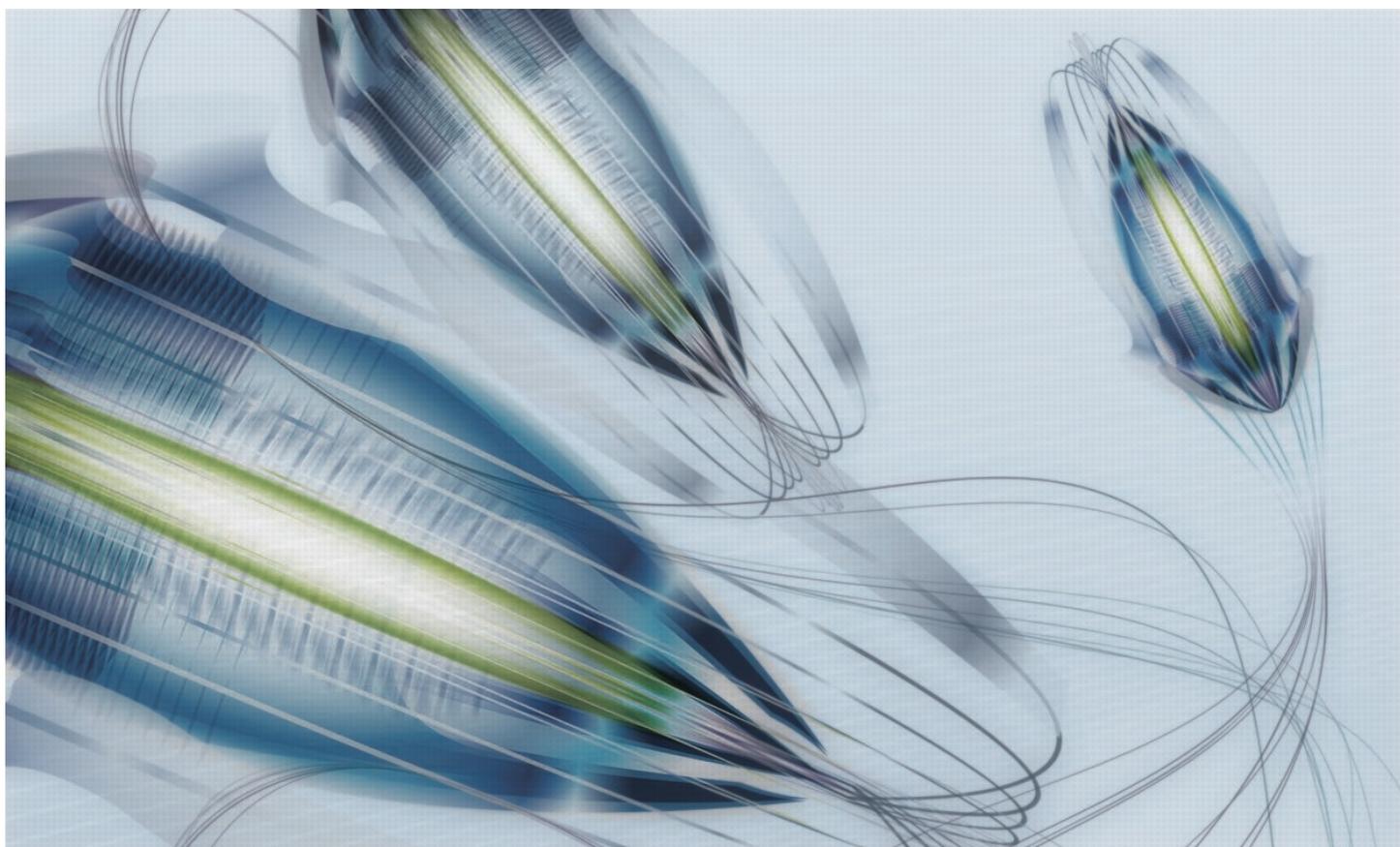


# Multilingual Cross-domain Perspectives on Online Hate Speech

Tom De Smedt, Sylvia Jaki, Eduan Kotzé, Leïla Saoud,  
Maja Gwóźdź, Guy De Pauw and Walter Daelemans



# Multilingual Cross-domain Perspectives on Online Hate Speech

Computational Linguistics & Psycholinguistics Research Center

CLiPS Technical Report Series (CTRS)

September 10, 2018

CTRS-008

## Authors

### **Dr. Tom De Smedt**

CLiPS Research Center, University of Antwerp (BE)  
tom.desmedt@uantwerpen.be

### **Dr. Sylvia Jaki**

Dept. of Translation & Specialized Communication, University of Hildesheim (DE)  
jakisy@uni-hildesheim.de

### **Dr. Eduan Kotzé**

Dept. of Computer Science and Informatics, University of the Free State (ZA)  
KotzeJE@ufs.ac.za

### **Leïla Saoud**

Dept. of Computer Science, University of Leuven (BE)  
ls@leilasaoud.com

### **Maja Gwózdź**

Dept. of Mathematics, University of Munich (DE)  
maja.gwozdz.mkg33@gmail.com

### **Dr. Guy De Pauw**

Textgain  
guy@textgain.com

### **Prof. Dr. Walter Daelemans**

CLiPS Research Center, University of Antwerp (BE)  
walter.daelemans@uantwerpen.be

# Multilingual Cross-domain Perspectives on Online Hate Speech

Tom De Smedt,<sup>1</sup> Sylvia Jaki,<sup>2</sup> Eduan Kotzé,<sup>3</sup> Leïla Saoud,<sup>4</sup>  
Maja Gwózdź,<sup>5</sup> Guy De Pauw,<sup>6</sup> Walter Daelemans<sup>1</sup>

<sup>1</sup> University of Antwerp, <sup>2</sup> University of Hildesheim, <sup>3</sup> University of the Free State,  
<sup>4</sup> University of Leuven, <sup>5</sup> University of Munich, <sup>6</sup> Textgain

**Abstract.** In this report, we present a study of eight corpora of online hate speech, by demonstrating the NLP techniques that we used to collect and analyze the jihadist, extremist, racist, and sexist content. Analysis of the multilingual corpora shows that the different contexts share certain characteristics in their hateful rhetoric. To expose the main features, we have focused on text classification, text profiling, keyword and collocation extraction, along with manual annotation and qualitative study.

**Keywords:** hate speech, social media, text analysis, text profiling, forensic linguistics

## 1 Introduction

Hate crimes have been on the rise,<sup>1,2,3</sup> and especially online social media are believed to act as a propellant for **polarization** and **radicalization**. Previous studies have argued that online social media can function as “echo chambers” (Colleoni, Rozza & Arvidsson, 2014), lending themselves to the expression of more radical views than face-to-face interaction. As such, hate speech is perceived to infiltrate various types of (mainly) political discourse online.

Hate speech is defined in the Encyclopedia of the American Constitution as “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). It is illegal according to Article 20 of the International Covenant on Civil and Political Rights.<sup>4</sup> However, in the US, freedom of speech is protected by the First Amendment of the United States Constitution and there is no exception that prohibits hate speech, unless it constitutes incitement to imminent crime (*Brandenburg v. Ohio*, 395 U.S. 444, 1969). In the EU, the Code of Conduct on countering illegal hate speech online<sup>5</sup> defines hate speech as “the public incitement to violence or hatred directed to groups or individuals on the basis of certain characteristics, including race, colour, religion, descent and national or ethnic origin”.

