Table 1 shows a number of tweets in our corpus that qualify for a discussion in terms of the above-mentioned articles, and NetzDG would demand their expedited removal:

<table>
<thead>
<tr>
<th>CASE</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>INCITEMENT (CRIME)</td>
<td>DE: Findet diese Drecksau und entmannt sie an Ort und Stelle.</td>
</tr>
<tr>
<td></td>
<td>EN: Find the pig and kill him.</td>
</tr>
<tr>
<td>INCITEMENT (MASSICES)</td>
<td>DE: Wir müssen uns wehren sonst sind wir bald Fremde.</td>
</tr>
<tr>
<td></td>
<td>EN: We have to fight back now or we will soon be foreigners in our own country.</td>
</tr>
<tr>
<td>INSULT</td>
<td>DE: Verpiss dich aus Deutschland du scheiss Kreatur.</td>
</tr>
<tr>
<td></td>
<td>EN: Piss off and leave Germany, you nasty creature.</td>
</tr>
<tr>
<td>SLANDER</td>
<td>DE: Steigt Motumbo aus der Bahn, ist er sicher schwarz gefahnt.</td>
</tr>
<tr>
<td></td>
<td>EN: Mutombo will surely have no train ticket.</td>
</tr>
<tr>
<td>SLANDER (POLITICIAN)</td>
<td>DE: Schulz muss erst mal das Saufen aufgeben, zum Wohle seiner Partei.</td>
</tr>
<tr>
<td></td>
<td>EN: Schulz needs to give up drinking for the sake of his party.</td>
</tr>
<tr>
<td>INTIMIDATION</td>
<td>DE: Dir hau ich sowas von gepflegt in die Fresse.</td>
</tr>
<tr>
<td></td>
<td>EN: I’ll smash you right in the face.</td>
</tr>
</tbody>
</table>

Table 1. Examples of hate speech compared to German criminal law.
The difference between legal and illegal, as well as distinguishing between different legal forms, is often ambiguous in right-wing hate speech, especially with incitement of the masses, defamation, and slander. While the StGB clearly differentiates between the simple spreading of defamatory information and knowing that this information is incorrect (slander), the distinction is hard to make in reality when it comes to hate speech on Twitter.

The tweets we examined do not contain examples of coercion. Many cases involve a tendentiously framed dissemination of news, since users interpret news by connecting events to personal experience and world views (Maireder & Ausserhofer 2014:307). By framing we understand ‘select[ing] some aspects of perceived reality and mak[ing] them more salient in a communicating text’ (Entman 1993:52). For example, a crime committed by one refugee may be framed as something that all refugees do all the time. This interpretation is often explicitly stated, as in “Dieser Mehrfach-Straftäter ist der Beweis: Multikulti funktioniert... Nicht!” (‘this multiple offender proves that multiculturalism does not work’), where one criminal act is used to insinuate that multiculturalism automatically leads to an increased crime rate. The reported event may be real news or fake news.

**Real news.** At a glance, many tweets appear to be factual, as in “Dieser Mann schnitt seiner zweijährigen Tochter die Kehle durch” (‘man cuts throat of baby girl’), referring to a news article about a Pakistani man. Citing crime reports is not hate speech of course, but this user and many others in the corpus do so repeatedly and exclusively about people of foreign origin. On average, a profile in our dataset posts a new tweet every three hours. By comparison, German politicians posted a new tweet every six hours during their electoral campaign. The users in our dataset provide a continuous stream of negative tweets that implicate foreigners in criminal and violent activities, which can be called propaganda.

**Fake news.** This stream of crime reports often includes hearsay and prejudice. In the April 2018 Münster attack, a man drove a van into the crowd in front of a restaurant, killing 4 and injuring more than 20. One AfD politician swiftly responded by tweeting “Wir schaffen das! 🚖”, suggesting that the incident involved refugees, but the perpetrator turned out to be a German with psychiatric problems. Many users in the corpus tweet about non-existent jihadist attacks, in this case for example: “Das Wetter lockt die Menschen raus, da lohnen sich #MuslimAttacken mit Autos wieder mehr. Deutsche wacht endlich auf!” (‘nice weather lures people out, and Muslim car attacks become lucrative again, wake up, Germans!’).
Another user responds: “Islam Terror unterstützt von Ferkel Vasallen und Gutmenschen. Die paar Toten machen ja nichts, hauptsache der Islam wird geschützt” (‘Islamic terror supported by Merkel pigs and moralists, a few dead people is fine as long as Islam is protected’). The main message here is that crime is directly attributable to Merkel’s welcoming policy.

