# Linguistic Threat Assessment: Understanding Targeted Violence through Computational Linguistics

Isabelle Wen-Li Johanna van der Vegt

A dissertation submitted in partial fulfilment of the requirements for the degree of

**Doctor of Philosophy**

in Security and Crime Science

University College London

2021

# Student declaration

I, *Isabelle Wen-Li Johanna van der Vegt*, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Signed ______________

Date 4 January 2021

# Abstract

Language alluding to possible violence is widespread online, and security professionals are increasingly faced with the issue of understanding and mitigating this phenomenon. The volume of extremist and violent online data presents a workload that is unmanageable for traditional, manual threat assessment. Computational linguistics may be of particular relevance to understanding threats of grievance-fuelled targeted violence on a large scale. This thesis seeks to advance knowledge on the possibilities and pitfalls of threat assessment through automated linguistic analysis.

Based on in-depth interviews with expert threat assessment practitioners, three areas of language are identified which can be leveraged for automation of threat assessment, namely, linguistic content, style, and trajectories. Implementations of each area are demonstrated in three subsequent quantitative chapters. First, linguistic content is utilised to develop the Grievance Dictionary, a psycholinguistic dictionary aimed at measuring concepts related to grievance-fuelled violence in text. Thereafter, linguistic content is supplemented with measures of linguistic style in order to examine the feasibility of author profiling (determining gender, age, and personality) in abusive texts. Lastly, linguistic trajectories are measured over time in order to assess the effect of an external event on an extremist movement.

Collectively, the chapters in this thesis demonstrate that linguistic automation of threat assessment is indeed possible. The concluding chapter describes the limitations of the proposed approaches and illustrates where future potential lies to improve automated linguistic threat assessment. Ideally, developers of computational implementations for threat assessment strive for explainability and transparency. Furthermore, it is argued that computational linguistics holds particular promise for large-scale measurement of grievance-fuelled language, but is perhaps less suited to prediction of actual violent behaviour. Lastly, researchers and practitioners involved in threat assessment are urged to collaboratively and critically evaluate novel computational tools which may emerge in the future.

# Impact statement

This thesis holds implications for future research, as it illustrates how computational linguistics can be used to automate the study of targeted violence. Unlike many previous applications of computational linguistics to targeted violence, this thesis builds on consultation with expert practitioners in order to define the most fruitful and relevant areas for linguistic automation. The methods and tools presented in this thesis may possibly contribute to further theory testing and formation within the linguistic study of grievance-fuelled targeted violence. The Grievance Dictionary presented in this thesis provides an opportunity for researchers across different fields to linguistically measure violence and grievances in the same way. Furthermore, this thesis makes use of several unique datasets, and provides code and study materials for future replication. By adopting this open science approach, this thesis contributes to the capability of other researchers to conduct similar studies, or for other (crime) assessment procedures to be automated in a similar, transparent way.

The studies within this thesis have been disseminated at several academic conferences, such as the Society for Terrorism Research conference (Liverpool, 2018 and Oslo, 2019), VOX-Pol (Amsterdam, 2018), Terrorism and Social Media Conference (Swansea, 2019), European Computational Social Science conference (Zürich, 2019), POLTEXT (Tokyo, 2019), and Social Informatics (Pisa, 2020). Results have also been presented during invited talks at UCL, the University of Amsterdam, and Tilburg University. Chapters in this thesis have been published or are under review in leading journals.

Outside academia, findings from this thesis have been presented and discussed with government and police practitioners in the UK and the Netherlands. This thesis stresses the importance of data-sharing and collaboration between practitioners and academics. It also demonstrates that the primary practical potential of linguistic automation lies in measurement of large-scale data, and that violent incidents are difficult to predict using linguistic information, if at all. In short, insights from this thesis have the potential to guide policymakers and practitioners in making decisions about the automation of large-scale threat assessment.

---

Certain aspects of this thesis were published or submitted for publication in the following outlets:

van der Vegt, I., Kleinberg, B., Gill, P. (2021). Linguistic analysis and lone-actor violence. In: *Lone-Actor Terrorism: An Integrated Framework*. Oxford University Press.

van der Vegt, I., Mozes, M., Kleinberg, B. & Gill, P. (2021). The Grievance Dictionary: Understanding Threatening Language Use. *Behavior Research Methods.*

van der Vegt, I., Mozes, M., Gill, P., & Kleinberg, B. (2020). Online influence, offline violence: Language use on YouTube surrounding the ‘Unite the Right’ rally. *Journal of Computational Social Science*.

## Funding declaration

Aspects of this research were supported by The European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement 758834, PI: Paul Gill).

# Acknowledgements

Paul, thank you for always looking out for me and for the constant trust in me and my work. Coming to London for the PhD was one of the best decisions I've made and I am very grateful for the unique opportunity I was given by you to pursue the projects I've always dreamt of.

Bennett, I've learned an incredible amount from you, both as a researcher and as a person. To this day I am very grateful I accidentally ended up in your deception detection module, and I have thoroughly enjoyed all the years working together. Thank you for all that you do.

I would also like to extend my thanks to the practitioners with whom I collaborated, as well as those who shared their insightful thoughts in interviews or surveys. I also owe much to the inspirational conversations about open science shared with Sandy and other JDI Open participants, as well as the fascinating discussions in our data science research group.

To my fellow colleagues at the JDI, including the Secret Nacho Club and my GRIEVANCE teammates, thank you for making this an unforgettable experience. Felix, thank you for the great conversations and for putting up with my compulsive baking during lockdown. Bettina, thank you for being the sweetest and most supportive friend I could wish for. Norah, I am so glad to have gained you as a friend (and housemate!) through this experience and I can't wait to see what the future holds for us.

To the girls I befriended in London, I am forever grateful that you were there for some much-needed distraction. Yaloe, Lisa S., Daphne, Lydia, I could not have done this without you. Thank you Maartje, Tessel, Diba, Izzy, Bea, and Emma for all the laughs and for always welcoming me back to Amsterdam with open arms. Thank you Lisa B. and the Yolofishes for your visits to London, your friendship means a lot to me.

Thank you to Koen for always being there on the other side of the phone. I am so glad I got to share this PhD experience with you, even though we were separated by the North Sea. Our mini vacations were a true highlight of the past years and I love all the time we get to spend together, now in the same city ☺

Finally, thank you to my family for their endless encouragement. Oma, Hein, and Hong, your continued support and interest in my work should not go unnoticed. Phing and Jaap, I owe all of this to you and your ever-available advice, the wonderful meals, and knowing you are always there for me.

P.S.

For those looking for musical accompaniment to this thesis, this playlist was carefully curated during the past three years of work: https://sptfy.com/9hzF

# Table of Contents

# List of Tables

# List of Figures

# Introduction

In October 2008, ‘Year2183’ posted a message on the anti-Muslim website ‘Gates of Vienna’, arguing that Muslims should be forcibly deported from Norway (Townsend & Traynor, 2011). The same user was also simultaneously active on the white supremacist forum Stormfront (where he declared Britain would soon be faced with “a civil war due to Muslim immigration”) and frequented a Norwegian neo-Nazi forum, where members would for example discuss the blast power of fertiliser and diesel (Townsend & Traynor, 2011). Finally, on the 22nd of July 2011, the user e-mailed a 1500-page document to thousands of people and also posted it on the Stormfront forum. The document, titled ‘2083: A European Declaration of Independence’, described the user’s ideology and the extensive preparations that he, Anders Breivik, made before killing 77 people in his attacks later that day.

This sequence of events is not unique to the Breivik case. Several other lone-actor terrorists and mass murderers have also made their beliefs and intentions known via the internet. Examples include threats sent via e-mail by the 2010 Stockholm bomber (Nyberg, 2010), videos recorded by the 2018 Parkland high school shooter (James, 2018), and the livestream and manifesto of the 2019 Christchurch Mosque attacker (Ma, 2019). In other cases, the grievances of lone-actor terrorists are directed at public figures. The murders of UK politician Jo Cox in 2016 (Cobain & Taylor, 2016) and German politician Walter Lübcke in 2019 (BBC News, 2020) are just two examples. In the latter case, the right-wing extremist held responsible was also known to be active on online forums, where he is said to have made explicit threats (Der Spiegel, 2019). Other verbal and physical attacks on public figures may not be motivated by an extremist ideology, but are motivated by a serious grievance nonetheless. Examples of this include the vast amount of threats and abuse directed at politicians reported in several European countries (James et al., 2007), the UK (James et al., 2016; Perraudin & Murphy, 2019), Norway (Bjørgo & Silkoset, 2018), and New Zealand (Every-Palmer et al., 2015).

### *Grievance-fuelled targeted violence*

What unites aforementioned cases is that each online threat or violent extremist post has the potential to result in an act of grievance-fuelled targeted violence. Although the base rate of such incidents is low (Corner et al., 2018), it is an incredibly difficult task to identify which aggrieved individuals will eventually resort to violence. Grievance-fuelled targeted violence is a term that is increasingly used to refer to acts of planned and premeditated violence which are fuelled by an identifiable grievance or the ‘perception of having been wronged or treated unfairly or inappropriately’ (Silver et al., 2019, p.15). The term covers acts of violence perpetrated by lone mass murderers and lone actor terrorists (Corner et al., 2018; Silver et al., 2019), in addition to school shooters (see e.g., Vossekuil, 2004) and attackers of public officials (see e.g., James et al., 2007; Meloy & Amman, 2016). An identifiable grievance has also been raised as a characteristic of fixated individuals (e.g., individuals with pathological fixations on politicians, royalty or other public figures; Corner et al., 2018). A grievance is subjective, which also implies that it can be the result of either real or imagined causes (e.g., due to mental illness, Silver et al., 2018). In the threat assessment literature, it has further been raised that grievances often result in 'a desire, even a sense of mission, to right the wrong' (Silver et al., 2018; see also Calhoun & Weston, 2017). In most cases, the violent act may be seen as the only way to resolve a grievance (Calhoun & Weston, 2017). Importantly, it has been shown that in 82.4% of cases of lone-actor terrorist violence, other people were aware of the perpetrators grievance (Gill et al., 2014). In 63.9% of lone-actor terrorism activity, the perpetrator verbally told others of their plans and in 58.8% of cases the lone-actors produced some form of public statement about his or her beliefs prior to their plot. These communications happened both on- and off-line (Gill et al., 2017).

Nevertheless, the overwhelming majority of extremist and threatening posts will not lead to violence. Still, the spread of these messages is worrying due to the difficulties associated with predicting exactly which posters plan to actualise their threat or who will further radicalise towards violence. Moreover, threats and abuse not only cause fear and distress in targets but may also add to general unrest in society. The problem is not only constrained to niche corners of the web (e.g., forums such as Stormfront and 8chan), but is also continuously battled on mainstream social media platforms (e.g., on Facebook and Twitter). As a result, security professionals and tech companies are increasingly faced with the issue of understanding, countering and preventing the spread of these messages. Essentially, these parties are continuously engaged in threat assessment. This process is generally defined as 'the process of gathering information to understand the threat of violence posed by a person' (Meloy, Hart, & Hoffman, 2013, p. 4). The field of threat assessment is largely concerned with assessing possible acts of grievance-fuelled targeted violence, and is becoming increasingly relevant to such threats in the online domain (Simons & Tunkel, 2013).

### *A computational approach to threat assessment*

The large scale of online abusive language, threats, and extremist messaging impose a significant workload on both law enforcement (e.g., identifying public figure threateners) and tech companies (e.g., moderating social media platforms). Threats, abuse, and extremist content are often communicated through language. Therefore, a way in which to tackle the problem of scale is by means of computational linguistics. This field broadly analyses language by quantifying it using computer software, resulting in numerical features on which statistical operations can be performed. Research increasingly uses computational linguistics to gain insight into psychological processes. Applications of computational linguistics include detecting deception (e.g., Mihalcea & Strapparava, 2009; Bond et al., 2017; Newman et al., 2003), measuring emotion and gauging social relationships (see Tausczik & Pennebaker, 2010), as well as predicting traits of text authors (e.g., gender and age; Burger et al., 2011; Goswami et al., 2009). The relative success of linguistic analysis within the field of psychology and related fields suggests that a similar approach is worthwhile within the specific domain of studying targeted violence.

There are several benefits associated with a computational approach to threat assessment. First, automated procedures process large amounts of data in a short amount of time. Second, computer software can identify patterns in data that might be indistinguishable to a human analyst. Third, automatic procedures can be combined with human judgments, so that an algorithm serves as an initial filter, removing low-risk (irrelevant) texts, and thus reducing the workload for an expert threat manager who makes a final decision about whether an individual warrants attention. In recent years, academic research using computational linguistics to tackle threat assessment and grievance-fuelled violence in general has begun to emerge. Researchers increasingly study online extremist forums, lone-actor terrorist manifestos, and extremist content on social media platforms or messaging apps (e.g., Baele, 2017; Clifford & Powell, 2019; Scrivens et al., 2020). In general, these studies constitute applications of existing computational linguistics methods to the domain of grievance-fuelled violence. However, we know little about how threat assessment practitioners working in the field of grievance-fuelled violence approach cases from a *linguistic* perspective. We also do not know if and how computational methods will fit into this workflow, and what such methods should look like. The current thesis serves to address these issues.

### Aims and thesis outline

This thesis aims to discover whether and how threat assessment can be automated using computational linguistics. In order to address this aim, we first examine in which areas of threat assessment computational linguistics methods may be of use. In Chapter 1, a literature review describes the extant literature on applications of computational linguistics within the domain of grievance-fuelled targeted violence. This chapter shows that the computational capability is available, but reveals little about whether and how these methods are relevant to the work of threat assessment practitioners. In Chapter 2, we therefore present a qualitative interviewing study of expert threat assessment practitioners in order to provide insight into the current (non-automated) approach to threat assessment, particularly from a linguistic perspective. These results help to further delineate which areas of threat assessment can be automated, improved, or supplemented using computational linguistics.

The second way in which we address the overarching aim of this thesis is by developing and testing computational linguistics methods which are specifically attuned for use in threat assessment. Chapters 3 to 5 each address a relevant linguistic area of threat assessment identified through the expert interviews in Chapter 2. In Chapter 3, we examine *linguistic content* by developing and testing the Grievance Dictionary. This psycholinguistic dictionary can be used to measure 22 concepts related to grievance-fuelled targeted violence in text. Chapter 4 examines *linguistic style*, where we assess whether features of abusive language can be used to predict author characteristics such as age, gender, and personality traits. Chapter 5 deals with *linguistic trajectories*, or changes in language over time. By examining language use amongst the alt-right on YouTube surrounding the 2017 Charlottesville rally, we are able to model the online responses of extremist communities to offline events over time.

Finally, we provide a future outlook of what is needed to further improve our ability to automatically assess threats through language. In Chapter 6, we discuss the limitations and practical feasibility of the methods proposed in this thesis and outline necessary future work. We conclude with recommendations for research and practice on how to leverage linguistic data to better understand and counter-act grievance-fuelled targeted violence.

# Chapter 1: Literature review

## 1.1 Introduction

This chapter discusses the application of computational linguistics to grievance-fuelled targeted violence. Such methods have, for example, studied texts produced by lone-actor terrorists, school shooters, and extremist populations. Various approaches quantified texts, resulting in measures which can subsequently be used to perform statistical comparisons or machine learning predictions. This review of methods is by no means exhaustive of all methods in computational linguistics, but depicts the most common approaches used in the study of grievance-fuelled violent communications. In general, studies use open-source data collected through web-scraping or documents made available after violent attacks (by journalists, law enforcement, or researchers). The focus has largely been on identifying violent or radical individuals from within a larger sample through language (Kaati, Shrestha, & Sardella, 2016; Neuman et al., 2015; Scrivens et al., 2018; Smith et al., 2020), or comparing samples of (non-) violent documents obtained from different populations (Baele, 2017; Egnoto & Griffin, 2016; Jaki et al., 2019; Kaati, Shrestha, & Cohen, 2016).

This chapter broadly categorises methods utilised to study grievance-fuelled violence into top-down or bottom-up approaches. In this work, top-down approaches refer to methods where the linguistic measures taken within a sample of documents are pre-defined. Typically, a top-down method uses wordlists (dictionaries) in order to quantify text data, which means measurements are constrained to the words that appear in the pre-defined list. In contrast, a bottom-approach is data-driven in the sense that language is quantified based on the natural occurrence of words in the documents. That is, all words can be used for quantification and they are not pre-defined before measurement. Table 1.1 depicts the methods discussed in the sections hereafter. Each method can be categorised according to the top-down or bottom-up distinction, but due to variation within methods some may fall into both categories.

Table 1.1: Text quantification approaches

| Method | Approach category |
|---|---|
| Psycholinguistic dictionaries | Top-down |
| Sentiment analysis | Top-down or bottom-up |
| Abusive language detection | Top-down or bottom-up |
| Bag-of-words models | Bottom-up |
| Topic models | Bottom-up |
| Word embeddings | Bottom-up |

## 1.2 Psycholinguistic dictionaries

Psycholinguistic dictionaries are commonly used to study a range of psychological and social constructs through counting word occurrences (Pennebaker et al., 2015), and are widespread in studying grievance-fuelled language (e.g., Akrami et al., 2018; Baele, 2017; Figea et al., 2016; Kaati, Shrestha, & Cohen, 2016; Kaati, Shrestha, & Sardella, 2016). Psycholinguistic dictionaries count as a top-down approach, in that the specific words (and constructs) are pre-defined prior to quantifying documents. The assumption behind a psycholinguistic dictionary is that specific words in a document reflect the author's emotions and cognitive processes (Tausczik & Pennebaker, 2010). Dictionary software typically processes each word in a document and determines whether the same word appears in a proprietary dictionary, often organised into categories representing different concepts or constructs. A large majority of research utilising psycholinguistic dictionaries makes use of the LIWC, a standardised tool developed to measure a wide variety of psychological constructs (Pennebaker et al., 2015). Other researchers have also developed custom dictionaries which measure very specific terminology or constructs. Therefore, we discuss both approaches in turn.

### *1.2.1 LIWC software*

The LIWC is one of the most commonly used psycholinguistic dictionaries. It consists of almost 6,400 words measuring 90 constructs (Pennebaker et al., 2015). The LIWC organises words into descriptive categories (e.g., words per sentence, words longer than six characters), grammatical categories (e.g., pronouns, articles), psychological concepts and processes (e.g., power, positive emotion), personal concern categories (e.g., family, money), informal language (e.g., swearing, filler words), and punctuation (e.g., periods, commas; see Pennebaker, et al., 2015). LIWC output consists of proportions for each category, thereby providing an indication of the presence of certain constructs and processes in a text. In addition to the scores reported in terms of proportions, LIWC includes four summary variables (analytical thinking, clout, authenticity, and emotional tone). Scores for summary variables are represented in terms of percentiles, which are based on standardised scores on the four measures in large comparison samples obtained in previous research (Pennebaker et al., 2015).

In the context of understanding targeted violence, studies frequently use LIWC software to compare texts written by violent individuals to those written by non-violent individuals, in order to discern the characteristics of texts written by violent individuals (e.g., Baele, 2017; Kaati, Shrestha, & Cohen, 2016). For example, Baele (2017) examined a sample of texts written by lone-actor terrorists (*N* = 11) for various psycho-social variables. The study had two aims. First, it assessed whether texts written by lone-actor terrorists were characterised by higher levels of anger and negative emotion than texts written by non-violent individuals. Second, it investigated how lone-actor terrorists function cognitively, and measured LIWC categories 'cognitive processes' (a summary category including e.g., 'insight', 'causality', 'certainty', 'tentative'), words with more than six letters, and a separate measure of cognitively complex language (see Pennebaker et al., 2014). The paper compared scores for the lone-actor texts to those written by non-violent activists (e.g., Martin Luther King, Nelson Mandela), as well as samples of standard control writings and emotional writings ('baseline' texts provided by the LIWC developers expressing low and high emotionality, respectively). Lone-actor texts contained higher proportions of negative emotion words (including anger) than the non-violent activist texts, standard control texts, and emotional texts. Furthermore, the lone-actor texts showed similar scores to non-violent activists on cognitive processes, but scored slightly higher on cognitively complex language. Compared to control and emotional writings, terrorist writings scored higher on cognitive processes, as well as words with more than six letters and prepositions. In short, Baele (2017) argues that the psycho-social characteristics of lone-actor terrorist texts support the notion that perpetrators exhibit higher levels of anger than non-violent individuals and are characterised by high cognitive complexity.

In a similar vein, Kaati, Shrestha, & Cohen (2016) compared a sample of lone-actor terrorist writings to 'control' texts retrieved from personal blogs. They aimed to identify the drives and emotions preceding an attack using the LIWC. First, they demonstrated lone-actor texts contain significantly higher levels of negative emotion as well as significantly lower levels of positive emotion and friendship-related words than the control texts. Second, lone-actor texts scored higher in terms of power-related language as well as anger when compared to the control writings. The lone-actor texts also showed higher proportions of the LIWC category 'certainty' than the control writings, which was used to measure the extent of cognitive flexibility in the texts. The category denoting 'big' words longer than six letters measured psychological distancing from a violent act, and was found more often in the lone-actor group. Finally, the lone-actor texts contained more third person pronouns than the control texts, which was considered a response to outgroup threat and 'us versus them' thinking (Kaati et al., 2016).

In an exploration of an 'incel' (i.e., 'involuntary celibate') forum, several linguistic analyses assessed whether the forum fostered radicalisation (Jaki et al., 2019). This misogynistic online community is united by their perceived injustice of women not being attracted to them (Baele et al., 2019; Hoffman et al., 2020). Some acts of lone-actor violence have been committed by individuals with similar beliefs (Elliot Rodger, Alek Minasian; see Baele et al., 2019). Comparing 50,000 messages from an incel forum to 50,000 'neutral' control texts extracted from Wikipedia articles and random English tweets via LIWC software, Jaki et al (2019) found that incel messages contained more swear words, personal pronouns, adverbs, and negative adjectives, but fewer positive adjectives than the control texts. The incel texts also scored higher on the sub-categories of negative emotion (anger and uncertainty categories) and social inhibition (avoidance and anxiety categories). In terms of personal concerns, the incel texts discussed relationships and sexuality to a larger extent than the control texts, but family, work, hobbies, goals, and beliefs to a lesser extent (Jaki et al., 2019).

A similar approach compared 'legacy tokens' of spree killers (*N* = 21, including journal entries, suicide notes, video transcripts and manifestos) to a comparison sample of student writing (*N* = 20,000, Egnoto & Griffin, 2016). Similar to previous work on lone-actor terrorists, LIWC categories for negative emotion and anger particularly distinguished between samples.

The studies discussed thus far have largely reported statistical comparisons between different samples on LIWC measures. In contrast, studies discussed from here onwards have used LIWC software in the context of machine learning. Here, the aim often is to discriminate between texts written by violent individuals and texts written by non-violent individuals (i.e., predicting whether a text was written by a violent individual). These studies use the LIWC dictionary categories as features for a supervised machine learning algorithm. In short, the algorithm learns to what extent certain LIWC categories are present in violent-author and non-violent author texts (the training set), then considers the presence of these categories in unseen texts (the test set) and classifies the texts in this test sample as violent-author or non-violent author based on this information.

For example, Kaati, Shrestha, & Cohen (2016) examined ten lone-actor terrorist manifestos for psychological warning signs of targeted violence using LIWC software. Kaati, Shrestha, & Sardella (2016) followed the same procedure for texts written by non-violent activists, texts from personal blogs, forum postings on Stormfront (a white supremacy forum), and personal interest forum postings. Classification algorithms then distinguished between the terrorist texts and the non-offender texts. All LIWC categories were used as features. In one of the experiments where the aim was to distinguish between terrorist texts and Stormfront posts, important features for classification were LIWC categories relating to negative emotion (e.g., 'sad', 'angry'), time (e.g., 'before', 'often), seeing (e.g., 'appear', 'show'), differentiation (e.g., 'but', 'without'), biological processes (e.g., 'eat', 'blood'), and cognitive processes (e.g., 'imagine', 'admit'), in addition to linguistic categories such as articles (e.g., 'a', 'the'), personal pronouns (e.g., 'he', 'me'), prepositions (e.g., 'above', 'onto'), and quantifiers (e.g., 'bunch', 'more').

A synthesis of these efforts forms the basis of a tool for risk assessment in written communication called the Profile Risk Assessment Tool (PRAT; Akrami, Shrestha, Berggren, Kaati, et al. 2018). Although the tool does not aim at classification per se, it has a predictive aim in that it can be used to compute the similarity between a text of interest and texts in the comparative samples it offers, namely texts extracted from jihadist forums, right-wing extremist forums, texts written by school shooters, and lone-actor terrorist manifestos and communications. A normative comparison sample is also provided in the form of personal blog posts. The PRAT computes intra-class correlations between the unseen text and the comparison samples based on a dictionary-based approach similar to the LIWC, measuring concepts including social processes, leakage, and fixation behaviour (Akrami et al., 2018). PRAT measures warning behaviours by assessing references to killing, power, weapons, military terms, as well as mentions of well-known previous lone offenders and school shooters. PRAT additionally measures custom dictionary categories covering Judaism, migration, Islamisation, Islamic State terminology, and ‘involuntary celibate’ terminology (Akrami et al., 2018). The PRAT developers also describe constructing a personality profile for the text authors, based on findings from previous research stressing the importance of personality factors to behaviour in general as well as political extremism in particular (Thomsen et al., 2014). It is suggested that the PRAT assesses the ‘Big Five’ personality factors through language, resulting in scores for neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Unfortunately, specific linguistic features for the personality factors are not explicitly described, but LIWC dictionary categories such as anxiety, negative emotions, causality, friends, work, and social words are reportedly measured. The tool is also used to measure sentiment in texts, which is described in the respective section hereafter.

Another study focuses on racism, aggression, and worries on a white supremacy forum (Figea et al., 2016). Three independent human annotators scored 300 posts from the forum on a scale from 0 to 7 for each of the aforementioned affects. Thereafter, several different linguistic variables were extracted from the dataset to serve as features: all LIWC categories, the number of misspellings, the number of words, part-of-speech tags (e.g., nouns, verbs), and words from three ‘expert knowledge dictionaries relating to worries, racism and aggression’, as well as the 100 most frequent words in posts with a high affect (i.e., level 6-7), and the 100 most frequent words that differed between high and low affect posts. Important linguistic characteristics for classifying racism posts were LIWC categories for religion (e.g., ‘Muslim’, ‘church’), seeing (e.g., ‘view’, ‘saw’) and third person pronouns (e.g., ‘they’, ‘them’). The LIWC categories for anger (e.g., ‘hate’, ‘kill’) and an expert dictionary category for aggression were important for recognizing both worries and aggression in the posts.

### *1.2.2 Custom dictionaries*

Due to the often highly specialised language among extremists, several studies also developed custom dictionaries to apply a word-count based approach. Typically, word lists are created that include terms specific to an ideology, such as racist slurs and hate symbols (e.g., WPWW: white pride worldwide; 88: Heil Hitler, where ‘8’ is a representation of the eighth letter of the alphabet ‘H’) for the extreme right (Kleinberg, van der Vegt, & Gill, 2020; Scrivens, Burruss, et al., 2020). An example of this approach modelled language use on right-wing extremist forum Stormfront, in order to assess the effect of the 2008 Obama and 2012 Trump elections using ARIMA timeseries intervention modelling (Scrivens, Burruss, et al., 2020). The total number of posts, as well as the posts that included right-wing extremist terms, and posts that referred to firearms were measured 120 days before and after each election. While firearm posts did not change as the result of either election, both the total number of posts and right-wing extremist posts increased after both events. However, the volume of all three types of posts was markedly higher during the 2008 Obama election, suggesting that 'political defeat' has a bigger effect on online behaviour of the extreme right than 'political victory' represented by the Trump election (Scrivens, Burruss, et al., 2020).

A similar approach to modelling language use over time on Stormfront made use of a combination of profane language and racial slurs as custom dictionary for extremist language (Kleinberg, van der Vegt, & Gill, 2020). The analysis of language use between 2002-2015 showed that extremist language on the forum increased in a stepwise manner until approximately mid-2011, followed by a decrease. Analyses of individual users also showed that a small percentage of users (10%) accounted for the overwhelming majority (90%) of both forum activity and extremist language (Kleinberg, van der Vegt, & Gill, 2020).

In a study of Twitter users who voiced support for Daesh[1], both LIWC categories and a custom dictionary of Daesh vernacular were measured (Smith et al., 2020). The authors do not share the word list for ethical reasons, but state it included derogatory terms directed at non-Daesh supporters, as well as English and Arabic transliterations of Daesh-relevant terms. First, the Daesh-tweets were classified from neutral baseline Tweets using the standard and custom dictionaries, resulting in 89% accuracy. Second, a linear mixed model was used to predict within-user changes in conformity to 'extremist linguistic style' represented by the use of Daesh-vernacular and Daesh-specific use of function words. Daesh-supporters indeed showed an increase in conformity to extremist language across time represented by their account age in days (Smith et al., 2020).

### *1.2.3 Limitations of psycholinguistic dictionaries*

A widely acknowledged limitation of a dictionary approach to automated linguistic analysis is the constrained (i.e., top-down) nature of any given dictionary. Because of the difficulties associated with creating a fully comprehensive dictionary, the possibility exists that important linguistic markers of certain concepts go missing. Furthermore, the meaning of certain words in a dictionary can be highly context-dependent (Akrami et al., 2018; Chen, 2008). Some words may be incorrectly considered simply because they appear in a dictionary. For example, the word 'ape' may be considered a derogative slur (de Gibert et al., 2018), but may also occur in a context where the animal is described. Furthermore, some words may not be picked up due to misspellings or spelling variations, especially when studying noisy online text data. Although solutions to this problem exist, such as automatic spelling correction or substitution based on a dictionary of word variations (see e.g., Han & Baldwin, 2007; Clark & Araki, 2011), this problem may not be adequately circumvented in all cases. Moreover, even though some custom dictionaries can be used to extract specific terms

[1] Daesh is an alternative, pejorative name for the Islamic State jihadist group (Irshaid, 2015)

used by right-wing or jihadi extremists (e.g., Abbasi & Chen, 2007; Figea et al., 2016; Kleinberg et al., 2020; Scrivens, 2020), the specific jargon might be highly sensitive to linguistic adaptation where users change or introduce new terms to evade filters on platforms (see e.g., van der Vegt et al., 2019).

A further matter to consider is that humans often construct dictionaries. The LIWC documentation (Pennebaker et al., 2015) notes that categories and word instances were constructed through brainstorm sessions and crowd-sourced initiatives on discovering word associations. Such a procedure is highly sensitive to human biases that may influence which words are included in the word categories and dictionary. Likewise, the dictionary critical to the PRAT tool was developed through consultations with domain experts on the topic of risk (Akrami et al., 2018). Experts suggested themes for the PRAT dictionary which were then supplemented with distributional semantic models, and consequently manually verified by experts again. Furthermore, Abbasi & Chen (2007), Figéa et al. (2016) and Smith et al. (2020) also report relying on expert annotations or suggestions to construct custom dictionaries for concepts such as violence and aggression.

Whilst consultation of domain experts might promote the validity and applicability of a tool, the exact procedure of this consultation remains opaque. That is, characteristics of the experts are not described, therefore the reader cannot verify the quality of the judgments given by the experts in question. In addition, little information is given as to how and why experts selected certain words and concepts for inclusion in the dictionary. It is highly likely that various custom dictionaries that may have been intended to measure the same construct (e.g., extremism), vary in terms of validity and scope, because the procedure of constructing the dictionaries varies across research groups. Unfortunately, the exact procedures of development for the dictionaries are sparsely documented and the content of dictionaries are rarely made publicly available (see e.g., Smith et al., 2020).

### 1.3 Sentiment analyses

Sentiment analyses aim to measure the polarity of a document, specifically whether the language in a document is positive, neutral, or negative in nature (Mohammad, 2016). There are several ways to conduct sentiment analysis, which fall into either top-down or bottom-up approaches. The simplest way is to measure the frequency (or proportion) of negative and positive polarity words in a document. This method can for example be conducted with the LIWC, which includes categories for positive and negative emotions words. A somewhat more advanced and widely known approach to sentiment analysis is similarly dictionary-based (see Pang & Lee, 2008), but uses lists where words are assigned a weight signifying polarity. Weight-based sentiment analysis provides an average polarity score instead of (two) proportion scores. Specifically, this top-down approach uses sentiment dictionaries in which a large number of words' polarity scores (e.g., ranging from -1 for highly negative words to +1 for highly positive words) are represented. When a sentiment analysis is conducted on a text, the words appearing in the sentiment dictionary are extracted and assigned the according weight. After correcting for text length, an average sentiment score is typically reported for a piece of text.

An alternative approach to sentiment analysis does not make use of a dictionary. Instead, an algorithm learns positive and negative polarities based on texts annotated for polarity (e.g., Duwairi & Qarqaz, 2014). Due to the data-driven nature of this approach, this method would constitute a bottom-up approach. However, the majority of work in grievance-fuelled communications has used a top-down, weight-based dictionary approach, thus the current section largely focuses on that method.

Sentiment analysis can sometimes also refer to the analysis of other affectual states, such as anger, surprise, sadness, and joy (Mohammad, 2016). Again, emotions can be measured with a dictionary approach, where words are scored for the extent to which they represent a specific emotion. For example, the PRAT tool measures positive sentiment, negative sentiment, and anger, arguing that emotions are an important predictor of violent extremism (Akrami et al., 2018). Further studies discussed below similarly apply sentiment analyses to extremist, radical, grievance-fuelled and lone-actor texts.

Sentiment analysis identified radical users of Islamic web forums (Scrivens et al., 2018). The 100 most frequent nouns across four forums (Gawaher, Islamic Awakening, Islamic Network, and Turn to Islam) were identified in order to define a list of common keywords in the dataset. Then, a sentiment analysis on the context surrounding each keyword in a forum posting, resulted in an average sentiment score for each post. Scrivens et al (2018) considered four components to identify radical forum users: 1) the average sentiment score across all posts from a single author, 2) the volume of negative posts, consisting of the number of posts with a negative polarity value, as well as the number of negative posts in proportion to all posts), 3) severity of negative posts, a score based on the number of very negative posts and the number of very negative posts in proportion to all posts, and 4) the duration of negative posts, referring to the time difference (date) between the first and last negative post. All components result in a score between 1 and 10. This resulted in a ‘radical score’ ranging between 1 to 40 when summed. Based on the radical score for 26,171 users across all four forums, Scrivens et al. (2018) found the most radical users were concentrated within two forums (Islamic Awakening and Gawaher), with the most radical poster achieving a score of 39.03 out of 40. The authors emphasise that no single profile or behaviour pattern was found that describes a radical author. Instead, the results stress the importance of considering multiple factors when analysing online radical behaviour.