---

<sup>1</sup> <http://www.reuters.com/article/u-s-hate-crimes-rise-for-second-straight-year-fbi-idUSKBN1DD2BA>

<sup>2</sup> <http://www.reuters.com/article/uk-hate-crimes-surge-on-brexite-and-militant-attacks-idUSKBN1CM15E>

<sup>3</sup> <http://www.reuters.com/article/us-europe-rights-idUSKBN1672AB>

<sup>4</sup> <http://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx>

<sup>5</sup> [http://www.europa.eu/rapid/press-release\\_IP-16-1937\\_en.htm](http://www.europa.eu/rapid/press-release_IP-16-1937_en.htm)

Hate speech laws are now being considered in several countries such as the UK and France,<sup>6</sup> Germany (Netzwerkdurchsetzungsgesetz)<sup>7</sup> and South Africa (B9–2018).<sup>8</sup> Due to the pervasive nature of hate speech, such laws are often notoriously vague,<sup>9,10</sup> which can lead to potential problems with the infringement on freedom of speech of private citizens. The response by IT companies has also been rather cautious, for example with Facebook CEO Mark Zuckerberg stating before US Congress that the problem can be solved in “5-10 years”.

Hate speech online presents a new challenge for Natural Language Processing (NLP). The aim of this study is to identify common features of hate speech across domains, and to advance automatic detection. We collected 8 social media text corpora for a multilingual and cross-domain study of hate speech, composed of jihadism, right- and left-wing extremism, racism, and sexism (section 2). We provide an overview of the NLP techniques that we have used to gain insight from these resources, in particular using text classification, keyword extraction, collocation extraction (section 3) and stylometry and sentiment analysis (section 4).

## 2 Methods and Materials

### 2.1 Context

Table 1 provides an overview of our target domains, followed by a summary of the social and political context of each domain in respective countries:

DOMAIN	CONTEXT	LANGUAGE
JIHADISM	Islamic State (ISIS) propaganda	EN / AR
EXTREMISM	far-right activism in Germany	DE
EXTREMISM	far-right activism in Belgium & the Netherlands	NL
EXTREMISM	political debate in the UK / US / Canada	EN
SEXISM	incel subculture (male supremacy)	EN
RACISM	far-left activism in South Africa	EN
RACISM	far-right activism in France	FR
RACISM	racist comments in Belgium	NL

**Table 1.** Overview of target domains.

<sup>6</sup> <http://www.gov.uk/government/news/uk-and-france-announce-joint-campaign-to-tackle-online-radicalisation>

<sup>7</sup> [http://www.buzer.de/NetzDG\\_Netzwerkdurchsetzungsgesetz.htm](http://www.buzer.de/NetzDG_Netzwerkdurchsetzungsgesetz.htm)

<sup>8</sup> <http://www.justice.gov.za/legislation/hcbill/hatecrimes.html>

<sup>9</sup> <http://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law>

<sup>10</sup> <http://www.hrw.org/news/2017/02/21/south-african-move-hate-speech-step-too-far>

## JIHADISM

During the Syrian Civil War and the rise of the Islamic State (IS, ISIS, ISIL, Daesh) in Iraq and Syria, Salafi jihadists were able to set up an effective propaganda machine on online social media such as Twitter and YouTube (Tomé, 2015), spreading fundamentalist views on Islam, war reporting, execution videos, and terrorist manuals, to incite fear and hatred and to recruit new members (Klausen et al., 2012). For example, on August 19, 2014, a video was released on YouTube showing the beheading of abducted US journalist James Foley, to be met with worldwide condemnation. The EU subsequently announced a new Code of Conduct, and Twitter reported suspending over 125,000 user profiles related to ISIS.<sup>11</sup> Insurgents then moved to anonymous messaging apps such as Telegram (Weimann, 2016).

## EXTREMISM (EU)

In the wake of the Syrian crisis, high numbers of refugees arrived in the EU. For example, the number of refugees arriving in Germany more than doubled in 2015 (Eurostat, 2017). During this time, the country also witnessed a number of violent incidents with refugees, such as the 2015 New Year's Eve sexual assaults and the December 2016 Berlin truck attack. These events sharply polarized the public opinion (YouGov, 2016), which became clear during the 2017 German federal election, with a 12.6% victory for the far-right AfD. Since then, there has been a surge of German far-right propaganda on social media (Davey & Ebner, 2017), correlating with increased violence towards refugees (Müller & Schwarz, 2017). For example, after a stabbing involving an Iraqi and a Syrian in Chemnitz on August 27, 2017, hundreds of far-right protesters took to the streets, "fueled by fake news", "hunting down" immigrants.<sup>12</sup>

In Belgium, authorities have struggled with the prominent role of Belgian foreign fighters in Syria and Iraq, and their involvement in the November 2015 Paris attacks and the 2016 Brussels bombings (Van Ostaeyen, 2016), incidentally with US presidential candidate Donald Trump calling Brussels a "hellhole".<sup>13</sup> In 2010, the Belgian federal election already revealed a deep divide between the left wing and the right wing, which led to a year-long government formation crisis. The immigration debate, in particular concerning Muslims, is perceived to further polarize opinions (see Torrekens, 2015: 161). To illustrate this, on August 20, 2018, Belgian security services issued a remarkable press statement warning that far-right militias patrolled the streets "to protect citizens" and that "violent incidents could not be ruled out".<sup>14</sup>

---

<sup>11</sup> [http://blog.twitter.com/official/en\\_us/a/2016/combating-violent-extremism.html](http://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html)

<sup>12</sup> <http://www.dw.com/en/german-state-official-fake-news-fueled-chemnitz-riots/a-45263589>

<sup>13</sup> <http://www.politico.eu/article/donald-trump-brussels-is-like-a-hellhole>

<sup>14</sup> <http://www.sudinfo.be/id70744/article/2018-08-20/selon-la-surete-de-letat-lextreme-droite-se-renforce-en-belgique-plusieurs>

The EU Terrorism Situation and Trend Report (Europol, 2018) notes that violent right-wing extremism is expanding, “partly fuelled by fears of a perceived Islamization of society and anxiety over migration”, particularly in the case of the far-right Identitarian movement, which currently has branches in Germany, Austria, France, the Netherlands, and the US.

#### EXTREMISM (US)

In the US, a nation deeply divided by liberal and conservative views (Westfall et al., 2015), new President Donald Trump has successfully tapped into the concerns of conservative “angry white men” (Kimmel, 2017), for example by linking immigration to crime, globalization to unemployment, and politics to elitism.<sup>15</sup> His “post-truth” Twitter messages have been called sexist, racist, and contagious (Ott, 2017: 64), and have contributed to a strained relationship with the press, each accusing the other of proliferating fake news. A recent study has shown that, on Facebook, fake news stories tend to favor Trump (Allcott & Gentzkow, 2017).

Trump has repeatedly framed the press as an “Enemy of the People” (Figure 1), drawing criticism but also inflaming public sentiment. To illustrate this, in August 2018, the New York Times received a threatening voicemail declaring: “You’re the problem. You are the enemy of the people. And although the pen might be mightier than the sword, the pen is not mightier than the AK-47”. The Boston Globe received another call,<sup>16</sup> saying: “You’re the enemy of the people, and we’re going to kill every fucking one of you. [...] Still there faggot? [...] I’m going to shoot you in the fucking head later today, at 4 o’clock.”

The man charged with threatening the Boston Globe stated that he will continue to do so until the newspaper stops its “treasonous” attacks on Trump.