3.2 Targeted Groups

Hate speech is disparaging by nature, but disparaging tweets do not always use disparaging language, e.g., der friedliche Islam (‘peaceful Islam’) is used as sarcasm. However, they do often resort directly to disparaging words, such as insults or highly negatively connoted words. These linguistic cues are very useful to expose who is targeted by politically motivated hate tweets, namely immigrants, political opponents, and other German voters or Germany as a whole. The dataset includes more targets than these, for example women, especially feminists, and homosexuals, but those tweets are far less frequent. The following overview gives a non-exhaustive list of designations for the three main groups:

Immigrants and specifically refugees are considered to be the major source of danger to Germany’s safety, which is why hate tweets targeting this group are abundant (for related work on online comments, see Geyer 2017). Immigrants are often designated as Nafri,3 Invasoren (‘invaders’), Asyltouristen (‘asylum tourists’), Merkel-Gäste (‘Merkel’s guests’), Mob, and occasionally also as Illegale (‘illegal residents’), Wohlstandsflüchtlinge (‘fortune seekers’), Bunte (‘multi-colored people’), or Zudringlinge (‘intruders’). They are described as being kriminell (‘criminal’), unterentwickelt (‘primitive’), Abfall (‘trash’), Müll (‘garbage’), Abschaum (‘scum’), Pack (‘vermin’), Parasiten (‘parasites’), and Gesindel (‘rabble’). Expressions such as Musel, Islamesianer, Hamudi, Salafistenschwester, Kampf muslimas, Burka-Frauen, or Vollbärte (‘full beard’) are used in relation to Muslims, Mutombo, Bongo, or Kloneger (toilet nigger) when it comes to African immigrants.

Political opponents of the users include individuals from the entire political spectrum, but most notably the SPD politicians Martin Schulz (Arschkriecher ‘brown-noser’), Heiko Maas (Vollpfosten ‘idiot’), and Ralf Stegner (Einzeller ‘single-cell organism’), the CDU politician Stanislaw Tillich (Bodensatz ‘dregs’) and, with little surprise, Chancellor Angela Merkel (Volksverräterin ‘betrayer of the people’, Bauerntrampel ‘yokel’, blöde Kuh ‘stupid cow’).

---

3 Nafri is a modern short form for Nordafrikaner (North African) or Nordafrikanischer Intensivtäter (North African intensive offender), which was initially used as an abbreviation in police slang, but then started spreading to general language.
More broadly, opponents include all left-leaning politicians and political parties (Verbrecher ‘criminals’, Sozi Clowns ‘socialist clowns’, SPD Heuchler ‘Social Democrat hypocrites’, linkes Faschistenpack ‘left-wing fascist vermin’, Grünfaschisten ‘green fascists’), which are described with adjectives like dummi (‘dumb’), gehirnamputiert (‘brainless’), and neokolonial (‘neocolonial’). They spread Gelaber (‘nonsense’), Idiotie (‘idiocy’), and Lügenpropaganda (‘propagandist lies’) in concert with the Lügenpresse (‘fake news media’).

Voters who have a positive attitude towards refugees, and mainstream voters in general, are called Gutmenschen (‘starry-eyed idealists’) and Traumtänzer (‘dreamers’). By consequence, Germany is in decline and referred to as a Buntland (‘rainbow nation’), Dummstaat (‘idiot nation’), PlemPlemLand (‘nation of fools’), Schandland (‘nation of shame’), and Berlin as a Bundeskloake (‘cesspit’) and Dreckloch (‘shithole’). Its citizens are perceived to be Idioten (‘idiots’), Schwächlinge (‘weaklings’), and verblödet (‘stupid’).

From a linguistic point of view, it is interesting to observe the creativity with which new derogatory political terms are coined: die Bundeskuh und ihr Idiotenstall (‘the Cowcellor and her stable of stupids’), Politmaden-Bürgermeister (‘maggot mayor’), Maas durchfallgesetz (‘Heiko Maas’ NetzDiarrhea law’), Erdowahn (‘Erdogoon’), Religioten (‘religiots’), etc. This is in line with findings for Italian data by Assimakopoulos, Baider, & Millar (2017:88).

3.3 Speech Acts

How do users try to persuade? An analysis of tweets in terms of speech acts allows for a focus on the intention of what is uttered, and there have been attempts for automatic speech act classification with Twitter data (e.g., Zhang, Gao, & Li 2011; Vosoughi & Roy 2016). Speech Act Theory (Searle 2008) distinguishes between different illocutions, i.e., meanings of utterances, and establishes assertives (e.g., ‘We are working on hate speech detection’), declarations (‘I hereby fine you 1500 Euros for illegal hate speech’), directives (‘Remove this profile!’), commissives (‘We promise to remove the profile’), and expressives (‘We wish we had less hate speech on our platform’). This variety of illocutions is also represented in hate speech (Sponholz 2018:65). The analysis of our data shows that hateful tweets are marked by some predominant types of speech acts, combinations of speech acts, and indirect speech acts, i.e., speech acts in the form of other speech acts.
Expressive speech acts are often the most aggressive and disparaging ones, as “wenn ich diese Kasperle in Ihren roten Clownkostümen sehe kommt mir das kotzen” (‘when I see these clowns in their red costumes, I could vomit’). They are often accompanied by a multitude of emojis, as in “mehr grüne 😖👍 solllen nicht sprechen dürfen! 👎👍😍” (‘no more talk by greens’), or by multiple emojis of the same kind to indicate distress, for example, “DANKE FRAU MERKEL 🙄👎💔👍”. Expressive speech acts are used to vent negative emotions about politicians, political events, crime, etc. The majority of them seems to be motivated by fear of change in Germany. Some are curses, like “Hoffentlich verrottet er in irgendeiner Ecke” (‘hope he rots in some dark corner’).