A similar approach studied three types of radical right-wing posting behaviour on the Stormfront forum, which focuses on users whose radical posting behaviour can be characterised as high-intensity, high-frequency, or high-duration (Scrivens, 2020). A similar sentiment analysis procedure was conducted as in Scrivens et al. (2018). The authors conducted further qualitative analyses for samples of the 100 most radical posters for each of the three groups. Results showed that high-intensity posters shared highly negative posts for a short amount of time (155 days), often using alarming words (e.g., ‘bomb’, ‘kill’) and advocating violence against Jews in particular. High-frequency posters generally shared a high volume of posts (513 on average) and highly negative messages in particular over a longer period of time (170 days), discussing adversary groups but not necessarily advocating violence. Lastly, high-duration users generally posted over a long period of time (2,864 days), with very negative messages similar to high intensity posters (155 days). Scrivens (2020) suggested that the long duration of general posting possibly illustrates high commitment to the forum.

Some earlier work also applied sentiment analyses to extremist and violence-related texts. Abbasi & Chen (2007) measured the intensity of hate and violence on American and Middle Eastern dark web forums. The authors utilised a custom lexicon containing words and phrases from the forums related to violence and hate, each manually scored for intensity (on a scale from 1-20). They compared messages from 16 U.S. supremacist and Middle Eastern extremist group forums against the hate and violence lexicons. Results indicated that Middle Eastern forums scored significantly higher than American forums in terms of violent affect intensity. Forums from both regions did not differ in terms of hate affect intensity. In addition, they found a correlation between hate and violence across messages and forums from the Middle East. In a similar vein, Chen (2008) proposed an automated method for analysing affect within two jihadist dark web forums. Up to 909,039 messages were collected from the forums, of which 500 were utilised to manually construct a custom lexicon for violence, anger, hate and racism affects. One of the forums, anecdotally known to be more radical, was indeed found to contain higher levels of violence, anger, hate, and racism than the other (Chen, 2008). A simple sentiment analysis indicated that the entire radical forum could be classified as having a negative sentiment polarity, while the moderate forum was found to be neutral in terms of sentiment polarity.

### *1.3.1 Limitations of sentiment analysis*

A dominant characteristic of sentiment analysis is to report sentiment in a static manner. That is, a single score is computed representing the polarity of a document, or even across a whole sample of documents. For example, Scrivens et al. (2018) compute a cumulative ‘radical score’ based on average sentiment across all forum postings of a user, as well as the volume, severity and duration of negative posts. Chen (2008) computed a single average sentiment score across all posts of two entire forums. A downside to this approach is that shifts in sentiment may be obscured when a text is represented by a single score. If a text author is highly negative in one section of text, but highly positive in another, the polar scores may average each other out, resulting in a neutral sentiment score for the text (see also, Jockers, 2015a; Kleinberg et al., 2018). In such cases, the resulting single static score is not informative, especially when the purpose is to detect highly negative language use within or across texts or pathways towards negative language use across texts. It has been suggested that sentiment deterioration or escalation may be useful linguistic signals of threat (Spitzberg & Gawron, 2016), a process that can only be linguistically measured if language use over time is considered. Indeed, within the context of mitigating extremist violence, it may be of particular interest to model trajectories towards more radical language or detect bursts of positive or negative sentiment prior to a violent attack. Examples of this include Kleinberg et al. (2020) and Scrivens et al. (2020). This trajectory modelling can be performed within texts or across multiple texts. Modelling the trajectories of sentiment within texts has already been applied in other domains, including novels (Jockers, 2015a; Reagan et al., 2016), Ted Talks (Tanveer et al., 2018) and YouTube videos (Kleinberg, Mozes, et al., 2018; Soldner et al., 2019). Besides sentiment progression, the trajectory of LIWC categories has also been modelled within texts to represent the narrative structure of texts (Boyd et al., 2020).

### 1.4 Abusive language detection

The study of abusive and hateful language may be of particular interest to linguistic threat assessment, considering the likelihood of grievances being communicated through such language. Similar to sentiment analysis, abusive language can be measured with a top-down or bottom-up manner. The former approach uses a dictionary of abusive words, whereas the latter uses posts annotated for abusiveness to learn which words constitute abusive language. In the top-down approach, some research can be considered weight-based (like sentiment analysis) when the *intensity* of abusive language or hate speech is measured (Davidson et al., 2017; de Gibert et al., 2018; Mondal et al., 2017; Sahlgren et al., 2018).

Recent years witnessed a sharp increase in research aimed at detecting abusive or hateful language, mainly on social media platforms. Previous work defined ‘abusive language’ as an overarching term for hate speech, profanity, and derogatory language (Nobata et al., 2016). Attempts at measuring and detecting abusive language and hate speech frequently make use of weight-based approaches. The website Hatebase[2] is a common resource for this. Hatebase records hate speech terms and an associated rating (weight) for offensiveness, as well as their sightings across the world. The website categorises hate speech terms into nationality, ethnicity, religion, gender, sexual discrimination, disability, and class. The repository has for example been used to measure hate speech on Twitter and Whisper (Mondal et al., 2017), as well as to classify hate speech and offensive language (Davidson et al., 2017). De Gibert et al (2018) extracted 10,568 sentences from the Stormfront forum that were then manually annotated for containing hate speech or not (defined as: ‘a deliberate attack directed

[2] https://hatebase.org/

at a specific group motivated by aspects of the group's identity', de Gibert et al., 2018). For this particular dataset of Stormfront posts, words with the highest hate score include 'ape', 'scum', and 'savages' (de Gibert et al., 2018). A classification error analysis demonstrated the importance of context sentences to correctly identify hate speech, as well as the difficulty for classifiers to possess 'world knowledge' (e.g., knowing that 'hoax' may refer to the Holocaust in white supremacist communities).

A special case of sentiment analysis has been used in the context of abusive language. Words with a negative polarity from the 'Subjectivity Lexicon' (Wilson et al., 2005) were rated for their abusiveness through crowdsourcing (Wiegand et al., 2018). A sample of 500 nouns, adjectives, and verbs with a negative polarity were judged for abusiveness (yes/no) by five annotators, and words were only considered abusive if four out of five annotators agreed. In addition to this base lexicon, other linguistic features were also considered for later detection of abusive language. Wiegand et al. (2018) determined a finer-grained measure of intensity by extracting the words from online reviews (including abusive reviews of persons[3]) and computing the weighted mean of the star ratings in which the word occurred (Wiegand et al., 2018). Among other things, further linguistic features included affect categories, sentiment views (the perspective of the opinion holder of a polar expression: Wiegand et al., 2016), semantic associations, and word embeddings (see the respective section below). Wiegand et al. (2018) then applied the initial lexicon extended by the linguistic features to detect abusive posts in several different datasets (Twitter, Wikipedia comments). They found that the lexicon supplemented by aforementioned features performed better at detecting abusive online posts than other lexicons considered, such as the Hatebase lexicon.

### *1.4.1 Limitations of abusive language detection*

Some limitations pertaining to abusive language detection need to be considered. First, several different definitions of abusive language exist, and it is important (but difficult) to distinguish between hate speech, derogatory language, and profanity (Davidson et al., 2017). Specifically, hate speech has been defined as 'language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, sexual orientation, or gender identity' (Nobata et al., 2016, p. 149). In other words, the target of hate speech is typically a disadvantaged group in society. In contrast, derogatory language is considered to attack an individual or group, but is not considered part of the hate speech categories above (e.g., 'yikes.. another republiCUNT weighs in..', Nobata et al., 2016, p.149). Profane language contains sexual remarks or profanity (Nobata et al., 2016). The importance of distinguishing between hate speech and offensive language has also been raised in an experiment aimed at classifying tweets containing these forms of language (Davidson et al., 2017).

[3] The underlying assumption is that 1-star review of persons (e.g., celebrities, politicians) are abusive comments. These were found on the website http://www.rateitall.com/

The authors describe instances in which offensive terms are used without being hate speech. If no distinction between hate speech and offensive language is made, such instances may be erroneously classified as hate speech.

Besides definitional issues, abusive language similar to sentiment and affect, is typically measured with single scores. In the context of grievance-fuelled language, the field may particularly benefit from measurement of abusive language over time (across texts), potentially signifying radicalisation or escalation towards violence. A notable example includes the custom dictionary including used to measure the temporal evolution of the Stormfront forum, which also included profane language (Kleinberg, van der Vegt, & Gill, 2020).

### 1.5 Bag-of-words models

In a bag-of-words (BoW) model, a piece of text is represented as an *unordered* vector of terms, hence the term ‘bag’. This approach would be considered a bottom-up method, in that it is fully driven by the terms that appear in the document(s) of interest. Each term in the vector can be represented in terms of its frequency, but other, more informative measures also exist. A common approach is to weigh terms in the bag-of-words vector for how informative they are about the text in question (Manning et al., 2008; Sahlgren et al., 2018). That is, terms in the BoW model are represented with a TF*IDF (term frequency * inverse document frequency) weighting, where TF represents the frequency of a term and IDF often represents the inverse of the number of documents in which the terms appears (Manning et al., 2008; Salton & Buckley, 1988)[4]. BoW models can be built for different types of *n*-grams, where *n* represents the number of terms, such as unigrams (e.g., ‘terror’, ‘violence’), bigrams (e.g., ‘donald trump’, ‘threat assessment’) and trigrams (‘war on terror’, ‘middle eastern country’).

A BoW approach is frequently chosen in the context of machine learning, where a BoW model can for example be used as features to classify a text as being abusive versus non-abusive, or as terrorism-related or not. Often, a BoW model will serve as a baseline model to test a classifier, after which other models are tested that include additional or different sets of features. In one such example, tweets were classified as belonging to a pro-ISIS dataset versus a non-pro ISIS dataset (e.g., journalists reporting on ISIS), where the baseline model consisted of a TF*IDF weighted unigram BoW model (Fernandez & Alani, 2018). This was compared to a model where BoWs were enhanced by semantic contexts; unigrams were annotated for categories (e.g., ‘unrest, conflicts and war’, ‘religion and belief’), topics (e.g., ‘terrorism’, ‘Taliban’), named entities (e.g., ‘geopolitical entity’, ‘organisation’), and entity types (an ID associated with the named entity, see Fernandez & Alani, 2018). Indeed, semantic contexts enhanced classification performance when compared to the BoW model

[4] Various other forms of document and term weighting exist, see Manning et al. (2008) for an overview.

alone (Fernandez & Alani, 2018), achieving 0.82 classification accuracy (0.82 precision, 0.80 recall) for the BoW model and 0.85 (0.86 precision, 0.84 recall) for the enhanced model.

In another example of a BoW approach unigrams alone were used as features to classify whether posts from a jihadist website had a recruitment purpose or not (Scanlon & Gerber, 2014). A recruitment post would, for example, describe 'a golden chance to join Jihad in Somalia' or a 'pitch for new overseas recruits' (Scanlon & Gerber, 2014). TF*IDF weighted unigrams were used as features to classify posts with different statistical models. Highly discriminating unigrams included 'nigeria', 'hamas', and 'jihad', among others.

### *1.5.1 Limitations of bag-of-words models*

A notable limitation of BoW models is that they are highly data-dependent. A BoW model represents the words in a corpus of interest, and is difficult to generalise to other corpora. For example, a BoW model trained on jihadi Tweets will have limited applicability in classifying texts from an extreme-right corpus, due to differential *n*-gram use. Bag-of-words models also do not take into account co-occurrence of words (context) and word order (Wallach, 2006). As has been raised previously, context can be highly important to the interpretation of (extremist) terms.

## 1.6 Topic modelling

Another way to represent text data from a bottom-up perspective is by means of topic modelling, a method that extracts underlying, 'latent' topics in (a collection of) documents. Topic models rely on co-occurrences of common words, in that the co-occurring words in a corpus 'president', 'white house', and 'trump' may form a topic which could subsequently be labelled as 'presidency'. Since topic models rely on co-occurrences, they can address the problem of contextual ambiguity present in dictionary and BoW approaches. Indeed, to understand the use of the word 'jihad' in a text it may make more sense to look at the words that co-occur with it, where words such as 'de-radicalisation' and 'prevent' in the same topic may hint at a condemning or reporting context, and words such as 'infidel' and 'war' may hint at a jihadist context. A common approach to topic modelling is Latent Dirichlet Allocation (LDA), a probabilistic model which is based on the assumption that a piece of text consists of a mix of topics, which in turn are a mix of probabilities of words more likely to co-occur together under the topic (Blei et al., 2003; see also Grimmer & Stewart, 2013; Saif et al., 2017). Topic modelling usually starts with a term-document matrix, in which terms (often words) are represented on each row, with columns representing each document, and each cell representing the frequency of a term in a given document. LDA is then used to produce a topic distribution for each word, which shows the degree to which a word belongs to the different topics identified (Schmidt & Wiegand, 2017). Next, a topic-document matrix is produced, which showcases the strength of each topic for each document, thereby showcasing differences between documents. In short, this approach allows researchers to automatically discover the topics in a large volume of texts, which may be of particular use for large-scale linguistic threat assessment.

Examples of a topic model approach to grievance-fuelled language include a case study of a Tumblr blog of a young woman, examining the process of radicalization (Windsor, 2018). The study analyses the changes of specific topics in the woman's language use in the period leading up to and during her travel to ISIS-territory in Syria (Windsor, 2018). Topic modelling showed self-references (e.g., 'I', 'me') decreased over time, whereas other-references (e.g., 'them', 'they') increased. Religion topics were also prevalent at the onset of radicalisation. In addition, positive emotion reflected in her language use seemed to increase after her emigration to Syria, whereas negative language decreased (Windsor, 2018).

Topic modelling has also forecasted online recruitment of violent extremists on a jihadist discussion forum (Scanlon & Gerber, 2015). Specifically, it was hypothesised that recruitment does not happen randomly, but that recruiters target communities or time periods in which individuals would be vulnerable to propaganda (e.g., following an attack, Scanlon & Gerber, 2015). Known recruitment posts were analysed with LDA, then the topics were used to forecast cyber-recruitment activity on a given day (number of recruitment posts). Indeed, topics related to conflict (including terms 'attack', 'kill', 'jihad') often preceded recruitment posts on the forum.

Another application of topic modelling includes categorizing extreme right videos on YouTube (O'Callaghan et al., 2013). A dataset of YouTube channels and videos referred to by known extreme right Twitter accounts was created, and a topic model was created for the metadata of the videos (titles, descriptions, and keywords). Thereafter, they performed a manual categorization of the topics (e.g., a topic containing 'hitler', 'adolf', and 'reich' was categorised as Neo-Nazi). In addition, they examined which channels best represented each topic and the corresponding category (e.g., anti-Islam, anti-Semitic, conspiracy theory), based on topic assignment weights (O'Callaghan et al., 2013).

### *1.6.1 Limitations of topic modelling*

Importantly, topic models involve subjectivity: interpreting and assigning labels to topics must be done by humans, a process which is highly sensitive to bias (e.g., confirmation bias: interpreting topics so that they are in line with the researcher's hypothesis: Nickerson, 1998). Furthermore, even though various methods have been proposed (Mimno et al., 2011; Roberts, Stewart, & Tingley, 2014), there is also no single agreed upon way to define the "right" number of topics for a corpus (Grimmer & Stewart, 2013).

## 1.7 Word embeddings

A final way to quantify language, which aims to approximate semantic aspects of language, is known as word embeddings. Word embeddings are vector representations of words derived by learning word co-occurrences in a (large) corpus. The resulting embedding is a representation for each word in a document as a continuous vector in an *n*-dimensional space (see e.g., Pilehvar & Camacho-Collados, 2020). Importantly, words that are semantically similar tend to have vector representations that are closer to each other in the vector space than words that are semantically less similar. For example, 'man' and 'woman' have vector representations close to each other, as do 'cat' and 'kitten'. The proximity between two vectors in that space is defined by a distance measure, such as the Euclidean or cosine distance. Word embeddings are characterised as a bottom-up, unsupervised approach in that they rely fully on the data itself (Pilehvar & Camacho-Collados, 2020).

The introduction of two models has popularised the use of word embeddings, namely Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The former relies on a neural network to iterate through training data, whereas the latter is count-based in that it fits a model on a co-occurrence matrix of the whole corpus (for a detailed explanation of both models, see also: Pilehvar & Camacho-Collados, 2020). Embeddings can be obtained in two ways. The first is to generate corpus-specific embeddings, in that word co-occurrence statistics are computed using the corpus of interest. In the case of word2vec, this is done for individual words via local context (Mikolov et al., 2013), whereas GloVe makes use of global co-occurrences in the whole training corpus (Pennington et al., 2014). When using word embeddings, one can either re-train models on the data of interest or use pre-trained word embeddings[5] where words in a corpus of interest are represented by embeddings that have been learned from a very large corpus. Using pre-trained embeddings has proven particularly popular within the field of computational linguistics, because generating one's own embeddings is computationally intensive and requires very large amounts of data (Pilehvar & Camacho-Collados, 2020). However, within the domain of grievance-fuelled language, researchers frequently opt for corpus-specific embeddings, for example due to the domain-specific language use in extremist groups (Simons & Skillicorn, 2020; Voué et al., 2020).

Once embeddings are obtained for individual words in a document, they can be combined into a vector that represents the whole document. A common method used to achieve this simply averages across the vectors for each word in a document (i.e., an average for each of *n* dimensions per word vector, see also Arora et al., 2016; Pilehvar & Camacho-Collados, 2020). Other alternatives also exist, including TF-IDF weighting of averages, simple addition of vectors, or the use of neural networks (Arora

[5] Generated in the same way as for corpus-specific embeddings, locally (Mikolov et al., 2013) or globally (Pennington et al., 2014) depending on the model used.

et al., 2016; Pilehvar & Camacho-Collados, 2020). The subsequent representations of documents can then be used as features for machine learning. Such an approach has for example been applied in the context of detecting hate speech, where word embeddings as features for classifying racist and sexist tweets outperformed methods where simple bag-of-words were used (Badjatiya et al., 2017). Other work represented online comments (on the Yahoo website) with embeddings to classify comments as ‘hateful’ or ‘clean’ (Djuric et al., 2015; Nobata et al., 2016). Again, the embeddings approach outperformed simple bag-of-words approaches.

Vector representations also measured the presence of personality and other mental health disorders in texts written by school shooters and a sample of writings from non-offender males (Neuman et al., 2015). Word embeddings for various personality traits and disorders were used (e.g., 'depressed', represented by the words 'anxious', 'angry', 'suicidal', 'sad', etc.), and their distance to vectors representing texts written by school shooters were computed (a special model for comparing vectors of words and texts was used, see also Neuman & Cohen, 2014; Turney, 2012). The assumption is that this measure of distance represents the extent to which a certain personality trait or disorder is present in the text (i.e., the closer a vector of a text is to a vector of interest, the more similar it is to it). When the school shooter texts were compared to male blogger texts, school shooter texts had smaller distances to the vectors representing 'revengeful' and ‘humiliated', and those associated with narcissistic personality disorder (e.g., 'arrogant', 'egocentric'). From there, different statistical models predicted whether certain texts originated from a school shooter, producing a ranking of texts that needs to be screened in order to identify all school shooters. The authors state that this ranking and prioritization method requires one to search only 3% of the entire corpus in order to identify all school shooter texts. Neuman et al. (2015) argue that such an automated method could provide practitioners with 'red flags' for a small amount of texts that express potential danger, reducing the amount of texts that required attention.

Simons and Skillicorn (2020) applied a notable use of word embeddings to sentences from three extreme-right corpora including Stormfront, the Iron March forum, and the manifesto of the New Zealand Mosque attacker. First, templates of language expressing intent were defined, which generally include a first-person pronoun (‘I’), a desire verb (‘will’) and an action verb (‘kill’). These templates inferred whether sentences included language expressing intent (1) or not (0) by bootstrapping, using both an *n*-gram based model and a deep learner. Second, the presence of abusive language was predicted for the same segments based on pre-existing training data annotated for abusive language from Stormfront, a dataset of insults, and Wikipedia comments. Here, word embeddings were used as features to predict whether the segment included abuse (1) or not (0). As a result of both approaches, each text segment was assigned a score between 0 and 1 for both intent and abusive language, where the product of both scores was presented as “abusive intent”. The outcome of these predictions were compared to human judgments of abusive intent, resulting in

80% agreement between computational and human judgments (Simons & Skillicorn, 2020).

### *1.7.1 Limitations to word embeddings*

Like all methods reviewed in this chapter, the use of word embeddings is also met with various limitations. Word embeddings disregard polysemy, where a word with multiple meanings (e.g., 'arms' can represent weapons or the body part) is represented by a single vector, although some solutions to this issue have been suggested (e.g., Tian et al., 2014). Word embeddings also encode existing social and cultural biases, where for example certain occupations ('doctor') are associated with specific genders ('he') (Khattak et al., 2019). In the special context of grievance-fuelled communications, one limitation is of particular importance. Word embedding models are considered uninterpretable to humans, because of the semantic representation of words in low-dimensional space. Because of this, the method is sometimes referred to as a "black box" (Şenel et al., 2018). This is particularly problematic in a practical context such as threat assessment, were EU and UK regulations dictate that algorithmic decision making should be explainable (Goodman & Flaxman, 2017; Oswald et al., 2018)

## 1.8 Conclusion

The automatic linguistic analysis of grievance-fuelled texts has received some attention recently. With an increasing presence of aggrieved, radical, and extremist populations in the online sphere, it has become even more important to understand, detect, and counter their messages. The body of research reviewed thus far has highlighted the various ways in which this can be approached, covering both bottom-up and top-down methods. Although considerable strides have been made at discovering linguistic features that are of interest when studying potentially violent populations, some work remains to be done before such methods can be considered for implementation in practice.

Thus far, grievance-fuelled targeted violence has served as an interesting use case for applications of computational linguistics. However, we still know little about how actual threat assessors deal with cases from a linguistic perspective in practice. The development of computational approaches could potentially be informed by expert insight and thereby strongly improved. This thesis serves to unite the approaches of threat assessment practitioners with the available technical capabilities, and thus it is important to first examine current approaches to threat assessment in practice from a linguistic perspective. The next chapter presents a qualitative interview study of expert threat assessment practitioners, aimed at revealing how current (non-automated) linguistic threat assessment is performed. The purpose of this investigation is to further define which areas of threat assessment can be automated, improved, or supplemented using the computational linguistics methods reviewed in this chapter.

# Chapter 2: Assessment procedures in written threats of harm and violence

## 2.1 Introduction

Assessing threats of grievance-fuelled targeted violence often involves examining a written component, for example when cases concern repeated social media threats directed at public figures, contamination threats to businesses, or bomb threats sent to public places (Calhoun & Weston, 2017; Simons & Tunkel, 2013). Anonymous cases are particularly relevant to the domain of linguistic threat assessment because the language in the communication itself is often the only evidence available. With the rise of the internet, anonymously communicating threats became easier, partly due to the ease of access to targets through social media or e-mail and (perceived) untraceability, for example (Simons & Tunkel, 2013). These circumstances have also increased the scale at which threat assessors need to work, suggesting that a computational approach, particularly focussed on language, may reduce workload and aid in identifying patterns in large-scale data. However, little is known about how threat assessors approach anonymous written cases, particularly from a linguistic perspective. Knowledge about these procedures is necessary in order to identify how threat assessment can be automated through computational linguistics.

This chapter elucidates different procedures by which experienced threat assessment professionals approach an anonymous threatening communication (ATC). Through semi-structured interviews built around a real-life case, we examine the differences and similarities in threat assessment procedures across professionals. Based on a thematic analysis of responses, we highlight the areas of ATC assessment where linguistic information is considered in order to inform future automation of these processes.

## 2.2 The nature of (non-)anonymous threats

When assessing a threat from a known source, threat assessors will typically gather and evaluate personal, historical, contextual, and clinical information about the threatener (Monahan et al., 2001). In the case of an anonymous threat, such information is typically lacking. In some cases, the communication itself will be the only information available. A further complicating factor is that an anonymous threatener may include misleading information in their threat to conceal their intent or identity (e.g., an individual posing as a group of threateners).

A significant body of research concerning threats and assessment procedures in the case of a known source exists (e.g., Borum et al., 1999; Calhoun & Weston, 2017;

Meloy & Hoffmann, 2014), but little has been written about anonymous threats and how to conduct assessments in such cases (Simons & Tunkel, 2013). One possible explanation for this is that threatening communications (studied) in the past often were *not* anonymous. For example, in a 1991 study of threatening letters to US Congress members, 81% of writers gave their full name, 74% gave an address, and 86% gave some (other form of) identifying information, whereas just 14% of letter writers remained anonymous (Dietz et al., 1991). Similar patterns were observed in letters directed at celebrities (*N* = 214), with just 5% of writers maintaining anonymity and the remainder giving some form of identifying information in their (first) communication (Dietz et al., 1991).

Although threat authors may continue to identify themselves in some cases, the internet has significantly facilitated the process of anonymously sending threatening communications. Simons & Tunkel (2013) predicted that anonymous threats would become more common and time-consuming, noting at the time that the majority of threats handled by the FBI's Behavioural Analysis Unit were anonymous in nature. Although the (previously) low rate of anonymous threatening communications may explain the lack of research in this area thus far, we acknowledge that this is an increasingly common issue exacerbated by online communication, and thus warrants further attention (Simons & Tunkel, 2013, 2021; Wallace, 2015). Moreover, it has been previously suggested that computer-mediated communication further increases disinhibition and lowers behavioural constraints (see e.g., Wallace, 2015), potentially leading to an increased prevalence and severity of anonymous threats.

### 2.3 Previous work on anonymous threat assessment

Simons & Tunkel (2013) described the procedure of assessing anonymous threats adhered to by the North-American Federal Bureau of Investigation's (FBI) Behavioural Analysis Unit (BAU). It is emphasised that a team of evaluators is needed to carry out this process. The BAU will also designate a lone assessor who assesses the ATC in isolation from the rest of the team in order to manage confirmation bias. The lead member of the team will gather and distribute information on the threat. This includes seeking to answer triage questions relating to the delivery of the threat, the feasibility of the threat, a possible relationship between the victim and offender, the characteristics of the target, and other contextual information (Simons & Tunkel, 2013). In contrast, the lone assessor focuses on possible linguistic staging (e.g., the use of plural pronouns by a possible single author), possible motives, cues to deception (e.g., level of detail in the threat), as well as the resolution to violence and its imminence. Although the lone assessor also has access to information on the ATC's mode of delivery, they are not provided with further contextual information regarding the target, again in an attempt to avoid confirmation bias. After the threat is independently reviewed by the other team members, they will reconvene to reach a group assessment. Thereafter, the lone assessor presents their assessment to the team which was achieved without access to background and victim information. The full team will then re-evaluate their conclusions in light of the new assessment. The final resulting assessment typically discusses threat-enhancing and mitigating factors, as well as a recommended threat management strategy (Calhoun & Weston, 2017; A. Simons & Tunkel, 2013).  Although the described procedure using a lone assessor may not be applicable or possible in all circumstances, the authors importantly state that threat assessment teams need to have an established, consistent procedure in place for the effective management of anonymous threat (Simons & Tunkel, 2013).

A notable example of an SPJ tool applicable to ATCs is The Communications Threat Assessment Protocol-25 (CTAP-25), which has been specifically developed for the assessment of problematic communications (James et al., 2014). The CTAP-25 was not developed for anonymous cases specifically, but due to its focus on risk factors that can be derived from the communication itself, it may be of particular relevance to anonymous cases where a threat manager does not have much information besides the communication. Although the developers state that the instrument makes use of empirically defined risk factors, an evaluation has yet to be published  (Geurts et al., 2017).

There have also been some efforts made at the automatic assessment of violent threats, which includes ‘Threat Triage’ (Smith et al., 2013). This software is aimed at coding threatening messages based on language, by assessing eight content indicators. In the study of language use, linguistic *content* generally refers to the language use under the conscious control of the author (e.g., use of specific nouns and verbs) whereas linguistic *style* is assumed to not be under one’s conscious control (e.g., use of pronouns and spelling mistakes; see e.g., Goswami et al., 2009; Holmes, 1998; Pennebaker, 2011). The content categories assessed by Threat Triage include conceptual complexity (considering different factors vs. seeing things in simple terms), paranoia, naming or identifying the victim, mention of love, marriage or romance, polite language, specification of harm, and whether or not the threatener has contacted the victim before (the latter is a non-linguistic indicator). Each indicator is seen as either increasing or decreasing the risk of targeted violence, based on the indicator values found in previous research on 89 cases for which the outcome (violence or non-violence) was known. Together, the indicators are used to predict whether an unseen threatening message should be considered high, moderate, or low risk. Although the tool is based on a limited sample size, the authors note that they are continuously updating their database with new cases (Smith et al., 2013).

All in all, even though a procedure in assessing ATCs has been documented (Simons & Tunkel, 2013) and some tools (James et al., 2014; Smith et al., 2013) aimed at assessing worrying communications have emerged, no single standardised procedure for the assessment of ATCs exists. Moreover, we presently know little about how different threat assessors approach such cases, and whether they use aforementioned procedures or tools at all. The present chapter serves to address this issue.

## 2.4 The current study

The present research aims to gain a better picture of assessment procedures in the case of anonymous threatening communications, in order to inform future endeavours of automated linguistic threat assessment. Research on assessment procedures is prevalent in the case of non-anonymous cases, but little is known about how experts approach a case based merely on communications and very little contextual information. In a semi-structured interview, expert practitioners are asked about their general approach to ATCs, as well as their thoughts regarding a specific real-life case. Responses are qualitatively coded, in order to reveal common themes and divergent approaches. Thereafter, we highlight the main areas of linguistic threat assessment raised by experts in the interviews, and discuss implications for possible automation.

## 2.5 Method

### *2.5.1 Transparency statement*

Materials (interview questions) for this study are available on the Open Science Framework: https://osf.io/5twzu/

### *2.5.2 Participants*

Threat assessment experts were approached through the professional contacts of the involved authors, some of which conduct research on threat assessment or are threat assessment professionals with several years of experience in the public and private sector. Interviewees were further approached at the 2019 *Association for European Threat Assessment Professionals* conference (Rotterdam, The Netherlands), as well as the 2019 *Grievances and Grudges* conference (Cambridge, UK). The majority of interviews was conducted at the conferences, or shortly after via video calls. Through snowball sampling further participants were identified, resulting in a total sample of 13 interviewees.

### *2.5.3 Procedure*

The study procedure was approved by the Ethics Committee of the UCL Department of Security and Crime Science. After participants read the study information sheet and signed a consent form, the interviews were started. The semi-structured interviews took about one hour to complete (see the Open Science Framework page for all interview questions). Participants were first asked about their background within the field of threat assessment and their experience with anonymous threats in particular. Then they read two copies of anonymous threat letters provided by law enforcement. Both letters were handwritten on paper and contained 134 and 159 words, respectively. Identifying information of the case (names of persons or places) were redacted. The participants were provided with some contextual information, namely that the first letter was sent to a supermarket branch and the second to a local church approximately two months later. Although both letters were presented as part of the case, participants were not informed whether the two letters were connected. They only knew that the hypothetical security manager of the supermarket who contacted them to assess the case was in possession of both letters. That is, the supermarket and church were in the vicinity of each other and both received worrying communications, and the latter party shared their communication with the manager. In the first letter, the author threatened to contaminate products in the supermarket and demanded a large amount of cash in order to evade police. In the second letter, the author threatened to torture a child and bomb the church, in addition to demanding the church stop its services. The interviewers were not aware of the outcome of the case or of any contextual information beyond that provided to the interviewees.

Participants had up to twenty minutes to read the communications, but most did not take the full amount of time. They were then asked about various aspects of the case, including their thoughts on possible characteristics of the author, the level of risk associated with the case, and appropriate security measures. Finally, interviewees were asked about their confidence in assessing the case. Participants were not directly asked about linguistic factors that stood out in the communication, in order to assess which themes emerged naturally.

### *2.5.4 Analysis*

All interviews were transcribed (*M* = 3,577 words, *SD* = 1,739) and subsequently analysed for recurring themes using NVivo software (QSR Internation Pty Ltd., 2014), which can be used to qualitatively organise and analyse interview data. A theoretical thematic analysis was performed, in that the interview questions drove the coding process (in contrast to an inductive approach that is independent of interview questions, see: Braun & Clarke, 2006). Each interview question was considered as a separate node, and responses categorised according to themes. For example, the node ‘professional background’ may contain themes such as ‘psychology’, ‘psychiatry’, and ‘police’, with sections of interview transcripts highlighted for the theme they belong to (e.g., ‘I attended the police academy’ would be coded as part of the theme ‘police’). Consequently, percentages can be calculated for each theme (i.e., the proportion of interviewees who mentioned the theme). It must be noted that percentages for each theme do not always add up to 100% because a single participant may have mentioned several themes in their response. For example, a participant may have obtained a degree in psychology *and* attended the police academy. All percentages reported in the results were derived from this thematic analysis using NVivo.

## 2.6 Results

### *2.6.1 Background of sample*

On average, the thirteen participants had 18.10 years of experience with threat assessment, ranging between 6 months to 30 years. A large number of interviewees worked for or owned a private threat assessment company (46.15%). Others worked for the police (30.77%), government (7.69%), mental health services (7.69%), or a university (7.69%). Their work concerns a wide variety of cases within the domain of grievance-fuelled targeted violence, such as threats to public figures, school violence, workplace violence, and counter-terrorism.