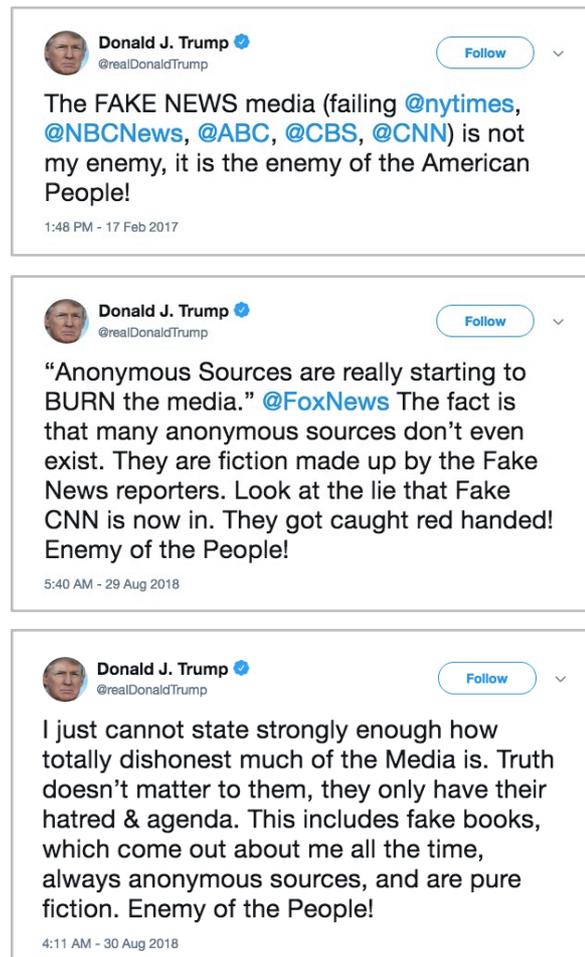


Figure 1. Example tweets by Donald Trump.

<sup>15</sup> <http://time.com/4566304/donald-trump-revenge-of-the-white-man>

<sup>16</sup> <http://time.com/5383632/donald-trump-editorial-death-threat>

## SEXISM

Two recent killing sprees in the US and Canada, the 2014 Isla Vista killings and the 2018 Toronto van attack, have brought the online male supremacist movement to the attention of the media. In particular, both perpetrators self-identified as *incels* or involuntary celibates, a subculture of “angry white men” that blame women for depriving them of sexual contacts, using violent, misogynist, homophobic, and racist vernacular (Ging, 2017). For example, one user on the Incels.me forum states: “Even if it’s shooting up a school, don’t sit back and do nothing”. This shows how hate speech is not necessarily always politically motivated, even though ties between male supremacy and the far-right have been observed (Nagle, 2017).

## RACISM (EU)

In France, anti-immigration and far-right political parties have steadily gained in popularity (Golder, 2016). In 2017, the Rassemblement National (formerly Front National) qualified for the second round of the presidential election and received 33.9% of the votes. The party has advocated against immigration, linking it to terrorism and crime (Davies, 2012) and fueling xenophobic sentiment in the public opinion, which has been perceived to contribute to the rise of racism on social media (Benveniste & Pingaud, 2016).

In Belgium, the far-right Vlaams Belang rose steadily to 19% of the votes in the 2007 federal election, but dropped to 5.8% in 2014, with voters turning to the right-wing N-VA (32.2%), in part because the N-VA is not subject to *cordon sanitaire* (an agreement between all parties not to form a coalition with Vlaams Belang). On August 30, 2018, a message was plastered on a mosque in Leuven stating: “Vote N-VA, brown people out”, along with Nazi symbolism.<sup>17</sup>

## RACISM (AF)

In South Africa, the left-wing ANC has 62.15% of the votes and the far-left EFF 6.35% (third-largest party). Recently, the government has begun advocating for more stringent land reform regulations (i.e., expropriation without compensation), to redistribute land owned by white Afrikaners (~72%) to black farmers. Commentators have subsequently noted an increase of online hate speech targeting white people (e.g., Steward, 2016; Khoza, 2017). At the same time, the country has experienced a number of farm attacks that may have been racially motivated (see Kerkvliet, 2017), while new hate speech regulations (Bill B9-2018) have been criticized for violating freedom of expression.<sup>18</sup>

---

<sup>17</sup> <http://www.vrt.be/vrtnws/en/2018/08/31/racist-slur-on-leuven-mosque-alle-bruine-buite>

<sup>18</sup> <http://www.hrw.org/news/2017/02/21/south-african-move-hate-speech-step-too-far>

## 2.2 Corpora

Table 2 provides an overview of the collected corpora:

DOMAIN	REGION	LANGUAGE	PERIOD	GENRE	SAMPLE
JIHADISM	Iraq & Syria	EN / AR	2014-2016	Twitter	50,000
EXTREMISM	Germany	DE	2017-2018	Twitter	55,000
EXTREMISM	Belgium & the Netherlands	NL	2017-2018	Twitter	30,000
EXTREMISM	UK / US / Canada	EN	2017-2018	Twitter	7,500
RACISM	South Africa	EN	2017-2018	Twitter	10,000
RACISM	France	FR	2017-2018	Facebook	10,000
RACISM	Belgium	NL	2015-2016	Facebook	5,000+
SEXISM	-	EN	2017-2018	Incels.me	65,000

**Table 2.** Overview of collected corpora.

Combined, the corpora cover jihadism (~20%), extremism (40%), racism (10%) and sexism (30%), although not every text message in every corpus necessarily constitutes hate speech. Most text messages are written in English (~45%), German (25%) and Dutch (15%), which are linguistically similar, complemented with messages written in French (5%), Arabic (5%) and assorted other languages (5%). Following is a summary of our data collections methods.

### JIHADISM (EN/AR)

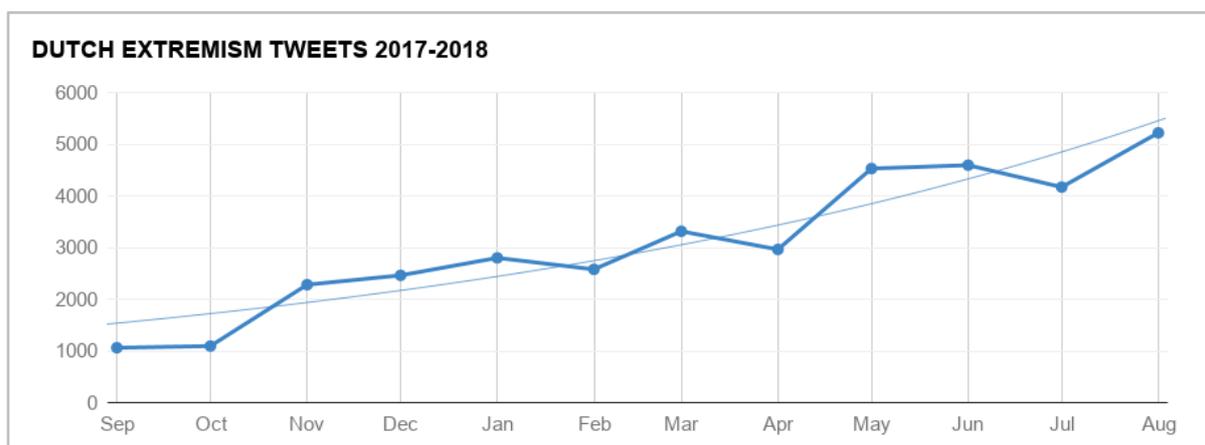
In the period from October 2014 to December 2016, we collected about 50,000 tweets posted by 350+ jihadists (De Smedt, De Pauw & Van Ostaeyen, 2018), in the wake of 10 terrorist attacks (e.g., Berlin, Brussels, and Paris attacks). For example: “May the streets of France be filled with the blood of these filthy kuffar [...]” (January 2015). The profiles were identified manually using a combination of search keywords (e.g., *kuffar*, unbelievers), cues in the profile usernames (e.g., *muhajir*, foreign fighter) and profile pictures (e.g., masks, weapons, or images of lions, which symbolize bravery). The tweets posted by these profiles were then automatically collected using the Pattern toolkit for the Python programming language (De Smedt & Daelemans, 2012a). About 40% of the tweets are in English, 30% are in Arabic and 5% are in French. For comparison, we also collected a companion corpus of 50,000 tweets by news agencies, imams, muslimas, etc., that talk about the Syrian crisis without overt bias.

