Routledge
Taylor & Francis Group

Journal of Policing, Intelligence and Counter Terrorism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rpic20

Proceed with caution: on the use of computational linguistics in threat assessment

Isabelle van der Vegt, Bennett Kleinberg & Paul Gill

To cite this article: Isabelle van der Vegt, Bennett Kleinberg & Paul Gill (2023) Proceed with caution: on the use of computational linguistics in threat assessment, Journal of Policing, Intelligence and Counter Terrorism, 18:2, 231-239, DOI: 10.1080/18335330.2023.2165137

To link to this article: https://doi.org/10.1080/18335330.2023.2165137

Published online: 18 Jan 2023.

JOURNAL OF POLICING, INTELLIGENCE AND COUNTER TERRORISM
2023, VOL. 18, NO. 2, 231–239
https://doi.org/10.1080/18335330.2023.2165137

Routledge
Taylor & Francis Group

COMMENTARY

Proceed with caution: on the use of computational linguistics in threat assessment

Isabelle van der Vegt[a,b], Bennett Kleinberg[b,c] and Paul Gill[b]

[a]Department of Sociology, Utrecht University, Utrecht, Netherlands; [b]Department of Security and Crime Science, University College London, London, UK; [c]Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands

ABSTRACT

Large-scale linguistic analyses are increasingly applied to the study of extremism, terrorism, and other threats of violence. At the same time, practitioners working in the field of counterterrorism and security are confronted with large-scale linguistic data, and may benefit from computational methods. This article highlights the challenges and opportunities associated with applying computational linguistics in the domain of threat assessment. Four current issues are identified, namely (1) the data problem, (2) the utopia of predicting violence, (3) the base rate fallacy, and (4) the danger of closed-sourced tools. These challenges are translated into a checklist of questions that should be asked by policymakers and practitioners who (intend to) make use of tools that leverage computational linguistics for threat assessment. The 'VISOR-P' checklist can be used to evaluate such tools through their Validity, Indicators, Scientific Quality, Openness, Relevance and Performance. Finally, some suggestions are outlined for the furtherance of the computational linguistic threat assessment field.

ARTICLE HISTORY
Received 9 November 2022
Accepted 1 January 2023

KEYWORDS
threat assessment; violence; extremism; computational linguistics

Introduction

Online extremism and threats of violence receive significant attention within academic research and security practice. Recently, solutions are sought from computational linguistics, in which large-scale language analyses aim to shed light on threats of (extremist) violence. We call this field: 'computational linguistic threat assessment.' Exponents computationally analyse linguistic data to understand and/or potentially predict extremism (e.g. Kleinberg, van der Vegt, & Gill, 2021; Scrivens, Davies, & Frank, 2018; Simons & Skillicorn, 2020), lone-actor terrorism (e.g. Baele, 2017; Kaati, Shrestha, & Cohen, 2016), and other threats of violence such as mass murder or school shootings (e.g. Egnoto & Griffin, 2016; Kop, Read, & Walker, 2021; Neuman, Assaf, Cohen, & Knoll, 2015). Large-scale linguistic data also increasingly features in the work of security

CONTACT Isabelle van der Vegt i.w.j.vandervegt@uu.nl Department of Sociology, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands; Department of Security and Crime Science, University College London, London WC1E 6BT, UK

practitioners such as those involved in moderating extremist social media posts (Torregrosa, Bello-Orgaz, Martínez-Cámara, Ser, & Camacho, 2022; van der Vegt, Gill, Macdonald, & Kleinberg, 2019), gathering and interpreting open-source intelligence (Pastor-Galindo, Nespoli, Gómez Mármol, & Martínez Pérez, 2020), or assessing threats to public figures (James et al., 2022; Pelzer, Kaati, & Akrami, 2018).

This article focuses on the use of computational linguistics to potentially predict the risk of *an individual* engaging in targeted violence, such as a terrorist attack or a school shooting. Others have leveraged computational linguistic methods to predict the risk of violence *at an aggregate level,* such as on a country or city level. For instance, Müller and Schwarz (2021) report that anti-refugee sentiment in Facebook posts can predict hate crimes against refugees in Germany. This finding was supported by the fact that Facebook outages were associated with a lower probability of anti-refugee incidents. Similarly, Williams, Burnap, Javed, Liu, and Ozalp (2020) linked hateful Twitter posts to police crime data on hate crimes in London. The authors found a positive temporal and spatial association between hate speech targeting race and religion and offline hate acts. Despite the significance and possible practical utility of these aggregate level analyses, we focus explicitly on threat assessment at the individual level.