The majority of the participants had a background in psychology (84.62%), with two participants having a psychiatry degree (15.38%) and some (additionally) attended a police or military academy (23.08%). Besides degrees in forensic psychology and psychiatry, no specific formal training (e.g., a degree or course) in threat assessment was mentioned. Participants mentioned they learned through experience, mentorship and attending conferences:

> *"I joined the Association for Threat Assessment Professionals in [YEAR] and started going to those meetings, presenting there. And that's really my training. And going to courses, going to people's lectures, teaching it."*

> *"And the model of learning how to do threat assessment, still today, but definitely back then, it was brand new as an art and science. So, it was taught through mentorship."*

Interviewees generally reported to have worked many anonymous threat cases (46.15%) ranging between a dozen to 100 a year. Some were unsure how many (7.69%) or reported to work just a few (30.77%):

> *"Very few, very few [anonymous cases]. Just the vast majority of the cases are being signed because people want to make their point. [...] The vast majority, they just sign and give their details and where they live, and telephone number and email accounts."*

The majority of interviewees stated they were confident when dealing with anonymous cases in general (53.85%). Others were moderately confident (7.69%), not confident (7.69%) or unsure about their confidence level (7.69%). On the topic of confidence, one participant added:

> *"I do not work on it myself so the confidence derives from the fact that I work in a team. So, it's like a team confidence – I'm very*

> *confident in the team. And in the ability of the team to change their hypothesis."*

*2.6.2 General approach to anonymous threats*

Before reading the case, participants were asked about their general approach to assessing anonymous communications. Several approaches were mentioned. The majority of interviewees mentioned they would want to seek further information besides the communication (76.92%):

> *"I would ask for any other information that they have. Did they have any other types of communications? Any other strange incident that they might worry about in the last couple of months."*

> *"I want to have information on possible suspicious activities around the victim or not, and I want to know about possible conflicts. I want to know more about the victim itself."*

Several interviewees also mentioned they would examine the content of the threat (46.15%); for example, to look for aggressive or threatening language. Some participants also mentioned they would consult other experts (61.54%) such as a linguist or a private investigator. Some participants mentioned they make use of a professional judgment tool, framework, or software (38.46%) such as the CTAP-25, the JACA principle[6] or threattriage.com, respectively.

*2.6.3 First thoughts on case*

After reading the two letters, the interviewees were asked to report their first thoughts on the case. The majority of interviewees made remarks about language use in the communications (92.31%). Of these remarks, 69.23% said something about linguistic content, and 46.15% said something about style (note these do not add up to 92.31% because some interviewees mentioned both content *and* style). Remarks about content included:

> *"He uses a lot of the same words [...] 'smashed', 'clairvoyant', 'torturing'."*

> *"They [the letters] follow a similar pattern. 'I'm going to do some really, really terrible things, and I've already done that terrible thing and I'm going to do another terrible thing.'"*

[6] Assessments based on searching for Justification, Alternatives, Consequences and Ability of a suspect (see: de Becker, 1998)

> *“He is expressing some violent ideas or maybe even fantasies, about what he will do, to a child, to a boy. And it’s not very specific, but it is very violent and concerning.”*

> *“Another word that got my attention is the word ‘can’, that's also a sign for me he is not really committed yet. It seems he hasn't decided yet what to do and not to do.”*

When discussing linguistic style, interviewees noted spelling and grammar, as well as possible characteristics of the author influencing writing style:

> *“Just the way they spell, the way the person has put their sentences together. Again, there’s a lack of punctuation.”*

> *“[It could be] the poor educational level of the writer, so the spelling mistakes, poor grammar, not very well thought out.”*

> *“There are some spelling mistakes. I wonder, but my English is not good enough to judge that, whether this is a native English speaker.”*

> *“Also, there is similarity in use of language, although this one is more coherent than this one, which could suggest that the person is either under the influence of drugs here and less here, or something.”*

Other interviewees also mentioned the handwriting in the communications (30.77%):

> *“I’m looking at kind of the graphic nature of the note. One of the things you would see is use of the… whenever he uses the letter ‘g’, you will see a tail on the ‘g’ moving from right to left in both letters. And it’s very… also with the ‘y’’s too. And it’s very consistent throughout the note.”*

When discussing first thoughts, several participants already mentioned they deemed the risk of violence to be low (84.62%). Several different possible motivations were mentioned when discussing first thoughts (69.23%), such as a financial motive or hoping to create fear. Participants also made remarks about possible indicators for a mental health disorder (61.54%). Almost half of the participants made remarks about wanting to seek further information (46.15%).

> *“I would ask around as well to see if other people have any letters of this individual and whether they have been acted upon.”*

> *“I would again wanna check delivery method, postage, was it put in an envelope, where are the postage stamps.”*

### *2.6.4 Author characteristics*

Two letters were presented to interviewees and no information was given regarding the relation between the two. The majority of participants expected the two letters to have been written by the same author(s) (53.85%). The majority also expected there to be a single person, rather than a group, behind the threats (84.62%). Most interviewees expected the author to be male (69.23%), but several also mentioned they didn't know for sure (46.15%).

> *"Well, I mean gender wise the tone of it is male."*
>
> *"Again, that used to be easier, my overall inclination would be a male. The graphic nature of the violence would be more male. Severe mental illness can compromise that, maybe gender becomes less important than other pathology. But I would probably guess a male."*
>
> *"Well, that would be based, for me, on the fact that most letter writers like this would be males. So, I would say the probability is a male writer. I would not try to draw a conclusion to gender from the writing itself, I just don't have the skills to be able to do that."*

With regards to the relation between the author and target, several interviewees noted the possibility of the author living in the neighbourhood (46.15%), some suspected they were a former employee (30.77%) or frequent visitor (23.08%) of the target. Three interviewees noted there was no relation between the author and target (23.08%).

### *2.6.5 Motivation*

Some participants already mentioned possible motivations when asked about their first thoughts (69.23%), but interviewees were also asked to specifically state what they thought was driving the author of the ATCs. Several possible motivations for sending the threats were mentioned, with the majority expecting it to be financial (53.85%). Other possibilities included creating fear (46.15%), harbouring a grudge (46.15%), resulting from mental illness (38.46%), or seeking attention (30.77%).

### *2.6.6 Personality and mental health*

Regarding the possible personality and mental health of the threat author, a large variety of possibilities were raised in the interviews. Delusions were a common theme in responses to questions about both personality and mental health. Although some participants did not want to make any claims about these issues, several possible diagnoses were mentioned by others, such as Autism Spectrum Disorder (ASD), psychopathy, and psychoticism. They are depicted in Table 2.1. Please note that participants were asked about personality and mental health separately, but certain themes appeared in response to both questions (e.g., delusions, narcissism)

Table 2.1 Personality and mental health themes mentioned

| **Personality** | | | **Mental health** | | |
|---|---|---|---|---|---|
| *Theme* | *%* | *n* | *Theme* | *%* | *n* |
| Don't want to say | 38.46 | 5 | Delusions | 53.85 | 7 |
| Delusional | 38.46 | 5 | Disordered (unspecified) | 30.77 | 4 |
| Inadequate | 23.08 | 3 | Psychotic | 30.77 | 4 |
| Unstructured | 23.08 | 3 | Paranoia | 15.38 | 2 |
| Rational | 15.38 | 2 | ASD | 15.38 | 2 |
| Antisocial | 15.38 | 2 | No disorder | 15.38 | 2 |
| Low intelligence | 15.38 | 2 | Don't want to say (specific) | 15.38 | 2 |
| Sadist | 15.38 | 2 | Personality disorder | 7.69 | 1 |
| Don't know | 7.69 | 1 | Histrionic personality disorder | 7.69 | 1 |
| Frustrated | 7.69 | 1 | Borderline | 7.69 | 1 |
| Narcissistic | 7.69 | 1 | Psychopathy | 7.69 | 1 |
| Lack of empathy | 7.69 | 1 | Narcissistic personality disorder | 7.69 | 1 |
| | | | Grandiose delusions | 7.69 | 1 |

When discussing personality traits of the author, some interviewees referenced language in the communications, particularly its content. Some also stated the information in the communications was not enough to draw any conclusions:

> *“’Hunted’ by the police is an interesting word, not being chased, hunted is persecutory. Again, that could be [...] because there are quite a lot of paranoid suspicious feelings about this.”*

> *“So, the person’s thoughts weren’t very organised in their mind and they would tend to change topics rapidly and would then do… associate different sentences in ways that are considered very loose and non-sequential. So, that kind of the thinking is very disorganised in the person, and you see this in both of the notes [...] It’s very tangential, and very non-sequential. And that’s an indication of a mental disorder for the person.”*

> *“I couldn't, based on two letters. Personality is something that is a very stable pattern in somebody’s lives, how he interacts with different people and how he lives. And there is too little information in just two letters to say something about his personality.”*

When discussing mental health, similar points about linguistic content emerged:

> "*Probably [suffers from a mental disorder]. For some of the reasons said before, this erratic nature to this, it is not fully rational. [...] So those references to those supernatural powers are not unusual. The reference to the violent themes, and the variety of violent themes [...]"*
>
> *"When you talk about psychotic people, you expect even more chaos, or more illogical sentences. So, in that sense, but I'm not a psychiatrist, it's still quite coherent. It's not that incoherent as you see with other psychotic people, but it doesn't have to be the case."*

### *2.6.7 Risk level*

Twelve out of thirteen interviewees expected the author to send similar letters to other targets (92.31%). Nevertheless, the majority of interviewees stated the risk level of the case was low (76.92%). One interviewee said it was moderate (7.69%), whereas two said it was not possible to determine (15.38%). Some interviewees (23.08%) also discussed whether and how the threat may escalate:

> *"It's becoming a lot more descriptive in terms of what he wants to do, it's becoming quite violent. I do think it'll become increasingly, as he unravels, he will become increasingly violent yes."*
>
> *"Well certainly, this [second letter] is more intense so sure it could get racked up. In the language, it's more inflammatory, which could indicate some deterioration in metal state."*

### *2.6.8 External expertise*

Participants were also asked what kind of external expertise they would involve, if any, to assess the case. Several participants said they would consult a handwriting expert (53.85%), a linguist (46.15%) or the police or private investigator (46.15%). Some interviewees would wish to discuss the case with other threat assessment professionals (23.08%).

### *2.6.9 Confidence in assessment*

When asked about their confidence level for the case they just assessed, the majority of participants noted they were confident (61.54%), moderately confident (23.08%), or that they had low confidence (15.38%):

> *"Absolute 100% confidence in saying here's what I can tell you based on the letter, here investigative leads we can run, here's my*

*assessment, here're some management strategies. I have full confidence in doing that because I do it every day."*

*"Well, at this stage I have to say everything is a hypothesis. So, it's not about me saying this is my judgement, goodbye. And that's a 60% confidence."*

## 2.7 Discussion

The interviews revealed several important points about the (linguistic) assessment of anonymous threatening communications, which are summarised in this discussion. We discuss linguistic content and style factors raised, linguistic trajectories, handwriting analysis, and inconsistencies in assessment.

### *2.7.1 Linguistic content and style*

While not explicitly prompted to do so, most interview participants made reference to linguistic information when discussing the process of assessing ATCs. This is perhaps not unexpected, as participants were merely presented with communications and very limited contextual information. However, responses can potentially reveal *which* linguistic factors are of particular interest in (automated) linguistic threat assessment. When asked about their first thoughts on the presented case, the majority of interviewees discussed language use, such as spelling and violent themes. When discussing characteristics of the author, such as gender, personality and mental health, factors such as linguistic 'tone', word use, and (in)coherence were mentioned. Moreover, several interviewees stated they would consult a linguist in further assessing the case.

The responses discussing language use can be broadly categorised as relating to linguistic content or style. This distinction is frequently made within the psychological and social study of language use (Goswami et al., 2009; Pennebaker & King, 1999; Schler et al., 2006) and seems to be equally relevant here. Linguistic content for example covers examining specific word use or topics (e.g., violence and delusions), whereas linguistic style would cover spelling and grammar, as mentioned by some interviewees in this study. Linguistic style is also related to remarks made about possible characteristics or (mental) states of the author that influence writing, such as their educational level, drug use, personality and mental health. These remarks relate to the notion that linguistic style is not under the conscious control of a speaker, and that this measure can be used as a marker of individual differences (Pennebaker & King, 1999). Indeed, there is some evidence for relationships between language and author characteristics such as gender (Newman et al., 2008), age (Pennebaker & Stone, 2003), and personality (Pennebaker & King, 1999). However, there has been little research thus far on the relationship between author characteristics and language use within the specific domain of threatening communications. All in all, even though interviewees raised several different linguistic indicators during their assessment, responses can be summarised as belonging to either content or style. Consequently, we suggest that future development of automated linguistic threat assessment focuses on these domains of language use, paying attention to the specific content and style factors that play a role in grievance-fuelled targeted violence. Therefore, subsequent chapters of this thesis will examine both content (Chapter 3) and style (Chapter 4) approaches.

### *2.7.2 Linguistic trajectories*

Another important matter that emerged from the interview responses is possible escalation in risk of violence, assessed by means of language by some interviewees. This matter was highlighted by the fact that some interviewees compared the language use at the beginning and end of the letters, as well as between the first and second communication (for those who believed both letters were written by the same source), with some stating the second communication was more intense or violent. Assessing possible escalation through language is particularly relevant to the study of grievance-fuelled targeted violence, which frequently deals with processes of radicalisation or repeated worrying communications. These processes can only be assessed through language if one considers the trajectory of language use over time.

The importance of trajectories of language has also been raised previously as an important measure within the context of targeted (extremist) violence (Spitzberg & Gawron, 2016). Considering trajectories was also raised as a possible solution to the limitations of static measures of sentiment and abusive language in Chapter 1. Trajectories of language use can generally be measured in two ways, both of which have been raised in the interviews. One can measure the changes in language use within a text, where content or style features are measured throughout the progression of a single document (Boyd et al., 2020; Jockers, 2015b; Kleinberg, Mozes, et al., 2018). Alternatively, one can measure the temporal trajectory of linguistic features across multiple texts (Kleinberg, van der Vegt, & Gill, 2020; Scrivens, Davies, et al., 2020), which for example applies to measuring extremist language across multiple messages. Although the current study only considered two communications (presented as possibly originating from the same person), it is imaginable that practitioners will be confronted with this issue at a larger scale, for example in the case of ‘superusers’ on extremist forums (Kleinberg et al., 2020) or public figure threateners who send several thousand messages (Simons & Tunkel, 2013). Therefore, in addition to linguistic content and style, we argue that trajectories of language use, as noted in the interviews conducted, are another important area to consider in linguistic threat assessment. For this reason, Chapter 5 of this thesis applies the measurement of linguistic trajectories to grievance-fuelled language.

### *2.7.3 Handwriting*

Besides linguistic content, style, and trajectories, several interviewees also made remarks about the handwriting in the communications, and some suggested enlisting the help of handwriting experts (graphologists). With the increasing threat of online violent communications, handwriting analysis may be of less relevance to large-scale threat assessment operations. More importantly, the validity of graphology has been called into question for decades and, within the scientific community, is widely regarded as debunked (e.g., Dazzi & Pedrabissi, 2009; King & Koehler, 2000; Neter & Ben-Shakhar, 1989). Therefore, it is also surprising that several interviewees still reported to rely on handwriting analysis. Although it is beyond the scope of this thesis, future research may examine to what extent practitioners still make use of graphology and promote awareness of its limitations.

### *2.7.4 Inconsistencies in assessment*

Although a large number of interviewees made use of linguistic information in their assessment, there was little consistency in terms of the specific factors that were considered. For example, while most interviewees agreed the communications were authored by a single person who was male, there were no specific linguistic variables that united participants in drawing these conclusions. Further incongruencies in the assessment of the case also emerged, for example concerning the suspect's personality and mental health. These matters provided grounds for disagreement, illustrated by the wide variety of possible traits and diagnoses raised by participants. Several interviewees agreed the threatener must suffer delusions (both when asked about personality and mental health), but an equal number of interviewees declined to speak on the suspect's personality based on the communications alone. When interviewees did speak on these matters, some mentioned specific word use or disorganised writing, but did not necessarily engage in a structured assessment of linguistic information. Furthermore, even when several interviewees mentioned the same linguistic variable, it is difficult to determine (based on the available data) if all participants defined the concept in the same way. For example, to some experts, incoherent writing may constitute tangential topics, whereas others may be referring to incorrect word use or spelling. Other interviewees also raised linguistic cues that are unspecific or lack specific (scientific) definitions, such as linguistic tone and 'intense' language.

All in all, both the linguistic factors considered and the conclusions drawn from this information were highly inconsistent between interviewees. In practice, this may lead to highly divergent assessments, management strategies, and profiles of the suspect, depending on the threat manager who is enlisted for a case. One way to combat this issue is by using a Structured Professional Judgment (SPJ) tool, in which an assessment protocol is laid out which covers possible risk factors for violence. For communicated threats, one such tool already exists in the form of the CTAP-25 (James et al., 2014). However, the manual nature of such ‘checklists’ makes it difficult to scale its use to a large amount of communications. For this reason, a computational approach may be of use. By using pre-defined linguistic factors, a computational system can also provide structure and promote consistency between assessments. As raised previously, the current thesis serves to explore the use of computational methods in the threat assessment context.

However, the interviews demonstrated that the majority of interviewees did not make use of threat assessment software. There are several possible explanations for this. Practitioners may not be aware of its existence, or may not see the necessity for using such software. Another possible explanation is that the variables measured in existing software do not align with those that are of interest to threat assessment practitioners. For example, Threat Triage (Smith et al., 2013) largely examines content categories, and does not consider linguistic style or trajectories. The content categories that it does examine may not be a comprehensive representation of indicators typically assessed by practitioners. The current thesis serves to expand on existing initiatives by examining content indicators that are derived from those considered by expert threat assessors, in addition to testing the feasibility of linguistic style and trajectory measures in the domain of grievance-fuelled targeted violence.

## 2.8 Limitations

There are important limitations to be noted with regards to this study. Most importantly, the circumstances in which the experts performed their threat assessment will have been strongly different from their day-to-day practice. In practice, experts may take more time for their assessment, work in teams, or make (more) use of tools or software. It is also a likely possibility that several claims made by interviewees in the study will not have been made if they were assessing the same case in real-life. The setup of this study will undoubtedly have allowed participants to speculate more, and thus we do not wish to claim that participants assess their real-life cases in precisely the same way. Nevertheless, we have sought to minimise the effect of the ‘artificial’ setup of this study on the interview responses and outcome of analyses. For example, participants received a lot of freedom when assessing the case (e.g., possibility to take notes) and were asked on several occasions which procedures and tools they would apply in a ‘normal’ situation. Furthermore, participants were never encouraged to make claims they were uncertain about and had ample opportunities to caveat their statements or refrain from responding to a question. Indeed, several participants did choose to do so, for example in response to questions about the author’s personality and mental health. All in all, it is important to appreciate both the possible limitations to the generalizability of this study, as well as the efforts made to minimise the effect of the study design on the results.

It is also important to note that the perpetrator in the real-life case used in this study was apprehended before they could actualise their threat. Therefore, it was not possible to code (or quantitatively assess), for example, the number of interviewees who correctly judged the risk of violence as low. For privacy reasons, it is also not possible to disclose the extent to which participants were right about characteristics of the perpetrator. However, the aim of this study was never to test the accuracy of threat assessment experts, but rather to demonstrate how practitioners come to their assessments using linguistic information. That is, the contribution of this chapter is that the assessments in itself differed strongly between practitioners, and that the linguistic indicators used to arrive at these assessments were similarly inconsistent.

Related to this, the sample size of this study does not allow us to draw any statistical conclusions. Future research may focus on gathering a larger sample size (e.g., through online recruitment) on which statistical analyses can be performed. If another case (for which the outcome is known) is used, the relationship between assessment accuracy and professional background or linguistic cues considered could be examined. However, we hope the qualitative patterns uncovered in the current study lay the groundwork for future quantitative study of ATC assessment. More importantly, we believe this qualitative investigation is necessary to assess current practice in threat assessment, and can serve as a starting point for automation initiatives.

## 2.9 Conclusion

Drawing from interviews with an expert group of threat assessors, we have highlighted several areas of language use which are considered in the assessment of ATCs. Similar to other domains of psychological linguistic study, factors can broadly be defined as belonging to linguistic content or style. In this thesis, we introduce linguistic trajectories (measuring language over time) as an additional important factor to consider in linguistic assessment of grievance-fuelled violent threats. Specific factors that made up assessments of linguistic content, style and trajectories were relatively inconsistent between interviews. This could hold worrying implications for subsequent assessments of violence risk, security strategies, and suspect profiles. Therefore, increased structure may be provided by means of SPJ tools or automation of linguistic assessments. The focus of this thesis is on the latter, due to the increasing prevalence and scale of online threats. In the three subsequent chapters, we examine computational assessment of linguistic content, style, and trajectories, respectively. While implementing these measures, we also take note of the limitations to each computational linguistic method raised in Chapter 1.

Chapter 3 examines *linguistic content* by translating cues used in traditional threat assessment to a psycholinguistic dictionary for automated analysis. This chapter describes the development of the Grievance Dictionary, which can be used to measure 22 content variables related to grievance-fuelled targeted violence in text. Chapter 4 investigates whether *linguistic style* can be used to obtain a profile of abusive text authors, in a study of the relationship between gender, age, and personality with abusive language use. Chapter 5 describes how *linguistic trajectories* can be measured within the context of grievance-fuelled violent texts. It specifically models online extremist language use over time in response to an offline event. Together, the subsequent computational chapters of this thesis address three areas of language that emerged from the analysis of interviews in this study.

# Chapter 3: Linguistic Content: Introducing the Grievance Dictionary

## 3.1 Introduction

The previous chapter demonstrated the importance of linguistic content of grievance-fuelled communications to threat assessors. In order to scale the measurement of linguistic content and deal with the limitations of manual efforts, computational methods are needed. One way in which to computationally measure linguistic content is by means of a psycholinguistic dictionary. As has been described in Chapter 1, the use of psycholinguistic dictionaries, particularly the LIWC, is already widespread in violence research (e.g., Akrami et al., 2018; Baele, 2017; Kaati, Shrestha, & Cohen, 2016; Kaati, Shrestha, & Sardella, 2016). Other studies have also developed custom dictionaries to measure right-wing extremist or jihadist content (Abbasi & Chen, 2007; Kleinberg, van der Vegt, & Gill, 2020; Scrivens, 2020; Smith et al., 2020). Beyond the field of grievance-fuelled violence, other general dictionaries (e.g., Wmatrix, Rayson, 2008; Empath, Fast et al., 2016; Moral Foundations Dictionary, Frimer et al., 2019; IBM Watson Tone Analyzer[7]) are also used and measure different concepts and categories. The current chapter discusses the development of the novel Grievance Dictionary, which is specifically designed for the study of linguistic content in grievance-fuelled communications. Before doing so, we expand on the limitations to psycholinguistic dictionaries discussed in Chapter 1, by highlighting issues that will be specifically addressed in this chapter.

## 3.2 Limitations to existing dictionaries

The psycholinguistic dictionaries frequently used in grievance-fuelled violence research are met with three important limitations. Firstly, standard psycholinguistic dictionaries have not been developed for the purpose of assessing grievance-fuelled language and therefore do not measure content that may be of interest to researchers and threat assessment practitioners (i.e., they may lack specificity). Although the LIWC provides categories such as anxiety and anger, we argue that key concepts for threat assessment and violence research are absent in this and other dictionaries. As a result, previous work on grievance-fuelled violence that used the LIWC (e.g., Baele, 2017; Kaati, Shrestha, & Cohen, 2016) may not have been specific enough in terms of the linguistic measures used to indicate potential violence.

Second, the content and construction procedure of existing (custom and standard) dictionaries is often unclear, because details on the motivations and procedures for

[7] https://www.ibm.com/watson/services/tone-analyzer/

the inclusion of certain words are lacking. Yet, it is vital to be transparent about the development of these dictionaries because of the far-reaching consequences of false positives and negatives within the context of threat assessment. In the UK, the ALGO-CARE framework suggests that algorithms used in the context of policing need to be explainable, in that decision-making rules and the impact of each factor on the outcome is available (Oswald et al., 2018). In short, it is highly important for practitioners and researchers to understand the capabilities and limitations of a given dictionary. Many available dictionaries and threat assessment software (e.g., PRAT: Akrami et al., 2018; Threat Triage: Smith et al., 2013) are not transparent. That is, the contents of wordlists or other 'under the hood' operations are not available to its users, and thus cannot be adequately evaluated or explained. This possibility is desirable and necessary if such systems are to be used in practice. Moreover, opaque dictionaries are not suitable for researchers aiming to engage in open science (see e.g., Masuzzo & Martens, 2017), a movement gaining traction in several fields including terrorism research (Schumann et al., 2019), because the precise operations performed to produce measurements are not available.

Third, custom dictionaries are often domain-dependent and non-transparent regarding the population of experts consulted. By consulting domain experts (e.g., in right-wing extremism, radical Islam) the dictionaries are specifically attuned to a specific type of violence or extremism. The nature of online communication in these populations is that language is community-specific and constantly changes (Farrell et al., 2020; Shrestha et al., 2017). Some fringe communities may also continuously adapt their language use to evade content moderation filters on social media platforms which automatically delete or flag posts with specific word use (van der Vegt et al., 2019). As a result of these phenomena, dictionaries would have to be continuously updated to capture the appropriate jargon. Furthermore, custom expert dictionaries are referenced in Abbasi & Chen (2007), Chen (2008), Figea et al. (2016) and Smith et al. (2020), but little is said about what the consultation process entailed and why those consulted can be considered experts. In short, readers are expected to trust the judgment of the researchers and experts without having access to the specifications of the tool.

### 3.3 The current study

To address the aforementioned limitations, this chapter describes the development of the Grievance Dictionary, which specifically aims to measure psychological and social content that is of interest in the context of grievance-fuelled violence threat assessment. First, the Grievance Dictionary is specifically aimed at measuring concepts that are of interest in threat assessment and violence research and practice. Its aim is to supplement measures obtained through dictionaries such as the LIWC with concepts that are specifically relevant to the threat assessment domain. Second, the Grievance Dictionary is transparent in terms of its construction and final format. All data collected are made available freely (e.g., for researchers and practitioners), including the words that are included in the final dictionary as well as background characteristics of consulted experts. Third, the dictionary is not restricted to a specific type of violence or extremism (such as custom dictionaries often are). Any threat, abuse, or violent writing fuelled by a grievance can be assessed with the Grievance Dictionary. This would apply to a wide spectrum of phenomena, including right and left-wing extremism, religious extremism, and (in many cases) threats directed at public officials. Resultingly, dictionary terms will not necessarily need to be continuously updated as is the case for other domain-specific dictionaries.

In Part 3.4 of this chapter, we discuss how the Grievance Dictionary was developed through expert consultation, human and computational word list generation, and crowdsourced annotations. We also perform a psychometric evaluation for each dictionary category. In Part 3.5, we present empirical results using the final dictionary. The dictionary is validated by performing statistical comparisons as well as classification tasks on several datasets. We conclude with a discussion of the dictionary development, validation and intended use, as well as possible future avenues.

### *3.3.1 Transparency statement*

The approach to developing the Grievance Dictionary was fully pre-registered before data collection: https://osf.io/szvm7. All data and materials used for development and validation are available on the Open Science Framework: https://osf.io/3grd6/. A user guide for the dictionary can be found there too.

## 3.4 Dictionary development

The dictionary development consisted of five phases. (1) Threat assessment experts suggested dictionary categories. (2) Human subjects generated seed terms for each category. (3) Computational linguistics methods augmented the word list. (4) Human annotators rated word candidates on their fit into a set of categories. (5) The internal reliability for each dictionary category is assessed and their correlation with LIWC2015 categories is computed.

### *3.4.1 Phase 1: Expert survey*

An online survey was sent out to experts within the field of threat assessment. Participants were professional contacts of the involved researchers in the field of threat assessment and terrorism research. Participants were asked the following:

*Imagine you are tasked with assessing whether a piece of text signals a threat to commit violence against a designated area, individual, or entity. It may be a physical letter or an online message that you are asked to examine. In short, you are trying to judge whether the person who wrote the text will act on their threat.* ***What do you look***

***for in the text to assess its threat level?*** *Please mention all relevant factors that come to mind.*

The response to this question was an open text box, with no word limit. Following this, participants could add any other relevant factors that came to mind (again with an open answer response) and were asked about their professional experience in threat assessment (in years) and with linguistic threat assessment (on a 10-point scale, 1 = no experience, 10 = a lot of experience).

In total, 21 responses were gathered. On average the participants had 16 years of experience with threat assessment (*SD* = 8.84, range: 2-30 years). Overall, the participants indicated they had significant experience with threat assessment based on language, with a mean score of 8.17 (*SD* = 2.04, on a scale from 1-10).

Based on the survey responses gathered, it became clear that assessing the threat of violence through language relies on a wide variety of factors. In order to adequately measure these factors, they need to be condensed into psycholinguistic content categories (e.g., similar to the LIWC). The lead author categorised free text responses. For example, the concepts ‘preparation’, ‘rehearsal’, ‘developing capacity’, ‘refining method’, or ‘developing opportunity’, were all coded as a single category relating to ‘planning’. In total, this resulted in 79 categories (available on the OSF). The categories could broadly be defined to relate to the content of a communication (e.g., direct threat, violence, relationship), emotional processes (e.g., anger, frustration, desperation), mental health aspects (e.g., psychosis, delusional jealousy, paranoia), the communication style (e.g., unusual grammar, politeness, incoherence), and meta-linguistic factors (e.g., number of communications, font, use of graphics). Lastly, the lead author selected categories that could feasibly be represented as a psycholinguistic wordlist, serving as an overarching category (e.g., including ‘weaponry’ but excluding ‘mentioning target’ because it is too situation-specific). This resulted in a final selection of 22 categories (Table 3.1).

Table 3.1 Dictionary categories with example words (defined in later steps)

| Category | Examples | Category | Examples |
|---|---|---|---|
| planning | long-term, tactic, organise | deadline | time run out, due date, upcoming |
| violence | bloodshed, fight, bullet | murder | kill, stab, fatal |
| weaponry | AK-47, ammo, fire arm | relationship | marry, romantic, love |
| help | support, SOS, save | loneliness | disconnected, nobody, abandon |
| hate | enemy, loathe, hatred | surveillance | spy, CCTV, monitor |
| frustration | annoyed, problem, powerless | soldier | fighter, battle, patriot |
| suicide | die, overdose, last resort | honour | integrity, hero, brave |
| threat | warn, danger, unsafe | impostor | impersonate, fraudulent, undercover |
| grievance | wrong, disappoint, injustice | jealousy | cheat, resent, bitter |
| fixation | obsess, possess, watch | god | pray, holy, almighty |
| desperation | sorrow, last chance, urgent | paranoia | suspicious, conspiracy, suspect |

### *3.4.2 Phase 2: Seed word generation*

Human subjects generated seed words for each category from Phase 1. A total of 13 participants suggested words for the categories in an online survey. Participants were all PhD students at English-speaking universities (full details of the sample are reported in the supplementary materials on OSF). For each category, participants were asked to write down all the words that came to mind, considering the category as an over-arching concept for the words they noted down. This resulted in a total of 1,951 seed words across categories. Instructions for the word generation task as well as the resulting words for each category are available in the online materials.

### *3.4.3 Phase 3: Word list extension*

Two processes extended the word list. First, WordNet (Fellbaum, 1998) provided semantic associations for each seed word. This tool provides a lexical database of English words, grouped into 'cognitive synonyms' of meaningfully related words, which are added to the wordlists (e.g., 'knife' is supplemented with 'dagger', 'machete', and 'shiv'). All words related to the initial seed words were added to the list of the respective category.

Second, we obtained pre-trained word embeddings for each candidate word using GloVe, an unsupervised learning approach trained on a 6 billion word corpus (Pennington et al., 2014). GloVe represents words as a vector in multi-dimensional space (embeddings) which aim to encode semantic relationships between individual words based on the contexts in which they appear. This means that words which are similar in meaning have vector representations that are close to each other (based on a similarity measure) in the resulting vector space (e.g., a word embedding for 'gun' appears close to 'handgun', 'pistol', 'firearm', etc. in the learned vector space). For the dictionary, each seed word across all categories was supplemented with its ten nearest neighbour words in terms of cosine similarity. After removing duplicates obtained through WordNet and the embeddings, the final resulting wordlist across all categories contained 24,322 words. These words may appear in multiple categories (e.g., 'knife' may appear in both the weaponry and murder category).

### *3.4.4 Phase 4a: Word list rating*

Human annotators rated all 24,322 words obtained through Phase 3 for the extent to which they fit within their respective category. An online task was developed where participants were presented with a category, a word, and the option to select, on a scale, 'how well the word displayed fits into the above category' (0 = does not fit at all, 10 = fits perfectly). They also had the option to select 'I do not know this word'. After reading instructions and consenting to participating, a total of 100 words (i.e., a random sample of 100 word-category pairs, with words shown for their associated category only) were rated by each participant. Participants were recruited through the crowdsourcing platform Prolific Academic and remunerated for their time. Human workers were only eligible to participate if their first language was English. Interspersed between normal items, four attention checks were included (e.g., ‘This is an attention check. Rate this word with 9 to continue’).

In sum, the 24,322 words of the extended wordlist were rated by 2,318 online participants. A total of 238,366 ratings were obtained, with each word receiving at least 7 ratings, with an average of 9.42 ratings per word. All ratings from participants who failed at least one of the attention checks were removed (1.81%). Words for which the majority (50% or more) of participants indicated that they did not know the word, were also removed from the dictionary (0.39%). Following this, all dictionary words were stemmed and the ratings averaged per word stem (e.g., the ratings for ‘friendship’, ‘friendly’, and ‘friends’ were combined into a single score for the stem ‘friend’). This resulted in a final list of 20,502 words each of which could appear in more than one category.

### *3.4.5 Phase 4b: Scoring methods*

Departing from the rated word list, several versions of the Grievance Dictionary can be used. First, it is important to note that in all versions the words in the dictionary are stemmed (e.g., ‘friendship’ and ‘friends’ are equated to the word stem ‘friend’) in order to find more possible matches. Word stemming is done with Porter’s stemming algorithm (Porter, 2001) using the *quanteda R* package (Benoit et al., 2018).