## EXTREMISM (DE)

In the period from August 2017 to April 2018 we collected about 55,000 tweets posted by 100+ German right-wing extremists (Jaki & De Smedt, 2018). For example: “Das Pack gehört in Straflager” (the rabble belongs in the camps, March 2018). The profiles were identified manually using a combination of search keywords (e.g., *Neger*, *nigger*), cues in the profile usernames (e.g., *59 = EI*, shorthand for *Eil Itler*) and profile pictures (e.g., Hitler moustache parodies, knight crusaders, blue-eyed wolves). The tweets posted by these profiles were then automatically collected using Pattern. For comparison, we also collected a companion corpus of 20,000 tweets by well-known German politicians, who we expect not to post hate speech, and 30,000 random German tweets. About 35% of the data is publicly available in the POLLY corpus (De Smedt & Jaki, 2018a), along with other (anonymized) political German tweets.<sup>19</sup>

## EXTREMISM (NL)

In the period from September 2017 to August 2018, we used a manual set of high-precision Dutch keywords (e.g., *makakken*, offensive word for Moroccan immigrants) to continuously collect tweets and track trends in the data. This resulted in a corpus of about 30,000 tweets, and trending insights such as the spread of new offensive words such as *kansparel* (roughly: pearl of opportunity). For example: “Dank u Merkel, alweer een kansparel van u” (thanks Merkel, another one of your immigrants, August 2018). Figure 2 shows a timeline of tweets collected per month. In general, Dutch hate speech is on the rise. We can also correlate spikes in the data to real-world events. For example, on March 16, 2017, we collected 3x as many hate tweets as on other days in March, after it became known that a convicted Belgian foreign fighter could not be extradited because of human rights concerns.<sup>20</sup>



**Figure 2.** Timeline of collected Dutch extremism tweets.

<sup>19</sup> [https://docs.google.com/spreadsheets/d/1c5peNMjt24U0FcEMSj8gD\\_JjzumqXTWbPWa\\_yb2nNt0](https://docs.google.com/spreadsheets/d/1c5peNMjt24U0FcEMSj8gD_JjzumqXTWbPWa_yb2nNt0)

<sup>20</sup> <https://www.tijd.be/politiek-economie/belgie/federaal/francken-mag-veroordeelde-syriestrijder-niet-uitwijzen/9992783.html>

#### EXTREMISM (EN)

During Google Summer of Code 2018, we collected about 7,500 tweets (Gwóźdź, 2018) with political debate in the US (35%), the UK (35%), and Canada (20%), where each tweet has been manually annotated with keywords, speech acts (~10% directives), profanity (~15%), polarity (45% negative vs 10% positive), mood (15% angry vs 5% happy), gender (50/50 men vs women), political bias (50/50 left-wing vs right-wing), and metaphors (<1%; see also Sulis et al., 2016). For example, “@realDonaldTrump: You are a pathetic, disgusting excuse for a human being” is labeled as left-wing, negative, and agitated, with *pathetic* and *disgusting* marked as profanity. The data is publicly available in Google Sheets as the MAGA corpus.<sup>21</sup>

#### RACISM (FR/NL)

In February 2018, we collected about 60,000 public posts from Facebook newspaper pages such as La Dernière Heure and Le Figaro, using the Graph API (Saoud, 2018). For example: “Ton Coran tu te le mets ou je pense, vous n’etes que des barbares” (put your Quran where I’m thinking, you are nothing but barbarians). Over 10,000 of these were labeled by multiple annotators for racism (high agreement,  $K=0.92$ ), which constitutes about 15% of the subset. These were then used to extract a set of high-precision French racial slurs (e.g., *crouille*) for continuous monitoring. We also annotated over 5,000 posts in Dutch for racism ( $K=0.60$ ), collected from Facebook pages that are likely to attract racist content (Tulkens et al., 2016).

#### RACISM (EN/AF)

In the period from September 2017 to March 2018, we collected over 10,000 tweets with the hashtag #Orania (Kotzé & Senekal, 2018). For example: “That #Orania shit place must be burnt down, we want them back in Europe” (January, 2018). Orania is a minority community in South Africa that seeks to preserve Afrikaner culture and language amongst a majority of black African population. The community has been the subject of polarized political debate, often being accused by the far-left of upholding racial segregation.

#### SEXISM (EN)

In April 2018, we collected about 65,000 messages by 1,250 users in 3,500 discussion threads on the Incels.me forum (Jaki et al., 2018), after a 25-year-old incel killed 10 in the Toronto van attack. The corpus contains instances of misogyny (30%), homophobia (15%) and racism (3%). For example: “The slut DESERVED the murder 100%” (April, 2018). For comparison, we also collected a companion corpus of 40,000 paragraphs of Wikipedia texts (moderated for neutrality) and 10,000 random English tweets (to account for internet slang).

---

<sup>21</sup> [https://docs.google.com/spreadsheets/d/1mFV7uIEbMQ9LyaLRLQc-c0zVfKFn0CY\\_DakHSYWyPNg](https://docs.google.com/spreadsheets/d/1mFV7uIEbMQ9LyaLRLQc-c0zVfKFn0CY_DakHSYWyPNg)

## 2.3 Methodology

We applied a combination of known “superficial” NLP techniques (e.g., classification models based on character trigrams, keyword extraction by word frequency, sentiment analysis using dictionaries) on the collected corpora. The study was explorative, meaning that we did not apply all techniques to all datasets. Most of the work has been trial-and-error.

# 3 Results

## 3.1 Text Classification

Machine Learning (ML) is a subfield of AI that uses statistical techniques to train systems that “learn by example”. For example, given 10,000 HATE messages and 10,000 SAFE messages, the system will automatically discover that words or word combinations such as *filthy pigs* occur more often in hateful messages, which can then be used as cues to detect whether other (unknown) messages are hateful or not. In the process, each training example is mapped to a vector of features with weights. The features could be the words in a message for example, and the weights could be word count. In our case, we used character trigrams as features, where *filthy pigs* is mapped to { *fil, ilt, lth, thy, ...* }, so that the system can deal with spelling errors, word endings, etc., more efficiently (e.g, *filth* and *filthy* will have several overlaps). We usually also included character unigrams and bigrams, and word unigrams and bigrams, which often raises the predictive performance by a few percent.