Directive speech acts are also common in our data, usually in combination with hashtags. For example, #AntiKap refers to the Antikapitalistisches Kollektiv, a network of right-wing militant protesters, and is used to remind members to gather for protests to disrupt the system, as in: “Den nationalen Aufbau unterstützen! #NSjetzt #KapitalismusZerschlagen #AntiKap” (‘Help us build our nation! National Socialism now, crush capitalism’). A second type of directive speech acts calls to stop the flow of immigrants, or remove them: “Obdachloser lebendig begraben! Man muss nicht lang überlegen, welche ‘Kultur’ hier wieder zugange war! Abschaum finden und abschieben!” (‘Homeless man buried alive! Doesn’t take long to figure out which “culture” is at it again, find the scum and deport it!’).

Assertive speech acts are frequent in the data sample, but sometimes complicated to distinguish from expressive speech acts as the tweeted news always shows a connection to one’s world view and hence emotions. In cases like “Erzbischof beschreibt Migration als Waffe zur Islamisierung Europas” (‘archbishop describes migration as a weapon for the Islamization of Europe’), however, the distinction is quite clear as we are definitely dealing with a statement about the world here. What is common is a combination of assertive and expressive speech acts (although various other combinations come up in the data too, but less systematically), where a piece of news is presented and then commented on. This occurs in the Obdachlosen-example above, but also in “Familiennachzug: De Maizière schlägt ‘Vorab-Vereinbarung’ vor und will ‘Spaltung der Gesellschaft überwinden’ Wenn ich schon das Transparent am Boden sehe, schwillt mir der Hals” (‘De Maizière wants to reuniote families. I get furious when I see the banner on the floor’). An accompanying photo shows a group of refugees and a banner stating that no human being is illegal.
**Commissive speech acts** are rare and involve exclusively threats, for example, “Erwischt es nur eine Freundin oder Bekannte von mir, leg ich den Pisser noch vorm Richter um” (‘if this happens to a friend of mine, I will kill the bastard in front of the judge’).

**Indirect speech acts** occur particularly when there is a blurring of facts and opinions. The utterance “Almans werden auf öffentlichen Plätzen von Nafris abgezogen” (‘blacks rob Germans at public places’) is such an indirect speech act, having an expressive disguised as an assertive. It is not a real statement about the world or a news item, but part of a series of tweets by one person that exclusively spreads various racial stereotypes. Another striking case of an indirect speech act is “Ich wünsche gute Heimreise” (‘safe travels home’). At first sight, it seems to constitute an amiable wish (thus an expressive speech act) but the illocution is, in fact, a demand, i.e., calling on refugees to go back to Syria. This becomes clear in the preceding part of the tweet, which claims that Islamic State in Syria has been annihilated.

### 3.4 Figurative Language

Instead of attempting to conduct a comprehensive metaphor analysis, including lexicalized metaphors, we will restrict the following brief observations to two cases with the apparent aim to render a claim more perspicuous: less lexicalized or highly intensifying metaphors on the one hand, and sayings, either in their original or modified form, on the other.

Metaphors may be used to directly target Angela Merkel: “Ich denke das in Honeckers Versuchslabor nen Fenster angeklappt war. Und da kam die Merkel raus” (‘a window in Honecker’s lab was open and out came the Merkel’). Honecker and Merkel both originate from the poorer (and often derided) eastern part of Germany. During the 1970’s, with the Iron Curtain in place, Erich Honecker was the leader of East Germany (Deutsche Demokratische Republik, DDR), after being imprisoned as a communist during the Nazi regime. Conspiracy theories suggest a connection between the two politicians. Another user remarks: “Honeckers Ziehtochter! Genähert mit Muttermilch von der Stasi!” (‘Honecker’s stepdaughter! Nursed with Stasi milk’). Here, Merkel’s background is linked to the DDR’s notorious intelligence agency. Some tweets picture Merkel or SPD’s Martin Schulz as the incarnation of evil, as in “rechte Hand des Teufels” (‘devil’s right hand’) or “bewaffnet euch, denn das böse unter der sonne will euch vernichten” (‘arm yourselves, the evil wants to destroy you’).
Metaphors may also be used to describe the current immigration policy, as in “Dämlicher Deutsche Politiker verlangen praktisch von Deutsche Bürger; während Dein Haus verbrennt, DU sollst Löschaktionen nicht bei Dir, sondern bei Nachbarhaus führen.” This user compares Germany to a burning house, while the Germans, spurred by foolish politicians, put out fires in neighboring houses (a possible reference to other EU countries or the Middle East). The fact that the number of refugees now exceeds the original guideline also elicits comparisons to biological processes such as fermentation: “Meine Gute… 😃….Das Gährt so sehr, das sich schon der Deckel hebt!!!!!🤔🤔🤔” ('my goodness, the fermentation process is so advanced that the fumes are lifting the lid'). All in all, descriptions of politics and politicians are far more creative when it comes to metaphors than comments on the refugees’ behavior, which often use more widespread dehumanizing metaphors, like the ones listed in section 3.2, particularly the parasite metaphor (see also Musolf 2015).