With a practitioner audience in mind, this article highlights the challenges and opportunities associated with applying computational linguistics in the domain of threat assessment. First, we identify current issues with computational linguistic threat assessment in four sections, namely (1) the data problem, (2) the utopia of predicting violence, (3) the base rate fallacy, and (4) the danger of closed-sourced tools. Second, we translate these challenges into a checklist of questions, titled 'VISOR-P', that should be asked by policymakers and practitioners who (intend to) make use of recently available tools that leverage computational linguistics for the purpose of threat assessment. We conclude with a look into the future, outlining some suggestions for the furtherance of the computational linguistic threat assessment field.

Four problems in computational linguistic threat assessment

The data problem

One of the most challenging issues within the computational linguistic threat assessment field is access to appropriate datasets. Targeted violence is a low base rate phenomenon (Corner, Gill, Schouten, & Farnham, 2018) and the number of cases where the perpetrator produced linguistic material related to the incident is smaller still. This means that there are little linguistic data authored by individuals who resorted to violence, compared to the vast amount of data by individuals who have not done so. Due to this data scarcity, researchers within the field of grievance-fuelled targeted violence often sample on the dependent variable (Clemmow, Schumann, Salman, & Gill, 2020). That is, text data are selected from sources about whom it is known committed violence, often consisting of the writings of lone-actor terrorists (Baele, 2017; Kaati et al., 2016; Kaati, Shrestha, & Sardella, 2016; Kop et al., 2021). In order to discover linguistic 'markers' of violence, studies frequently compare texts from highly violent individuals (e.g. lone-actor terrorist manifestos, school shooter writings) to a large sample of linguistic data which are not violent in any way, such as neutral texts from blogs, online forums, social media platforms

or even Wikipedia pages (Baele, 2017; Jaki, Smedt, & Gwó, 2019; Kaati et al., 2016; Neuman et al., 2015; van der Vegt, Mozes, Kleinberg, & Gill, 2021). As a consequence, large linguistic differences between these two naturally dissimilar sources of data are frequently found, and classification algorithms generally perform well (e.g. Jaki et al., 2019; Kaati et al., 2016). However, these comparisons tell us little about the linguistic factors that can be used to distinguish between violence-actualisers and non-actualisers (e.g. an extremist planning to engage in violence, versus an extremist who will not), nor do they result in meaningful classification algorithms. Discovering these discriminating factors could perhaps be considered the 'holy grail' of computational linguistic threat assessment.

Another research approach compares lone-actor terrorist manifestos to other extremist texts such as those collected from online forums (Kaati et al., 2016; van der Vegt et al., 2021). When attempting to identify possible linguistic markers of extremist violence, this comparison can be considered preferable over a comparison to neutral data. However, this approach remains somewhat problematic due to two underlying assumptions, namely (i) that participation in an extremist forum equates to extremism, and (ii) that none of the forum members have engaged in extremist violence. While extremist forum data counts as a valuable resource in this field, its main shortcoming is that it lacks ground truth. A notable exception to this issue can be found in Scrivens et al. (2023, 2022), where ground truth behind an extremist forum dataset was ascertained by having a former right-wing extremist code posters on the Canadian sub-forum of Stormfront as violent or non-violent based on his experience with these members in real life. Analyses on these data have revealed different posting typologies (Scrivens et al., 2023) as well as negative attitudes of right-wing extemists towards black, jewish, and LGBTQ + communities as measured through language in the forum posts (Scrivens, Davies, Gaudette, & Frank, 2022). Although the dataset remains small and relies on the annotation of a single individual, this endeavour can be viewed as a step in the right direction. On the whole, however, it can be said that several important questions in the field of computational linguistic threat assessment remain unanswered due to challenges with data availability.