Table 3 provides an overview of the predictive performance for each domain, measured by recall (cf., quantity, how many hateful messages are discovered) and precision (cf., quality, how many messages predicted as hateful are *really* hateful), using 10-fold cross validation:

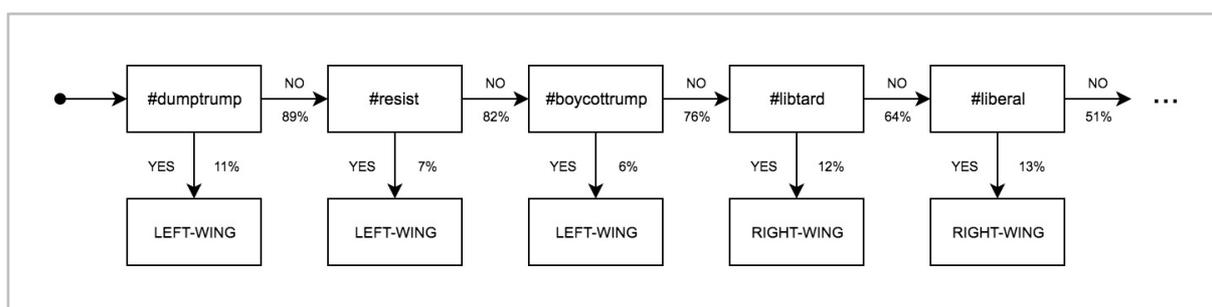
DOMAIN	CLASSIFICATION MODEL	TASK	LANGUAGE	PRECISION	RECALL
JIHADISM	support vector machine	HATE vs SAFE	EN / AR	82%	82%
EXTREMISM	single-layer perceptron	HATE vs SAFE	DE	84%	84%
EXTREMISM	decision trees	LEFT vs RIGHT	EN	82%	81%
RACISM	support vector machine	HATE vs SAFE	FR	84%	83%
RACISM	support vector machine	HATE vs SAFE	NL	39%	70%
SEXISM	single-layer perceptron	HATE vs SAFE	EN	93%	92%

**Table 3.** Overview of predictive performance (precision & recall).

For example, the LIBSVM algorithm (Chang & Lin, 2011) yields 82%  $F_1$  score (the mean of precision and recall) for jihadism detection, while Perceptron (Collins, 2002) yields 84%  $F_1$  for German right-wing extremism and 92.5% for sexism. We achieved the best results with a deep learning Convolutional Neural Network (CNN; Kim, 2014) for sexism: about 95%. This shows that it is useful to discuss automatic detection systems for hate speech. However, such systems are also problematic once deployed “in the wild”, because of their high error rate. For example, 80% accuracy means that 2 out of 10 SAFE messages are labeled HATE, or vice versa, raising ethical concerns. Two other main issues are scalability and interpretability:

**Scalability.** In a technical study on automatic detection of offensive language (De Smedt & Jaki, 2018b), we discuss how in-domain performance can be misleading when systems are applied to out-of-domain data. Classification models may be prone to overfitting, memorizing training examples instead of learning general trends, and consequently perform poorly during cross-domain evaluation. But cross-domain scalability is essential, since perpetrators will use countermeasures against identification (Berger & Perez, 2016), for example by adapting their language use. With the rising interest in addressing hate speech online, some systems have reported up to 99% accuracy.<sup>22</sup> This may be somewhat optimistic, since even policy makers and legal experts do not agree on what exactly constitutes hate speech across domains.

**Interpretability.** The best results are typically achieved using the most recent deep learning techniques. However, the decision-making process of these systems is often also more difficult to explain (“black box”), raising legal concerns. In this regard, systems built for deployment should adopt a multipronged approach, where the best classification models are accompanied by explanatory tools, such as decision trees to inspect the decision-making process, keyword dictionaries to highlight words for human moderators, and visualizations of language use like word clouds and word trees. Figure 3 shows a decision tree trained on the MAGA corpus:



**Figure 3.** Decision tree trained on the MAGA corpus (visualization).

<sup>22</sup> <https://i-hls.com/archives/81392>

Figure 4 shows the source code to train the decision tree on left-wing vs right-wing tweets:

```
# coding: utf-8
# http://github.com/textgain/grasp
from grasp import download, tmp, csv, DecisionTree, wc, tokenize, kfoldcv
# Google Sheets can be downloaded as a CSV-formatted string:
MAGA = 'https://docs.google.com/spreadsheets/d/%s/gviz/tq?tqx=out:csv&sheet=%s'
MAGA %= '1mFV7uIEbMQ9LyaLRLQc-c0zVfKFf0CY_DakHSYWyPNg', 'FULL'
# Each training example will be mapped to a {word: count}-dict:
def v(s):
    return wc(tokenize(s)) # "SAD!" => {'sad': 1, '!': 1}
data = []
with tmp(download(MAGA, cached=True)) as f:
    for r in csv(f.name)[1:]:
        if r[8].startswith(('left', 'right')):
            data.append((v(r[2]), r[8])) # (vector, label)
# Calculate precision & recall:
P, R = kfoldcv(DecisionTree, data, k=3, min=100)
print(P)
print(R)
```

**Figure 4.** Decision tree trained on the MAGA corpus (source code).

The source code uses the Grasp toolkit for the Python programming language.<sup>23</sup> It downloads the MAGA corpus from Google Sheets, maps each tweet to a word → count vector, each with a LEFT-WING or a RIGHT-WING label, and trains a decision tree with a minimum leaf size of 100. It then computes precision and recall, using 3-fold cross-validation. We get about 80%  $F_1$  score. As shown in Figure 3, nearly 50% of the labels was learned by hashtags. For example, any tweet that contains *#libtard*, but not *#dumptrump*, *#resist* or *#boycottrump*, will now be predicted as RIGHT-WING. The decision-making sounds plausible, however this model is not very scalable and may be outdated soon, since popular hashtags change all the time.

## 3.2 Keyword Extraction

Another way to interpret large volumes of text that worked well for us is keyword extraction, using feature selection techniques (Liu & Motoda, 2007). Given a HATE and SAFE corpus, we can use word count with a chi-square test to expose significantly biased words, and posterior probability to find out in what set a word occurs more often. For example, in the JIHADISM corpus, the word *kuffar* (unbeliever) is significantly biased ( $p < 0.05$ ) and occurs in about 3% of HATE tweets (1,500x) and 0.05% of SAFE tweets (25x). It is a useful cue for hate speech.

<sup>23</sup> <http://github.com/textgain/grasp>

By examining keywords across the collected corpora, we can identify a number of general mechanisms prevalent in all domains:

**Slang.** In-group expressions that are used to promote group identity by categorizing adversaries, such as *coconut* (Muslim apostate; brown outside, white inside), *gutmensh* (dogooder), *stacy* (attractive woman). Outsiders may not be aware of their meaning, but they will be used frequently and hence also stand out with keyword extraction.

**Slurs.** Disparaging expressions that are used to target adversaries by race, gender, sexual orientation, religion (e.g., *nigger*, *little bitch*, *infidel*, *faggot*), or to defame political opponents (e.g., *drunkard*, *dumbass*). Typically, a small variety is used repeatedly as a means of propaganda, defamation, or incitement.

**Dehumanization.** Disparaging expressions that are used to debase adversaries, to legitimize their removal or their extermination. Typically, this involves comparing individuals or groups to animals that are seen as unclean (e.g., *dog*, *pig*, *vermin*), describing them as waste (e.g., *scum*, *shit*, *trash*) or, predominantly in the case of misogyny, as objects (e.g., *cum dumpster*). Figure 5 shows a number of examples of dehumanization.

**Stereotyping.** Generalizing expressions that are used to frame heterogeneous individuals as a homogenous group, to argue that they are all accountable for reprehensible behavior based on a specific incident (e.g., a crime report). This includes using collective nouns (e.g., *horde*, *rabble*, *wave*) to describe immigrants, using social pejoratives (e.g., *bum*, *moron*, *whore*), and referring to historical conflicts (e.g., *barbarian*, *crusader*, *nazi*).

**Incitement** to crime and violence occurs less frequently, but it is also potentially illegal. It usually involves verbs expressing aggression (e.g., *fight*, *kill*, *rape*), directives (*arm yourself*), and indirect speech acts (*when will you learn?*) that are masked calls-to-action.



Figure 5. Examples of dehumanization.