Another persuasion strategy consists of employing phraseological units such as proverbs, sayings, or film titles, either in their canonical or a modified form. Proverbs lend themselves especially well to the aim of persuasion, as they captivate “wisdom, truth, morals, and traditional views in a metaphorical, fixed and memorizable form and which is handed down from generation to generation” (Mieder 1993:24). The same goes for (the related category of) famous quotes (Geflügelte Worte). For example, “Der Krug geht so lange zum Brunnen bis er bricht” (‘a jug goes to the well until it breaks’) encourages AfD voters to keep fighting the system. “Wer Wind sät wird Sturm ernten!” (‘you reap what you sow’) is a comment on a news article reporting that pro-refugee organisations often face threats. Proverbs enable users to argue with easily understandable imagery that sums up complex matters in a concise form.

Today, proverbs are used in various modified forms, as in “Verfassungsschutz beobachtet und wenn sie nicht gestorben sin, beobachten sie noch heute”, a modification of the classic fairy tale ending ‘Und wenn sie nicht gestorben sind, dann leben sie noch heute’ (‘they lived happily ever after’). Another example, “Warum in die Ferne schweifen, wenn das Fremde liegt so nah?”, has the canonical form ‘Warum in die Ferne schweifen, wenn das Gute liegt so nah?’ (‘all you need is right here’). Both examples are clear sarcasm. The advantage of modifications is that they have a highly associative potential due to the underlying original form, that they are attention-grabbing, allow for adaptations to specific contexts, are a concise form for complex information, and/or highlight the producer’s intelligence (Jaki 2014:18).
3.5 The Users

As mentioned in section 2, users in the corpus can broadly be identified as belonging to the political far-right, either by the content of their tweets, and/or by references in their profiles (username, profile picture, profile description). Profile descriptions are particularly revealing about the image users want to project about themselves. Some state their mission explicitly, either by proclaiming to be a member of (for example) the recently founded right-wing network Reconquista Germanica, or by messages like “Stoppt den linken Faschismus und die Islamisierung Deutschlands” (‘stop left-wing fascism and the islamization of Germany’) and “Kein Islam, keine Überfremdung und Sozialschmarotzer in meiner Heimat” (‘no islam, no immigration and social parasites in my country’). Some users are less explicit, but give hints about how to interpret their tweets: “Politisch inkorrekt. Tweets können Spuren von Sarkasmus enthalten” (‘politically incorrect, tweets may contain sarcasm’) or “Polizei- und Nachrichtenmeldungen über übergriffige Flüchtlinge und Migranten. Kein Generalverdacht” (‘Information about criminal refugees and immigrants, no general suspicion’). Others again highlight their personal preferences (“ungläubig, islamophob, populistisch” ‘irreligious, islamophobic, populist’) or demonstrate their education (“Student der Philosophie” ‘student of philosophy’, “Betriebswirtschaftler und Zahlenmensch” ‘business economist with a knack for numbers’). Some add famous quotes that can be interpreted as political statements, such as “Wenn Unrecht zu Recht wird, wird Widerstand zur Pflicht!” (‘when injustice becomes law, resistance becomes duty’; Bertolt Brecht).

A contrast to some users’ self-proclaimed education is the deficient language of others’ tweets. However, since many users in the corpus post actively, misspellings can often be attributed to typos. Some tweets do not employ capitalisation, and others do not use it systematically, but this is a common feature of the medium. In social media, it is also common to write without punctuation, especially when emotionally upset: “ihr beiden könnt euch vom Acker machen ihr holzkörpf verpisst euch ihr lutschen 😖” (‘both of you can get lost, you idiots, piss off you wimps’). But there are also numerous mistakes that indicate a foreign speaker, which becomes obvious in tweets with major syntactic mistakes, as in “Unter Burka jede möglichen Figur kann sich verstecken. Daher Burkaträgerinnen sind undefinierbare Objekte *g* wie UFO ½” (Burka is missing an article, and the verbs should come directly after the adverbial and the adverb). Another indicator are specific types of
grammatical errors, such as flaws in grammatical gender, e.g. “ICH wünsche Trump langes Leben und genau die Gegenteil wünsche ich zum Merkel und die politische Bande; Muss reserviert für Die ganze Mafia Bande !! Der Zeit wird kommen!!!! Brauchen nur neues Wahl! Mfg .eine Ausländer” (apart from another missing article before langes, an error with the passive in reserviert and the drop of the pronoun wir before brauchen, most errors here are due to a non-matching article). This finding is interesting in so far as it indicates that German right-wing hate speech on Twitter is not restricted to native speakers of German.

However, language errors are certainly not restricted to non-native speakers, but may simply indicate a low written language proficiency. This is apparent in minor grammatical flaws explicable by the transfer of grammatical structures from spoken non-standard varieties to a written context, as in “Wem wundert das?” (dative case instead of accusative case) or “Weil die in ihrem Land verfolgt werden wegen ihrer Straftaten!” (where it should be ihr or at least ihren, which becomes increasingly acceptable for less formal contexts). Another indicator are blatant spelling errors, e.g., in “Und solche Invasoren holt man ins Land, wo die Kinder radikal erzogen worden sind, dass sie kaum Skrupel haben, auf Christen jagt zu gehen?” (jagt is third person singular of the verb jagen, while the noun is written Jagd and should be merged with Christen) or “Deine scheidß Kölnern sind woll nicht der drei fach Belastung gewagsen #scheissKölen” (the correct spellings would be wohl, Dreifachbelastung, gewachsen and the city name Köln).