The utopia of predicting violence

While technical advances in machine learning have enabled us to predict several phenomena with increasing accuracy, practitioners should be aware this is not guaranteed for individual acts of violence. As mentioned, data resources in this domain are scarce. In other behavioural research areas, real life outcomes can still not be accurately predicted. For example, in a collaboration study of 160 research teams using data (12,943 possible variables) collected from 4,242 families over 15 years, life outcomes (e.g. material hardship, GPA, eviction) were not accurately predicted (Salganik et al., 2020). That is, the best performing models achieved an explained variance (R^2) in the test set data of 0.23 (on a scale from 0 to 1, where 1 equals perfect prediction). Targeted violence similarly concerns a real-life outcome that may be the result of hundreds of interacting variables, for which it is questionable if a dataset of the same magnitude and depth will ever be collected. The large-scale prediction effort of life outcomes raised an important point, namely that understanding phenomena perhaps does not mean we can accurately predict them (Salganik et al., 2020), nor that

predicting them implies that we understand them. In other words 'prediction is not a good measure of understanding and [..] understanding can come from description or causal inference' (Salganik et al., 2020, p. 8402).

It is worth putting these findings into a wider perspective: we have exceptional computational power at our disposal, more data than ever before, and have ready access to sophisticated machine learning algorithms, yet this does not provide any assurance that we can solve complicated problems. The introduction of machine learning into the social sciences as a prominent technique has created somewhat unrealistic expectations. While it is true that machine learning systems can solve mundane problems (e.g. determining whether a movie review is positive or negative) and in some cases have even pushed the boundaries of human knowledge (e.g. discovering protein structures; Jumper et al., 2021), there is a risk of making a category mistake. All of the advances made by machine learning rely on a seemingly negligible assumption, namely that patterns exist in the data. These patterns need not be obvious to the human eye and may even be inaccessible for many machine learning systems, but the assumption holds that there are some reliable patterns. The category mistake happens when we conflate advances in one area with those that we expect to be made in another. Put differently: if we expect machine learning to be able to predict human behaviour because we also used it to discover protein structures, we commit a category mistake seeing as these are two different categories of problems. Human behaviour may simply be too complex to accurately predict or 'may be subject to a predictability ceiling' (Garip, 2020, p. 8235). This may be unsatisfying for some, but the honest outlook on predicting human behaviour (including acts of violence) is that we may be facing an unsolvable problem.

The base rate fallacy

Suppose hypothetically, that linguistic threat assessment predictions somehow improved to the level of near 100% accuracy. In this unlikely scenario, the practical utility of these systems will remain limited. The so-called base rate fallacy is one of the main reasons for this. Due to the low base rate of individuals who resort to actual violence (Corner et al., 2018), the rate of possible false positives – even with very accurate systems – is alarmingly high. In other words, due to the strong imbalance towards non-violent individuals, a large number of individuals will be incorrectly classified as being potentially violent, thereby rendering the system's predictions meaningless.

The base rate fallacy can be illustrated as follows. Imagine a situation where practitioners have a sample of documents written by 100 million different individuals that need to be classified as violence-actualisers (e.g. they will commit a terrorist attack) or non-actualisers (e.g. they will not commit a terrorist attack). For simplicity, let us assume that the base rate of actual violence is 0.01% (i.e. 10,000 individuals will commit an attack), and our system is 99% accurate – a wholly overoptimistic accuracy given the current state of research. An accuracy rate of 99% means that the system can correctly identify both violence-actualisers and violence non-actualisers 99% of the time. Within the hypothetical sample of 100 million documents (each written by a different author), 10,000 documents will actually derive from violence-actualisers (the 0.01% base rate), of which 9,900 will be correctly classified (the 99%

accuracy). Within the documents from non-actualisers (the remaining 99.99%), this system will *incorrectly classify 999,900 individuals* as violence actualisers. This means that of all the documents classified as deriving from violence-actualisers (total 1,009,800), *less than 1% are correctly classified* (0.98% or 9,900 out of 1,009,800) as being truly written by a violence-actualiser. See Table 1 for a breakdown of this calculation (adapted from Kleinberg, van der Toolen, Arntz, & Verschuere, 2018; van der Vegt et al., 2019). It is important to recognise that even if prediction systems could be further developed to achieve high accuracy rates in violence prediction, the base rate fallacy will continue to persist. This is especially crucial in a field where the potential consequences of false negatives (e.g. an unexpected terror attack) and false positives (e.g. wrongful allegations or convictions) are far-reaching.

