Profanity & Offensive Words (POW)

Multilingual fine-grained lexicons for hate speech

Tom De Smedt, Pierre Voué, Sylvia Jaki, Melina Röttcher & Guy De Pauw

JUNE 2020

TEXTGAIN TECHNICAL REPORTS • TGTR3 • ISSN 2684-4842

AUTHORS

- **Tom De Smedt** (tom@textgain.com) has a PhD in Arts and is CTO at Textgain. He focuses on Natural Language Processing and was awarded the Research Prize of the Auschwitz Foundation in 2019.
- **Pierre Voué** is a data scientist at Textgain and focuses on online extremism and radicalization.
- **Sylvia Jaki** has a PhD in Linguistics and is an expert on hate speech and misogyny.
- **Melina Röttcher** is a linguist and an expert on abusive language in German.
- **Guy De Pauw** has a PhD in Linguistics and he is CEO of Textgain.

ABSTRACT

The POW lexicons are a steadily growing, interpretable NLP resource for online hate speech detection. They are currently available in English, German, French, Dutch and Hungarian, capturing thousands of verbal expressions of abusive, aggressive, dehumanizing, discriminatory, offensive and toxic language use, and have been field-tested in real-life applications: https://www.textgain.com/pow-lexicons.

Keywords: *natural language processing*, *hate speech*, *abusive language*, *social media*

1 INTRODUCTION

In their 2018 report on terrorism, Europol observed that online propaganda is increasingly becoming a central strategy for violent extremism and radicalization (Europol, 2018). Several authorities have also been introducing crackdown legislation. For example, in May 2016, the European Commission agreed with Facebook, Microsoft, Twitter and YouTube on a Code of Conduct on countering illegal hate speech online,[1] and in June 2017, the German government passed a Network Enforcement Act (NetzDG) to counter hate speech in social networks.[2] In May 2020, UN Secretary-General António Guterres called for an all-out effort against hate speech, after a "tsunami" of COVID-19 xenophobia.[3] Online hate speech, i.e., incitement of racism, sexism, violence, has also attracted academic interest, with a surge of new AI Machine Learning technology (ML) that can detect hate speech automatically. However, many of these systems raise ethical and legal concerns in real-life applications.

First, ML models are vulnerable to algorithmic bias, amplifying pre-existing prejudices in the training data (Hajian, Bonchi & Castillo, 2016). This raises ethical concerns. A well-known example is a risk assessment AI that predicts higher criminal risk for people of color, since its historical training data has people of color disproportionately targeted by law enforcement.[4] Second, ML models (especially deep neural nets) have also been called black boxes, whose decision-making process is difficult to explain and interpret in a straightforward way (Rudin, 2019). This raises legal concerns. Finally, it has been shown that hate speech detection systems can be misled, for example by adding the word "love" to a message (Gröndahl, Pajola, Juuti et al., 2018). Proposed solutions include techniques for explaining black boxes (Ribeiro, Singh & Guestrin, 2016) or more interpretable logic algorithms (cf. Rudin).

One highly interpretable approach is to use domain-specific lexicons handcrafted by domain experts. Words and word combinations from the lexicon can be **highlighted** in a given text, so that human reviewers can instantly see how and why a prediction was made. As a drawback, such lexicons can also become outdated quickly, particularly in the area of hate speech, where new toxic words emerge every day. Our method for collecting lexicons uses a human-machine feedback loop, with large-scale ML that continuously discovers new expressions and presents them to expert annotators. We scan social media using existing lexicons, find words that we didn't know about yet, annotate or discard these, and update the lexicons to stay on top of trends as they evolve in real-time.

2 METHODS AND MATERIALS

Our English lexicon has 3,000+ toxic words and word combinations extracted from our 4chan and 8chan embeddings (Voué, De Smedt & De Pauw, 2020) and 3,000+ expressions collected by hand from Twitter, Facebook, Gab.com, Incels.me and Iron March, a defunct neo-Nazi forum.[5] The German lexicon has 1,500+ expressions extracted from the German Twitter Embeddings (Ruppenhofer, 2018) and 3,500+ collected by hand from Twitter (De Smedt & Jaki, 2018; Jaki & De Smedt, 2019). The French lexicon has 2,000+ expressions from 8chan/dempart (Démocratie Participative). The Dutch lexicon has 10,000+ expressions from Twitter, Facebook and private groups not available anywhere else. The English lexicon was manually translated into Hungarian. It also has a specialized companion set of 2,000+ expressions that relate to anti-semitism. More lexicons are in development.