Three approaches to using the dictionary are discussed. The first two rely on proportional scoring, based on word counts. Following the LIWC, we may wish to only retain words which received a high rating for belonging to a specific category (Pennebaker et al., 2015). In this first version, we would retain only those words which received an average rating of 7 or higher, resulting in a dictionary with **3,643** words. This version is used for evaluation and validation in this paper. An alternative second version retains words with a score of 5 or higher, resulting in a dictionary with 7,588 words. In both of these versions, scoring the texts follows the same approach as the LIWC, which is based on word count. When the dictionary is applied to a text, the incoming text is first stemmed and lowercased using *quanteda* (Benoit et al., 2018) in the same way that has been done with the words in the Grievance Dictionary. The number of words in the texts are subsequently counted, and a warning is given if the word count is below 25 (with the option to remove texts that fall under this threshold). This procedure derives from the evaluation of the LIWC2015, which only included texts with a minimum word count of 25, and further instructions that results are more ‘trustworthy’ when the word count is higher[8]. We expect the same holds for the Grievance Dictionary. Therefore, we similarly recommend using the Grievance Dictionary on texts with 25 words or more.

[8] As stated on the LIWC website: https://liwc.wpengine.com/

Following this, each word in the dictionary is searched in the respective text and a document-feature matrix (i.e., the rows represent a document and the columns represent individual features/dictionary categories) is returned, based on which we can calculate the proportion of a text that belongs to each dictionary category (i.e., frequency of all word matches in category / all words in text) using *quanteda*. As an alternative to measuring proportions per category (22 features), documents could also be represented as a function of all words (3,643 or 7,588 features) in these versions of the Grievance Dictionary.

The third approach relies on average scoring, using the ratings assigned to each word through crowdsourcing. This version of the dictionary makes use of all 20,502 words and their associated average goodness-of-fit rating, assigning each word match in a text the appropriate weight. To measure each category for a text of interest, the average weight of all word matches per category is reported[9]. While the first version using proportional scoring of words with a mean score of 7 and higher is used in this paper, alternative versions are available on the Open Science Framework.

### *3.4.6 Phase 5: Psychometric dictionary evaluation*

To assess the quality of the dictionary, it is important to examine the internal consistency of each category by measuring whether the words in each category yield a similar score for the respective category. We compute Cronbach's alpha using the proportional occurrence of each word in the 22 categories for a total of 17,583 texts across four corpora (Table 3.2). Similar to the development of LIWC2015 we use a varied selection of texts to compute reliability, including texts from deception detection experiments (Kleinberg et al., 2019), novels (Lahiri, 2014), movie reviews (Maas et al., 2011), and Reddit posts (Demszky et al., 2020).

When assessing the internal reliability of psychological tests, typically a Cronbach's alpha score of 0.70 or higher is considered acceptable (Taber, 2018). Cronbach's alpha ranges between 0 to 1 and is based on the number of items and the correlation between them, where a score of 1 represents perfect inter-item correlation, such that the items adequately measure the same underlying concept. When computing internal consistency for wordlists, each word serves as an 'item' for the measurement of the overarching category. The proportional occurrence of each word in the 22 categories is thus computed for each of the four corpora, in order to compute the correlation between words in a category (i.e., the Cronbach's alpha score for the category). We report the average Cronbach's alpha across the four corpora for each category.

[9] When using the weighted dictionary, users need to be aware that mean scores in the middle of the scale may be a result of disagreement (high standard deviations) between raters, rather than a reflection of 'medium' fit into a category (see Pollock, 2018).

As raised in Pennebaker et al. (2015), assessing the reliability of dictionaries is somewhat more complicated. In language, similar concepts are typically not repeated several times; once something has been said it is generally not necessary to be said again. In contrast, similar concepts may be assessed repeatedly in psychological test items. Thus, it has been argued that an acceptable alpha score for dictionary categories will be lower than that for a psychological test (Pennebaker et al., 2015).

Table 3.2 Corpora used for internal consistency computation

| Corpus | *N* documents (*N* tokens) |
|---|---|
| Deception detection experiments* | 2,547 (454,217) |
| Novels (Lahiri, 2014) | 3,036 (247,142,420) |
| IMDB reviews (Maas et al., 2011) | 50,000 (13,934,687) |
| Reddit posts (Demszky et al., 2020) | 70,000 (1,081,539) |

*Note.* *Hotel reviews (Ott et al., 2011, 2013), descriptions of past and planned activities (Kleinberg et al., 2019)

A psychometric evaluation was performed for each version of the dictionary (words with a rating of 7 or higher, words with a rating of 5 or higher, weighted words). The results reported from here onwards concern the dictionary using words with a rating of 7 or higher, because this dictionary performed best (results for the other versions are available on the OSF). The average alpha scores across corpora are reported in Table 3.3. The highest reliability of 0.37 is achieved for the category ‘soldier’, followed by 0.36 for ‘violence’. The lowest scores (0.12, 0.16) were found for the categories ‘fixation’ and ‘grievance’, respectively, which possibly shows that these concepts are difficult to reliably measure with the current approach. The average reliability achieved across categories was 0.26 (*SD* = 0.07). This average reliability is somewhat close to the average reliability of 0.34 achieved with the LIWC2015. The alpha scores for the LIWC2015 ranged between 0.04 and 0.69, whereas ours range between 0.12 to 0.37.

Table 3.3 Internal consistency scores

| Category | Cronbach’s alpha | Category | Cronbach’s alpha |
|---|---|---|---|
| deadline | 0.27 | loneliness | 0.18 |
| desperation | 0.21 | murder | 0.35 |
| fixation | 0.12 | paranoia | 0.23 |
| frustration | 0.22 | planning | 0.31 |
| god | 0.35 | relationship | 0.33 |
| grievance | 0.16 | soldier | 0.37 |
| hate | 0.30 | suicide | 0.26 |
| help | 0.19 | surveillance | 0.25 |
| honour | 0.26 | threat | 0.30 |
| impostor | 0.19 | violence | 0.36 |
| jealousy | 0.21 | weaponry | 0.34 |

In addition to internal reliability, we also assessed whether and how the Grievance Dictionary categories correlated with existing LIWC categories. We correlated Grievance Dictionary scores with LIWC scores (using document-feature-matrices) for each dataset in Table 3.4, and report the mean correlation for each category. Although high correlations with a gold standard dictionary may illustrate that the Grievance Dictionary is comparable to the LIWC in terms of psychometric qualities, we do not expect such a pattern because the Grievance Dictionary categories were designed to *supplement* LIWC categories and not replace them. Reported correlations serve to illustrate which other psycholinguistic concepts measured through the LIWC are related to each respective Grievance Dictionary category. In short, the three highest correlating LIWC categories for each Grievance Dictionary category are depicted in Table 3.4 (full list of correlations available on OSF).

Overall, correlations were low (but statistically significant), suggesting that the Grievance Dictionary does not measure precisely the same constructs as the LIWC. Most Grievance Dictionary categories were correlated to LIWC categories which one might expect to be psychologically related. For example, several Grievance Dictionary categories such as frustration, grievance, hate, murder, paranoia, surveillance, violence, and weaponry were positively correlated to the LIWC category negative emotion. Hate, murder, surveillance, threat, and violence were also positively related to the LIWC's anger category. These results may suggest that some LIWC categories serve as 'umbrella categories' for some in the Grievance Dictionary. That is, the LIWC can provide measures of more general concepts such as negative emotion, whereas the Grievance Dictionary is suited to give more granular measures of psychological constructs (e.g., frustration, paranoia) which fall into this overarching category.

Table 3.4 Correlations (with confidence interval) Grievance Dictionary and LIWC

| Category | Strongest correlating LIWC categories | | |
|---|---|---|---|
| deadline | cause: 0.10 [0.06-0.13] | drives: 0.06 [0.03-0.09] | work: 0.11 [0.06-0.16] |
| desperation | discrep: 0.27 [0.15-0.4] | sad: 0.16 [0.08-0.25] | verb: 0.13 [0.09-0.16] |
| fixation | insight: 0.24 [0.15-0.33] | pronoun: 0.18 [0.08-0.29] | verb: 0.20 [0.12-0.27] |
| frustration | feel: 0.17 [0.07-0.26] | negemo: 0.13 [0.07-0.19] | sad: 0.09 [0.05-0.14] |
| god | affiliation: 0.21 [0.1-0.31] | posemo: 0.14 [0.11-0.18] | relig: 0.32 [0.12-0.52] |
| grievance | affect: 0.08 [0.07-0.09] | negemo: 0.16 [0.06-0.26] | sad: 0.12 [0.05-0.18] |
| hate | affect: 0.09 [0.06-0.12] | anger: 0.23 [0.12-0.34] | negemo: 0.15 [0.09-0.21] |
| help | affect: 0.17 [0.1-0.25] | posemo: 0.20 [0.14-0.26] | reward: 0.23 [0.12-0.35] |
| honour | affect: 0.18 [0.09-0.27] | drives: 0.16 [0.07-0.26] | posemo: 0.22 [0.12-0.32] |
| impostor | power: -0.03 [-0.04--0.02] | relativ: -0.05 [-0.09--0.02] | space: -0.04 [-0.07--0.02] |
| jealousy | cogproc: 0.11 [0.06-0.16] | discrep: 0.07 [0.05-0.1] | insight: 0.15 [0.07-0.23] |
| loneliness | discrep: 0.06 [0.03-0.1] | sad: 0.08 [0.03-0.13] | time: 0.06 [0.04-0.08] |
| murder | affect: 0.09 [0.04-0.13] | anger: 0.2 [0.1-0.31] | negemo: 0.17 [0.07-0.27] |
| paranoia | anx: 0.11 [0.05-0.17] | cogproc: 0.08 [0.04-0.13] | negemo: 0.11 [0.06-0.16] |
| planning | Authentic: 0.13 [0.05-0.21] | focuspres: 0.14 [0.08-0.19] | insight: 0.15 [0.07-0.23] |
| relation. | affiliation: 0.28 [0.12-0.43] | family: 0.23 [0.13-0.33] | social: 0.28 [0.1-0.46] |
| soldier | achieve: 0.12 [0.1-0.15] | drives: 0.15 [0.12-0.18] | power: 0.17 [0.09-0.25] |
| suicide | death: 0.16 [0.09-0.23] | health: 0.17 [0.07-0.28] | sad: 0.14 [0.08-0.21] |
| surveillance | affect: -0.05 [-0.07--0.02] | anger: -0.04 [-0.06--0.02] | negemo: -0.04 [-0.06--0.02] |
| threat | anger: 0.23 [0.13-0.33] | negemo: 0.17 [0.1-0.25] | Tone: -0.14 [-0.2--0.07] |
| violence | anger: 0.21 [0.1-0.32] | death: 0.2 [0.09-0.32] | negemo: 0.28 [0.1-0.45] |
| weaponry | negemo: 0.1 [0.05-0.15] | posemo: -0.07 [-0.11--0.04] | Tone: -0.11 [-0.16--0.05] |

*Note.* All correlations were statistically significant at the *p* < 0.0023 (0.05/22 categories) level.

## 3.5 Dictionary validation

The dictionary validation reported in this section serves to assess whether and how the Grievance Dictionary can be used to distinguish between different types of writing, for example neutral language and grievance-fuelled communications produced by terrorists or extremists. We first apply the Grievance Dictionary to different datasets to assess its external validity. Then, we test the performance of the dictionary in classification tasks.

### *3.5.1 External validity*

**Data**. We apply the dictionary to five different datasets to test its validity in the context of grievance-fuelled writings. The rationale for selecting each dataset is as follows:

- Lone-actor terrorist manifestos: previous research has frequently used texts written by known lone-actor terrorists to make statistical comparisons between documents written by violent (extremist) individuals and non-violent individuals

(Baele, 2017; Egnoto & Griffin, 2016; Kaati, Shrestha, & Cohen, 2016). For this sample, we draw sequential 100-word chunks from 22 lone-actor terrorist manifestos resulting in a total sample of 4,572 documents. This 'chunking' is performed so that the average word count for the terrorist manifestos is more comparable to that of the datasets described below.

- Neutral texts from blogs and forums: violent texts are often compared to a neutral 'control group' (Baele, 2017; Egnoto & Griffin, 2016; Kaati, Shrestha, & Cohen, 2016; Kaati, Shrestha, & Sardella, 2016). These samples often consist of publicly available datasets with texts collected from publicly accessible blogs and forums, such as the Blog Authorship Corpus (Schler et al., 2006) and the Boards.ie forum dataset[10].
- Stormfront posts: right-wing extremist forum posts have been compared in the past to both neutral texts and lone-actor terrorist manifestos (Kaati, Shrestha, & Cohen, 2016; Kaati, Shrestha, & Sardella, 2016). The assumption behind the use of this data is that it represents a 'non-violent extremist' sample. That is, no *known* acts of violence have been committed by the extremist text authors, in contrast to, for example, lone-actor terrorist manifestos. The Stormfront posts in this dataset include all posts between 2012-2015 from the dataset used in Kleinberg et al. (2020).
- Abuse directed at politicians: worrying or threatening communications directed at politicians are often motivated by a grievance, and thus the Grievance Dictionary may shed further light on this phenomenon. The full procedure documenting how this data was collected is described in Chapter 4 of this thesis.
- Stream-of-consciousness essays: this form of writing is often used in psycholinguistic research (e.g., Newman et al., 2008; Vine et al., 2020), and is collected by having participants write about their current thoughts, feelings, and sensations. This sample is included because it is derived from the same participants who wrote the abuse directed at politicians (above), and thus can be used for a within-subject comparison of Grievance Dictionary category use.

**Descriptive statistics.** All datasets and associated descriptive statistics are reported in Table 3.5. In order to demonstrate the extent of dictionary matches, Table 3.6 shows the mean proportion of word matches per dataset for each category. The last row of Table 3.6 also shows the mean proportion of words in the documents which were not matched with any word in the dictionary. These results show that most matches with the Grievance Dictionary were found in the lone-actor terrorist manifestos (44%) and the least matches were found in the in the neutral texts from blogs and forums (18%).

[10] From the 2008 SIOC Semantic Data Competition: https://semantic-web.com/2008/08/27/boardsie-sioc-semantic-data-competition-starts-september-1st/

Table 3.5 Corpora used for statistical tests

| Corpus | *N* documents | Mean word count (*SD*) |
|---|---|---|
| Lone-actor terrorist manifestos | 4,572 | 100 (4) |
| Neutral texts from blogs and forums | 680,792 | 243 (503) |
| Stormfront posts | 461,950 | 95 (229) |
| Stream-of-consciousness (SOC) | 789 | 121 (35) |
| Abuse directed at politicians | 789 | 121 (38) |

Table 3.6 Mean dictionary matches per dataset

| Category | Lone-actor manifestos | Neutral texts | Stormfront posts | SOC | Abusive writing |
|---|---|---|---|---|---|
| deadline | 0.03 | 0.01 | 0.01 | 0.02 | 0.02 |
| desperation | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 |
| fixation | 0.00 | 0.01 | 0.01 | 0.02 | 0.01 |
| frustration | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 |
| god | 0.03 | 0.01 | 0.01 | 0.01 | 0.01 |
| grievance | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 |
| hate | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 |
| help | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| honour | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| impostor | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| jealousy | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 |
| loneliness | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 |
| murder | 0.03 | 0.01 | 0.01 | 0.01 | 0.01 |
| paranoia | 0.01 | 0.00 | 0.00 | 0.02 | 0.01 |
| planning | 0.04 | 0.02 | 0.02 | 0.03 | 0.02 |
| relationship | 0.04 | 0.02 | 0.02 | 0.03 | 0.02 |
| soldier | 0.04 | 0.01 | 0.01 | 0.01 | 0.01 |
| suicide | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| surveillance | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| threat | 0.03 | 0.00 | 0.01 | 0.01 | 0.01 |
| violence | 0.04 | 0.01 | 0.01 | 0.01 | 0.01 |
| weaponry | 0.03 | 0.00 | 0.01 | 0.00 | 0.01 |
| no match | 0.56 | 0.82 | 0.80 | 0.68 | 0.77 |

**Statistical tests.** In total, three statistical tests are performed on the proportional matches (category matches per document / total number of words per document) shown in Table 3.6. First, following previous work on violent language use (Kaati, Shrestha, & Cohen, 2016), we make statistical comparisons between the lone-actor terrorist manifestos and the neutral 'control' texts retrieved from online forums and blogs.

Second, we perform a comparison between the lone-actor terrorist manifestos and the posts from right-wing extremist forum Stormfront. For both tests, mean dictionary outcome values of the lone-actor terrorist manifestos are compared to the means of the control samples with an independent samples *t*-test. The control samples are down-sampled through bootstrapping to match the *N* of the lone-actor manifestos, with outcome measures reported as an average across 100 bootstrap iterations. We report the effect size for the difference by means of Cohen's *d*[11], in addition to the Bayes Factor (*BF*). The Bayes Factor is a measure of the degree to which the data are more likely to occur under the hypothesis that there is a difference in the dictionary categories between samples, compared to the hypothesis that there is no difference (Ortega & Navarrete, 2017; Wagenmakers et al., 2010). For example, a *BF* between above 10 would constitute strong evidence for the alternative hypothesis that there is a difference (Ortega & Navarrete, 2017).

The third comparison is between abusive texts directed at politicians and neutral, stream-of-consciousness (SOC) essays (van der Vegt et al., 2020). For this comparison a dependent samples t-test is performed, because individual participants produced both types of text. Again, effect size *d* and *BF* are reported for the difference between the two samples (note that this comparison is not based on bootstrapping due to the smaller, equal sample sizes).

**Results.** The outcome of each comparison is reported in Table 3.7. Overall, statistically significant differences were found for the majority of categories in all comparisons. In the majority of cases, the lone-actor texts scored higher on Grievance Dictionary categories than the control texts. In the first comparison with neutral texts from blogs and forums, the lone-actor manifestos scored higher on all categories except 'fixation' and 'loneliness' (denoted by a negative effect size *d*). The evidence for a difference between samples was very strong (*BF* > 10) in all cases except 'desperation'. In the second comparison with Stormfront forum posts, the lone-actor manifestos scored proportionally higher on all categories except 'fixation' (strong evidence with *BF* > 10) and 'loneliness' (weak evidence *BF* < 10). For the comparison between abusive writing and stream-of-consciousness texts, differences in favour of SOC texts (*BF* > 10) were found (denoted by negative *d*) for the categories deadline, desperation, fixation, frustration, grievance, hate, jealousy, loneliness, paranoia, planning, relationship, and suicide. However, the abusive texts contained proportionally more references to honour, impostor, murder, violence, and weaponry (positive *d* and *BF* > 10).

[11] Cohen's *d* expresses the magnitude of the difference after correcting for sample size. A *d* of 0.20, 0.50 and 0.80 can be interpreted as a small, moderate and large effect, respectively (Cohen, 1988)

Table 3.7 Statistical test results (Effect size *d* with confidence interval and BF)

| **Category** | **Manifestos vs. neutral** | | **Manifestos vs. Stormfront** | | **Abuse vs. SOC** | |
|---|---|---|---|---|---|---|
| | *d (bootstrapped)* | *BF* | *d (bootstrapped)* | *BF* | *d* | *BF* |
| deadline | 0.71 [0.7;0.71] | **531.5** | 0.85 [0.85;0.86] | **759.93** | -0.43 [-0.52;-0.31] | **62.75** |
| desperation | 0.07 [0.06;0.07] | 1.69 | 0.16 [0.15;0.16] | **24.67** | -0.88 [-1.03;-0.78] | **221.2** |
| fixation | -0.47 [-0.48;-0.47] | **243.56** | -0.27 [-0.28;-0.27] | **78.67** | -0.68 [-0.82;-0.57] | **146.28** |
| frustration | 0.21 [0.2;0.21] | **42.79** | 0.34 [0.33;0.34] | **122.09** | -0.87 [-1;-0.74] | **215.71** |
| god | 0.87 [0.86;0.87] | **782.42** | 0.84 [0.84;0.84] | **735.08** | 0.1 [-0.0032;0.23] | 0.85 |
| grievance | 0.26 [0.26;0.26] | **74.05** | 0.22 [0.21;0.22] | **50.00** | -0.84 [-0.97;-0.73] | **205.73** |
| hate | 1.16 [1.16;1.17] | **$>10^3$** | 0.84 [0.84;0.84] | **735.28** | -0.32 [-0.43;-0.2] | **34.72** |
| help | 0.41 [0.41;0.42] | **186.5** | 0.36 [0.36;0.36] | **140.69** | 0.03 [-0.09;0.15] | -2.95 |
| honour | 0.85 [0.85;0.86] | **765.27** | 0.69 [0.69;0.7] | **513.27** | 0.53 [0.41;0.65] | **91.99** |
| impostor | 0.36 [0.35;0.36] | **141.45** | 0.23 [0.22;0.23] | **54.51** | 0.45 [0.38;0.55] | **69.52** |
| jealousy | 0.38 [0.37;0.38] | **153.22** | 0.36 [0.36;0.36] | **145.00** | -0.72 [-0.83;-0.61] | **160.35** |
| loneliness | -0.17 [-0.17;-0.16] | **27.41** | -0.05 [-0.05;-0.05] | -0.29 | -0.57 [-0.7;-0.47] | **105.38** |
| murder | 1.27 [1.27;1.27] | **$>10^3$** | 0.96 [0.95;0.96] | **929.79** | 0.33 [0.22;0.43] | **36.14** |
| paranoia | 0.39 [0.38;0.39] | **165.56** | 0.42 [0.42;0.43] | **194.29** | -0.99 [-1.11;-0.88] | **263.82** |
| planning | 0.94 [0.94;0.95] | **915.80** | 0.97 [0.97;0.98] | **968.30** | -0.41 [-0.53;-0.29] | **56.93** |
| relation. | 0.55 [0.55;0.56] | **334.21** | 0.52 [0.52;0.53] | **294.16** | -0.21 [-0.32;-0.09] | **13.69** |
| soldier | 1.57 [1.57;1.57] | **$>10^3$** | 1.27 [1.26;1.27] | **$>10^3$** | -0.03 [-0.14;0.07] | -2.82 |
| suicide | 0.74 [0.74;0.75] | **581.54** | 0.74 [0.74;0.75] | **585.71** | -0.22 [-0.34;-0.12] | **15.8** |
| surveillance | 0.71 [0.7;0.71] | **531.55** | 0.54 [0.54;0.55] | **326.47** | 0.17 [0.05;0.27] | 7.84 |
| threat | 1.46 [1.46;1.46] | **$>10^3$** | 1.11 [1.1;1.11] | **$>10^3$** | 0.16 [0.05;0.27] | 7.15 |
| violence | 1.55 [1.55;1.55] | **$>10^3$** | 1.16 [1.16;1.16] | **$>10^3$** | 0.36 [0.26;0.49] | **44.94** |
| weaponry | 1.39 [1.39;1.4] | **$>10^3$** | 1.05 [1.05;1.06] | **$>10^3$** | 0.39 [0.3;0.48] | **51.38** |

*Notes.* A positive *d* denotes a higher score on the category for the lone-actor terrorist manifestos (test 1 and 2) and abusive texts (test 3). A *BF* above 10 (in bold) constitutes strong evidence for the alternative hypothesis.

### *3.5.2 Classification*

Previous work classified terrorist or extremist texts from neutral ‘control samples’ using the LIWC. We investigate whether the Grievance Dictionary can achieve similar results, or increase prediction performance when used to supplement the LIWC. In three classification tasks, we examine whether the Grievance Dictionary and the LIWC can distinguish between:

1) Texts written by known terrorists vs. non-violent individuals
2) Texts written by known terrorists vs. non-violent extremists
3) Abusive vs. neutral texts (non-violent within-subject comparison)

**Method.** All classification tasks were performed using a multinomial Naïve Bayes classifier, a linear SVM, and a random forest model. We report the results for the best performing model. All analyses were performed in *R*, using the *quanteda textmodels* (Benoit et al., 2020) and *randomForest* (Liaw & Wiener, 2002) packages.

In Classification Task 1, we classify lone-actor terrorist manifesto excerpts (*N* = 4,572) versus neutral posts from blogs and forums (*N* = 680,792). The majority class of neutral posts is down-sampled to the same *n* as the manifesto sample by means of bootstrapping (100 times), to allow for a balanced classification task. For each bootstrapped sample, we perform a five-fold cross validation using 80% of the sample as training data, and the remaining 20% as test data. Classification results are reported as an average across each of the 5 cross-validations across the 100 bootstrapped samples. In Classification Task 2, we classify lone-actor terrorist manifesto excerpts (*N* = 4,572) versus Stormfront posts (*N* = 461,950). Following the same procedure as in Task 1, the majority class of Stormfront posts is down-sampled 100 times and cross-validated five times with an 80/20 split. In classification Task 3, we perform classification for abusive vs. neutral, stream-of-consciousness writing with data from van der Vegt et al. (2020), using 789 documents per sample. Note that due to the smaller sample size in Task 3 we do not perform bootstrapping, and instead opt only for a five-fold cross-validation with an 80/20 split.

**Feature sets.** Each classification task is performed using three different feature sets, to test the performance of the Grievance Dictionary, the LIWC and a combination of the two in classifying aforementioned datasets. The following feature sets are used:

a) All 22 Grievance Dictionary categories.

b) All psychological and social categories (*N* = 55) of the LIWC2015[12]. We exclude linguistic style (grammar) categories from the LIWC such as pronouns and verbs because we are interested in the predictive ability of content (psychological concepts) only, and grammatical categories do not appear in the Grievance Dictionary either.

*c)* A combination of the Grievance Dictionary and psycho-social LIWC categories (*N* = 77).

**Results.** Performance metrics[13] for the classification tasks are reported in Table 3.8. In all tasks, the random forest model performed best, with the exception of task 3b and 3c, where a linear SVM produced higher prediction performance. Classification Task 1 shows high performance for distinguishing between lone-actor terrorist texts and neutral texts. The Grievance Dictionary alone achieves 96% accuracy, which is further increased to 99% when using the LIWC. The combination of the LIWC and Grievance Dictionary does not provide a substantial improvement over the LIWC alone. Classification Task 2 similarly shows that the LIWC alone (and in combination with the

[12] Including the umbrella categories analytical thinking, clout, authentic language, emotional tone, affect words, social words, cognitive processes, perceptual processes, biological processes, core drives and needs, time orientation, relativity, personal concerns and informal speech (Pennebaker et al., 2015).

[13] 1) Classification accuracy: true positive + true negatives / true positives + false positives + true negatives + false negatives, 2) Kappa: observed accuracy – expected accuracy / 1 – expected accuracy, 3) Specificity: TN / TN + FP, 4) Precision: TP / TP + FP, 5) Recall: TP / TP + FN (see Sammut & Webb, 2011 for an overview).

Grievance Dictionary) achieves nearly perfect classification accuracy. In Task 3, the LIWC (alone and in combination with the Grievance Dictionary) similarly outperforms the Grievance Dictionary. Here, performance metrics are somewhat lower compared to Task 1 and 2, but the majority of cases are still accurately classified.

Table 3.8 Classification results

| Task | Feature set | Accuracy | Kappa | Specificity | Precision | Recall |
|---|---|---|---|---|---|---|
| 1. LA vs. neutral | a. Grievance | 0.96 | 0.92 | 0.97 | 0.97 | 0.96 |
| | b. LIWC | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |
| | c. Grievance + LIWC | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |
| 2. LA vs. Stormfront | a. Grievance | 0.94 | 0.87 | 0.94 | 0.94 | 0.94 |
| | b. LIWC | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |
| | c. Grievance + LIWC* | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| 3. Abuse vs. neutral | a. Grievance | 0.83 | 0.67 | 0.86 | 0.86 | 0.82 |
| | b. LIWC* | 0.96 | 0.92 | 0.98 | 0.98 | 0.94 |
| | c. Grievance + LIWC* | 0.96 | 0.92 | 0.98 | 0.98 | 0.94 |

**The best performing model for these tasks was a linear SVM, rather than a random forest model (best performing in all other tasks)*

All in all, classification accuracies were high, with several near ‘perfect’ performances. Therefore, we examined feature importance for each task in order to discover whether the model was biased towards some features. The five most important features for each task are reported in Table 3.9. Feature importance rankings are based on a ROC curve analysis, where a cut-off for each feature is defined that maximises true positives predictions, and minimises false positives; a larger area under the ROC curve implies larger variable importance (Kuhn, 2008). Tables with ROC values for each feature per task are available on the Open Science Framework.

Table 3.9 Feature importance per task (top five, full list of features on OSF)

| Task | Feature set | Important features |
|---|---|---|
| 1. LA vs. neutral | a. Grievance | soldier, weaponry, violence, impostor, threat |
| | b. LIWC | analytic language, present focus, power, differentiation, work |
| | c. Grievance + LIWC | analytic language, differentiation, present focus, soldier, violence |
| 2. LA vs. Stormfront | a. Grievance | soldier, relationship, impostor, threat, hate |
| | b. LIWC | differentiation, analytic language, present focus, tentative, discrepancies |
| | c. Grievance + LIWC | differentiation, analytic language, present focus, tentative, discrepancies |
| 3. Abuse vs. neutral | a. Grievance | paranoia, grievance, frustration, fixation, desperation |
| | b. LIWC | authentic language, social words, clout, feel, male |
| | c. Grievance + LIWC | authentic language, social words, clout, feel, male |

Features with high importance also showed stark differences in mean proportional dictionary scores between datasets. For example, the most important feature 'soldier' in Task 1a showed a mean score for lone-actor terrorist manifestos of 0.04 (*SD* = 0.03), whereas neutral texts and Stormfront posts scored 0.01 (*SD* = 0.009) and 0.01 (*SD* = 0.01), respectively. This was reflected in the results observed in aforementioned Bayesian *t*-tests, where a decisive difference (*BF* > $10^3$) was observed for 'soldier'. The second most important feature 'weaponry' (*BF* > $10^3$), had a mean of 0.03 (*SD* = 0.03) in lone-actor manifestos, in contrast to 0.004 (*SD* = 0.007) and 0.01 (*SD* = 0.01) in neutral texts and Stormfront posts, respectively. These large differences between datasets will have contributed to the high prediction performance in this (and other) task(s), in that the classifier learned to over-rely on these features. In contrast, classification Task 3 showed somewhat lower performance compared to Task 1 and 2, likely because smaller differences between samples were observed. Indeed, the most important feature 'paranoia' scored 0.02 (*SD* = 0.01) in the stream-of-consciousness essays and 0.01 (*SD* = 0.007) in the abusive texts, with the Bayes Factor demonstrating a smaller difference (*BF* = 263.82) than the differences observed for the most important features in Task 1 and 2 (*BF* > $10^3$). Therefore, the model was perhaps less able to strongly rely on these feature differences. It remains to be seen in future research how the Grievance Dictionary performs on datasets with even smaller statistical differences between texts (e.g., violent texts written by individuals who want to actualise their threat, vs. similarly violent texts written by those who do not plan to actualise).

## 3.6 Discussion

In this chapter, we introduced the Grievance Dictionary, a psycholinguistic dictionary for grievance-fuelled violence threat assessment. The aim of this study was to develop a dictionary which can specifically measure constructs relevant to threat assessment, and can be used for a wide variety of violence and extremism fuelled by a grievance. Furthermore, we aimed to address the limitations we identified pertaining to existing psycholinguistic dictionaries. In this section, we examine the results obtained through statistical tests and classification tasks. This is followed by a discussion of the intended use for the Grievance Dictionary, as well as its limitations and possible future work.

### *3.6.1 Linguistic differences*

Based on the validation results of the dictionary, we saw that the Grievance Dictionary can elucidate differences between threatening and non-threatening language. Differences in Grievance Dictionary categories were found between texts written by lone-actor terrorists, neutral writing, and extremist forum posts, as well as between abusive language and stream-of-consciousness writing. The evidence for these differences was strong.

It must be noted that a high score on Grievance Dictionary categories is not exclusive to threatening and violent texts. In our comparison between stream-of-consciousness essays and abusive writing, the former obtained significantly higher scores for categories such as desperation, fixation, and frustration. Therefore, it is important to note that high scores on single dictionary categories should not be interpreted as individual risk factors for violence, as they may also occur in non-violent texts. Instead, the measures should be interpreted jointly to gain an understanding of the content of a grievance-fuelled text, with particular attention paid to the highly 'violent' categories such as murder, violence, threats, and weaponry. Furthermore, the importance of Grievance Dictionary categories for distinguishing between different populations may also be context-dependent. For example, mentions of a (perceived) romantic relationship may positively predict violence in a threat directed at a public figure, while it may negatively predict violence (a 'linguistic protective factor') in an extremist text. Further research will be needed to establish and replicate differential meanings of Grievance Dictionary categories across contexts.

### *3.6.2 Classification with the Grievance Dictionary*

The dictionary categories were also used to classify different types of writing, including terrorist manifestos and extremist forum posts, neutral and extremist forum posts, as well as abusive and neutral writing. First, it is important to note that prediction was not the main objective for developing the Grievance Dictionary, because dictionary scores as features generally do not offer high prediction performance when compared to other features such as *n*-gram frequencies, parts-of-speech frequencies, or word embeddings (see e.g., Figea et al., 2016; Neuman et al., 2015). However, since related research on extremism and terrorism has previously used the LIWC to classify text samples (Figea et al., 2016; Kaati, Shrestha, & Sardella, 2016), we found it important to examine whether the Grievance Dictionary can achieve the same. One benefit of using the Grievance Dictionary for prediction is that the contributing features remain interpretable to humans, in contrast to methods such as word embeddings, which are difficult to interpret as features. Therefore, the Grievance Dictionary may be preferable in light of regulations such as the ALGOCARE framework (Oswald et al., 2018), but it is important to realise that other more sophisticated (but less explainable) methods exist.

Nevertheless, the classification accuracy achieved in this study did approximate or outperform previous work in the violence research domain. The Grievance Dictionary alone already outperformed previous research, for example in classifying lone-actor terrorist manifestos from Stormfront posts (here: accuracy of 0.96 vs. 0.90 in Kaati et al., 2016). However, performance was further improved (sometimes to 99% accuracy) when using the LIWC (alone and in combination with the Grievance Dictionary). These results imply that although the Grievance Dictionary can achieve adequate prediction performance, it does not necessarily offer enhanced prediction performance over the LIWC. However, as has been raised previously, this was not the primary objective for developing the Grievance Dictionary. Moreover, the potential for obtaining more nuanced (violence-specific) measures with the Grievance Dictionary remains.