The danger of closed-source tools

Besides academic research leveraging computational linguistics to study extremism and violence, several (commercial) tools claiming predictive ability for violent behaviour through linguistic analysis have also emerged. By claiming to predict (individual-level) violence, these tools ignore the issues raised previously concerning the lack of quality data and valid comparisons within this field, the difficulties associated with predicting behaviour, and the base rate fallacy. We deem it unlikely that these tools have outsmarted decades of research, with their results simultaneously flying under the radar of the academic community.

A crucial issue is that these commercial tools are not transparent, in that the source code underlying the tool is not available for scrutiny. In addition to this, it is often not reported what data have been used to develop the tool, which linguistic indicators were used, and what prediction performance can be expected from the tool. For instance, the tools may make use of a dictionary of violent words, searching for 'hits' in the text to be analysed, but the exact words are not made available. In addition, a resulting 'violence risk score' for the text may be the result of a (statistical) comparison to linguistic features of texts written by known perpetrators or extremists, but the contents of the comparison database and performed computations are not known to users of the tool. While concealing this information is perhaps understandable for commercial reasons, it raises the question whether practitioners should be using tools that produce outcomes that cannot be fully explained. It becomes even more worrying if the tool is not validated through scientific research or if the supporting studies do not adhere to basic scientific standards (e.g. being independently peer-reviewed, not containing a conflict of interest, and providing sufficient detail for an independent replication of the results). In short, we urge practitioners to be aware that the field

Table 1. Demonstration of base rate fallacy.

		Prediction		
		Violence	No violence	Total
Reality	Violence	9,900	100 (false negatives)	10,000
	No violence	999,900 (false positives)	98,990,100	99,990,000
	Total	1,009,800	98,990,200	100,000,000

of countering violence and extremism appears to be sensitive to tools emerging that are no more than snake oil.[1]

Evaluating computational linguistic threat assessment tools

In this section, we translate the four issues raised above into a checklist of questions that should be asked by professionals prior to implementing language-based threat assessment tools.[2] Individual questions are represented in the mnemonic 'VISOR-P'. We urge policymakers and practitioners to refrain from or seriously reconsider implementing tools if the answer to any of the questions is negative.

The VISOR-P checklist

- **V**alidity: Is the tool based on a valid comparison between violent and non-violent data?
- **I**ndicators: Is it clear what linguistic indicators are used, and how these are measured?
- **S**cientific quality: Has the tool been validated in empirical, peer-reviewed research?
- **O**penness: Is it clear which data were used to develop the tool, and how the model arrives at its final outcome decision?
- **R**elevance: Are the linguistic indicators considered relevant to the outcome that is to be assessed?
- **P**erformance: Is the prediction performance of the tool on out-of-sample data reported?

To our knowledge, a tool satisfying all of the above criteria does not exist. It also remains unclear whether this will be the case in the future. Nevertheless, we strongly believe these questions set out the minimal criteria for practical implementation. As a consequence, the above checklist may also serve as a guide for researchers and developers involved in creating computational linguistic threat assessment tools.

Moving forward

There are several ways in which the use of linguistic data can assist the work of security practitioners. While the data problem will likely continue to persist due to the low base rate of acts of extremist violence, this can be somewhat ameliorated through data sharing between practitioners and researchers, so that meaningful linguistic comparisons between violence-actualisers and non-actualisers can be carried out. However, it must be acknowledged that some problems in this field remain unsolvable, and that it is unlikely that highly accurate prediction of violent acts based on language will ever be possible, even when supplemented with other measures. The main takeaway of this article should be that the field of computational linguistic threat assessment needs to become comfortable with unsolvable problems and unpredictable behaviour.

Bearing in mind the impossibilities when it comes to predicting violence based on language use, it is helpful to focus on what, in contrast, *is* possible. Academic research will likely continue to enhance our understanding of extremism and violence through language. As a consequence, practitioners may, for example, be able to understand

which linguistic factors make up a worrying communication, as well as which populations may be more inclined to produce such content. Based on this knowledge, research and practice may also be able to develop novel strategies to adequately manage such threats. In short, it is worthwhile for both researchers and practitioners to continue to work towards understanding violence and grievances to the best of our ability, aided by computational tools.