[1] https://ec.europa.eu/commission/presscorner/detail/en/MEMO_19_806

[2] https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html

[3] https://www.un.org/press/en/2020/sgsm20076.doc.htm

[4] https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai

[5] https://www.bellingcat.com/news/2019/12/19/transnational-white-terror-exposing-atomwaffen-and-the-iron-march-networks

3 RESULTS

Each lexicon groups words and word combinations into 10+ different categories, each represented by an emoji and a four-letter alias that is easy to remember and concise in programming code. There are categories for anger and contempt (HATE, SHIT, FUCK), for discrimination (FOOL, SCUM, SLUT, GOOK), for incitement and propaganda (HELL, HEIL, PLOT), and for verbal aggression (KILL).

HATE	😠	Words that relate to **negativity** (e.g., *lame*, *worthless*), anger (*disgusting*, *hate*, *kick*, *rage*), cynicism (*we're doomed*) and sarcasm (*"very fine people"*).
SHIT	💩	Words that relate to **profanity** (e.g., *damn*, *piss*, *shit*), in particular vulgar name-calling (*damn sambo*, *pisslam = piss + Islam*) and swearing (*BS*, *WTF*).
FUCK	🐷	Words that relate to **pornography**. (e.g., *cunt*, *dick*, *fuck*), in particular with regard to sex crimes (*goat fucker*, *pedo*, *rapist*).
FOOL	🤡	Words that relate to **ridicule** (e.g., *deplorable*, *poor snowflake*, *tinfoil hat*), in particular insults of intelligence (*degenerate*, *dotard*, *retard*).
SCUM	🐸	Words that relate to **dehumanization** (e.g., *cum dumpster*, *rat*, *scum*, *thug*, *vermin*) or defamation (*fake news peddler, treasonous dog*).
SLUT	👱‍♀️	Words that relate to **sexism** (e.g., *gay*, *lesbian*), on the basis of sexual orientation (*fag*), sexuality (*slut*), gender (*bitch*) and gender stereotypes (*coward*, *cuck*).
GOOK	👳🏿	Words that relate to **racism** (e.g., *black bitch*, *white trash*), on the basis of race (*nigger*), ethnicity (*hebress*), nationality (*africoon*, *chexican*) and looks (*fatso*).
HELL	☝🏽	Words that relate to **religious ideology** (e.g., *Christians*, *Jews*, *Muslims*), in particular islamophobia (*hatebeard*), jihadism (*infidel*) and antisemitism (*lolocaust*).
HEIL	✊🏻	Words that relate to **political ideology** (e.g., *communist, fascist*, *traitor*), in particular activism (*Antifa*, *Pegida*), extremism (*Islamic State*) and propaganda (*Infowars*).
PLOT	👽	Words that relate to **conspiracy** (e.g., *fake news*, *hoax*), including government cover-up (*deep state, NWO*), doomsday (*lab virus*) and the occult (*Thule Society*).
KILL	💣	Words that relate to **conflict** (e.g., *civil war*, *riot*, *terror*), including violence (*kill*, *shoot*), threats (*kill you*, *shoot you*) and extortion (*dig up dirt*).

Each word or word combination in each lexicon also has a toxicity score (0–4) and may be related to more than one category, as shown in the example below. The scores and categories were assigned manually by multiple, diverse experts in linguistics, social and political sciences, and security.

SCORE	WORD(S)	😠	🐸	🤡	👱‍♀️	👳🏿	✊🏻	👽	💣
★★★★	*black bitch*	○	●	○	●	●	○	○	○
★★★☆	*kill crusaders*	○	○	○	○	○	●	●	●
★★☆☆	*disgusting*	●	●	○	○	○	○	○	○
★☆☆☆	*tinfoil hat*	○	○	●	○	○	○	●	○

The toxicity score can be 0 (neutral), 1 (tendentious), 2 (demeaning), 3 (offensive; low, biased, vulgar) or 4 (extremely offensive). We can use it to map words to numbers, and then use statistics with those numbers. For example: "Eat your tinfoil hat, you disgusting conspiracist person" would score 1 + 2 = 3. A large collection of texts can then be sorted by how toxic each text is.

This works even better if we count with exponential weights. The following weights essentially mean that a word with score 4 is as alarming as ten words with score 1:

SCORE ▸	☆☆☆☆	★☆☆☆	★★☆☆	★★★☆	★★★★
weight	0.01	0.10	0.25	0.50	1.00

A demonstration of a concise Python script that imports the English POW lexicon and analyzes texts:

Listing 1. Example Python code for loading and using the POW, using Grasp.py.

```
from grasp import csv, trie # https://github.com/textgain/grasp

pow = {}

for r in csv('pow-en.csv'): # (score,word,HATE, SHIT, ...)
    pow[r[1]] = {
        '0': 0.01,
        '1': 0.10,
        '2': 0.25,
        '3': 0.50,
        '4': 1.00,
    }[r[0]]

pow = trie(pow)

def toxicity(s):
    return sum(weight for i, j, word, weight in pow.search(s))

print(toxicity('Eat your tinfoil hat!'))
```

4 ANALYSIS

The performance of the lexicons can be measured statistically and in terms of trust. Statistically, the performance is about 75% (F_1 score), which is several percentages higher than many other available lexicons. In terms of trust, the lexicons are integrated into a dashboard app,[6] with dozens of users that continually monitor and fine-tune the performance with real-life checks & balances. Another crucial step is to establish a baseline truth: the average toxicity score in random data in comparison to data of interest, i.e., is a message more toxic than what a “normal” person would write?