4 Quantitative Analysis

4.1 Word Bias

We want to know exactly what words constitute hate speech. In statistics, the chi-square test examines how likely it is that an observed distribution is due to chance. For example, is it due to chance that the hate speech dataset contains a lot of derogatory words? In Computational Linguistics, the chi-square test is often used as a method for feature selection (Liu & Motoda 2007), to expose relevant keywords. More specifically, we can count the number of times that a word occurs in all tweets (50,000 hate + 50,000 safe), and then observe if it occurs more often in hateful tweets. Intuitively, we would expect function words such as der, die, das to occur equally often in any kind of text, since they are an essential grammatical requirement.
Contrarily, we would expect content words such as *Muslimenhorden* or *Kloneger* to be absent from the usual tweets that people post about their pets, their cooking skills, or their favorite TV stars. As it turns out, thousands of words (~4,500) are significantly biased (*p* < 0.05) and occur more often in the hate speech corpus, including racial slurs, ideological insults, and verbs expressing aggression.⁴

Table 2 shows a sample of biased words, with the total number of times that they appear (#), sorted by the number of times that they appear in hateful tweets (%), along with an example of use. Conversely, we can then also use these words as cues to automatically predict whether a text that we have not seen before is hateful or not.

<table>
<thead>
<tr>
<th>WORD</th>
<th>#</th>
<th>% HATE</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nafri</td>
<td>21</td>
<td>99%</td>
<td>nafri müll raus aus deutschland!!!</td>
</tr>
<tr>
<td>Salafisten</td>
<td>42</td>
<td>98%</td>
<td>Soso, die Salafistschwester ruf also zum Kampf auf 😊</td>
</tr>
<tr>
<td>sexuell</td>
<td>57</td>
<td>96%</td>
<td>islamische Prediger mit gewisser Neigung zu sexueller Gewalt</td>
</tr>
<tr>
<td>Horden</td>
<td>15</td>
<td>93%</td>
<td>bewaffnet euch vor den heranstürmenden Horden</td>
</tr>
<tr>
<td>Volk</td>
<td>518</td>
<td>90%</td>
<td>fahrt zur Hölle, das Deutsche Volk lässt sich nicht mehr verarzchen</td>
</tr>
<tr>
<td>Gutmenschen</td>
<td>160</td>
<td>90%</td>
<td>Gutmenschen ist es scheissegal wieviele Mädchen sterben</td>
</tr>
<tr>
<td>brutal</td>
<td>72</td>
<td>89%</td>
<td>brutale Rumänen überfallen Senioren</td>
</tr>
<tr>
<td>Patrioten</td>
<td>60</td>
<td>88%</td>
<td>Das ist eine Kampfansage gegen uns Patrioten. Es wird ernst.</td>
</tr>
<tr>
<td>töten</td>
<td>103</td>
<td>87%</td>
<td>ich würde Töten–bei Gott das schwöre ich!</td>
</tr>
<tr>
<td>Kopftuch</td>
<td>96</td>
<td>85%</td>
<td>Das habe ich auch schon gesehen.....Kopftuch in fettem X7-BMW.</td>
</tr>
<tr>
<td>Neger</td>
<td>27</td>
<td>85%</td>
<td>Was waren sie denn nun? Araber, Nafris, Neger?</td>
</tr>
<tr>
<td>😡</td>
<td>578</td>
<td>85%</td>
<td>Findet diese Drecksau und entmannt sie an Ort und Stelle. 😡😡😡</td>
</tr>
<tr>
<td>🖒</td>
<td>132</td>
<td>83%</td>
<td>im schutz seiner kettenhunde fühlt sich der stinker sicher! 😡 Hate</td>
</tr>
<tr>
<td>Muslime</td>
<td>438</td>
<td>82%</td>
<td>Muslime schlagen mit Stücken auf Zahörer ein</td>
</tr>
<tr>
<td>Abschaum</td>
<td>69</td>
<td>77%</td>
<td>Einwanderung von Religiösem Abschaum = Riesensproblem 😈���</td>
</tr>
<tr>
<td>Gefährder</td>
<td>128</td>
<td>75%</td>
<td>Gefährder und Kriminelle? Die sollte man überall hin abschieben.</td>
</tr>
<tr>
<td>sofort</td>
<td>344</td>
<td>69%</td>
<td>Sofort müßte ihr ein Blitz in den Arsch fahren dieser Volksverräterin</td>
</tr>
<tr>
<td>müssen</td>
<td>1264</td>
<td>66%</td>
<td>die Invasoren müssen lernen</td>
</tr>
</tbody>
</table>

Table 2. Sample keywords in hate speech.