The primary potential for the use of computational linguistics in threat assessment thus may not lie within prediction, but in measurement. For practitioners, computational tools could support the review of large amounts of written content and enable them to identify from within those data evidence for a range of features deemed relevant to their assessment (e.g. a practitioner seeking to find those documents in a vast corpus that often mention weaponry). Which features are relevant may vary on a case-by-case basis, and the possible meaning of the obtained outcome will need to be judged by the practitioner(s) involved. This approach is more closely aligned to the risk assessment principles of Structured Professional Judgement (SPJ) than to actuarial approaches. SPJ is focused on helping the user to consider the totality of circumstances that surround the individual being assessed. Actuarial approaches, on the other hand, are solely focused on prediction by comparing an individual's similarity to a group of people with a known rate of offending (Hart, Douglas, & Guy, 2016; Pedersen, Rasmussen, & Elsass, 2010). An actuarial approach to computational linguistic threat assessment would constitute a tool that compares an incoming text to, for example, a database of neutral, extremist, and terrorist texts, resulting in a violence risk judgment. We argue that professional expertise in combining and interpreting risk factors should be valued in arriving at a final risk judgement (Douglas, Ogloff, & Hart, 2003; Pedersen et al., 2010), hence advocate for a SPJ approach instead. This can be achieved through a hybrid approach to computational linguistic threat assessment, where automation enables practitioners to identify linguistic patterns usually imperceptible to the human eye, but where final risk judgments lie in the hands of professionals.

Conclusion

The field of computational linguistic threat assessment will continue to develop in the years to come. For professionals involved in these endeavours, it is important to critically evaluate the results and tools that will emerge, bearing in mind that predicting violent behaviour on an individual level is an exceptionally challenging – and likely impossible – endeavour. Moving away from prediction and investing in measurement and hybrid decision making may eventually turn out to be the most beneficial use of computational linguistics in threat assessment.

Notes

1. See also: https://crestresearch.ac.uk/resources/substance-or-snake-oil/ for a practitioner guide on how to evaluate a written claim of efficacy regarding a product or service.
2. These criteria are specific to linguistic threat assessment. For a more general framework on the role of algorithms in law enforcement, see the ALGOCARE guidelines (Oswald, Grace, Urwin, & Barnes, 2018).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Baele, S. J. (2017). Lone-Actor terrorists' emotions and cognition: An evaluation beyond stereotypes. *Political Psychology*, *38*(3), 449–468. doi:10.1111/pops.12365

Clemmow, C., Schumann, S., Salman, N. L., & Gill, P. (2020). The base rate study: Developing base rates for risk factors and indicators for engagement in violent extremism. *Journal of Forensic Sciences*, *65*(3), 865–881. doi:10.1111/1556-4029.14282

Corner, E., Gill, P., Schouten, R., & Farnham, F. (2018). Mental disorders, personality traits, and grievance-fueled targeted violence: The evidence base and implications for research and practice. *Journal of Personality Assessment*, *100*(5), 459–470. doi:10.1080/00223891.2018.1475392

Douglas, K. S., Ogloff, J. R., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment Among forensic psychiatric patients. *Psychiatric Services*, *54*(10), https://ps.psychiatryonline.org/doi/full/10.1176appi.ps.54.10.1372

Egnoto, M. J., & Griffin, D. J. (2016). Analyzing language in suicide notes and legacy tokens. *Crisis*, *37*(2), 140–147. doi:10.1027/0227-5910/a000363

Garip, F. (2020). What failure to predict life outcomes can teach us. *Proceedings of the National Academy of Sciences*, *117*(15), 8234–8235. doi:10.1073/pnas.2003390117

Hart, S. D., Douglas, K. S., & Guy, L. S. (2016). The structured professional judgement approach to violence risk assessment. In D. P. Boer (Ed.), *The Wiley handbook on the theories, assessment and treatment of sexual offending* (pp. 643–666). Wiley Blackwell. https://onlinelibrary.wiley.com/doi/abs/10.1002978111857400 3.wattso030

Jaki, S., Smedt, T. D., & Gwó, M. (2019). Online hatred of women in the incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Agrression and Conflict*, *7*(2), 30.