Table 4 provides an overview of significantly biased HATE words ( $p < 0.05$ , non-exhaustive):

DOMAIN	SLANG	SLURS	DEHUMANIZING	STEREOTYPING	INCITEMENT
JIHADISM	<i>brother</i>	<i>kafir</i>	<i>dogs</i>	<i>crusader</i>	<i>burn</i>
	<i>coconut</i>	<i>murtadd</i>	<i>pigs</i>	<i>hypocrite</i>	<i>destroy</i>
	<i>lion</i>	<i>rawafid</i>	<i>scum</i>	<i>prostitute</i>	<i>slaughter</i>
EXTREMISM (DE)	<i>gutmensch</i>	<i>muselmann</i>	<i>gesindel</i>	<i>barbaren</i>	<i>abschieben</i>
	<i>linksgurke</i>	<i>nafri</i>	<i>pack</i>	<i>horden</i>	<i>bewaffnen</i>
	<i>merkelgast</i>	<i>neger</i>	<i>parasiten</i>	<i>salafisten</i>	<i>kämpfen</i>
EXTREMISM (EN)	<i>brexiteer</i>	<i>bitch</i>	<i>maggot</i>	<i>clown</i>	<i>kick</i>
	<i>libtard</i>	<i>fucker</i>	<i>puppet</i>	<i>hobo</i>	<i>nuke</i>
	<i>snowflake</i>	<i>soyboy</i>	<i>trash</i>	<i>moron</i>	<i>resist</i>
RACISM (NL)	<i>bakfietsbobo</i>	<i>bananeplukker</i>	<i>beesten</i>	<i>geitenneuker</i>	<i>afknallen</i>
	<i>kansparel</i>	<i>bruine aap</i>	<i>gespuis</i>	<i>kamelenpoeper</i>	<i>afmaken</i>
	<i>zwette</i>	<i>makkak</i>	<i>kakkerlakken</i>	<i>kopvod</i>	<i>uitroeien</i>
SEXISM	<i>beta</i>	<i>bitch</i>	<i>cum dumpster</i>	<i>nerd</i>	<i>beat</i>
	<i>chad</i>	<i>faggot</i>	<i>femoid</i>	<i>white knight</i>	<i>rape</i>
	<i>stacy</i>	<i>slut</i>	<i>it</i>	<i>whore</i>	<i>shoot</i>

**Table 4.** Overview of significantly biased HATE words.

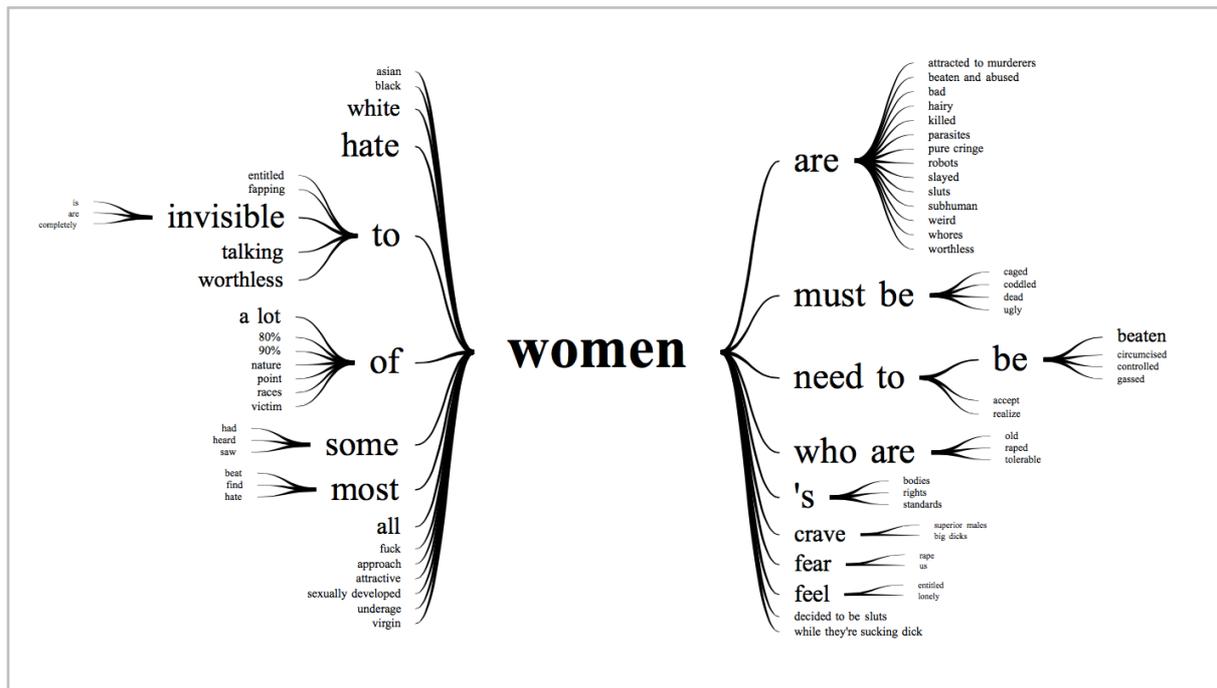
### 3.3 Collocation Extraction

Most words are not intrinsically hateful, but can become so in combination with other words (e.g., *bruine + aap*), and the variety of hateful word combinations that people can conjure up appears to be infinite. In this regard, collocation extraction (i.e., finding words that co-occur often) is a useful approach. One way to accomplish this is by using part-of-speech tagging, an NLP technique that identifies word types in context, and then counting what adjectives often precede which nouns (for example). In our corpora, this exposes hateful word combinations such as *white devil*, *black bitch*, *bearded carpet kisser* and *fat Muslim smurf*.<sup>24</sup>

A related approach, word embeddings (Mikolov, Yih & Zweig, 2013), is also useful. In brief, each word is mapped to a skip-gram vector that consists of the words that frequently precede or succeed it. This model can then be used to find semantically similar words. For example, we automatically expanded a manual dictionary of 100 hateful German words to 2,000 other hateful words (including word blends, e.g., *refutschies*  $\approx$  “*refusees*”) by querying the German Twitter Embeddings (Ruppenhofer, 2018). The resulting dictionary has 75%  $F_1$  score and is very easy to interpret, for example by highlighting words for human moderators.

<sup>24</sup> <https://docs.google.com/spreadsheets/d/1gVfkxOzLiv47WH506eDselji96vD2Q4ofFTiVNU14c>

Figure 6 shows the context of keyword *women* on Incels.me, as a word tree visualization:<sup>25</sup>



**Figure 6.** Context for keyword *women* in online misogyny.

Skip-gram vectors can also be combined with dimensionality reduction techniques, e.g., to reduce each vector to its two most significant features, after which they can be clustered and represented as a 2-D visualization. We experimented with spherical *k*-means clustering (Hornik et al., 2012) and t-SNE visualization (Maaten & Hinton, 2008). By visually inspecting such clusters, we observed that right-wing extremism mainly focuses on immigration, crime and politics, and that the incel community is ethnically highly heterogeneous, contrary to the popular belief that male supremacy primarily involves white men. Such insights can usually be verified by qualitative analysis, which is more reliable but also more time-consuming.

## 4 Analysis

Not all online hate speech necessarily also involves hateful language. Some instigators simply share news updates (mainly crime reports), but only if these reinforce their worldviews, and then they will do so perpetually. For example, on an average daily basis, German right-wing extremists post 2x more tweets than German politicians did during their election campaign, which can be seen as a form of propaganda. On the other hand, such profiles may also simply show up more in the search queries that we used to collect the corpus.

<sup>25</sup> <https://developers.google.com/chart/interactive/docs/gallery/wordtree>

In this section, we provide a summary of NLP techniques that we used to profile instigators and their state-of-mind. In general, there is no archetypical instigator, e.g., infighting about the community's belief systems is common. Instead, we argue that **unregulated social media** seems to enable angry persons to band together and stir one another (Sunstein, 2018: 9).