⁴ 1,000 biased words: https://docs.google.com/spreadsheets/d/1JCYspKagNx0PSqy5YZ03t63ks1eoE1p5KzZ905BUkQU
4.2 Word Co-occurrence

Not all of these words are necessarily hateful by themselves, for example *Flüchtlinge* is not a derogatory term. Rather, a combination of them is. One way to expose this is by looking at biased adjectives that co-occur with (e.g., precede) biased nouns, as in *kriminelle Flüchtlinge*. Thousands of biased adjective-noun pairs can be found almost exclusively in the hate speech corpus (~3,500), each of them usually occurring once or twice. A sample is shown in Table 3. We have divided the given examples into six categories that correspond to the hate speech parameters defined in section 3.1. Some examples may fit into multiple categories:

<table>
<thead>
<tr>
<th>ISLAMOPHOBIA</th>
<th>RACISM</th>
<th>XENOPHOBIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>stinkenden Hasspredigern</td>
<td>dunkelhautiger Räuber</td>
<td>bewaffnete Araber</td>
</tr>
<tr>
<td>zwangsbeufuchte Mädchen</td>
<td>primitives Nergersesindel</td>
<td>ominös umtriebigen Südländer</td>
</tr>
<tr>
<td>gewaltbereiten Islamisten</td>
<td>bärtigen Teppichknutscher</td>
<td>gruppen Syrern</td>
</tr>
<tr>
<td>brutale Islamisierung</td>
<td>behaarte Kanakenfotzen</td>
<td>antisemitische Barbaren</td>
</tr>
<tr>
<td>dressige Salafistenpack</td>
<td>fetten Muselstrümpfe</td>
<td>nordafrikanischen Horden</td>
</tr>
</tbody>
</table>

Islamophobia constitutes hate speech based on religion, particularly involving a fallacious and negative generalisation of Islam, for example, Muslims being portrayed as extremist, violent, misogynist, and unclean.

Racism constitutes hate speech based on physical characteristics. This includes relating skin color to crime, portraying dark-skinned people as savage, and negative caricaturizations of Muslims (e.g., bearded carpet kissers).

Xenophobia constitutes hate speech based on nationality or ethnicity, for example, people of African or Eastern origin (*Araber, Barbaren, Kanake*) being portrayed as menacing and/or violent mobs or hordes.

---

5 1,000 biased adjective-noun pairs: https://docs.google.com/spreadsheets/d/1gVfKxOzLiv47WH506eDse6j96vD2Q4oFVTiVNU14c
Sexism constitutes hate speech based on gender, for example, German women portrayed as being either naive or infirmed, or otherwise as being inappropriately promiscuous (blöde Schlampe).

Extremism constitutes hate speech based on political ideology. In general, anyone that is not a right-wing proponent (Patrioten) is portrayed as foolish. Left-wing proponents (Linksgurken) are seen as deceptive, dumb and violent.

Stereotyping constitutes hate speech by means of exaggeration. Examples include unemployed immigrants loitering on street corners wearing jogging pants, greedy politicians, “waves” of terrorists, and homosexual predators in prison.

### 4.3 Word Dependency Graph

Derogatory adjective-noun pairs are often combined with verbs that express aggression, such as greifen (1,000×), schlagen (1,000×), kämpfen (750×), stechen (500×), and klauen (250×), for example, in “Dunkelhäutiger Unterhosenmann geht plötzlich auf Renterin los, schlägt und tritt sie fast zu Tode” (‘black man in underpants suddenly attacks elderly woman and nearly beats her to death’). The news quickly proliferates in the user’s network. The event is factual, the perpetrator had psychiatric problems, but the formulation is opinionated, suggesting that colored people are unpredictably dangerous. Another user is quick to comment: “Jetzt ist er in Nervenheilanstalt weil wir nicht zugeben dürfen wie kriminell die Nafirs sind, die aus den Gefängnissen Nordafrikas günstigst entsorgt wurden” (‘he is in the psychiatric ward because we can’t admit how criminal the blacks that we got from the African prisons really are’).

---

6: [http://www.nordbayerischer-kurier.de/nachrichten/mann-unterhose-attackiert-rentnerin_628300](http://www.nordbayerischer-kurier.de/nachrichten/mann-unterhose-attackiert-rentnerin_628300)
Figure 1 shows 10 biased verbs that express aggression, in relation to another 100 biased words (nouns, adjectives, or quantifiers) that often co-occur in the same tweet as the verb. The size of each word represents its frequency in the hate speech corpus. Verbs are marked with a black background, adjectives are displayed in uppercase, and quantifiers in italics. Such network structures are useful to study the vernacular as a knowledge-representation task (e.g., De Smedt 2013) to answer questions such as ‘who does what to whom’. In this case, the answer is typically: ‘refugees attack women’.

**Figure 1.** Sample dependency of nouns, verbs, and adjectives in hate speech.

4.4 Word Clusters

As noted earlier in section 3, the users in the hate speech corpus mostly seem to post tweets about immigration, crime, and politics. To verify this observation, we explored techniques to automatically group words into meaningful categories. We used skip grams (Mikolov, Sutskever, Chen et al. 2013) and spherical k-means clustering (Hornik, Feinerer, Kober et al. 2012) to create 3 clusters of the 250 most biased keywords ($p < 0.05$, % hate speech > 80%).
A skip gram models the context of a word, i.e., those words that frequently precede or succeed it. Intuitively, a clustering algorithm can then group words that often occur in the same context into the same category (e.g., gewalttätigen Muslimen, gewalttätigen Terroristen, gewalttätigen islamofaschistoiden).