James, D. V., Allen, P., Wolfe Murray, A., MacKenzie, R. D., Yang, J., De Silva, A., & Farnham, F. R. (2022). The CTAP, a threat assessment tool for the initial evaluation of concerning or threatening communications: Development and inter-rater reliability. *Journal of Threat Assessment and Management*, *9*, 129–152. doi:10.1037/tam0000173

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 583. doi:10.1038/s41586-021-03819-2

Kaati, L., Shrestha, A., & Cohen, K. (2016). Linguistic analysis of lone offenders manifestos. *IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, 1–8.

Kaati, L., Shrestha, A., & Sardella, T. (2016). Identifying warning behaviors of violent lone offenders in written communication. *2016 IEEE 16th International Conference on Data Mining Workshops*, 1053–1060. doi:10.1109/ICDMW.2016.116

Kleinberg, B., van der Toolen, Y., Arntz, A., & Verschuere, B. (2018). Detecting concealed information and deception. In *Detecting concealed information and deception* (pp. 377–403). Elsevier. doi:10.1016/B978-0-12-812729-2.00016-1

Kleinberg, B., van der Vegt, I., & Gill, P. (2021). The temporal evolution of a far-right forum. *Journal Computational Social Science*, *4*(1), 1–23.

Kop, M., Read, P., & Walker, B. R. (2021). Pseudocommando mass murderers: A big five personality profile using psycholinguistics. *Current Psychology*, doi:10.1007/s12144-019-00230-z

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167. doi:10.1093/jeea/jvaa045

Neuman, Y., Assaf, D., Cohen, Y., & Knoll, J. L. (2015). Profiling school shooters: Automatic text-based analysis. *Frontiers in Psychiatry*, *6*(86), doi:10.3389/fpsyt.2015.00086

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the durham HART model and 'experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223–250. doi:10.1080/13600834.2018.1458455

Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., & Martínez Pérez, G. (2020). The Not Yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access*, *8*, 10282–10304. doi:10.1109/ACCESS.2020.2965257

Pedersen, L., Rasmussen, K., & Elsass, P. (2010). Risk assessment: The value of structured professional judgments. *International Journal of Forensic Mental Health*, *9*(2), 74–81. doi:10.1080/14999013.2010.499556

Pelzer, B., Kaati, L., & Akrami, N. (2018). Directed digital hate. 2018 *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 205–210. doi:10.1109/ISI.2018.8587396

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398-8403.

Scrivens, R., Davies, G., & Frank, R. (2018). Searching for signs of extremism on the web: An introduction to sentiment-based identification of radical authors. *Behavioral Sciences of Terrorism and Political Aggression*, *10*(1), 39–59. doi:10.1080/19434472.2016.1276612

Scrivens, R., Davies, G., Gaudette, T., & Frank, R. (2022). Comparing online posting typologies among violent and nonviolent right-wing extremists. *Studies in Conflict & Terrorism*, 1–23. doi:10.1080/1057610X.2022.2099269

Scrivens, R., Wojciechowski, T. W., Freilich, J. D., Chermak, S. M., & Frank, R. (2023). Comparing the online posting behaviors of violent and Non-violent right-wing extremists. *Terrorism and Political Violence*, *35*(1), 192–209. doi:10.1080/09546553.2021.1891893

Simons, B., & Skillicorn, D. B. (2020). A Bootstrapped Model to Detect Abuse and Intent in White Supremacist Corpora. *ArXiv:2008.04276 [Cs]*. http://arxiv.org/abs/2008.04276

Torregrosa, J., Bello-Orgaz, G., Martínez-Cámara, E., Ser, J. D., & Camacho, D. (2022). A survey on extremism analysis using natural language processing: Definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*, doi:10.1007/s12652-021-03658-z

van der Vegt, I., Gill, P., Macdonald, S., & Kleinberg, B. (2019). Shedding light on terrorist and extremist content removal | RUSI. *Global Research Network on Terrorism and Technology*. https://rusi.org/publication/other-publications/shedding-light-terrorist-and-extremist-content-removal

van der Vegt, I., Mozes, M., Kleinberg, B., & Gill, P. (2021). The grievance dictionary: Understanding threatening language use. *Behavior Research Methods*, doi:10.3758/s13428-021-01536-2

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Corrigendum to: Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 242–242. https://doi.org/10.1093/bjc/azz049