#### 4.1 Demographic Profiling

Stylometry refers to NLP techniques that can be used to identify the (anonymous) author of a text, based on the topics discussed or cues in writing style (Argamon, Koppel & Pennebaker, 2009). For example, Pennebaker (2011) has shown that women tend to use more personal pronouns (e.g., *I, my, we*) to talk about relationships, while men use more determiners (*the*) and quantifiers (*few, most*) to talk about objects and concepts. We used the Textgain API<sup>26</sup> to predict the age (75% F<sub>1</sub>), gender (75% F<sub>1</sub>), education (80% F<sub>1</sub>) and personality (60% F<sub>1</sub>) of the profiles in the collected corpora. While age may vary greatly, there is some consistency in that most users appear to be 20-40 years old. They are more often male, and tend to be somewhat less educated, but this may be an exponent of spelling errors made by non-native speakers, or simply a disregard for proper spelling. Consequently, spelling errors can be a useful predictive cue (e.g., *kakerlakke ~ cockroashes*) in combination with other features.

#### 4.2 Psychological Profiling

The Linguistic Inquiry and Word Count dictionary (LIWC; Tausczik & Pennebaker, 2010) can be used to identify psychological categories for common words. For example, *cried* is labeled as Affective, Negative, Past, Sadness, and Verb. We applied LIWC to the MISOGYNY corpus and found that incels tend to express more negative emotions (e.g., anger, uncertainty), display more social anxiety, and avoid topics such as family, work, hobbies, goals, and beliefs, which we substantiated with more time-consuming qualitative analysis (see Jaki et al., 2018). The incels feel left behind by society, or as one user states: “Loneliness has followed me my whole life, everywhere. In bars, in cars, sidewalks, stores, everywhere. There’s no escape. I’m God’s lonely man.” The in-group recategorization mechanism of the community is then to blame the “corrupt system” for their lack of self-esteem, social status, job opportunities, and so on. We can wonder about whether this generalizes to instigators in other domains. For example, in Belgium, news coverage about job discrimination based on race or gender surfaces every now and then.<sup>27</sup> In this regard, Wrench (2016) offers more in-depth perspectives.

---

<sup>26</sup> <https://textgain.com>

<sup>27</sup> <http://www.brusselstimes.com/opinion/11647/the-integration-process-in-belgium-bureaucracy-inefficiencies-and-incompetence>

### 4.3 Sentiment Analysis

Many users that post hate speech are angry, for one reason or another. This is manifested in a negative intensity in their language use, e.g., adversaries are not simply *parasites*, but *fucking parasites*, and what they say or do is not just *nonsensical*, but *UTTER NONSENSE*, and *enraging*, *infuriating*, *unforgivable*, etc., reinforced with exclamation marks, all caps, and emojis such as the angry face, the vomit face, or the fist punch (see Figure 7).

Sentiment analysis refers to NLP approaches that detect whether a text is objective (facts) or subjective (opinions), and if it is subjective, if it is positive or negative (Liu, 2012). This is often accomplished with a dictionary of adjective scores (*good* +0.5, *bad* -0.5) or models trained on reviews and star ratings. Since negative adjectives occur frequently in hate speech, the sentiment score predicted by these systems is usually negative as well, which can be a good predictive cue. Figure 8 shows the average polarity in random Wikipedia articles and tweets vs hate speech.

Using open source sentiment dictionaries that report about 80%  $F_1$  score (e.g., De Smedt & Daelemans, 2012b), we found that jihadist hate speech is more negative than random tweets, that German right-wing extremism tweets are even more negative, as well as the majority of incel misogyny (~65%) and Dutch racism (~75%). In the South African domain, sentiment analysis was less accurate, for example because many open source dictionaries assign a negative score to *apartheid*, which occurs frequently in neutral South African political debate (e.g., “We explain what is Orania, how it was established after the end of *apartheid*” is mislabeled as negative). We did note that negative sentiment is less common in tweets that take a defensive view (i.e., pro Orania) than in tweets with and offensive view (against Orania).

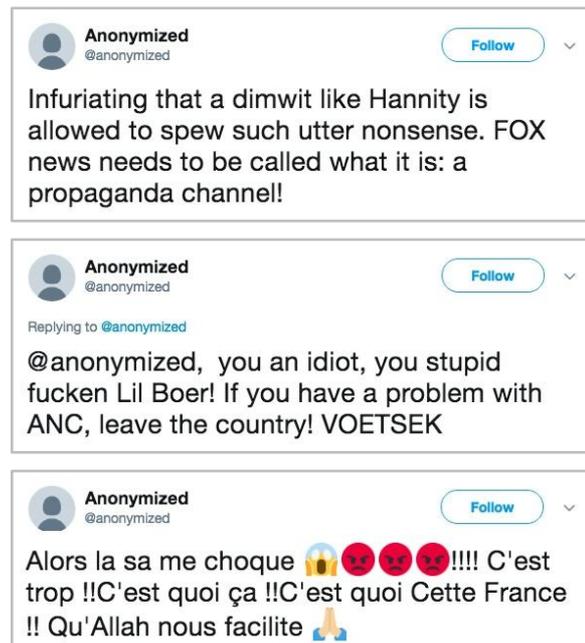


Figure 7. Examples of negativity markers.

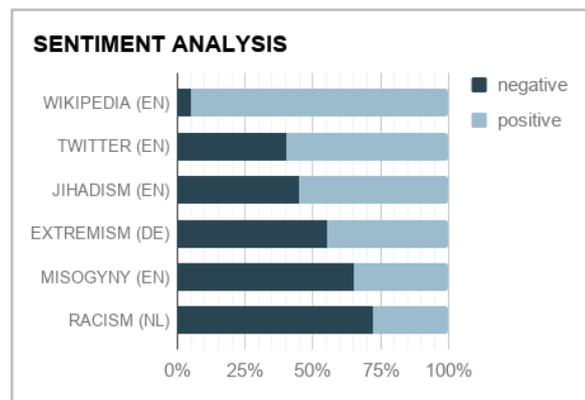


Figure 8. Sentiment polarity comparison.



## 5 Discussion

In this report, we have outlined general findings on online hate speech in different languages and domains. We demonstrated how detecting such content is theoretically viable, with an in-domain accuracy of about 80-95%. We believe that our study of the varied, yet often similar corpora can be relevant to future research on online hate speech, and may also have practical benefits, such as providing training material for detection and de-escalation of conflicts online. However, automatic detection of hate speech is challenging for a number of legal and ethical reasons. For one, hate speech laws often tend to be vague, in an attempt to capture a wide range of problematic content without infringing on the freedom of expression of private citizens. Without a well-defined legal framework, what exactly are we detecting? What is the threshold from sharing a personal opinion in a fit of anger to defamation and incitement? Also, the new GDPR<sup>28</sup> regulations now curtail the kind of data that can be collected in the EU (“right to be forgotten”).<sup>29</sup> Countering online hate speech using automated techniques may well be outside the legal realm, despite all good intentions.

Hate speech can target women and men alike, and any ethnicity, ideology or religion, so it is hard to come up with a linguistic definition. There is no standardized “list of bad words”, and if there is, then perpetrators are very creative in coining new offensive terminology. Online machine learning algorithms may be useful to continuously learn to detect evolving rhetoric, but even if such approaches are successful, perpetrators can always argue that they were only joking. Finally, developing useful technology tools does not mean that IT companies will adopt it, e.g., Facebook will not remove Holocaust denial.<sup>30</sup>

## Bibliography

- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2): 211-36.
- Shlomo Argamon, Moshe Koppel, James W. Pennebaker and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2): 119-123.