### *3.6.3 Usage of the Grievance Dictionary*

All things considered, the Grievance Dictionary shows promising results for demonstrating differences between different types of (non-)grievance-fuelled language. Even though mean scores on dictionary categories were low (i.e., the majority of words across different datasets were not matched), values still elucidated strong differences between several (non) threatening texts. These results also suggest that the categories elicited from expert threat assessment practitioners hold value in understanding violent from non-violent language.

Perhaps, the most important academic use case for the Grievance Dictionary is to gain a general picture of language use in a (large) corpus, and to make (statistical) comparisons between different corpora. Because of the context-specificity of the dictionary, it may be especially suited to testing theories within the violence domain. Certain questions (e.g., Are right-wing extremists more paranoid than left-wing extremists? Do jihadists discuss weaponry more than right-wing extremists?) were previously not testable. Additionally, the Grievance Dictionary may also be used to gain a broad understanding of large-scale online social media data on a user or platform level, or to compare an incoming threatening message to a (police) database of existing communications.

## 3.7 Limitations and future work

Some limitations to the Grievance Dictionary need to be considered. The first pertains to the construction of the dictionary. The seed words on which the dictionary categories are based were produced by human annotators who, to our knowledge, do not have violent ideations. Therefore, it may have been difficult for participants to produce words about attack planning and weaponry if they have little knowledge on the topic. We tried to somewhat ameliorate this problem by including word candidates obtained through automatic methods. Nevertheless, future improvements to the Grievance Dictionary may include seed words that are obtained by means of a data-driven approach. That is, we may extract words from texts which are known to have been written by lone-actor terrorists or other violent individuals to serve as seed words. A further limitation relates to the internal consistency of Grievance Dictionary categories. Although low internal consistency is generally expected for language-based measures (compared to self-report questionnaires, for example), the average reliability of Grievance Dictionary categories was lower than those observed for the LIWC (Pennebaker et al., 2015). This is somewhat surprising since LIWC categories were never intended to be semantically cohesive or comprehensive (Boyd & Schwartz, 2020), whereas our hope was to provide (somewhat) comprehensive linguistic measures of threat assessment concepts. These results potentially demonstrate the difficulty of cohesively measuring latent psychological concepts. Indeed, categories that can perhaps be considered as more abstract or difficult to interpret (grievance, fixation, impostor) scored lower on reliability than more concrete categories (soldier, weaponry), a factor dictionary users should also be aware of. It remains to be seen whether alternative (data-driven) wordlist generation procedures will result in higher internal consistency of categories.

### 3.8 Conclusion

The purpose of the Grievance Dictionary is to serve as a resource for threat assessment practitioners and researchers aiming to gain a better understanding of grievance-fuelled language use. Initial validation tests of the dictionary show that differences between violent and non-violent texts indeed can be detected using the dictionary. All information regarding the construction and specifications of the dictionary is available to researchers and practitioners, so that the capabilities *and* limitations of the Grievance Dictionary can be adequately scrutinised. We hope the current work serves as an impetus to gain a better understanding of grievance-fuelled language by automatic means.

# Chapter 4: Linguistic style: Predicting Author Profiles from Online Abuse

## 4.1 Introduction

In Chapter 2 and 3, it became clear that threat assessors search for certain content in a text in order to determine the associated risk of violence. However, several experts also referred to stylistic features of language, often intended to determine what kind of person the author of an anonymous text may be. Indeed, some scientific research also aims to 'profile' the authors of text, examining language use to estimate, for example, the age, gender, and personality of the author (e.g., Newman et al., 2008; Oberlander & Nowson, 2006; Pennebaker & Stone, 2003). In addition to content differences associated with author characteristics, this research domain also makes use of linguistic style. The underlying assumption of assessing linguistic style (e.g., pronouns, articles, and other function words) is that this dimension of language is not under the conscious control of a writer, and thus can reveal information about a text author or individual differences between different authors (Pennebaker & King, 1999).

In this chapter, we supplement our examinations of linguistic content in grievance-fuelled communications with measures of linguistic style. This combination of linguistic content and style is common in author profiling research, an application of computational linguistics that is of particular relevance to threat assessment, especially in the case of anonymous communications. One example of author profiling within the domain of grievance-fuelled violence includes the PRAT tool, which measures personality through language for its aim of assessing risk of violence in written communication (Akrami et al., 2018). While such an approach may be of particular interest to law enforcement agencies to triage online threats, we argue that author profiling in this domain requires further testing before it can be successfully deployed in practice.

Although a large body of research examined the relationship between writing and personality (Pennebaker & King, 1999), age (Pennebaker & Stone, 2003), and gender (Newman et al., 2008), the majority have obtained small effects (Azucar et al., 2018; Qiu et al., 2012). A few have used linguistic variables to predict author characteristics, but accuracy varies widely, for example from 45% to 92% (Argamon et al., 2005; Burger et al., 2011; Nguyen et al., 2013; Preotiuc-Pietro et al., 2016). This raises the question whether the small, yet statistically significant, relationships between language and author demographics can be adequately translated into prediction systems that are accurate enough for stakeholders. Importantly, it has yet to be examined whether there is a relationship between personality, age and gender when individuals write *abusive* text. For instance, do highly extraverted or narcissistic individuals write an abusive message differently than people who score low on these traits? It also remains to be tested whether aspects of abusive language can be used to infer one's personality, age, or gender. In order to investigate this, the current chapter focuses on author profiling within a sub-type of grievance-fuelled communications, namely abuse and threats directed at politicians. In the next sections, we discuss related research on author profiling.

### *4.1.1 Linguistic correlates of author characteristics*

Early studies using automated approaches to studying language departed from the assumption that linguistic content and style differ between individuals (Pennebaker & King, 1999). Specific traits such as the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) were correlated with certain linguistic characteristics, such as the use of negative emotion words, negations, and present tense (Pennebaker & King, 1999). The LIWC was applied to a sample of psychology students' writing samples ($N$ = 1203), who wrote a 'stream of consciousness' essay describing current thoughts, feelings, and sensations. Results showed small positive correlations between neuroticism and negative emotion words ($r$ = 0.16), and a positive correlation between positive emotion words ($r$ = 0.15), social references ($r$ = 0.12) and extraversion. Other endeavours (Hirsh & Peterson, 2009) showed correlations of $r$ = 0.23 between personality traits and LIWC categories.

A possible effect of age on language has also been examined. In a large-scale study, references to the self and others decreased with age, as well as an increase of present- and future-tense over past-tense verbs with age (Pennebaker & Stone, 2003). Ageing was also associated with an increase in positive emotion words ($r$ = 0.05) and a decrease in negative emotion words ($r$ = -0.04) (Pennebaker & Stone, 2003). Gender differences in language emerged in a study of 14,324 text samples including stream of consciousness essays (Newman et al., 2008). Women more often used LIWC dimensions such as pronouns (Cohen's $d$ = 0.36) and social words ($d$ = 0.21).

### *4.1.2 Predicting author characteristics from language*

Besides the study of linguistic correlates of author profiles, linguistic information has also been used to predict personality traits, age, and gender using a machine learning approach. In one prediction example, participants completed a personality questionnaire and wrote stream-of-consciousness essays (i.e., expressing their current thoughts and feelings), after which the traits neuroticism and extraversion were predicted (Argamon et al., 2005). A binary classification task was performed, where participants were either high (top third) or low (bottom third) scorers on the traits. Various psycholinguistic measures (such as the LIWC) were used as features, and the average classification accuracy was 58% (Argamon et al., 2005). In a similar effort, *n*-grams (i.e., word occurrences) were used as features to predict Big Five scores in several binary and multiclass prediction tasks (Oberlander & Nowson, 2006).

Accuracies ranged from 45% to 100% depending on the task, personality trait and feature set (Oberlander & Nowson, 2006).

Importantly, personality traits are considered more accurately conceptualised as continuous constructs rather than as binary or categorical variables (Haslam et al., 2012). Some prediction efforts estimated traits on a continuous scale, using a regression approach. This has, for example, been done for Big Five personality impressions (i.e., third-person annotations) of YouTube vlogger videos using the LIWC (Farnadi et al., 2014). The best performance was achieved for conscientiousness (*RMSE* = 0.64 on a scale of 1-7, $R^2$ = 0.18). Another study predicted Dark Triad traits (narcissism, Machiavellianism, and psychopathy) from Twitter data including unigrams, LIWC categories, and profile picture features, with ground truth established through a self-report survey (Preotiuc-Pietro et al., 2016). The best model showed a correlation of 0.25 between predicted and observed values (Preotiuc-Pietro et al., 2016). In another study, both regression and classification tasks were used for Big Five and Dark Triad prediction with LIWC measures of Twitter profiles as features (Sumner et al., 2012). Prediction performance was poor for both tasks, even though the authors identified correlations between personality traits and LIWC categories in the Twitter data (Sumner et al., 2012).

Various studies also worked on predicting age and gender. In the PAN[14] 2016 shared task on this topic, the best performance for predicting five age classes was 58.97% using stylistic features and vector representations of terms and documents (Rangel et al., 2016). Gender was correctly classified 75.64% of the time using stylometric features (e.g., pronouns and adjectives) and *n*-grams (Rangel et al., 2016). Age has also been predicted on a continuous scale using unigrams, with a mean absolute error of approximately four years (Nguyen et al., 2013). Furthermore, gender classification on Twitter using *n*-grams achieved 91.80% accuracy when all tweets from a profile were used (Burger et al., 2011).

### *4.1.3 Author profiling grievance-fuelled communications*

Importantly, author profiling is also gaining traction within violence threat assessment, for example when the source of an abusive, threatening, or extremist text posted online needs determining. The Profile Risk Assessment Tool (PRAT), which is intended for risk assessment of violent written communications, constructs a personality profile of a text author (Akrami et al., 2018). The profiles are constructed by means of IBM Watson Personality Insights, which predicts Big Five traits with models trained on word embeddings (i.e., words represented by vectors of other semantically close words) from a large dataset for which personality traits of text authors were known. IBM Personality Insights has also been used to study the texts of ‘pseudocommando mass murderers’, defined as individuals who ‘are obsessed with

[14] Plagiarism analysis, Authorship identification, and Near-duplicate detection: https://pan.webis.de/

weapons and meticulously plan their attack' (Kop et al., 2019). Personality traits measured in the mass murderer texts were compared to population medians, with the former scoring higher on openness, but lower on extraversion and agreeableness (Kop et al., 2019). In a study on profiling the texts of school shooters, personality profiles were constructed by means of word embeddings (Neuman et al., 2015). Distances were computed between vectors for each school shooter text and vectors representing traits such as narcissism, but also for disorders such as paranoid personality disorder, schizotypal personality disorders, and depression. The same was done for control samples of neutral writing. After ranking all texts on these measures, school shooter texts could be identified by examining 3% of the entire corpus (Neuman et al., 2015).

## 4.2 The current study

Since author profiling is increasingly applied within the domain of understanding (potentially) violent individuals and threat assessment, we recognise the importance of testing 1) whether there are statistical relationships between author characteristics (personality, age, and gender) and abusive language use (content and style), 2) whether author characteristics can indeed be predicted from abusive texts. We focus on personality due to its increased popularity in violence research (Akrami et al., 2018; Kop et al., 2019; Neuman et al., 2015), whereas age and gender may be of particular interest in practice to determine the source of an anonymous threatening communication. We examine both neutral and abusive texts written by the same participants, in order to examine whether the two types of writing are different in terms of statistical relationships and prediction performance.

## 4.3 Method

### *4.3.1 Transparency statement*

Data, code, and supplemental materials are publicly available on the Open Science Framework: https://osf.io/ag8hu/

### *4.3.2 Sample*

800 participants were recruited through the crowdsourcing platform Prolific Academic. Only adult UK citizens with English as their first language were eligible. Participants who failed the attention checks[15] were excluded, resulting in a sample of 789.

### *4.3.3 Procedure*

Participants wrote both a stream-of-consciousness (SOC) essay about current thoughts and feelings, and an abusive text directed at a politician. Each task lasted for

[15] Two questions asking participants to select a specific response (e.g., 'strongly disagree') to continue

at least three minutes and participants had to write at least 100 words. For the abusive writing task, participants rated eight UK politicians from most to least favourite, then were assigned to write about their negative thoughts and feelings about their least favourite politician. They were told they could be as insulting, abusive, and offensive as they wanted. Lastly, the participants completed two personality questionnaires and were asked for their gender and age.

### *4.3.4 Personality measures*

In order to assess personality, two tests were used. The HEXACO-60 (Ashton & Lee, 2009) measures honesty-humility, emotionality, extraversion, agreeableness versus anger, conscientiousness, and openness to experience, on a scale from 1 = strongly disagree to 5 = strongly agree, with 10 questions per trait (i.e., resulting in a score between 1-50 per person). The Short Dark Triad (SD3) (Jones & Paulhus, 2014) measures Machiavellianism, narcissism, and psychopathy on a Likert scale from 1 = strongly disagree to 7 = strongly agree, with 9 questions per trait (i.e., a score of 1-63 per person).

### *4.3.5 Writing examples*

Below, we provide a writing example (original wording, anonymization added) for both the stream-of-consciousness and abusive writing tasks.

**Stream-of-consciousness.** *I feel content and I am reasonably happy at this present moment in time. It may be a challenging few months for me and I am looking forward to the time ahead. Some times I do feel at times that things get on top of me and find it hard to get going in the morning. I think that the future is bright for me and I fight on with perseverance and determination even though I have had some setbacks. I overall feel more confident and determined than ever even though at times I doubt myself for a brief moment.*

**Abusive writing.** *[POLITICIAN] you are a liar, a cheat, an abhorrent person, your arrogance is beyond repair, you are determined to drag the country into the gutter, you are a complete shit with total disregard for women, I hope you die in regret of what you have dragged our country into, we are now the laughing stock of [redacted], I hope you rot, shame on you, you are possibly the worst politician that we have ever had, you deserve a long and hard punishment for what you've done, you utter prick, please rot in hell for a long long time I hope*

### *4.3.6 Statistical tests*

Prior to performing the prediction tasks, we test for statistical relationships between author characteristics (personality, age, and gender) and LIWC2015 variables (Pennebaker et al., 2015) as well as Grievance Dictionary categories. We compute correlations for personality traits, applying a Bonferroni-corrected threshold of 0.05 / (89*9) = 0.000062 for 89 LIWC categories and 9 personality traits, and 0.05 / (22*9) = 0.00025 for 22 Grievance Dictionary categories and 9 personality traits. A Bonferroni correction accounts for the possibility of inflated false positives as a result of conducting multiple tests (for each linguistic category and personality trait).

Multivariate regression assessed the effect of age (and quadratic age, here: the absolute difference from age 40) on all LIWC2015 and Grievance Dictionary categories, while controlling for gender, following (Pennebaker & Stone, 2003). To examine gender and language, we assess whether there is a multivariate effect of gender in a MANOVA for all LIWC2015 categories following (Newman et al., 2008). We do the same for Grievance Dictionary categories. For both analyses, we report Pillai's Trace, a test statistic (ranging between 0-1) that increases if the (gender) effects are contributing more to the model. Thereafter, we conduct univariate ANOVAs to demonstrate the direction and magnitude (reported using Cohen's *d* effect size) of gender differences in LIWC and Grievance Dictionary categories.

### *4.3.7 Prediction tasks*

All prediction and classification tasks below are performed for stream-of-consciousness and abusive writing separately. In addition to the LIWC and Grievance Dictionary measures of linguistic content, we also examine prediction performance for stylistic features (e.g., grammatical categories in the LIWC, parts-of-speech, number of words). For each task, we test each of the following feature sets:

1. Number of words (baseline model)
2. Stemmed uni- and bi-grams (with stop words removed): single words (e.g., 'kingdom' and word pairs (e.g., 'united kingdom')
3. Parts-of-speech (universal POS tags from the *spacyr R* package (Benoit & Matsuo, 2020): grammatical categories such as nouns, verbs, and pronouns.
4. All 89 LIWC2015 categories: includes both content and style variables. In the abusive writing condition, we also include the proportion of abusive language[16] words as a feature.
5. All 22 Grievance Dictionary categories.
6. Composite feature set: all of the above features.
7. Filtered feature set: a selection of features from the composite feature set, filtered using a General Additive Model (Chouldechova & Hastie, 2015), and included if there is a functional relationship ($p < 0.05$) between the feature and outcome variable, during ten resampling iterations (Kuhn, 2010).
8. Pre-trained word embeddings, using the GloVe 6B corpus (Pennington et al., 2014): each word is represented as a vector of the cosine distance with 100 semantically similar words from the corpus. These measures are then averaged in order to represent each text as a function of 100 distances.

[16] A composite measure of abusive language following Kleinberg, van der Vegt, & Gill (2020), measuring profane and racist language from various dictionaries.

9. Pre-trained BERT language model (base uncased model with 12 layers and 768 hidden nodes): similarly represents words as a vector, but takes into account contextual relations between words through bi-directional training (Devlin et al., 2019).

All tasks are performed with a 10-fold cross validation on the training set. The training set consisted of 80% of the data, and the remaining 20% of the sample was used as a hold-out test set. This means that the model is trained ten times on ten different random samples of the training data. Then, the optimal model is chosen to perform test set predictions on the test set (the held out 20% of data). We can then evaluate the prediction performance by comparing the predictions for the test set to the actual observed values of this sample. The prediction analysis included the following steps:

1. Predicting the HEXACO and Dark Triad traits in isolation on a continuous scale (a regression model using a Support Vector Machine algorithm). We report the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), which represent the average prediction error across all iterations in absolute and proportional form, respectively. For instance, a MAE of 10 on a scale of 1-100 means predictions by the algorithm were 10 points off on average, which translates to a MAPE of 10%.
2. Predicting partitioned personality traits (binary classification with a Naïve Bayes algorithm). Following (Celli et al., 2013) we perform a median split on each personality trait. We report classification accuracy, a measure representing the number of correct classifications divided by all classifications performed.
3. Predicting author age (regression with an SVM algorithm). Performance metrics reported are MAE and MAPE.
4. Predicting author gender (male or female; binary classification with a Naïve Bayes classifier). Again, we report classification accuracy.

## 4.4 Results

### *4.4.1 Descriptive statistics*

Mean age of the participants was 37 years (*SD* = 12.73; 63.75% female). The average word count for SOC writing was 120.51 words, and 120.62 for abusive writing, with no significant order effect found for word count. We observed differences between SOC and abusive writing (i.e., manipulation check) for 60 out of 89 LIWC categories (adjusted *p*-value of 0.05/89 LIWC categories). Furthermore, the average number of abusive words[1] in abusive writing was 4.03, with a mean of 2.05 in stream-of-consciousness writing, representing a difference of $t(788) = 16.992$, $p < 0.001$, Cohen's $d = 0.60$. The order in which participants wrote texts did not affect the number of abusive words written in the abusive text, $t(781.88) = -1.67$, $p > 0.05$. Participants who wrote the SOC essay after the abusive text, used somewhat more abusive words, *t*(745.86) = 4.12, $p < 0.001$, albeit with a small effect size $d = 0.29$.

### *4.4.2 Personality*

**Correlations.** In Table 4.1, we present significant correlations ($p < 0.00056$) between HEXACO and Dark Triad traits with LIWC2015 and Grievance Dictionary variables. Note that no significant correlations were found for honesty, agreeableness, conscientiousness, narcissism, and Machiavellianism with any of the traits and in neither of the writing conditions. In short, for stream-of-consciousness writing we found significant relationships for only three out of nine personality traits, and 11 out of 89 LIWC categories and 5 out of 22 Grievance Dictionary categories. For abusive writing, we saw effects for four out of nine traits with 7 out of 89 LIWC categories and 3 out of 22 Grievance Dictionary categories. The effects ranged between $r = 0.14$ to $r = 0.24$ for stream-of-consciousness writing, and $r = 0.14$ and $r = 0.20$ for abusive writing.

Table 4.1 Correlations LIWC and Grievance Dictionary with personality traits

| **Stream-of-consciousness** | | | **Abusive writing** | | |
|---|---|---|---|---|---|
| *Dictionary* | *Category* | *r* ($R^2$) | *Dictionary* | *Category* | *r* ($R^2$) |
| | **Emotionality** | | | **Emotionality** | |
| LIWC | per. pronouns | 0.19 (0.04) | LIWC | function words | 0.15 (0.02) |
| LIWC | 1st pers, sing. | 0.20 (0.04) | LIWC | pronouns | 0.17 (0.03) |
| LIWC | neg. emotion | 0.14 (0.02) | LIWC | verbs | 0.15 (0.02) |
| LIWC | anxiety | 0.18 (0.03) | | **Extraversion** | |
| GD | desperation | 0.24 (0.06) | GD | hate | 0.14 (0.02) |
| GD | grievance | 0.14 (0.02) | | **Openness** | |
| GD | loneliness | 0.15 (0.02) | LIWC | verbs | -0.15 (0.02) |
| GD | paranoia | 0.15 (0.02) | LIWC | cogn. processes | -0.15 (0.02) |
| GD | suicide | 0.14 (0.02) | LIWC | comma | 0.18 (0.03) |
| | **Extraversion** | | GD | murder | 0.20 (0.04) |
| LIWC | tone | 0.15 (0.02) | GD | violence | 0.17 (0.03) |
| LIWC | negation | -0.15 (0.02) | | **Psychopathy** | |
| LIWC | cogn. proc. | -0.16 (0.03) | LIWC | sexual words | 0.15 (0.02) |
| LIWC | differentiation | -0.16 (0.03) | | informal language | 0.15 (0.02) |
| LIWC | seeing | 0.14 (0.02) | | | |
| LIWC | leisure | 0.15 (0.02) | | | |
| | **Openness** | | | | |
| LIWC | commas | 0.19 (0.04) | | | |

**Prediction.** Next, we report personality prediction performance for stream-of-consciousness (Table 4.2) and abusive writing (Table 4.3). On average, honesty, emotionality, extraversion, agreeableness, conscientiousness and openness (i.e., HEXACO traits) were predicted in SOC writing with an error margin of 9.62 points on a scale from 1-50 (*MAPE* = 19.24%), and 9.46 points for abusive writing (*MAPE* = 18.93%). The lowest average error in SOC writing was 7.60 points (*MAPE* = 15.20%) for predicting conscientiousness, with equal performance using the baseline model, parts-of-speech, the Grievance Dictionary, the filtered feature set, or word embeddings. For abusive writing this was the case for conscientiousness using the filtered feature set (average error 7.20 points, *MAPE* = 14.40%).

For Dark Triad predictions, the average error rate was 17.40 points on a scale of 1-63 (*MAPE* = 27.61%) for SOC writing and 17.07 points (*MAPE* = 27.10%) for abusive writing. The best performance in SOC writing was obtained for predicting Machiavellianism, using either the baseline model or the Grievance Dictionary (*MAE* = 11.15, *MAPE* = 17.70%). In abusive writing, Machiavellianism was best predicted using word embeddings (*MAE* = 11.09, *MAPE* = 17.60%). Importantly, a baseline model using only number of words often outperformed other feature sets. In both conditions, *n*-grams, parts-of-speech, LIWC, the composite and filtered feature sets, and the BERT language model did not perform best for any of the traits.

Table 4.2 SVM prediction performance for SOC writing (Mean Absolute Percentage Error)

| **Model** | **HEXACO** | | | | | | **Dark Triad** | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Hon.* | *Emot.* | *Extr.* | *Agr.* | *Consc.* | *Open.* | *Narc* | *Mach* | *Psych.* |
| Baseline | 18.9 | 16.3 | 21.3 | 18.8 | **15.2** | 17.4 | 29.6 | 17.7 | 30.3 |
| *n*-grams | 23.1 | 18.0 | 24.3 | 21.4 | 18.0 | 19.9 | 33.7 | 21.9 | 34.8 |
| POS | 19.1 | 16.4 | 21.6 | 18.8 | **15.2** | 17.3 | 30.5 | 18.0 | 29.7 |
| LIWC | 21.0 | 16.8 | 22.6 | 20.0 | 15.8 | 18.8 | 31.5 | 19.5 | 32.2 |
| Grievance | 19.0 | 16.2 | 21.3 | 18.8 | 15.3 | 17.1 | 29.6 | 17.7 | 30.2 |
| Composite | 23.8 | 22.3 | 25.1 | 23.1 | 20.2 | 22.5 | 38.2 | 23.6 | 37.2 |
| Filtered | 20.1 | 15.8 | 22.4 | 19.8 | **15.2** | 17.8 | 30.6 | 18.3 | 29.5 |
| Embeddings | 19.1 | 16.1 | 21.1 | 18.6 | **15.2** | 17.2 | 29.3 | 17.5 | 29.5 |
| BERT | 21.4 | 18.2 | 25.7 | 19.3 | 16.5 | 19.3 | 30.4 | 20.5 | 34.5 |

Table 4.3 SVM prediction performance for abusive writing (Mean Absolute Percentage Error)

| **Model** | **HEXACO** | | | | | | **Dark Triad** | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Hon.* | *Emot.* | *Extr.* | *Agr.* | *Consc.* | *Open.* | *Narc* | *Mach* | *Psych.* |
| Baseline | 19.0 | 16.3 | 21.3 | 18.8 | 15.1 | 17.0 | 29.3 | 17.8 | 29.2 |
| *n*-grams | 21.1 | 18.2 | 27.0 | 19.7 | 17.3 | 19.9 | 32.9 | 21.2 | 32.2 |
| POS | 19.5 | 15.9 | 21.7 | 19.3 | 14.7 | 15.6 | 29.5 | 17.9 | 29.4 |
| LIWC | 19.9 | 16.6 | 23.2 | 19.0 | 16.3 | 17.1 | 31.2 | 18.9 | 30.7 |
| Grievance | 19.0 | 16.3 | 21.2 | 18.8 | 15.1 | 16.8 | 29.3 | 17.8 | 29.4 |
| Composite | 22.9 | 20.3 | 27.7 | 21.0 | 17.9 | 20.6 | 37.0 | 22.5 | 34.7 |
| Filtered | 19.8 | 16.2 | 22.6 | 19.7 | **14.4** | 17.1 | 30.8 | 19.4 | 31.1 |
| Embeddings | 19.0 | 16.0 | 21.3 | 18.5 | 15.1 | 16.2 | 29.2 | 17.6 | 27.7 |
| BERT | 19.79 | 19.44 | 22.96 | 19.60 | 17.58 | 19.22 | 34.50 | 19.89 | 30.73 |

We also performed binary classifications for each personality trait (based on median splits on each trait), using the same features. In SOC writing (Table 4.4), the highest accuracy (0.62) was achieved for predicting openness (random baseline = 0.50) using BERT. For abusive writing (Table 4.5), the highest accuracies (0.62) were achieved in predicting openness using either word embeddings. The baseline feature set was never the top performer in either prediction task.

Table 4.4 Classification results stream-of-consciousness writing (accuracy)

| **Model** | **HEXACO** | | | | | | **Dark Triad** | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Hon.* | *Emot.* | *Extr.* | *Agr.* | *Consc.* | *Open.* | *Narc* | *Mach* | *Psych.* |
| Baseline | 0.52 | 0.49 | 0.50 | 0.52 | 0.46 | 0.46 | 0.49 | 0.50 | 0.53 |
| *n*-grams | 0.54 | 0.46 | 0.58 | 0.48 | 0.52 | 0.52 | 0.49 | 0.49 | 0.49 |
| POS | 0.50 | 0.52 | 0.52 | 0.51 | 0.50 | 0.52 | 0.45 | 0.52 | 0.56 |
| LIWC | 0.57 | 0.48 | 0.56 | 0.62 | 0.50 | 0.53 | 0.47 | 0.52 | 0.55 |
| Grievance | 0.56 | 0.50 | 0.49 | 0.56 | 0.47 | 0.52 | 0.48 | 0.48 | 0.49 |
| Composite | 0.50 | 0.46 | 0.54 | 0.49 | 0.51 | 0.58 | 0.55 | 0.49 | 0.51 |
| Filtered | 0.54 | 0.55 | 0.54 | 0.50 | 0.51 | 0.52 | 0.55 | 0.52 | 0.50 |
| Embeddings | 0.58 | 0.51 | 0.55 | 0.55 | 0.48 | 0.60 | 0.41 | 0.58 | 0.48 |
| BERT | 0.52 | 0.52 | 0.49 | 0.61 | 0.58 | **0.63** | 0.50 | 0.51 | 0.46 |

Table 4.5 Classification results abusive writing (accuracy)

| | **HEXACO** | | | | | | **Dark Triad** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | Hon. | Emot. | Extr. | Agr. | Consc. | Open. | Narc | Mach | Psych. |
| Baseline | 0.52 | 0.54 | 0.54 | 0.52 | 0.52 | 0.56 | 0.52 | 0.53 | 0.49 |
| *n*-grams | 0.51 | 0.50 | 0.49 | 0.53 | 0.55 | 0.52 | 0.53 | 0.48 | 0.51 |
| POS | 0.53 | 0.54 | 0.53 | 0.52 | 0.52 | 0.57 | 0.48 | 0.48 | 0.52 |
| LIWC | 0.47 | 0.51 | 0.48 | 0.45 | 0.52 | 0.47 | 0.58 | 0.49 | 0.56 |
| Grievance | 0.52 | 0.49 | 0.52 | 0.52 | 0.52 | 0.61 | 0.51 | 0.51 | 0.51 |
| Composite | 0.55 | 0.50 | 0.48 | 0.54 | 0.56 | 0.60 | 0.53 | 0.48 | 0.47 |
| Filtered | 0.54 | 0.56 | 0.54 | 0.55 | 0.52 | 0.58 | 0.56 | 0.61 | 0.48 |
| Embeddings | 0.56 | 0.54 | 0.51 | 0.48 | 0.52 | **0.62** | 0.45 | 0.53 | 0.49 |
| BERT | 0.60 | 0.46 | 0.59 | 0.52 | 0.52 | 0.60 | 0.54 | 0.51 | 0.52 |

### *4.4.3 Age*

First, we tested for possible statistical relationships between age with LIWC and Grievance Dictionary categories. In both writing conditions, no significant effect of age or quadratic age (while controlling for gender) on any of the LIWC2015 categories was found (all $p > 0.00056$, alpha-level adjusted 89 LIWC categories) nor on any of the Grievance Dictionary categories ($p > 0.0028$, alpha-level adjusted for 22 Grievance Dictionary categories).

The results of the age prediction task are presented in Table 4.6, which shows that the different models predicted age with an average error of about ten years. For the prediction of age in SOC writing, the best performing model using the filtered feature set achieved a *MAE* of 9.15 years (*MAPE* = 24.61%). For abusive writing, best performance was achieved using word embeddings as features achieving a *MAE* of 10.01 years (*MAPE* = 27.04%).

Table 4.6 Results age prediction (Mean Absolute Error)

| **Model** | **Stream-of-consciousness** | **Abusive writing** |
|---|---|---|
| Baseline | 10.10 | 10.23 |
| *n*-grams | 10.57 | 11.25 |
| POS | 9.29 | 10.04 |
| LIWC | 9.67 | 10.44 |
| Grievance Dictionary | 10.21 | 10.22 |
| Composite | 11.11 | 12.28 |
| Filtered | **9.15** | 10.16 |
| Embeddings | 9.67 | **10.01** |
| BERT | 10.13 | 10.70 |

### *4.4.4 Gender*

We observed a significant multivariate effect of gender on LIWC2015 variables in SOC writing, Pillai's Trace = 0.30, *F*(178, 1398) = 1.37, *p* < 0.001. Significant differences between genders (*p* < 0.00056), on individual categories were also found, where a positive Cohen's *d* value means the category was used more by women. Differences were found for *analytical language (d = -0.34), pronouns (d = 0.27), personal pronouns (d = 0.30), first person singular (d = 0.28), verbs (d = 0.35), discrepancies (d = 0.27), focus on the present (d = 0.26)*, and *apostrophes (d = 0.28)*. We also observed a significant multivariate effect of gender on all Grievance Dictionary categories, Pillai's Trace = 0.07, *F*(22, 764) = 2.79, *p* < 0.001. Significant differences between genders were found for the categories desperation (*d* = 0.38), grievance (*d* = 0.24), and soldier (*d* = -0.30).

For abusive writing we also found a multivariate effect on LIWC categories, Pillai's Trace = 0.32, *F*(178, 1398) = 1.47, *p* < 0.001. Significant differences between genders (*p* < 0.00056) were found for *analytical language (d = -0.44), function words (d = 0.41), pronouns (d = 0.47), personal pronouns (d = 0.47), first person singular (d = 0.31), articles (d = -0.32), auxiliary verbs (d = 0.33), verbs (d = 0.51), social words (d = 0.33), sexual words (d = -0.24), present focus words (d = 0.45)*, and *apostrophes (d = 0.26)*. We also observed a significant multivariate effect of gender on all Grievance Dictionary categories, Pillai's Trace = 0.07, *F*(22, 764) = 2.79, *p* < 0.001. Significant differences between genders were found for desperation (*d* = 0.38), grievance (*d* = 0.24), and soldier (*d* = -0.30).

Results for the gender classification task are presented in Table 4.7. For the prediction of gender in SOC writing, the highest accuracy of 0.64 was achieved using parts-of-speech as features. For abusive writing, best performing prediction accuracy was 0.70, again using parts-of-speech. It must be noted that the proportion of females in the dataset was 0.64, therefore there is practically no improvement over a model which always predicts the majority class.

Table 4.7 Results gender classification (accuracy)

| Model | Stream-of-consciousness | Abusive writing |
|---|---|---|
| | *Observed proportion of females: 0.64* | |
| Baseline | 0.62 | 0.59 |
| *n*-grams | 0.55 | 0.56 |
| POS | **0.64** | **0.70** |
| LIWC | 0.63 | 0.63 |
| Grievance Dictionary | 0.54 | 0.54 |
| Composite | 0.58 | 0.63 |
| Filtered | 0.62 | 0.60 |
| Embeddings | 0.56 | 0.66 |
| BERT | 0.60 | 0.55 |

## 4.5 Discussion

The current chapter examined the feasibility of author profiling through normal and abusive language, supplementing linguistic content with stylistic features of text. We looked at statistical relationships between linguistic variables and authors' personality traits and demographics (age, gender), and performed prediction experiments.