Figure 2 shows a word cloud visualization of the three resulting clusters. The size of each word represents how often it appears in the hate speech corpus. With a little imagination, we can roughly label the respective clusters as immigration, crime, and politics. There is a large overlap between the clusters. In effect, many tweets in the hate speech corpus attempt to defame refugees, implicate them in crimes, and blame politicians, all in a single utterance.
4.5 Word Tree

Figure 3 shows the detailed context in which the word *Moslem* frequently appears:

![Sample word tree]

**Figure 3.** Sample word tree.

4.6 Word Polarity

Sentiment Analysis (Pang & Lee 2008) pertains to automatically detecting whether a text is factual or subjective, and if it is subjective, whether it is positive or negative (i.e., polarity). This task has a long history in Computational Linguistics, and many systems have been developed for many languages, usually with predictive accuracies that range between 75-85%. We used a system trained on 20,000 German tweets that contain emojis to predict polarity (e.g., ❤ = positive, 😡 = negative) with an accuracy of 86% (De Smedt & Jaki 2018). Overall, about 55% of hateful tweets are classified as negative, against 45% of non-hate tweets being classified as negative, or a difference of 10%. We also tested with the SentWS lexicon (Remus, Quasthoff, & Heyer 2010) that assigns scores to words (e.g., *gut* = +0.37, *schlecht* = −0.77), which yields comparable results with a 7% difference.
Broadly, we can argue that negative rhetoric is an intrinsic property of hate speech. For supportive evidence, see also De Smedt, De Pauw, & Van Ostaeyen (2018).

The 10 nationalities most often mentioned in hateful tweets are German, African (in large part Malinese, Somali, and Nigerian), Turkish, Israeli, Syrian, Russian, Afghan, Saudi, and Austrian, closely followed by Romanian, Iraqi, and Moroccan.

Nearly all coverage of Somalis is negative (93%), most coverage of Afghans (79%), Malinese (75%), Nigerians (73%), and also Turks (70%). The least negative content is about Austrians (48%) who are hailed for their right-wing FPÖ victory in 2017.

Hateful tweets contain significantly more references (a 2:1 ratio) to women and girls (Frauen, Mädchen) than to men and boys (Männer, Jungen). References to women as well as to men tend to occur more often in negative tweets than in positive tweets.

5 Automatic Detection

5.1 In-domain Evaluation

Using Machine Learning techniques, we can take advantage of word bias and word co-occurrence to train a system that automatically detects hate speech. Machine Learning is a field of Artificial Intelligence that uses statistical approaches to develop algorithms that can ‘learn by example’. For example, when shown a 1,000 German texts and a 1,000 English
In texts, a machine learning algorithm will infer that $\ddot{u}$ and $\beta$ seldomly occur in English texts. Such cues can then be used to predict whether another text is written in German or in English. In the same way, we can use examples of hate speech and examples of safe discourse to train a system to automatically spot the differences.

Machine learning algorithms expect their input to be vectorized, i.e., given as a set of vectors, where each vector is a set of feature-weight pairs. In this case, each tweet could be a vector, the features could be words, and the weights could be word count. We use character trigrams as features. A character trigram consists of three consecutive characters, e.g., Flüchtling $= \{ \text{Flö, lüc, üch, cht, htl, tli, lin, ing} \}$. Character trigrams efficiently model linguistic information such as spelling errors, word inflections, function words, punctuation flooding, and so on. For example, modelling Flüchtling and Flüchtlinge as character trigrams ensures that they have several overlaps, all except on the -nge trigram.

Character trigrams (CH3) give us about 82% predictive accuracy. We added additional features such as character bigrams (CH2), character unigrams (CH1 = punctuation marks, emojis), word unigrams (W1 = word lexicon), and word bigrams (W2 = word context) to boost the accuracy by 2% (Table 4). Other interesting features could be polarity and tweet frequency. We also anonymized usernames (@) to prevent overfitting. Overfitting means that the system starts to memorize the training data instead of discovering general patterns. For example, including usernames raises accuracy by 2%, but it also means that the system may turn into a blacklist of known profile names instead of a language analysis tool.

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH3</td>
<td>CH2</td>
</tr>
<tr>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4. Overview of precision and recall for different features.
We trained the single-layer averaged Perceptron algorithm (Collins 2002) with an overall accuracy of \textbf{84.21\%} (F1-score).\footnote{Proof-of-concept in Python code: https://gist.github.com/tom-de-smedt/9e9d9b9168ba703e0c336ee0128ebae5} The F1-score is the harmonic mean of recall and precision. Recall (R) is an estimate of how many hateful tweets the system is able to detect. Precision (P) is an estimate of how many tweets predicted as hateful are really hateful. For example, if the system marks every tweet that contains the word Flüchtling as hateful, its recall would be high but its precision would also be low, since many people talk about refugees without spreading hate speech. Recall and precision are obtained by training the system on a set of tweets, and then statistically testing its predictions on a different set. The results will include true positives, true negatives, false positives, and false negatives, where recall = \( \frac{TP}{TP+FN} \) and precision = \( \frac{TP}{TP+FP} \). We used 10-fold cross validation with 50,000 hateful tweets and 50,000 safe tweets, meaning that 10 tests were performed with a different 9/10 of training data and 1/10 of testing data each, then averaging the results.