---

<sup>28</sup> <http://en.wikipedia.org/wiki/GDPR>

<sup>29</sup> [http://en.wikipedia.org/wiki/Right\\_to\\_be\\_forgotten](http://en.wikipedia.org/wiki/Right_to_be_forgotten)

<sup>30</sup> <http://www.reuters.tv/v/hzL/2018/07/18/zuckerberg-gets-heat-for-holocaust-comment>

- Matthew C. Benigni, Kenneth Joseph and Kathleen M. Carley. 2017. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLOS One*, 12.
- Annie Benveniste and Etienne Pingaud. 2016. Far-right movements in France: The principal role of Front National and the rise of islamophobia. In *The Rise of the Far Right in Europe*, 55-79, Macmillan, London.
- J. M. Berger and Heather Perez. 2016. *The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters* (technical report). George Washington University.
- Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3): 27.
- Elanor Colleoni, Alessandro Rozza and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2): 317-332.
- Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1-8, Stroudsburg, Pennsylvania.
- Jacob Davey and Julia Ebner. 2017. *The fringe insurgency: Connectivity, convergence and mainstreaming of the extreme right* (technical report). Institute for Strategic Dialogue, London.
- Peter Davies. 2012. *The National Front in France: Ideology, Discourse and Power*. Routledge, London.
- Tom De Smedt and Walter Daelemans. 2012a. Pattern for Python. *Journal of Machine Learning Research*, 13: 2063-2067.
- Tom De Smedt and Walter Daelemans. 2012b. "Vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 3568-3572, Istanbul.
- Tom De Smedt, Guy De Pauw and Pieter Van Ostaeyen. 2018. *Automatic Detection of Online Jihadist Hate Speech* (technical report). University of Antwerp.
- Tom De Smedt and Sylvia Jaki. 2018a. The Polly corpus: Online political debate in Germany. In *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*, Antwerp, Belgium.

- Tom De Smedt and Sylvia Jaki. 2018b. Challenges of automatically detecting offensive language online: Participation paper for the GermEval Shared Task 2018 (HaUA). In *Proceedings of GermEval Task 2018*.
- Europol. 2018. European Union terrorism situation and trend report.  
<http://www.europol.europa.eu/activities-services/main-reports/european-union-terrorism-situation-and-trend-report-2018-tesat-2018>
- Eurostat. 2017. Asylum and first time asylum applicants by citizenship, age and sex: Annual aggregated data (rounded). Retrieved July 2018.  
[http://ec.europa.eu/eurostat/en/web/products-datasets/-/MIGR\\_ASYAPPCTZA](http://ec.europa.eu/eurostat/en/web/products-datasets/-/MIGR_ASYAPPCTZA)
- Debbie Ging. 2017. Alphas, Betas, and Incels: theorizing the masculinities of the Manosphere. *Men and Masculinities*, 5.
- Matt Golder. 2016. Far right parties in Europe. *Annual Review of Political Science*, 19: 477-497.
- Maja Gwóźdz. 2018. The MAGA corpus of anglophone political debate online.  
[http://drive.google.com/file/d/1BGyCO\\_CT0ax9Ox9b6i47\\_Z5t7B3py5sk](http://drive.google.com/file/d/1BGyCO_CT0ax9Ox9b6i47_Z5t7B3py5sk)
- Kurt Hornik, Ingo Feinerer, Martin Kober and Christian Buchta. 2012. Spherical k-means clustering. *Journal of Statistical Software*, 50(10): 1-22.
- Sylvia Jaki and Tom De Smedt. 2018. Right-wing German hate speech on Twitter: Analysis and automatic detection. Manuscript submitted.
- Sylvia Jaki, Tom De Smedt, Maja Gwóźdz, Rudresh Panchal, Alexander Rossa and Guy De Pauw. 2018. Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. Manuscript submitted.
- R. Kerkvliet. 2017. *Redefining Violence: The Securitisation and Desecuritisation of Farm Attacks in post-Apartheid South Africa* (dissertation). Utrecht University.
- Amanda Khoza. 2017. New social media research finds xenophobia rife among South Africans. *news24*. Retrieved June 2018.  
<http://www.news24.com/SouthAfrica/News/new-social-media-research-finds-xenophobia-rife-among-south-africans-20170404>
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751, Doha, Qatar.

- Michael Kimmel. 2017. *Angry White Men: American Masculinity at the End of an Era*. Hachette, United Kingdom.
- Jytte Klausen, Eliane Tschaen Barbieri, Aaron Reichlin-Melnick and Aaron Y. Zelin. 2012. The YouTube Jihadists: A social network analysis of Al-Muhajiroun's propaganda campaign. *Perspectives on Terrorism*, 6(1), 36-53.
- Eduan Kotzé and Burgert Senekal. 2018. Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa. *The Journal for Transdisciplinary Research in Southern Africa* (in press).
- Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1): 1-167.
- Huan Liu and Hiroshi Motoda. 2007. *Computational Methods of Feature Selection*. CRC Press, Boca Raton, Florida.
- Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 746-751, Atlanta, Georgia.
- Karsten Müller and Carlo Schwarz. 2017. *Fanning the Flames of Hate: Social Media and Hate Crime* (technical report). University of Warwick.
- Angela Nagle. 2017. *Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right*. John Hunt Publishing, Alresford.
- John T. Nockleby. 2000. Hate speech. In *Encyclopedia of the American Constitution*. Macmillian, New York.
- Brian L. Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1): 59-68.
- Larry Page, Sergey Brin, Rajeev Motwani and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web* (technical report). Stanford InfoLab.
- James W. Pennebaker. 2011. *The Secret Life of Pronouns*. Bloomsbury, New York.
- Josef Ruppenhofer. 2018. German Twitter embeddings.  
<http://www.cl.uni-heidelberg.de/research/downloads>
- Leïla Saoud. 2018. *Detecting racism in online French text* (dissertation). University of Leuven.

- Dave Steward. 2016. Anti-white racism has turned virulent. *Politicsweb*. Retrieved June 2018. <http://www.politicsweb.co.za/politics/antiwhite-racism-has-turned-virulent--fw-de-klerk->
- Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108: 132-143.
- Cass R. Sunstein. 2018. *#Republic: Divided democracy in the age of social media*. Princeton University Press.
- Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1): 24-54.
- Luis Tomé. 2015. The Islamic State: trajectory and reach a year after its self proclamation as a Caliphate. *e-journal of International Relations*, 6(1).
- Corinne Torrekens. 2015. Islam in Belgium: From formal recognition to public contestation. In *After Integration. Islam, Conviviality and Contentious Politics in Europe*, 153-169, Springer, Wiesbaden.
- Stéphan Tulkens, Lisa Hilte, Elyne Lodewyckx, Ben Verhoeven and Walter Daelemans. 2016. The Automated Detection of Racist Discourse in Dutch Social Media. *Computational Linguistics in the Netherlands Journal*, 6: 3-20.
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579-2605.
- Pieter Van Ostaeyen. 2016. Belgian radical networks and the road to the Brussels attacks. *CTC Sentinel*, 9(6): 7-12.
- Gabriel Weimann, 2016. Terrorist migration to the dark web. *Perspectives on Terrorism*, 10(3): 40-44.
- Jacob Westfall, Leaf Van Boven, John R. Chambers and Charles M. Judd. 2015. Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, 10(2): 145-158.
- John Wrench. 2016. *Diversity management and discrimination: Immigrants and ethnic minorities in the EU*. Routledge, New York.
- YouGov. 2016. German attitudes to immigration harden following Cologne attacks. <http://yougov.co.uk/news/2016/01/12/germans-attitudes-immigration-harden-following-col>