### *4.5.1 Statistical relationships*

First and foremost, some statistical relationships between (abusive) writing and author characteristics were observed. Language use in abusive texts were related to emotionality, openness, and psychopathy scores, whereas neutral writing showed relationships with emotionality, extraversion, and openness. We also observed gender differences in both types of text, but no significant effect of age on writing was found. Interestingly, our results seem to confirm that neutral and abusive writing are differently related to personality traits. Of particular interest is the fact that differences in language use based on differences in psychopathy can be measured in abusive writing, but did not emerge in neutral writing. Of further interest is the fact that differential gender differences emerged in abusive writing when compared to SOC writing, with men for example using more sexual words, and women using more social words.

It is important to note that the majority of LIWC categories and personality traits did not seem to be related to abusive or neutral writing. We also observed fairly low correlations with personality traits, with an average of $r$ = 0.14 for stream-of-consciousness writing, and $r$ = 0.12 for abusive writing. These values are smaller than the average correlation of $r$ = 0.23 found elsewhere (Hirsh & Peterson, 2009), and also do not reach the average of $r$ = 0.32 for language-based studies in particular (Azucar et al., 2018). Results were also qualitatively different from previous research: we do not observe relationships between agreeableness and conscientiousness with any linguistic variable in either writing condition, whereas previous research does report such effects (Azucar et al., 2018; Qiu et al., 2012). These disparities largely are due to the more stringent statistical criteria applied in the current study, but it can be argued that these corrections should have also been applied in previous studies in the first place. For instance, none of the correlations reported in Hirsh & Peterson (2009), a widely cited study on LIWC and personality traits, would have been considered statistically significant if corrections for the number of traits and LIWC categories had been performed[17].

In some cases, the relationships that emerged between author traits and LIWC categories are seemingly straightforward to interpret. For example, it is perhaps not surprising that participants who scored higher on the trait Emotionality used more words from the emotional LIWC categories negative emotion and anxiety, as well as similar (negative) Grievance Dictionary categories such as desperation, grievance, loneliness, paranoia, and suicide. The positive correlation between Extraversion and 'leisure' words could also have been anticipated, since it also replicates previous research (Nguyen et al., 2011). The result showing that individuals who scored higher on Psychopathy used more sexual words (in the abusive writing condition only) is interesting in light of previous research on the relationship between psychopathy and sexual deviance (Olver & Wong, 2006). For other relationships, particularly those with style categories, it is more difficult to explain why certain effects emerged (e.g., why higher openness was related to more use of commas or why high emotionality is related to more use of function words and pronouns). Of particular interest are the positive relationships between extraversion and hate, as well as those between high openness and murder and violence (all Grievance Dictionary categories). These results suggest that extraverted and open individuals are more inclined to write more 'violent' abuse, an effect that has not previously been shown. However, it is important to replicate this study in future in order to test whether these relationships persist. This chapter served as an exploratory study assessing possible relationships with abusive writing. In future replication studies, direct hypotheses on these relationships can perhaps be tested.

It must also be noted that the small effects obtained in this study would only be of practical significance for the specific purpose of author profiling (e.g., to identify sources of threats), if the linguistic variables can also serve as features for predicting demographic traits. For example, when converting correlations for personality traits to explained variance ($R^2$), on average the significantly related LIWC categories would explain just 0.01 percent of the variance in each of the traits. This means that the vast majority of variance cannot be explained by the LIWC or Grievance Dictionary, and we must explore further explanatory variables. In the next section, we discuss our

[17] The largest *r* in Hirsh & Peterson (2009) is 0.29 (for neuroticism and LIWC sadness), which equates to a $p = 0.0046$ (based on the reported $N = 94$) , which is above the threshold of $p = 0.00026$ if corrections for 5 traits and 39 LIWC categories are applied.

implementation of content features supplemented by additional stylistic features for author profiling.

### *4.5.2 Prediction tasks*

On average, the continuous prediction of personality traits was approximately 19% off in both neutral and abusive writing. Baseline models (using number of words) performed surprisingly well, whereas feature sets that showed success in previous studies (Golbeck et al., 2011; Preotiuc-Pietro et al., 2016) performed poorly in the current study. When personality prediction was simplified into a binary classification task, accuracy was also markedly lower than in previous research (Argamon et al., 2005; Oberlander & Nowson, 2006). The statistical tests showed that the LIWC and Grievance Dictionary alone explain little variance in personality, and even when supplementing these measures with a mixture of additional content and style variables (*n*-grams, parts-of-speech, embeddings, language models) we were not able to reach high regression or classification performance. Importantly, performance between writing conditions did not follow the same patterns, further illustrating the difference between abusive and neutral writing.

When predicting age, we observed an error margin of approximately ten years in both conditions. This stands in stark contrast with previous research, which used the same or fewer features and achieved an error of four years (Nguyen et al., 2013), potentially because a larger amount of data (in terms of text and participants) was available. However, approximating someone's age based on their language to plus or minus ten years may be helpful in a context where there is a wide range of possible ages.

Although we achieved an accuracy of gender classification of 70%, this is only marginally superior to a model which always predicts the majority class. Previous attempts achieved accuracy levels in the range of approximately 75% (with a 0.56 random baseline) to 92% (with a 0.55 random baseline) with similar feature sets as in the current work (Burger et al., 2011; Rangel et al., 2016). Again, even though we observed gender differences for various LIWC categories, these effects did not seem to transfer into high prediction performance.

There are several possible explanations for why the current results differ substantially from previous work on author profiling. First of all, our writing task involved instructed online writing, which is arguably different from handwritten stream-of-consciousness essays (Hirsh & Peterson, 2009; Pennebaker & King, 1999) or more natural, uninstructed social media posts on Twitter or Facebook (Azucar et al., 2018; Preotiuc-Pietro et al., 2016). In addition, the fact that participants were instructed to write abusive text when they normally may not be inclined to do so, may have lowered the external validity of the study. On the other hand, the highly anonymous nature of our task may have enabled some participants to be even more abusive than they would be in an online setting where messages can be traced back to a user profile. Lastly, the number of words (120 on average) may have impacted on our ability to adequately predict author traits from language. Nevertheless, online writing is generally short in nature, and therefore testing the ability to make predictions on short texts seems especially relevant for applying these methods to online contexts.

### *4.5.3 Practical significance*

Whether the error rates for personality, age, and gender obtained in this study are problematic, is a matter of perspective. One could argue that a prediction of personality within 10% of the actual value is useful if a general profile of a text author is desired. The same holds for the prediction of age and gender. However, if such an author profiling system were deployed in a threat assessment or law enforcement context, where decisions based on such a system may have far-reaching consequences, these inaccuracies may be highly problematic. For example, an inaccurate profile may lead to the identification or arrest of an innocent individual, and vice versa, the true source of a threat may be missed. However, to adequately evaluate the practical potential of an automatic system such as that utilised here, we would need to know what the 'accuracy rates' of human judgment of author profiles are. If the accuracy of human judgment is lower or equivalent to an automatic system, the benefits of an automatic system (scalability, reliance on measurable features) may be preferable.

The results of this study illustrate another important point: statistical significance does not equate to practical significance. Even though we observed significant statistical relationships between author demographics and language, these effects do not translate into accurate predictions, even when supplementing them with additional linguistic features. Increasingly, research focusing on violent individuals examines author characteristics through language, for example in terrorist manifestos and extremist forums (Akrami et al., 2018; Kop et al., 2019; Neuman et al., 2015). Oftentimes, these studies refer back to original research that has 'established' a link between language and personality (Hirsh & Peterson, 2009; Pennebaker & King, 1999), assuming that this relationship generalises to other types of language, such as that in violent or threatening texts.

The current study is the first to test this assumption in a context of abusive language, and found that these relationships are markedly different from neutral language, but of little importance in constructing accurate personality profiles. As such, our study suggests that the empirical body underpinning many studies on linguistic examinations of threats and terrorism, may be weaker than how it is portrayed. While the current study demonstrates that such predictions are currently inaccurate for the type of (abusive) writing tasks performed here, further research is necessary to explore if indeed there are other conditions where predictions are more successful. One future avenue may include using non-linguistic information (e.g., social media meta-data) as additional features in prediction algorithms. Other author characteristics may also be considered for prediction, such as education level or language proficiency (e.g., whether English is the first language of the author). The focus on age and gender in this study is straightforward because of its relevance to (criminal) investigations, for example those involving threateners of public figures, whereas personality prediction was chosen due to its increased popularity in threat assessment and offender profiling (Akrami et al., 2018; Neuman et al., 2015). All in all, regardless of which author characteristics and language features are used, it remains important to realise that these predictions are highly complex. Therefore, it is crucial to consider the limitations (i.e., error margins) of these systems before they are implemented in practice.

### 4.6 Conclusion

This chapter tested whether there are relationships between author personality, age and gender and the way in which texts are written, with specific attention paid to abusive texts, particularly those directed at public figures. We then used both content and style features from the (abusive) texts to predict personality, age and gender. Statistically significant relationships between author demographics and linguistic measures were found. For instance, individuals who scored high on extraversion and openness wrote more violently abusive texts. However, these statistical effects did not result in high prediction performance when compared to previous author profiling research. The results illustrate that statistical significance does not necessarily translate into practical significance. Therefore, we urge researchers and practitioners to exercise caution in author profiling based on (abusive) language, specifically in contexts were potentially dangerous individuals are the subject of interest.

# Chapter 5: Linguistic Trajectories: Language Use on YouTube surrounding the ‘Unite the Right’ rally

## 5.1 Introduction

Within the domain of grievance-fuelled violent language, it is of particular importance to study trajectories of language use, due to the relevance of processes of radicalization, escalation towards violence, and other changes in (extremist) language over time. Indeed, the importance of changes in language use was raised by threat assessment experts in Chapter 2, and receives increasing attention within violence research in general (Kleinberg et al., 2020; Scrivens et al., 2020; Smith et al., 2020; Spitzberg & Gawron, 2016). However, the preceding chapters examined ‘static’ language, in that content and style were measured for single rather than a sequence of texts. In contrast, this chapter fits within the growing trend in violence research of modelling linguistic changes over time (Kleinberg et al., 2020; Scrivens et al., 2020; Smith et al., 2020; Windsor, 2018).

The specific contribution of this chapter is that it implements a special use case for linguistic trajectories, namely its use in studying the effects of external events on language use. This is particularly relevant to grievance-fuelled violence, in the case of both small- (e.g., interpersonal conflict, adverse life events) and large-scale (e.g., elections, protests) events. When examining linguistic trajectories, it may for example be of particular interest to examine if certain events within the trajectory led to an increase or decrease in extreme or violent language. The 2017 Charlottesville rally and online activity of the alt-right offer an opportunity to study the effect of an offline event on the online behaviour of an extremist movement. In this chapter, we study language use among the alt-right on YouTube, in a time window surrounding the 2017 rally. By doing so, we demonstrate how computational linguistics can be used in threat assessment to model language over time, while assessing potential impact of an external event.

This chapter sheds light on the alt-right as a social movement by studying its language use over time in a unique dataset of YouTube video transcripts. We examine whether the Charlottesville rally functioned as a critical juncture in the online behaviour of the alt-right, and additionally contrast this with language use in a progressive sample of YouTube channels. In the next section, we discuss the alt-right and the Charlottesville rally. Thereafter, we outline the wider social movement literature as well as previous work on the effect of offline trigger events for online behaviour. Following this, we introduce our empirical examination of differences in language use on YouTube within and between alt-right and progressive channels, shortly before and after the Charlottesville rally.

## 5.2 The alt-right and Charlottesville

On 11 and 12 August 2017, dozens of alt-right, white supremacist and neo-Nazi individuals descended on Charlottesville, Virginia. The event, known as the 'Unite the Right' rally, turned fatal on the second day when a white supremacist deliberately drove into a crowd of counter-protestors, resulting in the death of one person and leaving several others injured (Hughes, 2018; Yan & Sayers, 2017). In recent years, the rise of the alt-right has been accompanied by several other acts of violence and terror attacks motivated by white supremacist ideologies, with 18 out of 34 extremist-related deaths in 2017 attributed to this group (Anti-Defamation League, 2017). In 2019, 90% of all 42 extremist murders in the United States were linked to right-wing extremism (Anti-Defamation League, 2018). At the same time, alt-right ideologies have become widespread online. Their content is easily accessible through social media platforms, and ideas are amplified on websites such as 4chan (Hine et al., 2017) and Gab (Zannettou et al., 2018). YouTube, in particular, has been described as a breeding ground for the alt-right (Ellis, 2018; Lewis, 2018).

The alt-right is not defined by a central organisation (Hodge & Hallgrimsdottir, 2019), nor does it 'offer a coherent or well-developed set of policy proposals' (Hawley, 2018). Instead, it has been referred to as a 'mix of rightist online phenomena' (Nagle, 2017) with white identity at its core (Hawley, 2017). The alt-right is variously characterised as anti-political correctness, anti-immigration, anti-Semitist, and anti-feminist (Hawley, 2017), ideologies which are commonly spread online through irony and dark humour. Scholars have begun to study the alt-right as a social movement, following the definition of 'a cluster of performances organised around a set of grievances or claims' (Hodge & Hallgrimsdottir, 2019; Tilly, 1993). It has been argued that the alt-right, mainly through online activity, engages in promoting a shared identity, fostering commitment to a common cause, and proclaiming the 'worthiness, unity, and size' of its movement (Hodge & Hallgrimsdottir, 2019).

The presence of the alt-right on social media has been particularly salient on YouTube (Ellis, 2018; Lewis, 2018). A 2018 report described an 'Alternative Influencer Network' on YouTube consisting of content creators 'who range in ideology from mainstream libertarian to openly white nationalist' (Lewis, 2018). It was found that alternative political influencers on YouTube adopt strategies of mainstream popular Youtubers to gain popularity, engaging in tactics for search engine optimisation and cultivating a relatable 'underdog' image (Lewis, 2018). Further research on this network argued that a "supply-and-demand framework" is needed to understand the popularity of alternative influencers, where the ease of uploading and monetizing fringe political videos on YouTube enables a supply that is in demand for viewers who feel alienated from mainstream media (Munger & Phillips, 2019). It has also been noted that the audience of alt-right YouTube videos is highly engaged with the content, displaying more likes and comments per view than other less extreme or mainstream media videos (Munger

& Phillips, 2019). While some videos or channels of extreme influencers may have been demonetised because advertisers do not want to be associated with the content, many have adopted alternative strategies to raise revenue. This includes the use of crowdfunding platforms such as Patreon or so-called "super-chats" where viewers make a donation for their message to be read out on a livestream (Munger & Phillips, 2019).

After the Charlottesville rally, various media outlets declared that 'white nationalists are winning' (Serwer, 2018) and 'the genie is out of the bottle' (Hughes, 2018). In addition, President Trump stated that there was 'blame on both sides' (Shear & Haberman, 2018), which prompted the suggestion that his claims 'reinvigorated' the alt-right movement (Shear & Haberman, 2018). In the aftermath of the rally, various reports also noted that white nationalists have entered mainstream conversation (Atkinson, 2018; Hughes, 2018) and some say they were aided in doing so by the Trump administration (Atkinson, 2018). Drawing from the study of protests by extreme right-wing groups and other social movements, one might argue that that the rally was not only important for the effects it had outside of the movement (e.g., in the media and politics), but also within the movement itself (Caiani et al., 2012; della Porta, 2018), which could be assessed by studying its YouTube videos. In the next section, we outline some of the social movement literature in order to better understand the alt-right and possible effects of the Charlottesville rally.

### 5.3 Social movement theory

Social movements have been studied for decades (Ackland & O'Neil, 2011; Calhoun & Weston, 2017; Diani, 1992; Langman, 2005; Polletta & Jasper, 2001; Porter, 2001). One definition states that a social movement is a group containing 'a plurality of individuals, groups and/or organizations, engaged in political or cultural conflicts, on the basis of shared collective identities' (Diani, 1992). Within these movements, the collective identity of the group can be actively emphasised through distinguishing between "us" and "them" (Hunt & Benford, 1994; Snow, 2001). These identities crucially need to be "framed" to mobilise supporters, where the frame generally serves to identify an injustice which can be addressed through a collective agency (Polletta & Jasper, 2001). Importantly, it is consistently shown that social movements make extensive use of the internet for communication and organisation (Ackland & O'Neil, 2011; Langman, 2005). Indeed, further definitions of social movements state that resources are generally shared through informal networks (Ackland & O'Neil, 2011; Diani, 2003). This phenomenon has also been studied within the context of white supremacist groups, for which the internet serves to reinforce their sense of collective identity, where white supremacy and difficulties faced by white people are emphasised (Adams & Roscigno, 2005).

Elements of social movement theory propose that people engage in social identity performance, which refers to behaviour that serves to express the norms of the social group one aims to belong to (Klein et al., 2007; Simon et al., 2008). Such behaviour includes affirming ones social identity, conforming to a social movement, strengthening ones identity, or mobilising others (Klein et al., 2007). Within the context of the alt-right, social identity performance may, for example, include using community-specific language (Hine et al., 2017) or memes online (e.g., Pepe the Frog, a popular internet meme appropriated by the alt-right (Hawley, 2018; Hine et al., 2017)), or to publicly adopt symbols related to white supremacism.

Research on the effect of media coverage and public discourse on social movements might explain the potential effect of the Charlottesville rally on the alt-right (Koopmans & Muis, 2009; Koopmans & Olzak, 2004). For example, research on right wing violence in Germany suggests that both positive and negative reactions from public figures to violent events may help to lend prominence to the movement (Koopmans & Olzak, 2004). That is, even if one aims to condemn a violent movement's message, the message is (at least partially) reproduced (Koopmans & Olzak, 2004). By studying newspaper sources, this line of research suggested that discursive opportunities, summarised as public visibility, resonance, and legitimacy affected the behaviour of right-wing movements, measured in terms of violent events against different target groups (Koopmans & Olzak, 2004). Public visibility refers to the number of outlets reporting on the movement and the prominence of the movement's message within those outlets (Koopmans & Olzak, 2004). Resonance is defined as the (positive or negative) reaction from public figures to the movement's message as well as the associated ripple effect in the media (Koopmans & Olzak, 2004). Legitimacy involves the general public's support of a message (Koopmans & Olzak, 2004). Similar discursive opportunities were also studied in relation to the rise in popularity of right-wing populist Pim Fortuyn in the Netherlands (Koopmans & Muis, 2009). In a similar vein, visibility (e.g., the extensive media coverage), resonance (e.g., responses to the rally from President Trump and other politicians), and legitimacy (e.g., subsequent protests and vigils denouncing the rally (Peltz, 2017)) can be observed in the context of the alt-right and Charlottesville rally.

The effect of discursive opportunities has yet to be examined for the specific case of the alt-right and the Charlottesville rally. If indeed the visibility of the alt-right increased following the rally, the message of the movement resonated in the media and public discourse, and the alt-right gained legitimacy through acknowledgement from opponents and the general public, we may expect to see changes in behaviour within the movement. Within the context of social identity performance, one may expect to see strengthened social identity consolidation within the alt-right movement as a result of the rally, President Trump's comments, and the media coverage of the rally. After the rally, we might expect increased expression of norms from the alt-right movement, for example in the form of stronger endorsement or more extreme expressions of in-group ideology. As has been raised previously, such behaviour may serve to further strengthen the movement or mobilise others to join.

In summary, the Charlottesville rally may have had an effect both within and beyond the alt-right, namely on their social identity performance and visibility, respectively. Similar theories have been proposed within the study of protests by extreme right-wing groups and other social movements. Large (sudden) protests are sometimes said to not only have important effects outside a social movement, but also within the movement itself by further radicalising or mobilising (non) members (Caiani et al., 2012; della Porta, 2018). Within this context, it is said that protests sometimes can trigger critical junctures that bring about abrupt and lasting changes both within and beyond a social movement (della Porta, 2018). In the next section, we further examine the effect of offline trigger events on online behaviour, as well as the interaction between the two domains.

### 5.4 Reactions to 'trigger' events

A large body of research has examined the interplay between online activity and offline events, particularly how both domains may influence each other. Early work in this area already suggested that the internet was transforming collective action by having a mobilizing influence on its users (Postmes & Brunsting, 2002). For instance, it has been argued that the online discussion within social movements influences the politicised identity of individuals (e.g., identification with a movement), which in turn influences their intentions to engage in collective action (e.g., attending a rally) (Alberici & Milesi, 2016). Similar claims have been made in light of the Arab spring, where online activity has been said to enable the formation of a new social identity (i.e., opposing the government) and mobilised people to engage in mass protests (McGarty et al., 2014). Besides political contexts, it has for example also been shown that online interactions in addiction recovery support groups (e.g., affirmation through likes, identification with the recovery community expressed in language) predicted offline retention in the program (Best et al., 2018; Bliuc et al., 2017). Besides offline (collective) action, online activity also seems to have an effect on offline media. For example, a 'symbiotic' relationship has been identified between Twitter feeds and top newspapers. Examining tweets from 2016 US presidential candidates and issue agendas in five US newspapers, it was found that tweets (e.g., on employment, immigration, national security) frequently predicted news agendas, and vice versa (Conway-Silva et al., 2018). In another study, it was suggested Tweets can be used to infer voter preferences (Tumasjan et al., 2011). Political party mentions and tweet sentiment were said to reflect actual election results in Germany (Tumasjan et al., 2011).

Of particular interest is the measurement of (hate) crimes in response to specific 'trigger' events, such as terrorist attacks (Williams & Burnap, 2016). Several studies have reported spikes in hate crimes following 9/11 or the 7/7 London attacks (Hanes & Machin, 2014; King & Sutton, 2013). In the online sphere, similar patterns can be observed. A survey conducted between 2013-2015 also showed that young people in Finland witnessed increased hate online shortly after the 2015 Paris attacks (Kaakinen et al., 2018). In the aftermath of the 2013 Woolwich terrorist attack, researchers observed hate directed at black and minority ethnic groups in Tweets directly related to the attack (Williams & Burnap, 2016). In another study on information flows on Twitter in the aftermath of the Woolwich attack, it was found that tweet sentiment was predictive of the number retweets and their timespan (time between first and last retweet). Offline news reports also (positively) predicted the number of retweets on the same topic (Burnap et al., 2014). Another study examined the effect of jihadist terrorism and Islamophobic attacks on hate speech on Twitter and Reddit, measured over a period of 19 months shortly after 13 extremist attacks (Olteanu et al., 2018). It was found that, following jihadist terrorist attacks, hate speech targeting Muslims, particularly those advocating violence, increased more after terror attacks compared to a counterfactual simulation (Olteanu et al., 2018). An increase in hate speech targeting Muslims was not found following Islamophobic attacks, with the exception of messages posted after the 2017 Finsbury Park Mosque attack (Olteanu et al., 2018). Hate speech and white nationalist rhetoric have also been measured during the 2016 US elections on Twitter, using a dictionary approach with Hatebase, the Racial Slur Database, and the Anti-Defamation League's database of white-nationalist language (Siegel et al., 2018). Tweets were examined by means of an interrupted time series analysis, showing a spike in hate speech in the Trump dataset following the imposed travel ban in early 2017 (Siegel et al., 2018).

A small number of studies have looked at the specific effect of the Charlottesville rally on online behaviour. In a qualitative study of Twitter accounts of two alt-right and one far-left organisation in the six weeks leading up to the Charlottesville rally, it was observed that the two sides frequently targeted each other, framing the opposing group as the enemy (Klein, 2019). Manual examination of the tweets showed that the alt-right accounts frequently referred to 'the left' and 'liberals' as unpatriotic and communist. At the same time, the far-left accounts dubbed the alt-right 'suit and tie Nazis'. Furthermore, both the alt-right and far-left groups incited violence in the weeks leading up to the Charlottesville rally and called for action among their supporters. A tweet from one of the alt-right groups read 'The left is preparing lynch mobs to descend on the Unite the Right rally in Charlottesville, VA... This is going to be fun.' (Klein, 2019). Other research has shown that anti-Semitic memes and rhetoric increased after the 2016 US elections and the Charlottesville rally (Zannettou et al., 2019). Several million posts and images from 4chan and Gab were studied for racial slurs and anti-Semitic terms, with a case study of a specific anti-Semitic meme showing that such content also spreads to mainstream platforms such as Twitter and Reddit (Zannettou et al., 2019).

### 5.5 The current study

Taken together, the theoretical lines discussed would suggest that the Charlottesville rally functioned as a critical juncture for the alt-right, engendering changes in online social identity performance and visibility of the social movement. In order to empirically examine this claim, the present study takes a closer look at a network of alternative political influencers (Ellis, 2018) (hereafter, 'alternative group'), by examining YouTube video transcripts extracted from channels by these individuals. These video transcripts are compared to those from YouTube channels whose political orientation can be considered more progressive (hereafter, 'progressive group'), in order to assess whether the Charlottesville rally also had an effect outside of the alt-right movement itself.

This study has two aims. First, we compare language use *between* the alternative and progressive group in a sixteen-week timeframe surrounding the Charlottesville rally. Second, we assess whether the rally had an effect on language use *within* the two groups. For the alternative group, we do not postulate any directional hypotheses about changes in language use. Nevertheless, in light of the social movement and social identity performance literature, we expect to see changes in social identity performance after the rally reflected in language use on YouTube. For the progressive group, we do not claim that the channels studied act as a social movement, and thus we have no expectations of social identity performance. However, we are interested in seeing whether the channels lend any discursive opportunities to the alt-right through language use in their videos, thereby potentially contributing to the increased visibility of the alternative group.

The first aim is addressed through structural topic modelling, in order to compare the prevalence and content of topics between the two groups. The second aim is addressed using a word frequency approach, in which we examine the frequency of common phrases before and after the Charlottesville rally, searching for sudden increases or decreases as a result of the rally.

## 5.6 Method

### *5.6.1 Transparency statement*

Supplemental materials, data and code to reproduce the analysis are available on the Open Science Framework: https://osf.io/yedt7/

### *5.6.2 Channel selection*

YouTube channels were selected for analysis from two main sources. First, we drew from the list of 65 YouTube users referred to as the 'Alternative Influence Network' in the 2018 Data & Society report on political influencers (Lewis, 2018). Based on this list, we searched for a designated YouTube channel for each individual. If an individual did not have a designated YouTube channel or their channel was no longer available, we searched for the individual's name through the YouTube search function. For example, videos featuring Alex Jones (who was banned from YouTube so no longer has a designated channel) were obtained through the search query 'alex jones full show'. The group of alternative YouTube channels consisted of 56 channels and search queries used for transcript retrieval. Because data collection was done retrospectively, some channels appearing in the Data & Society report may not have been available (also when searched) because they were banned or deleted (total of 9 channels, 13.85%). Second, for the comparison group of progressives, we drew from two online lists of progressive YouTube channels[18]. Since the lists referred to specific existing channels, search queries for specific persons were not necessary. In total, 13 progressive channels and 56 alternative channels were used for transcript retrieval. For all channels and search queries, we retrieved the URLs for all available videos on 1 October 2018.

### *5.6.3 Transcript retrieval*

The method for retrieving YouTube video transcripts follows the procedure of related research (Kleinberg, Mozes, et al., 2018; Soldner et al., 2019). In order to retrieve the transcripts, a Python script was written using www.downsub.com to obtain XML-encoded transcripts. The transcripts were either automatically generated by YouTube or manually added by the YouTube user. In some cases, no transcript was available, because users disabled the transcript availability. XML-tags and time-stamps were removed, resulting in a single, non-punctuated string for each video transcript.

### *5.6.4 Data cleaning*

Videos that contained fewer than 100 words were not considered for analysis, following previous work on YouTube transcripts (Kleinberg, Mozes, et al., 2018; Soldner et al., 2019). Using R software, each video was checked for English language, and was excluded if it contained fewer than 50% English words. Videos were also excluded if they contained fewer than 90% ASCII characters. The video transcript strings were lower-cased and stopwords, unnecessary whitespace or punctuation were removed using the R packages *tidytext* (Silge & Robinson, 2019), *tm* (Feinerer & Hornik, 2018) and *qdap* (Rinker, 2019).

### *5.6.5 Sample*

To capture the immediate and continuing effects of the rally we sampled video transcripts up to approximately two months after he rally, as well as an equal timeframe preceding the rally. Previous works assessing the online effects of offline events have examined timeframes ranging from two weeks (Burnap et al., 2014), a month (Kaakinen et al., 2018; Tumasjan et al., 2011; Williams & Burnap, 2016), to one or several years (Bliuc et al., 2019; Hanes & Machin, 2014; Olteanu et al., 2018). Because no consensus seems to exist in the literature, we opted for a middle ground of two months pre- and post-event (data from a longer timeframe is available on request). This resulted in a total sample of videos spanning sixteen weeks (eight weeks pre- and post-rally). Descriptive statistics for this sample are given in Table 5.1.

[18] http://the2020progressive.com/top-13-progressive-news-shows-youtube/ and https://medium.com/@tejazz89/top-5-youtube-channels-to-follow-if-you-are-a-true-progressive-ee2abc78d58f

Table 5.1 Descriptive statistics video sample

| | **Alternative** | **Progressive** |
|---|---|---|
| Total videos | 2,684 | 4,458 |
| Total *N* words | 3,868,744 | 2,804,703 |
| Word count | Mean: 1,448 (SD = 1,612)<br>Min: 34, Max: 12,085 | Mean: 632 (SD = 860)<br>Min: 33, Max: 11,645 |
| View count | Mean: 120,639 (SD = 234,513)<br>Min: 2, Max: 17,340,303 | Mean: 24,787 (SD = 47,468)<br>Min: 12, Max: 2,475,766 |

### *5.6.6 Structural topic model*

To assess the differences in language use between the alternative and progressive groups, we construct a structural topic model. This method can be used to automatically extract underlying latent topics in a corpus (Blei, 2003; Roberts, Stewart, & Tingley, 2014). Common approaches are Latent Dirichlet Allocation (Blei, 2003) and Correlated Topic Models (Blei & Lafferty, 2007), probabilistic models which are based on the assumption that a piece of text consists of a mix of topics, which in turn are a mix of words with probabilities of belonging to a topic (Blei, 2003; Roberts, Stewart, & Tingley, 2014). A structural topic model is a type of Correlated Topic Model, with the added benefit that one can incorporate document-level covariates (e.g., document author, political orientation, date) and assess whether these covary with topic prevalence (i.e., the degree to which documents in a corpus are assigned a specific topic) and content (i.e., the terms in a topic; Roberts, Stewart, & Tingley, 2014).

We first define a document-frequency-matrix with both unigrams (e.g., ‘president’) and bigrams (e.g., ‘donald trump’) in the corpus, which is then used to construct the structural topic model. We include group (alternative vs. progressive) as a covariate for topic prevalence and content. Topic models are fit with a varying number of topics[19], after which we select the best fitting model based on the trade-off between semantic coherence and exclusivity (Mimno et al., 2011; Roberts, Stewart, & Tingley, 2014), two metrics frequently used to assess whether a topic is semantically useful (Roberts, Stewart, & Tingley, 2014; Roberts, Stewart, Tingley, et al., 2014). Semantic coherence is a measure of the co-occurrence of highly probable words in a topic, and has been shown to correlate with expert judgments of topic quality (Mimno et al., 2011). It has been proposed that a measure of exclusivity of words to topics is needed to further determine topic quality, otherwise several topics may be represented by the same highly probable words, if one relies on semantic coherence alone. Exclusive topics are made up of words that have a high probability under one topic, but a low probability under other topics (Roberts, Stewart, Tingley, et al., 2014).

After selecting a model, we present topics for which a significant effect of the covariate group was found for topic prevalence (i.e., a significant difference in between

[19] 20, 40, 50, 60, 70, 80, and 90 topics

alternative and progressive channels), in order of total expected topic proportion for the corpus. Based on manual inspection of frequent and exclusive topic words (Bischof & Airoldi, 2012; Roberts, Stewart, & Tingley, 2014), we assign labels to topics. We also present a selection of three topics along with the words which differed between the alternative and progressive groups, in order to illustrate how the alternative and progressive groups talk about the same topic in different ways.

### *5.6.7 Word frequency*

To examine possible changes in word frequency within the alternative and progressive group as a result of the rally, we compute the frequency of all bigrams for each week in both groups separately. By dividing these values by the total number of bigrams for each day, we obtain the daily proportion for each bigram. Thereafter, we can assess whether there is a structural breakpoint in the proportion of each bigram as a result of the rally. This is done by means of the Chow test (Chow, 1960; Zeileis et al., 2002), with which we determine whether a breakpoint in the intercept and slope occurred at the time of the rally. In order to do so, we test for the equality between a model of bigram proportions before the Charlottesville rally, and a model of bigram proportions after the rally. In both models, the proportion of each bigram is represented as a function of Date (day on which the proportion was measured, between 15 June and 7 Oct 2017). We compute an F-value for the equality between the two models for each bigram, and report those which are found to differ significantly pre- and post-rally. In addition, we present associated intercept and slope changes.

## 5.7 Results

### *5.7.1 Structural topic model*

We decided on a structural topic model with 40 topics based on examination of semantic coherence and exclusivity (see supplemental materials for results with different numbers of topics). Thereafter, we found that the covariate Group was significant for the prevalence of 30 topics. Figure 5.1 shows the topics for which Group significantly covaried with topic prevalence. We assigned labels (e.g., ‘Obamacare’) based on examination of highly probable as well as frequent and exclusive words. The alternative group discussed more of the topics which were labelled as swearing, filler words, future focus, economy & business, race, immigration, women, free speech, internet, Fox News, police, social justice, mainstream media, personal concerns, radical Islam, and gay marriage. The progressive channels focused more on Donald Trump, taxes, healthcare, YouTube, the presidency, party politics, hate, law, media investigations, presidential candidates, Obamacare, voting, foreign affairs and Asia/nuclear weapons.