In related work, Hartung, Klinger, Schmidtke et al. (2017) report 95\% recall and 25\% precision for classifying right-wing extremist hate tweets, with a system trained on about 16,000 tweets from 20 extremist profiles and 30,000 tweets from 17 non-extremist profiles. The discrepancy between their recall and precision likely stems from the unbalanced data. However, unbalanced data can also be useful, since more safe tweets and less hateful tweets reflect real-life. In the wild, 15\% of hateful tweets would slip through undetected by our system versus 5\% in Hartung et al.. But in Hartung et al., 75\% of flagged tweets would also need to be green-lit by human moderators versus 15\% in our system.

To offer an example of a false positive misclassification: our system was trained on multiple negative tweets about \#Diätenerhöhung (Diet increase, where Diet means the parliament) and appears to have learned that Diät (diet, as in weight-watching) is related to hate speech.

\subsection*{5.2 Cross-domain Evaluation}

In Machine Learning, domain adaptation refers to the problem that a system trained on one kind of data may perform poorly on other kinds of data. For example, our system is trained on tweets and might perform poorly on longer texts (e.g., subversive essays) because no training examples were ever provided for this kind of data. To assess the scalability of our system, we tested its performance on a number of out-of-domain resources:
• A hold-out set of a 1,000 tweets in the hate corpus was assessed by human annotators as the most offensive hate speech. This set was not used for training. We tested the trained system on the hold-out set and about 92% of the tweets are predicted as hateful.

• A hold-out set of a 1,000 tweets in the hate corpus was assessed by human annotators as not hateful upon careful inspection. About 76% of these tweets are predicted as safe.

• We collected a 1,000 web pages from a far-right conspiracy website.8 About 98% of this content is predicted as hateful.

• We collected a 100 random articles from the German Wikipedia, which is moderated for neutrality (NPOV).9 About 90% of this content is predicted as safe.

• The Brothers Grimm’s well-known Kinder- und Hausmärchen collection of fairy tales (1812) is predicted as 94% safe. The 6% hate might account for some of the fairy tales’ violent nature (Tatar 2003).

5.3 Applications

Without doubt, the accuracy of our single-layer neural network can be improved by adopting the most recent multi-layer neural networks (i.e., Deep Learning), especially when combined with word embedding techniques, and we invite other scholars to obtain our data to do so. However, deep neural nets also require more hardware resources and more training time.

Our system can be trained in less than 10 minutes using off-the-shelf consumer electronics. It is able to classify hundreds of tweets per second. It can be deployed today, to support human moderators of social networks. The system provides a confidence score for each prediction (e.g., 75% hate) which can be useful for sorting. Human moderators would see a list of content to review, with the most worrisome content at the top, not unlike a triage. Finally, our system uses an online learning algorithm, meaning that it can improve itself in real-time by observing the decisions made by the moderators.

---

8 http://wien.orf.at/news/stories/2901924
6 Discussion

The aim of this paper has been to examine online right-wing German hate speech from a linguistic perspective. We found a small, tight-knit community that spreads political propaganda, implicating refugees in criminal activities, using disparaging language based on race, ethnicity, and ideology. As demonstrated, a considerable part of this content is vitriolic and deceptive, and some of the tweets are illegal according to German law. Using Machine Learning approaches, we can accurately detect such hate speech, which can aid IT companies in enforcing laws like the German NetzDG.

However, at this point we come to an ethical dilemma that we, as linguists, must leave to policy makers and society to answer. Although most of the content in our corpus is morally questionable and underpins a worrying trend (i.e., a growing ‘us versus them’ narrative), others might argue that it falls within the boundaries of freedom of expression. Many people are also afraid that AI will restrain their communication, that any opinion that does not fit ‘the norm’ will be censored. Hence, a debate about the limits of freedom of speech ensues. A clear legal definition of what constitutes hate speech will certainly be helpful, both in this debate and in the execution of the German NetzDG.

In this explorative study we have focused on the verbal aspects of hate speech. A next step involves taking into account the multimodal nature of tweets, such as the relation between text and emojis and images, since these combinations are essential to fully capture the character of online political discourse. In future work, we aim to annotate the 5,000+ images that we have collected, employing computer vision techniques to boost our predictive accuracy by looking at combinations of texts and images.

We invite other scholars to obtain our data for further examination. A part of the data is already freely available, in anonymized form, in the POLLY corpus.
Bibliography


Conover, Michael D., Ratkiewicz, Jacob, Francisco, Matthew R., Gonçalves, Bruno, Menczer, Filippo, & Flammini, Alessandro (2011). Political polarization on Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, 89-96. AAAI.