In a case study with the Dutch lexicon, we compared the average toxicity score of 100K random Dutch messages in 2015 to 100K random messages in 2020. Both sets have 50K Facebook messages from news pages such as HLN, DM, DS, VRT NWS, and 50K random Twitter messages. The average toxicity score has doubled from 0.05 in 2015 to 0.1 in 2020. One explanation is the ongoing polarization between left-wing and right-wing in European societies, in the aftermath of Islamic State terrorism and the Iraqi-Syrian refugee crisis (ca. 2015–2017). In 2015, we see 125 messages with a score of ≥1.0. In 2020, we see 500 of those, often with ideological expressions (*domme linkse*, *bende rechtse*), insults of intelligence (*clown*, *idioot*) and dehumanization (*linkse rat*, *rechts gespuis*, *achterlijke debielen*). In 2020, we find 3x more racist and sexist messages (about 2%) and 2x more verbal aggression (1%). In relative terms, the average toxicity score of messages on the neo-Nazi forum Iron March (data source: Bellingcat)[7] is 0.65, with 8.5% racist messages.

[6] https://projectgrey.eu/technology/?lang=en

[7] https://www.bellingcat.com/news/2019/12/19/transnational-white-terror-exposing-atomwaffen-and-the-iron-march-networks

The timeline below shows an overview of the amount of Dutch racist messages between 2017 and 2019, linked to news stories such as world-wide terrorist attacks or populist leaders speaking out:

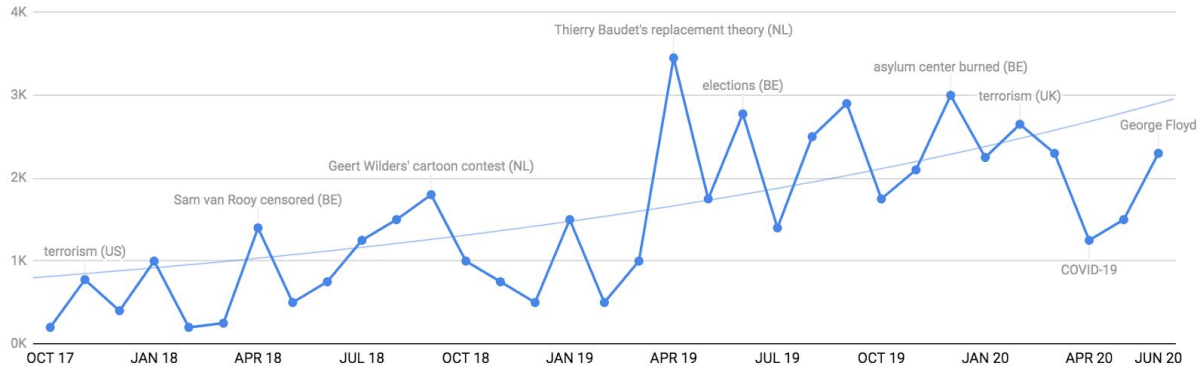

Figure 1. Timeline of Dutch racist messages vs related news stories.

5 DISCUSSION

The multilingual, fine-grained POW lexicons are a new NLP resource for toxic language use that can be integrated into Explainable AI systems for hate speech detection. They are not available for free. This would expose them to reverse engineering for unforeseen purposes. Academic research groups, societal NGOs and democratic government organizations can submit a request to get access.

ACKNOWLEDGEMENTS

Co-funded by the Rights, Equality and Citizenship Programme of the European Union (REC).[8]

REFERENCES

De Smedt, T., & Jaki, S. (2018). Challenges of automatically detecting offensive language online. In *Proceedings of the GermEval 2018 Workshop*, 27-32.

Europol (2018). EU Terrorism Situation & Trend Report (TESAT). https://www.europol.europa.eu/sites/default/files/documents/tesat_2018_1.pdf

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2-12.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of ACM SIGKDD 22*, 2125-2126.

Jaki, S., & De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint arXiv:1910.07518.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD 22,* 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215.

Ruppenhofer, J. 2018. German Twitter embeddings. https://www.cl.uni-heidelberg.de/research/downloads/

Voué, P., De Smedt, T., & De Pauw, G. (2020). 4chan & 8chan embeddings. *Textgain Technical Reports*, *1*, 1-4.

[8] The contents of this document are the sole responsibility of Textgain and cannot be taken to reflect the views of the European Commission.