Figure 5.1 Topic prevalence per group

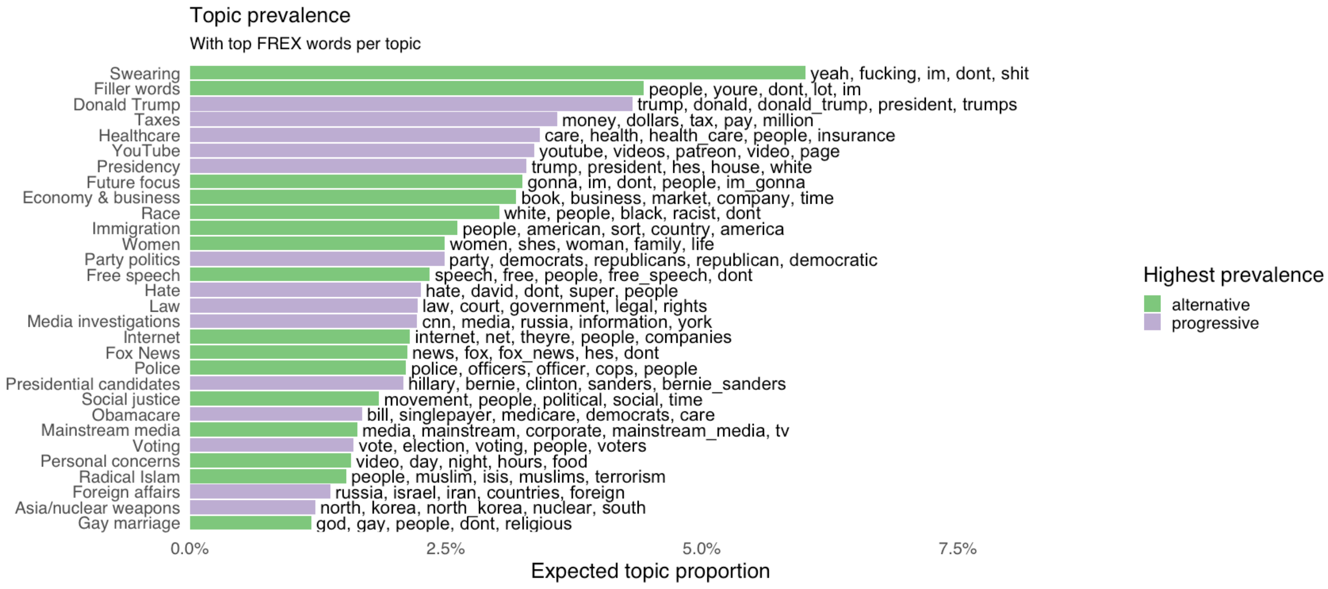


Table 5.2 shows how three topics are discussed differently by alternative and progressive channels (full list of topics available in supplemental materials). By including topic content as a covariate, we are able to see which words are more associated with each group per topic. For example, the ‘social justice’ topic is discussed as a ‘movement’ and ‘resistance’ by progressive channels, whereas the alternative group uses the term ‘identity politics’. The topic of ‘women’ is discussed with terms referring to sexuality by both groups, but the progressive group also includes terms referring to family. Although both topics discuss ‘race’ with terms relating to racism, the progressive group uses terms such as ‘white supremacist’ and ‘nazis’.

Table 5.2 Topic content per group

| Topic Label | Group | Common terms |
|---|---|---|
| Social justice | Alternative | diversity, social justice, google, differences, identity politics |
| | Progressive | movement, conference, resistance, Africa, organization |
| Women | Alternative | women, female, sexual, male, rape |
| | Progressive | women, sex, father, child, family |
| Race | Alternative | white people, black people, racist, blacks, racism |
| | Progressive | charlottesville, white supremacist, racism, racist, nazis |

### *5.7.2 Word frequency approach*

We show the ten bigrams for which the Chow test *F* statistic, indicative of a joint breakpoint in intercept and slope, was largest[20], in the alternative group (Table 5.3) and progressive group (Table 5.4). We also show the direction and magnitude of intercept and slope changes after the rally; please note that slope changes were very minimal (albeit statistically significant) and therefore have been multiplied by 10,000

[20] Further bigrams that exhibited breakpoints are available in the supplemental materials.

for interpretability. Among the top ten bigrams with breakpoints for both groups, the majority relate to the rally itself, such as ‘white nationalist’ and ‘happen charlottesville’. Note that some bigrams showed a breakpoint in both groups, namely, ‘white nationalist’, ‘happen charlottesville’, ‘charlottesville virginia’, and ‘neonazi white’. In the alternative group, several bigrams unrelated to the rally (e.g., ‘hit bell’, ‘video bitcoin’) also exhibit strong breakpoints. In the progressive group, only one bigram with a strong breakpoint in the top ten seems to be unrelated to the rally, namely ‘hurricane maria’. In order to further illustrate the bigram proportion breakpoints, we show the progression of the first three (based on the magnitude of the Chow test *F*) bigrams for the alternative group (Figure 5.2) and the progressive group (Figure 5.3). In both groups, the proportion of the bigrams depicted significantly increases in terms of intercept, with slight (negative) changes in slopes.

Table 5.3 Ten bigrams with largest Chow test (*F*) statistic in alternative group

| **Bigram** | **Chow test (*F*)**[a] | **Effect size *d*** | **Intercept change** | **Slope change**[b] |
|---|---|---|---|---|
| white nationalist | 42.84 | 0.89 | 1.12 | -0.64 |
| happen charlottesville | 38.09 | 0.84 | 0.20 | -0.11 |
| hit bell | 35.47 | 0.81 | 0.26 | -0.15 |
| video bitcoin | 31.85 | 0.76 | 0.20 | -0.11 |
| subscribe hit | 28.92 | 0.73 | 0.24 | -0.14 |
| charlottesville virginia | 28.10 | 0.72 | 0.11 | -0.06 |
| nazi flag | 26.86 | 0.70 | 0.19 | -0.11 |
| descript patreon | 26.70 | 0.70 | 0.19 | -0.11 |
| neonazi white | 26.00 | 0.69 | 0.11 | -0.06 |
| begin video | 25.90 | 0.69 | 0.33 | -0.19 |

*Notes.*[a]For all coefficients (*F*, intercept and slope changes)*:* $p < 0.001$
[b]Slope change estimates have been multiplied by 10,000 for interpretability

Table 5.4 Ten bigrams with largest Chow test (*F*) statistic for progressive group

| **Bigram** | **Chow test (*F*)[a]** | **Effect size *d*** | **Intercept change** | **Slope change[b]** |
|---|---|---|---|---|
| white supremacist | 50.32 | 0.96 | 5.30 | -3.04 |
| happen charlottesville | 38.16 | 0.84 | 0.38 | -0.22 |
| charlottesville virginia | 37.24 | 0.83 | 0.65 | -0.38 |
| white supremacy | 34.67 | 0.80 | 1.75 | -1.00 |
| robert lee | 33.47 | 0.78 | 1.04 | -0.59 |
| neonazi white | 31.28 | 0.76 | 0.42 | -0.24 |
| counter protest | 27.45 | 0.71 | 0.90 | -0.52 |
| white nationalist | 27.01 | 0.70 | 1.89 | -1.09 |
| confederate monument | 24.88 | 0.68 | 0.72 | -0.41 |
| hurricane maria | 24.88 | 0.68 | -0.64 | 0.37 |

*Notes.* [a]For all coefficients (F, intercept and slope changes): p < 0.001
[b]Slope change estimates have been multiplied by 10,000 for interpretability

Figure 5.2 Observed proportion of three bigrams with largest F-value in alternative group

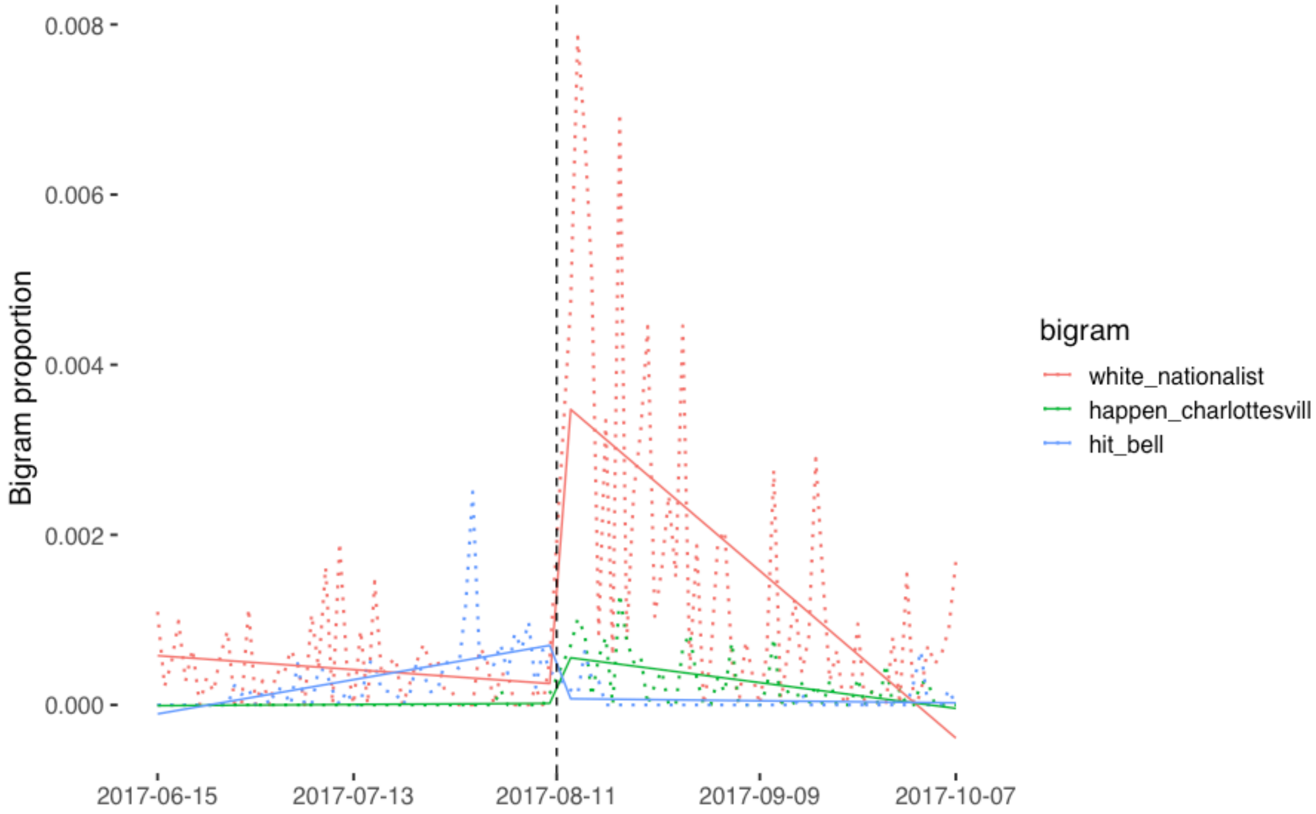
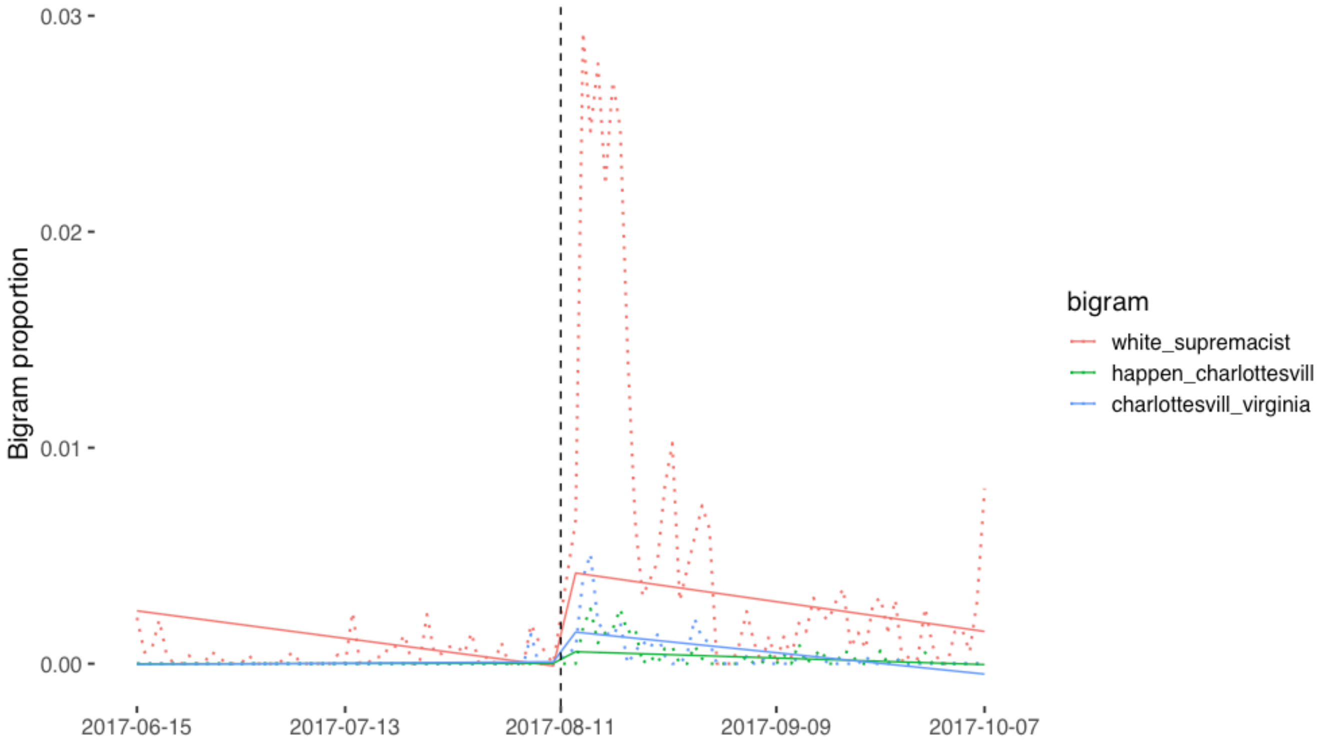


Figure 5.3 Observed proportions of three bigrams with largest with largest F-value in the progressive group



## 5.8 Discussion

The current study examined language use for alternative and progressive YouTube channels around the time of the Charlottesville 'Unite the Right' rally. Considering the wider aim of improving automated linguistic threat assessment, the purpose of this chapter was to demonstrate a way in which trajectories of language use over time can be measured, particularly in response to an external event. Both factors have been raised previously as holding particular relevance to grievance-fuelled targeted violence, both in this thesis as well as in the wider literature (Kleinberg, van der Vegt, & Gill, 2020; Scrivens, Davies, et al., 2020; Spitzberg & Gawron, 2016). This chapter compared language use between alternative and progressive YouTube channels surrounding the 2017 rally, and assessed whether the event had an effect on language use within the two groups. We examined language use in both groups in terms of structural topic models, and searched for structural breakpoints in a change of content as a result of the rally. We consider the outcome of both approaches in turn, followed by an interpretation of the results in light of social movement theory.

### *5.8.1 Differences between alternative and progressive channels*

The first line of inquiry examined whether there were structural differences in the prevalence and content of topics between groups. This analysis illustrates the matters discussed in videos throughout this period in the two groups. Perhaps unsurprisingly, several topics in both groups related to politics and current events (e.g., taxes, healthcare and economy). We found that the prevalence of the majority of topics covaried with the political orientation of channels (alternative or progressive). For instance, topics that may be loosely associated with the 'ideology' of the alt-right were found to be used more by the alternative group, such as race, immigration, radical Islam, gay marriage, and free speech (Hawley, 2017; Nagle, 2017). Indeed, the concept of free speech has frequently been linked to the alt-right and white nationalism, where the right to free speech is used to "advance racist and sexist ideas" (Mayer, 2018). In a similar vein, discussions relating to women's and LGBT rights as well as social justice which appeared in our corpus have also been linked to the far right (Lewis, 2019), a further potential indicator of expressing social norms within this group. The topic of so-called mainstream media was also discussed more by alternative channels, as well as Fox News in particular. In contrast, the progressive channels discussed Donald Trump to a larger extent, as well as other more general current affairs, such as the Democratic and Republican parties, legal matters, Obamacare, and foreign politics. Interestingly, we also observed a difference in prevalence of swearing, which was significantly higher for alternative influencers. Swearing may be a way of conforming to a social group, and our results suggest that this kind of language is more common among alternative than progressive YouTube channels. The content of topics further elucidated differences between groups, for example the way in which the alternative and progressive channels discussed the topic of race with differential terms, with the latter using terms that seem to condemn racism (e.g., 'white supremacist', 'nazis'). In short, the structural topic models indeed show that there are differences in topics between alternative and progressive YouTube channels. Some of these patterns in topics may support previous claims that the alt-right behaves as a social movement (Hodge & Hallgrimsdottir, 2019).

### *5.8.2 Effects of the rally within alternative and progressive channels*

The word frequency approach showed the rally had an effect on language use within the two groups, illustrated by several breakpoints in bigram proportions that coincided with the Charlottesville rally. Unsurprisingly, the use of words relating to the rally (e.g., confederate monument, white nationalist, white supremacist) increased at this point. While the proportions of these bigrams all exhibited sudden increases, the mentions did decrease over time in the post-rally timeframe. This possibly reflects a 'natural' descending trend for discussions of an event as time progresses, which potentially adds to the justification of measuring bigram proportions over time to assess reactions to events in language.

Although there was some overlap between groups in bigram use, it also appears that both groups discussed the events in a different light. The progressive group increasingly mentions 'white supremacists' after the rally, whereas the alternative group increasingly mentions 'white nationalists'. These differences in terminology seem to reflect a more general divide between groups. Indeed, 'white supremacists' is a term preferred by people who study or condemn the movement, but the term is not preferred among the extreme right itself (Hawley, 2017). Among the alt-right, the preferred term is ‘white nationalist’, which indeed emerges from our data (Hawley, 2017). This preference relates to the wish to establish separate white *nations*, in contrast to multiracial nations where whites are the dominant (‘supreme’) group (Hawley, 2017). One could argue that this difference in terminology may reflect increased expressions of in-group (alt-right) norms, an aspect of social identity performance.

Further breakpoints observed in the progressive group refer to several details related to the rally, such as the confederate statue of Robert Lee, the removal of which gave rise to the Charlottesville rally (McCausland, 2017). A strong increase within progressive post-rally videos was observed for the mention of counter-protestors, highlighting potential condemnation of the rally and the violence that ensued against counter-protestors (Reuters, 2019). Interestingly, none of these details appear in the top ten of breakpoints for the alternative group. We do not propose that these patterns in language use provide evidence for social identity performance on part of the progressive group, as we studied a user-generated and highly heterogenous list of channels, for which, in contrast to the alternative group, no claims have been made that they form a specific social movement. However, mentions of the rally on part of the progressive group may have lent further discursive opportunities and resulting visibility to the alternative group (Koopmans & Muis, 2009; Koopmans & Olzak, 2004).

Interestingly, a large number of the top ten bigrams  in the alternative group for which a breakpoint was observed did not relate to the rally, but to the promotion of YouTube channels, for instance urging viewers to subscribe to a channel, enable notifications, or donate to Patreon, a platform where content creators can crowdsource donations (Regner, 2020). This behaviour lends further support to previous findings that the alternative YouTubers promote their channels like mainstream influencers, and monetise their videos through donations to create a devoted fanbase (Lewis, 2018; Munger & Phillips, 2019). In short, the examination of bigram proportion breakpoints showed that the Charlottesville rally did seem to have an effect on language use in both groups separately.

### *5.8.3 The alt-right as a social movement*

Both the language differences between and within the progressive and alternative video transcripts can be interpreted in light of social movement theory, and also add to our understanding of the effect of offline events on online behaviour. First, we observed several topics prevalent among alternative channels that could be seen as in line with the social identity of the alt-right. Swearing, distrust in mainstream media, white nationalism, and an emphasis on free speech distinguished the alternative group from the progressive group. Second, we saw marked changes in language after the Charlottesville rally. The alternative YouTubers not only discussed the rally but seemingly also promoted their channels more. While further examination of the contexts in which these calls are made will be needed, the fact that (positive) breakpoints (in intercept) appear at the time of the rally may be a sign of mobilising others, urging viewers to show their support for the alternative channels and related movements. Indeed, if these calls are a direct result of the rally, the event may be viewed as a critical juncture for the alt-right movement, where the rally served as a triggering event for increased social identity performance and mobilisation, aimed at strengthening the movement. Furthermore, the progressive group was also shown to lend resonance and visibility to the alt-right by discussing the rally, even if condemning language (e.g., ‘white supremacist’ over ‘white nationalist’) was used. These discursive opportunities may in turn have fuelled social identity performance on part of the alt-right (Koopmans & Muis, 2009; Koopmans & Olzak, 2004). That is, by discussing and even condemning the alt-right rally, the progressive group lends further resonance and visibility to the movement (Koopmans & Olzak, 2004). All in all, results of this study may support the notion that the alt-right behaves as a social movement and that the (offline) Charlottesville rally had an effect on online social identity performance within the alt-right on YouTube, and possibly also outside of the movement as demonstrated by analyses of progressive YouTube video transcripts.

### 5.9 Limitations and future work

The current study is not without limitations. First, data selection and subsequent operations may have impacted the results of our analysis. For example, the sources that we have drawn on for the YouTube videos were unbalanced in nature, with the progressive sample consisting of more videos than the alternative sample. Furthermore, the two groups also differed in terms of view counts and video length, both factors which may have impacted on language use. In addition, while the alternative channels were drawn from a research report, the list of progressive channels were drawn from user-generated online lists. Future research may be aimed at curating an expert-verified or crowd-sourced dataset of channels with different political biases[21]. Other search strategies to identify alt-right channels that do not rely on keyword searches, for example using hyperlinks posted on alt-right forums (Mariconti et al., 2018), should also be considered in future work. Furthermore, when we selected videos for analysis only transcripts with more than 100 words and a pre-specified percentage of English words were retained. These decisions were guided by previous research (Kleinberg, Mozes, et al., 2018; Soldner et al., 2019) and our aim to retain only high quality transcripts suitable for topic modelling. A full dataset without these filters applied is made available for other researchers to experiment with other constraints. In a similar vein, researchers may be interested in examining longer or shorter timeframes surrounding the Charlottesville rally or even other events, and further data from our transcript retrieval (all videos available until 1 October 2018) is available on request. Lastly, transcript quality may have varied based on whether they

[21] Similar to https://mediabiasfactcheck.com/ and https://www.allsides.com/media-bias/media-bias-ratings but for YouTube channels

were generated through automatic speech recognition or manually reviewed and/or added to a video. YouTube notes that automatic captions may be inaccurate due to mispronunciations, accents, or dialect[22]. Nevertheless, relying on the provided captions was the most straightforward way to obtain transcripts, and future work may examine what the effect is of different automatic speech recognition technologies on linguistic analyses.

Topic modelling involves several decisions on part of the researcher. For instance, various approaches exist for selecting the number of topics for a model, with no consensus in the research community (Roberts, Stewart, & Tingley, 2014). Furthermore, assigning labels to topics is based on the interpretation of the researcher, with decisions highly sensitive to human bias. Nevertheless, we provide alternative models (with different numbers of topics) and further terms associated with topics in the supplemental materials, for the reader to examine the outcome of our analyses, giving way to alternative explanations. Furthermore, some topics were difficult to interpret (e.g., ‘Filler words’ and ‘Future focus’), mostly because they were composed of parts-of-speech with little meaning, or because the words did not form a coherent topic, and merely consisted of words that were used in the same way.

A bag-of-words approach utilised in both the topic modelling and the word frequency approach also holds its limitations. Specifically, bag-of-words models disregard word order and context. Furthermore, when measuring the prevalence of bigrams, polarity words or adjectives (e.g., ‘not’, ‘very’, ‘super’) that preceded each bigram may not have been captured. This issue may be solved in future by using trigrams, although relevant *n*-grams that occur even further away from the keyword will still not be captured and further context will still be disregarded. As has been raised in the discussion, the breakpoints we observed only show that there was a change in frequency (proportion) of a bigram, and say nothing about the context in which bigrams occurred. For example, mentions of ‘white nationalist’ may have appeared in a negative context in the progressive group, and a positive context in the alternative group, but further analyses will be needed to make such claims. A further noteworthy solution to this problem is the use of word embeddings, an approach used to learn vector representations for individual words that aim to capture semantic relationships between words based on the contexts in which they appear. This approach has already been used within the context of the Charlottesville rally, showing that US media associated African-Americans (e.g., the term ‘black’) less with negative character traits (e.g., ‘silly’, ‘extreme’) after the rally (Leschke & Schwemmer, 2019).

It can be argued that understanding of changes in language use of potentially violent groups on social media may be of particular interest to policy makers and security officials aiming to prevent or de-escalate violence. Future research may focus on extending the present approach to measuring changes in language over time on other

[22] https://support.google.com/youtube/answer/6373554?hl=en

social media platforms where alt-right supporters are active, such as 8Kun and Gab. It may also be of interest to measure concepts other than topics and *n*-gram frequencies, such as hate speech and abusive language, in response to the Charlottesville rally and perhaps other events of interest. Although it is beyond the scope of the current chapter, a follow-up study of the specific contexts in which certain topics and *n*-grams occur may be interesting. For example, is the sentiment regarding 'white people' or 'feminism' negative or positive in polarity?

## 5.10 Conclusion

Following the violent rally in Charlottesville, the alt-right received significant attention in the media and public discourse. As a result, we expected to see differences in social identity performance and visibility of the alt-right movement, which was measured through examining trajectories of language use. Contrasting a unique dataset of YouTube video transcripts from alternative, right-leaning channels to progressive, left-leaning channels, the present investigation indeed observed differences in language within and between the alternative and progressive groups. Results potentially reflect changes in social identity performance and visibility after the rally, as well as differences between the two groups more generally. This chapter demonstrates how linguistic trajectories and associated responses to external events can be measured on a large scale, thereby contributing to our understanding of grievance-fuelled movements through language.

# Chapter 6: General discussion and conclusion

The challenge associated with the vast amount of threatening and violent extremist messages found on the internet is not knowing which message has the potential to result in an actual act of violence. Making sense of large amounts of text data is a problem that continues to face security practitioners. This thesis sought to contribute to this ongoing effort, by exploring an automated approach to threat assessment using computational linguistics.

## 6.1 Main findings

The aim of this thesis was to discover whether and how threat assessment can be automated using methods from computational linguistics. Together, the chapters in this thesis illustrate that it is indeed possible to gain further insight into grievance-fuelled violence through computational linguistics. Through qualitative and quantitative analyses, we have also demonstrated which areas of language can be leveraged in order to automate threat assessment.

Chapter 1 highlighted the various computational linguistics methods that have already been applied to grievance-fuelled targeted violence. Approaches could be broadly categorised into using top-down or bottom-up approaches, including tools and methods such as psycholinguistic dictionaries, sentiment analyses, bag-of-word models, and word embeddings. Although a growing body of research has demonstrated which technical capabilities are suited to studying grievance-fuelled violence, the extant literature has not revealed how these methods would fit into the work of threat assessment practitioners. Therefore, the subsequent chapter examined the role of language in threat assessment procedures in practice.

Chapter 2 presented a qualitative analysis of the use of linguistic information by expert practitioners in approaching anonymous threat assessment. The domain of anonymous threatening communications is particularly suited to studying linguistic factors in threat assessment due to the fact that language (in the threat) is often the only evidence available. The results indicated that threat assessment practitioners broadly consider linguistic content, style, and trajectories. Overall, expert judgments of a specific case were inconsistent, a problem which may be partially remedied by increased automation. That is, automation may result in increased reliability and consistency between judgments, since the same indicators are considered across cases. The subsequent chapters each examined a domain of language raised in Chapter 2, namely content, style, and trajectories.

Chapter 3 examined *linguistic content* by translating cues used in traditional threat assessment to a psycholinguistic dictionary for automated analysis. The Grievance Dictionary was developed through consultation with expert threat assessors, as well as human and computational wordlist generation and annotation. The dictionary was subsequently validated by assessing statistical differences on dictionary categories between violent and non-violent text samples, which demonstrated strong differences on the majority of categories (e.g., violence, hate, weaponry). Further classification experiments on the same datasets demonstrated high prediction performance, likely due to the strong statistical differences that were elicited with the highly specific comparison of samples.

Chapter 4 supplemented content with measures of *linguistic style*, in order to examine the feasibility of author profiling for abusive and grievance-fuelled texts. First, statistical tests aimed at assessing the relationship between linguistic content and style with author personality, gender, and age. These analyses showed small, but statistically significant effects. Thereafter, prediction experiments were performed using various feature sets. Personality traits were predicted within approximately 19% of their actual value, whereas age was predicted with an error margin of +/- ten years. Gender classification achieved an average accuracy of 70%, which was only a negligible improvement over a model which always predicts the majority class. All in all, these results were poor when compared to previous work on author profiling.

Chapter 5 described how online *linguistic trajectories* in response to an offline event can be measured and applied within the context of grievance-fuelled violent texts. We assessed alt-right language use on YouTube surrounding the 2017 Charlottesville 'Unite the Right' rally, and compared this to progressive, left-leaning channels. Using structural topic models, qualitative and quantitative differences between channel samples were highlighted. By examining the trajectories of common bigram frequencies over time, this chapter showed that rally-related content increased for both groups at the time of the event. Interestingly, the alt-right also exhibited an increase of channel promotion which coincided with the rally.

### *6.1.1 Comparing findings*

On the whole, some implementations of computational linguistics were more successful than others. Automatically deducing content from large volumes of text using the Grievance Dictionary appears to hold promise for distinguishing between different populations of authors as well as measurement in large-scale datasets. In contrast, the implementation of linguistic style in addition to content did not achieve high prediction performance in author profiling of abusive texts, particularly when comparing results to previous work on neutral language. Therefore, at present we do not regard this method as fit for practice, because the error margins are too large. Additional research, for example using larger datasets, will be needed to further evaluate this research method, even if just to confirm it is not suitable for threat assessment in practice. Lastly, a dynamic approach to measuring linguistic concepts successfully shed light on social processes within an extremist movement, and thus seems to hold further promise for the implementation in large-scale threat assessment settings. In the next section, we discuss limitations that apply as a whole to the chapters in this thesis.

## 6.2 Limitations

One of the most challenging issues within the field of linguistic threat assessment is access to appropriate datasets. Targeted violence is a low base rate phenomenon (Corner et al., 2018) and the number of cases where the perpetrator produced linguistic material related to the incident will be even smaller. This means that there is little linguistic data authored by individuals who resorted to violence, compared to the vast amount of data by individuals who (to our knowledge) have not done so. Therefore, understanding grievance-fuelled linguistic data (e.g., finding indicators of possible violence) is more challenging than studying phenomena for which natural language data is more readily available (e.g., consumer reviews, news reports). Due to this data scarcity, researchers within the field of grievance-fuelled targeted violence often sample on the dependent variable (Clemmow et al., 2020). That is, text data is selected from sources about whom it is known they committed violence. This procedure has been followed in several studies described in this work (Baele, 2017; Kaati, Shrestha, & Cohen, 2016; Kaati, Shrestha, & Sardella, 2016; Kop et al., 2019; Neuman et al., 2015) as well as in Chapter 3 of this thesis. One specific dataset used in Chapter 3 is the only sample for which we know the authors engaged in actual (extremist) violence, and consists of lone-actor terrorist manifestos with a total sample size of 22. To discover meaningful differences between violence actualisers and non-actualisers, similar data (from the same context, e.g., all flagged as worrying) from non-actualisers would offer the most ideal comparison data. However, such comparisons are rarely performed due to data scarcity. Instead, studies more frequently compare texts from highly violent (terrorist) individuals to a large sample of linguistic data which are not violent in any way, such as neutral texts from (non-extremist) blogs, forums, and social media platforms (Baele, 2017; Kaati, Shrestha, & Cohen, 2016; Neuman et al., 2015). An extreme example of this suboptimal comparison includes previous work where “incel” forum posts were compared to Wikipedia pages (Jaki et al., 2019). This procedure poses problems for statistical comparisons due to unequal sample sizes, in addition to possible confounds (e.g., document format, specialised word use) that may explain linguistic differences.

Indeed, with such research designs it is difficult to disentangle whether statistical differences between violent and non-violent samples emerge based on indicators for violence and non-violence, or due to differences in topic or text type. It is arguably not difficult for the human eye *or* computer software to distinguish between a violent manifesto about attack planning and a blogpost about someone’s hobby. As a consequence, these are not ‘fair’ comparisons and thus the large statistical differences or high classification performances are not surprising. These suboptimal study procedures highlight the importance of performing linguistic comparisons between violent texts written by individuals who enact violent deeds, and violent texts written by individuals not planning to act violently. When comparing two sources of equally violent texts, it becomes easier to test whether the independent variable (enacted violence or not) is indeed the main explanation for linguistic differences between samples. Although such data may not become readily available, it is of the utmost importance to move towards such comparisons.

A second limitation that relates to the low prevalence of targeted violence is the base rate fallacy. A common misperception about machine learning is that high accuracy rates will equate to ‘perfect’ prediction of (non-)violence. As a result, practitioners may expect that human judgment will be rendered obsolete once highly accurate systems are developed. However, even with a highly accurate system, the rate of possible false positives is still alarmingly high. Due to the low base rate of individuals who actually resort to an act of violence (i.e., there is a strong imbalance towards non-violent individuals), a large number of individuals will be incorrectly classified as being violent. Imagine a situation where practitioners have a sample of documents written by 100 million different individuals that need to be classified as violence-actualisers or non-actualisers. We assume that that the base rate of actual violence is 1%, and our system is 95% accurate – an overoptimistic accuracy given the current state of research. An accuracy rate of 95% means that the system can correctly identify both violence-actualisers (i.e., sensitivity) and violence non-actualisers (i.e., specificity) 95% of the time. Within the hypothetical sample of 100 million documents (each written by a different author), 1 million documents will actually derive from violence-actualisers, of which 950,000 will be correctly classified (95%). Within the documents from non-actualisers, this system will *incorrectly classify 4,95 million individuals* as actualisers. This means that of all the documents classified as deriving from violence-actualisers, only 16.10% (950,000 of 5,900,000) are correctly classified as being truly written by a violence-actualiser (i.e., precision). See Table 6.1 for a breakdown of this calculation (adapted from Kleinberg et al., 2018; van der Vegt et al., 2019).

It is crucial for researchers and practitioners to recognise that even if measures such as those proposed in this thesis could be further developed to achieve high accuracy rates in violence prediction, the base rate fallacy will persist. A possible solution to this issue is that prediction systems merely serve as a filter system to reduce (rather than remove) necessary human review of documents, particularly if several cascading filters are applied (Kleinberg, van der Toolen, et al., 2018). However, in this hypothetical situation, it is necessary that the indicators used in each level of the filter system are independent of each other, in order to reduce the amount of data at each level (Kleinberg, van der Toolen, et al., 2018).

Table 6.1 Demonstration of base rate fallacy

| | | **Prediction** | | |
|---|---|---|---|---|
| | | Violence | No violence | Total |
| **Reality** | Violence | 950,000 | 50,000 | 1,000,000 |
| | No violence | 4,950,000 | 94,050,000 | 99,000,000 |
| | Total | 5,900,000 | 94,100,000 | 100,000,000 |

Another limitation in this thesis is the reliance on threat assessment experts. A large part of this work derived from linguistic indicators that were suggested by expert practitioners. This included the focus on content, style, and trajectories, as well as categories in the Grievance Dictionary. The rationale for this approach stems from the aim of optimising threat assessment practice through computational techniques, such that the same constructs used in practice could be measured efficiently and at scale. While interviews and surveys conducted with domain experts served as a good starting point, the possibility exists that experts were wrong about the relevance of indicators. Therefore, we acknowledge that the suggested linguistic variables derived from expert consultation may need to be supplemented with variables derived from a bottom-up approach (e.g., where an algorithm learns the language use associated with violence). Furthermore, future applications of the indicators proposed in this thesis may also show that some are not valuable in distinguishing between populations from other contexts not studied in this work (e.g., far-left extremism or misogynist extremism). Then, the indicators suggested by experts will need to be re-evaluated or supplemented.

The final limitation relates to the fact that we know little about whether the indicators measured with computational methods in this thesis translate to the equivalent measure when they are assessed through (expert) human judgment. For example, a text which scores high on violence according to the Grievance Dictionary, may not be judged as violent by a practitioner. Future research is needed to assess the agreement between human and computational judgments of linguistic indicators (such as those in the Grievance Dictionary), possible author characteristics (personality, age, gender), and trajectories (breakpoints or escalation) of language use. Within the context of linguistic threat assessment, it is also unknown whether and how expert and layperson judgments differ. Lastly, it is also unclear whether the linguistic indicators measured in this thesis indeed translate to real-life behaviour or processes. This particularly holds for psychological processes measured through LIWC or Grievance Dictionary categories, such as hate, jealousy, and paranoia. That is, did the authors of texts studied in this thesis actually experience the psychological processes that the linguistic measures indicated as present?

### 6.3 Contributions to the literature

The contributions of this thesis to the literature are threefold. First, this work describes a pathway for translating procedures in practice (threat assessment) to automatic linguistic systems. We view this as an academic endeavour, in that the validity of linguistic tools should be accompanied by thorough testing of theories and assumptions, for example through statistical inference. By departing from qualitative examinations of threat assessment procedures, we were able to identify computational linguistic methods aimed at supplementing or improving manual procedures. This same procedure could for example be implemented in other crime and security problems besides targeted violence, where expert analysts are interviewed in order to define and subsequently test the relevant computational methods. Of course, the possibility exists that linguistic measures considered in other domains also broadly fall into linguistic content and style categories, as these measures are widely used in computational linguistics in general. However, this thesis further contributes to the literature by highlighting the importance of considering linguistic trajectories and the effect of external events on this measure. Previous research has shown that content and style measures are typically measured in a static manner (i.e., one score per document or an average score per sample), but with this work we hope to have demonstrated the importance of measuring language over time, either throughout or across several texts.

Second, the methods and tools presented in this thesis may possibly contribute to further theory testing and formation within the linguistic study of grievance-fuelled targeted violence. Due to the emerging nature of this field, there are little to no widely-acknowledged theories that relate specifically to the language of violence-actualisers. The Grievance Dictionary presents an opportunity for researchers across different fields to measure violence and grievances in the same way. For the purpose of theory testing, the possible consensus that may emerge from this is perhaps favourable over both the different custom dictionaries that have been developed in the past (Kleinberg, van der Vegt, & Gill, 2020; Smith et al., 2020) and the widespread use of the LIWC (Pennebaker et al., 2015) which is not specific to violent behaviour. Alternative routes include further application of general social psychological theories to grievance-fuelled language, as was the case in Chapter 5 of this thesis, which applied social identity theory to the language of an extremist group. The proposed content and trajectory measures of grievance-fuelled language may also be used to apply theories used in the study of (deceptive) language, such as Reality Monitoring (Johnson et al., 1988; Johnson & Raye, 1981; Vrij, 2015) and Construal Level Theory (Trope & Liberman, 2010), since threatened violence can be considered a form of deceptive intent in some cases (Geurts et al., 2017).

Third, this thesis has emphasised open science practices and transparency within the violence research domain. Each chapter included a transparency statement, detailing whether and how data, code, and materials could be accessed. By relying on statistical analyses, this thesis also fits within the growing trend of the empirical study within terrorism research (Schuurman, 2020). However, some open science practices are still uncommon within this field, even though most terrorism researchers regard the movement as favourable (Schumann et al., 2019). This thesis has demonstrated that data sharing within a sensitive domain is possible. Terrorism researchers often cite the sensitivity of the data (e.g., protection of victims or offender privacy) as an argument against data sharing (Schumann et al., 2019). However, linguistic data is particularly suited for sharing publicly or between researchers, for instance if only the linguistic features (quantitative representations) of a dataset are shared (e.g., dictionary scores, part-of-speech tags, word embeddings). In addition, this thesis has endeavoured to highlight the often opaque research methods in linguistic studies of grievance-fuelled violence and ‘under the hood’ operations in tools developed for automated linguistic study. For instance, the exact content of several psycholinguistic dictionaries remains unknown to this day (e.g., the LIWC2015: Pennebaker et al., 2015; dictionaries used in Smith et al., 2020), as well as the precise measures underlying the PRAT (Akrami et al., 2018) and Threat Triage (Smith et al., 2013). While this is perhaps understandable for commercial reasons, it raises the question whether researchers and practitioners should be using tools that produce outcomes that cannot be fully explained in the first place. By raising awareness of this issue and by demonstrating alternative, transparent approaches to developing automation initiatives, we hope to encourage other researchers to choose the same path.

### 6.4 Implications for practice

This section raises three points that hold relevance for threat assessment practice. First, it is important to consider whether statistical significance also translates to practical significance before any automated linguistic methods are implemented in practice. Within the linguistic study of targeted violence and extremism, new linguistic indicators and relationships between language and behaviour will probably continue to emerge. It is possible that these results will follow a similar pattern as those found in this thesis. That is, we may tend to find very large statistical effects when sampling on the dependent variable (Chapter 3) and very small effects when trying to detect “real world” variables (author demographics, external events) using linguistic variables (Chapter 4 and 5). Moreover, it is highly likely that when comparisons between two types of violent text samples (differing only on actualisation) are performed, these effects will be similarly small. These small effects will not always translate into features that perform well at prediction, as was demonstrated in Chapter 4 of this thesis. Therefore, it is very important to consider the difference between statistical significance and practical significance. Small effects, even when statistically significant, do not always result in the high prediction performance that may be desired by practitioners. Therefore, threat assessment practitioners should take newly identified relationships between violence and language use with a grain of salt before the same variables have been adequately tested as predictive features in the relevant context.

Second, the increasing focus on prediction within the context of targeted violence is another matter that warrants (re)consideration. While technical advances have enabled us to predict several phenomena with increasing accuracy, practitioners should be aware that it remains to be seen if this will ever be the case for targeted 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 several interacting variables, for which it is questionable if a dataset of the same magnitude and depth will ever be collected. This 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). Real life behaviour may simply be too complex to accurately predict or 'may be subject to a predictability ceiling' (Garip, 2020, p. 8235). Consequently, our understanding of behaviour and life outcomes should perhaps be defined by the extent to which causal relationships or statistical differences between samples have been identified (Salganik et al., 2020). Although others within psychology have called for more predictive modelling within the field (Yarkoni & Westfall, 2017), both camps agree that a clear distinction needs to be made between explanatory and predictive modelling. Another relevant factor raised by proponents of predictive modelling is that the best performing predictive models may not be as comprehensible and theoretically elegant as explanatory models (Yarkoni & Westfall, 2017). This is particularly relevant to automated threat assessment, a field in which practitioners are often required to operate under regulations that stress the importance of explainable decision-making (Goodman & Flaxman, 2017; Oswald et al., 2018).

In the case of grievance-fuelled communication, practitioners may indeed 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, they may also be able to develop novel strategies to adequately manage such threats. However, this does not mean that the linguistic variables identified in research can be used to adequately predict whether real-life threats will lead to violence. It is perhaps impossible to be aware of or control for all possible factors that may impact on the outcome of a potentially violent situation. Therefore, it may be 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 for large scale problems. However, one must be aware of the possibility that accurate prediction (particularly to the extent that may be desired in practice) may be an unachievable goal.

Bearing this issue in mind, the third point that warrants consideration by practitioners is that measurement may become increasingly important. The primary potential for tools such as the Grievance Dictionary and assessments of language over time thus may not lie within prediction, but in measurement. In this way, such tools are more closely aligned to the risk assessment principles of Structured Professional Judgement (SPJ) than to actuarial approaches. The former 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 et al., 2016). Examples of an actuarial approach include tools such as the PRAT (Akrami et al., 2018) and Threat Triage (Smith et al., 2013). In contrast, an approach that is more in line with SPJ could help support practitioners to review all available written content in automated form and identify from within those data evidence for a range of features deemed relevant to the outcome to be prevented (e.g., a practitioner seeking to find those documents in a vast corpus that score high on weaponry), while bearing in mind the base rate fallacy. The methods proposed in this thesis thus are intended as a decision-making tool towards risk management in the individual case, the output of which consists of a list of features that leaves room for the interpretation of front-line practitioners.

## 6.5 Outlook

Considering the findings and limitations raised in this thesis, a research agenda for the field of automated linguistic threat assessment can be formulated (summarised in Table 6.2). First and foremost, efforts should be made to make meaningful comparisons of language use between datasets. The data scarcity problem can perhaps be somewhat ameliorated by increased collaboration between law enforcement and researchers. It is likely that police or threat assessment practitioner databases contain a multitude of cases (communications) which did not lead to violence, but were initially seen as worrying. If data from violence actualisers and non-actualisers is more widely available, it will for example be of great interest to assess whether and how differences in Grievance Dictionary categories emerge, as well as how classification tasks perform. Ideally, it will be similarly useful to model linguistic trajectories in samples of actualisers and non-actualisers to see, for example, whether one group exhibits more pronounced increases of extremist language over time than the other, or whether groups respond to external events differently. For lack of such gold standard data, comparisons should at least be made between known violent extremists and comparably extremist individuals. An alternative way to remedy the data problem is by instructed writing, where study participants are asked to produce threatening or violent texts. Chapter 4 of this thesis presented an example of this. Of course, the challenge remains to design studies in such a way that results are externally valid, in that participants genuinely support the threat they produce. In Chapter 4, this was attempted by having participants write abusively about a politician they actually dislike. An alternative procedure was followed in one of the few experimental studies on verbal threats, where participants were instructed to imagine a hypothetical situation, utter a threat, and subsequently instructed to (not) actualise their threat (Geurts et al., 2016). However, researchers need to remain mindful of the possibility that some participants would not normally be inclined to express themselves in an abusive or threatening manner. This problem could be partially circumvented in future experimental research by, for example, approaching possible participants who are known (or self-report) to have posted abusively online.

Second, increased unsupervised (bottom-up) computational linguistics research may lead to the identification of additional relevant linguistic indicators that did not emerge in this thesis. Chapter 5 presented an approach in which linguistic features were derived in an unsupervised manner, showing that an understanding of (potentially) violent language can also be achieved through this alternative method. Through methods such as structural topic modelling and *n*-gram frequencies, we may be able to identify further discriminant variables in specific contexts, such as previously unstudied extremist or aggrieved groups. Alternatively, through unsupervised methods we may identify linguistic patterns that unite several grievance-fuelled violent populations, or improve our ability to measure latent psychological concepts through text.

Third, it is important to test the validity of linguistic measures. This holds for the validity of computational linguistic measures as a substitute for human (expert) judgments, as well as the validity of the linguistic concepts in measuring psychological processes. The former may be achieved by increased research on the agreement between human (expert or layperson) and computational judgments of risk through language. That is, do humans judge a text as equally (or comparably) violent as a computational system does? How accurate are humans when compared to a computational system? The validity of linguistic measures of psychological states, on the other hand, needs to be assessed by increasingly obtaining ground truth data. That is, psychological states (e.g., positive emotion, anger, paranoia) need to be measured through self-report or other means in order to verify that the linguistic measurement adequately reflects these states (Kleinberg, van der Vegt, & Mozes, 2020). This holds for the study of grievance-fuelled violence as well as the field of computational linguistics in general. An alternative way to test this form of validity is by manipulating or eliciting psychological states, then assessing the construct of interest through language (Kleinberg, 2020; Marcusson-Clavertz et al., 2019). In the context of grievance-fuelled language, one could for example build on the efforts in Chapter 4 of this thesis. A possible study may similarly instruct participants to write about a politician they dislike, but also attempt to elicit anger or hate by additionally showing participants videos, news reports, or images of the politician in question.

While embarking on the research avenues raised here, researchers and practitioners should remain mindful of the points raised in this discussion. It is important to continue to adhere to open science principles, particularly when methodologies and knowledge are generated that need to be accessible and explainable to practitioners. When relating linguistic measures to real life behaviour, human characteristics, or psychological states, the possibility exists that small effects will continue to be obtained. While such findings are still worthwhile to research and practice, they may not be fit for subsequent prediction. Moreover, prediction as a whole should not be the main objective of continued study into the linguistic factors of grievance-fuelled targeted violence.

Table 6.2 Future research agenda

| Research agenda item | Achieved through |
|---|---|
| 1. Conducting meaningful linguistic comparisons | - Data sharing law enforcement<br>- Violent (extremist) open data<br>- Experimental research |
| 2. Data-driven identification of linguistic violence indicators | - Unsupervised linguistic measurement of ground truth or expert-judged threat cases |
| 3. Testing validity of linguistic measures | - Measuring agreement between human and computational judgments<br>- Measuring/manipulating psychological state in addition to linguistic measure of state |

## 6.6 Conclusion

Computational linguistics holds promise for automating threat assessment and promoting our understanding of grievance-fuelled targeted violence. In order to contextualise the growing body of literature applying computational techniques to violent language, this thesis identified areas of language that are of particular use to threat assessment practitioners. Departing from this, we examined three implementations of computational linguistics within targeted violence addressing linguistic content, style, and trajectories, respectively. This thesis presented the Grievance Dictionary, a novel tool which showed success in understanding grievance-fuelled language at scale. Attempts in this thesis to predict author demographics from abusive language use require further research before possible implementation in practice. Lastly, this thesis demonstrated the utility of measuring language over time to assess the effects of external events on an extremist group.

The field of automated linguistic threat assessment will probably continue to develop in the years to come. For researchers and practitioners involved in these endeavours, it is important to critically evaluate the tools and effects that may emerge, particularly if they appear too good to be true. We also urge for continued collaboration between research and practice, ideally bearing open science principles in mind. Lastly, moving away from prediction and investing in measurement and comprehension may eventually be the most beneficial use of computational linguistics in threat assessment.

# References

Abbasi, A, & Chen, H. (2007). Affect Intensity Analysis of Dark Web Forums. *2007 IEEE Intelligence and Security Informatics*, 282–288. https://doi.org/10.1109/ISI.2007.379486

Ackland, R., & O'Neil, M. (2011). Online collective identity: The case of the environmental movement. *Social Networks*, *33*(3), 177–190. https://doi.org/10.1016/j.socnet.2011.03.001

Adams, J., & Roscigno, V. J. (2005). White Supremacists, Oppositional Culture and the World Wide Web. *Social Forces*, *84*(2), 759–778. https://doi.org/10.1353/sof.2006.0001

Akrami, N., Shrestha, A., Berggren, M., Kaati, L., Obaidi, M., & Cohen, K. (2018). *Assessment of risk in written communication: Introducing the Profile Risk Assessment Tool (PRAT)*. EUROPOL. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-367346

Alberici, A. I., & Milesi, P. (2016). Online discussion, politicized identity, and collective action. *Group Processes & Intergroup Relations*, *19*(1), 43–59. https://doi.org/10.1177/1368430215581430

Anti-Defamation League. (2017). *Murder and Extremism in the United States in 2017*. Anti-Defamation League. https://www.adl.org/resources/reports/murder-and-extremism-in-the-united-states-in-2017

Anti-Defamation League. (2018). *Murder and Extremism in the United States in 2018*. Anti-Defamation League. https://www.adl.org/murder-and-extremism-2018

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. *Proceedings of Joint Annual Meeting of the Interface and The Classification Society of North America*, *January*, 1–16. https://doi.org/10.2105/AJPH.50.1.21

Arora, S., Liang, Y., & Ma, T. (2016). *A Simple but Tough-to-Beat Baseline for Sentence Embeddings*. https://openreview.net/forum?id=SyK00v5xx

Ashton, M. C., & Lee, K. (2009). *The HEXACO–60: A Short Measure of the Major Dimensions of Personality: Journal of Personality Assessment: Vol 91, No 4*. https://www.tandfonline.com/doi/full/10.1080/00223890902935878

Atkinson, D. C. (2018). Charlottesville and the alt-right: A turning point? *Politics, Groups, and Identities*, *6*(2), 309–315. https://doi.org/10.1080/21565503.2018.1454330

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International*

*Conference on World Wide Web Companion*, 759–760. https://doi.org/10.1145/3041021.3054223

Baele, S. J. (2017). Lone-Actor Terrorists' Emotions and Cognition: An Evaluation Beyond Stereotypes. *Political Psychology*, *38*(3), 449–468. https://doi.org/10.1111/pops.12365

Baele, S. J., Brace, L., & Coan, T. G. (2019). From "Incel" to "Saint": Analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and Political Violence*, *0*(0), 1–25. https://doi.org/10.1080/09546553.2019.1638256

Benoit, K., & Matsuo, A. (2020). *spacyr: An R wrapper for spaCy*.

Benoit, K., Watanabe, K., Wang, H., Müller, S., Perry, P. O., Lauderdale, B., & Lowe, W. (2020). *quanteda.textmodels: Scaling Models and Classifiers for Textual Data* [R].

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Best, D., Bliuc, A.-M., Iqbal, M., Upton, K., & Hodgkins, S. (2018). Mapping social identity change in online networks of addiction recovery. *Addiction Research & Theory*, *26*(3), 163–173. https://doi.org/10.1080/16066359.2017.1347258

Bischof, J. M., & Airoldi, E. M. (n.d.). *Summarizing topical content with word frequency and exclusivity*. 8.

Bjørgo, T., & Silkoset, E. (2018). *Threats and threatening approaches to politicians*. 56.

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Technique...Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics*. https://projecteuclid.org/euclid.aoas/1183143727

Bliuc, A.-M., Best, D., Iqbal, M., & Upton, K. (2017). Building addiction recovery capital through online participation in a recovery community. *Social Science & Medicine*, *193*, 110–117. https://doi.org/10.1016/j.socscimed.2017.09.050

Bliuc, A.-M., Betts, J., Vergani, M., Iqbal, M., & Dunn, K. (2019). Collective identity changes in far-right online communities: The role of offline intergroup conflict. *New Media & Society*, *21*(8), 1770–1786. https://doi.org/10.1177/1461444819831779

Bond, G. D., Holman, R. D., Eggert, J. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., & Rustige, R. (2017). 'Lyin' Ted' , 'Crooked Hillary' , and 'Deceptive Donald': Language of Lies in the 2016 US Presidential Debates. *Applied Cognitive Psychology*, *677*(October), 668–677. https://doi.org/10.1002/acp.3376

Borum, R., Fein, R., Vossekuil, B., & Berglund, J. (1999). Threat assessment: Defining an approach to assessing risk for targeted violence. *Behav. Sci. Law*, 16.

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), eaba2196. https://doi.org/10.1126/sciadv.aba2196

Boyd, R. L., & Schwartz, H. A. (2020). Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field. *Journal of Language and Social Psychology*, 0261927X20967028. https://doi.org/10.1177/0261927X20967028

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. *EMNLP*, 1301–1309. https://doi.org/10.1016/j.polymdegradstab.2015.06.018

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., & Voss, A. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 206. https://doi.org/10.1007/s13278-014-0206-4

Caiani, M., della Porta, D., & Wagemann, C. (2012). The Action Repertoires of the Radical Right: Violence and Beyond. In *Mobilizing on the Extreme Right*. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199641260.001.0001/acprof-9780199641260-chapter-5

Calhoun, F. S., & Weston, S. W. (2017). *Threat Assessment and Management Strategies: Identifying the Howlers and Hunters, Second Edition*. CRC Press.

Celli, F., Pianesi, F., Stillwell, D., & Kosinski, M. (2013). *Workshop on Computational Personality Recognition: Shared Task*. 4.

Chen, H. (2008). Sentiment and affect analysis of Dark Web forums: Measuring radicalization on the internet. *IEEE International Conference on Intelligence and Security Informatics, 2008, IEEE ISI 2008*, 104–109. https://doi.org/10.1109/ISI.2008.4565038

Chouldechova, A., & Hastie, T. (2015). Generalized Additive Model Selection. *ArXiv:1506.03850 [Stat]*. http://arxiv.org/abs/1506.03850

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, *28*(3), 591–605. JSTOR. https://doi.org/10.2307/1910133

Clark, E., & Araki, K. (2011). Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. *Procedia - Social and Behavioral Sciences*, *27*, 2–11. https://doi.org/10.1016/J.SBSPRO.2011.10.577

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. https://doi.org/10.1111/1556-4029.14282

Clifford, B., & Powell, H. (2019). Encrypted Extremism: Inside the English-Speaking Islamic State Ecosystem on Telegram. *Program on Extremism: George Washington University*.

Cobain, I., & Taylor, M. (2016, November 23). Far-right terrorist Thomas Mair jailed for life for Jo Cox murder. *The Guardian*. https://www.theguardian.com/uk-news/2016/nov/23/thomas-mair-found-guilty-of-jo-cox-murder

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Conway-Silva, B. A., Filer, C. R., Kenski, K., & Tsetsi, E. (2018). Reassessing Twitter's Agenda-Building Power: An Analysis of Intermedia Agenda-Setting Effects During the 2016 Presidential Primary Season. *Social Science Computer Review*, *36*(4), 469–483. https://doi.org/10.1177/0894439317715430

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. https://doi.org/10.1080/00223891.2018.1475392

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. https://doi.org/10.1561/1500000001

Dazzi, C., & Pedrabissi, L. (2009). Graphology and Personality: An Empirical Study on Validity of Handwriting Analysis. *Psychological Reports*, *105*(3_suppl), 1255–1268. https://doi.org/10.2466/PR0.105.F.1255-1268

de Becker, G. (1998). *The Gift of Fear: And Other Survival Signals That Protect Us from Violence*. Bloomsbury Trade. https://www.amazon.com/Gift-Fear-Survival-Signals-Violence/dp/0440226198

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018, September 12). Hate Speech Dataset from a White Supremacy Forum. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*.

della Porta, D. (2018). Protests as critical junctures: Some reflections towards a momentous approach to social movements. *Social Movement Studies*, 1–20. https://doi.org/10.1080/14742837.2018.1555458

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. *ArXiv:2005.00547 [Cs]*. http://arxiv.org/abs/2005.00547

Der Spiegel. (2019). *Father, Neighbor, Killer: Germany's New Far-Right Terror*. https://www.spiegel.de/international/germany/father-neighbor-killer-germany-s-new-far-right-terror-a-1273689.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Diani, M. (1992). The Concept of Social Movement. *The Sociological Review*, *40*(1), 1–25. https://doi.org/10.1111/j.1467-954X.1992.tb02943.x

Diani, M. (2003). Networks and social movements: A research programme. In *Social Movements and Networks: Relational Approaches to Collective Action*. Oxford University Press.

Dietz, P. E., Matthews, D. B., Martell, D. A., Stewart, T. M., Hrouda, D. R., & Warren, J. (1991). Threatening and Otherwise Inappropriate Letters to Members of the United States Congress. *Journal of Forensic Sciences*, *36*(5), 13165J. https://doi.org/10.1520/JFS13165J

Dietz, P., Matthews, D. B., Duyne, C. V., Martell, D. A., Parry, C. D. H., Tracy, S., Warren, J., Crowder, J. D., & Duyne, V. (1991). Threatening and Otherwise Inappropriate Letters to Hollywood Celebrities. *Journal of Forensic Sciences*, *36*(1), 185–209.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate Speech Detection with Comment Embeddings. *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, 29–30. https://doi.org/10.1145/2740908.2742760

Duwairi, R. M., & Qarqaz, I. (2014). Arabic Sentiment Analysis Using Supervised Classification. *2014 International Conference on Future Internet of Things and Cloud*, 579–583. https://doi.org/10.1109/FiCloud.2014.100

Egnoto, M. J., & Griffin, D. J. (2016). Analyzing Language in Suicide Notes and Legacy Tokens: Investigating Clues to Harm of Self and Harm to Others in Writing. *Crisis*, *37*(2), 140–147. https://doi.org/10.1027/0227-5910/a000363

Ellis, E. G. (2018, September 19). The Alt-Right Are Savvy Internet Users. Stop Letting Them Surprise You. *Wired*. https://www.wired.com/story/alt-right-youtube-savvy-data-and-society/

Every-Palmer, S., Barry-Walsh, J., & Pathé, M. (2015). Harassment, stalking, threats and attacks targeting New Zealand politicians: A mental health issue. *Australian & New Zealand Journal of Psychiatry*, *49*(7), 634–641. https://doi.org/10.1177/0004867415583700

Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M., & Davalos, S. (2014). A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14*, 1–6. https://doi.org/10.1145/2659522.2659526

Farrell, T., Araque, O., Fernandez, M., & Alani, H. (2020). On the use of Jargon and Word Embeddings to Explore Subculture within the Reddit's Manosphere. *WebSci'20*. 12th ACM Web Science Conference 2020, Southampton, UK. https://websci20.webscience.org/

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657. https://doi.org/10.1145/2858036.2858535

Feinerer, I., & Hornik, K. (2018). *tm: Text Mining Package* (R package version 0.7-6) [Computer software]. https://CRAN.R-project.org/package=tm

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fernandez, M., & Alani, H. (2018). *Contextual Semantics for Radicalisation Detection on Twitter*. CEUR.

Figea, L., Kaati, L., & Scrivens, R. (2016). Measuring online affects in a white supremacy forum. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, 85–90. https://doi.org/10.1109/ISI.2016.7745448

Frimer, J., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). *Moral Foundations Dictionaries for Linguistic Analyses, 2.0 (MFD 2.0)*.

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

Geurts, R., Göteborgs universitet, & Psykologiska institutionen. (2017). *Interviewing to assess and manage threats of violence*. Department of Psychology, University of Gothenburg.

Geurts, R., Granhag, P. A., Ask, K., & Vrij, A. (2016). Taking Threats to the Lab: Introducing an Experimental Paradigm for Studying Verbal Threats. *Journal of Threat Assessment and Management*, *3*(1), 53–64. https://doi.org/10.1037/tam0000060

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes. *Criminology and Public Policy*, *16*(1), 99–117. https://doi.org/10.1111/1745-9133.12249

Gill, P., Horgan, J., & Deckert, P. (2014). Bombing Alone: Tracing the Motivations and Antecedent Behaviors of Lone-Actor Terrorists. *Journal of Forensic Sciences*, *59*(2), 425–435. https://doi.org/10.1111/1556-4029.12312

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*, 253. https://doi.org/10.1145/1979742.1979614

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric Analysis of Bloggers' Age and Gender. *Third International AAAI Conference on Weblogs and Social Media*. Third International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/208

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Han, B., & Baldwin, T. (2007). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. *Proceedings of the 49th Annual Meeting of the Association for*

*Computational Linguistics: Human Language Technologies - Volume 1*, 368–378.

Hanes, E., & Machin, S. (2014). Hate Crime in the Wake of Terror Attacks: Evidence From 7/7 and 9/11. *Journal of Contemporary Criminal Justice*, *30*(3), 247–267. https://doi.org/10.1177/1043986214536665

Hart, S. D., Douglas, K. S., & Guy, L. S. (2016). The Structured Professional Judgement Approach to Violence Risk Assessment. In *The Wiley handbook on the theories, assessment and treatment of sexual offending* (pp. 643–666). https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118574003.wattso030

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories *versus* dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903–920. https://doi.org/10.1017/S0033291711001966

Hawley, G. (2017). *Making Sense of the Alt-Right*. Columbia University Press.

Hawley, G. (2018). *The Alt-Right: What Everyone Needs to Know®*. Oxford University Press.

Hine, G., Onaolapo, J., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., & Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 10.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, *43*(3), 524–527. https://doi.org/10.1016/j.jrp.2009.01.006

Hodge, E., & Hallgrimsdottir, H. (2019). Networks of Hate: The Alt-right, "Troll Culture", and the Cultural Geography of Social Movement Spaces Online. *Journal of Borderlands Studies*, 1–18. https://doi.org/10.1080/08865655.2019.1571935

Hoffman, B., Ware, J., & Shapiro, E. (2020). Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism*, *43*(7), 565–587. https://doi.org/10.1080/1057610X.2020.1751459

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*(3), 111–117. https://doi.org/10.1093/llc/13.3.111

Hughes, T. (2018). *A year after Charlottesville rally, white nationalists enter mainstream conversation.* Usatoday. https://www.theatlantic.com/ideas/archive/2018/08/the-battle-that-erupted-in-charlottesville-is-far-from-over/567167/

Hunt, S. A., & Benford, R. D. (1994). Identity talk in the peace and justice movement. *Journal of Contemporary Ethnography*, *22*(4), 488–517. https://doi.org/10.1177/089124194022004004

Irshaid, F. (2015, December 2). Isis, Isil, IS or Daesh? One group, many names. *BBC News*. https://www.bbc.com/news/world-middle-east-27994277

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., Mullen, P. E., Meloy, J. R., Pathé, M. T., Farnham, F. R., Preston, L., & Darnley, B. (2007). The role of mental disorder in attacks on European politicians 1990–2004. *Acta Psychiatrica Scandinavica*, *116*(5), 334–344. https://doi.org/10.1111/j.1600-0447.2007.01077.x

James, David V, MacKenzie, R., & Farnham, F. R. (2014). *Communications Threat Assessment Protocol*. Theseus LLP.

James, David V, Sukhwal, S., Farnham, F. R., Evans, J., Barrie, C., Taylor, A., & Wilson, S. P. (2016). Harassment and stalking of Members of the United Kingdom Parliament: Associations and consequences. *The Journal of Forensic Psychiatry & Psychology*, *27*(3), 309–330. https://doi.org/10.1080/14789949.2015.1124909

James, M. (2018). *Parkland's Nikolas Cruz made chilling videos before school shooting*. USA Today. https://eu.usatoday.com/story/news/2018/05/30/parkland-killer-video-im-going-next-school-shooter/657774002/

Jockers, M. (2015a). *» Revealing Sentiment and Plot Arcs with the Syuzhet Package Matthew L. Jockers*. http://www.matthewjockers.net/2015/02/02/syuzhet/

Jockers, M. (2015b). *Syuzhet: Extract Sentiment and Plot Arcs from Text.* https://github.com/mjockers/syuzhet.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, *117*(4), 371–376. https://doi.org/10.1037/0096-3445.117.4.371

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A Brief Measure of Dark Personality Traits. *Assessment*, *21*(1), 28–41. https://doi.org/10.1177/1073191113514105

Kaakinen, M., Oksanen, A., & Räsänen, P. (2018). Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, *78*, 90–97. https://doi.org/10.1016/j.chb.2017.09.022

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. https://doi.org/10.1109/ICDMW.2016.116

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, *4*, 100057. https://doi.org/10.1016/j.yjbinx.2019.100057

King, R. D., & Sutton, G. M. (2013). High Times for Hate Crimes: Explaining the Temporal Clustering of Hate-Motivated Offending. *Criminology*, *51*(4), 871–894. https://doi.org/10.1111/1745-9125.12022

King, R. N., & Koehler, D. J. (2000). Illusory correlations in graphological inference. *Journal of Experimental Psychology: Applied*, *6*(4), 336–348. https://doi.org/10.1037/1076-898X.6.4.336

Klein, A. (2019). From Twitter to Charlottesville: Analyzing the Fighting Words Between the Alt-Right and Antifa. *International Journal of Communication*, *13*(0), 22.

Klein, O., Spears, R., & Reicher, S. (2007). Social Identity Performance: Extending the Strategic Side of SIDE. *Personality and Social Psychology Review*, *11*(1), 28–45. https://doi.org/10.1177/1088868306294588

Kleinberg, B. (2020). Manipulating emotions for ground truth emotion analysis. *ArXiv:2006.08952 [Cs]*. http://arxiv.org/abs/2006.08952

Kleinberg, B., Mozes, M., & van der Vegt, I. (2018). Identifying the sentiment syles of YouTube's vloggers. *Proceedings of the 2018 Conference on Empirical Methods of Natural Language Processing*.

Kleinberg, B., van der Toolen, Y., Arntz, A., & Verschuere, B. (2018). Detecting Concealed Information on a Large Scale. In *Detecting Concealed Information and Deception* (pp. 377–403). Elsevier. https://doi.org/10.1016/B978-0-12-812729-2.00016-1

Kleinberg, B., van der Vegt, I., & Gill, P. (2020). The temporal evolution of a far-right forum. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-020-00064-x

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring Emotions in the COVID-19 Real World Worry Dataset. *ArXiv:2004.04225 [Cs]*. http://arxiv.org/abs/2004.04225

Kleinberg, B., Vegt, I. van der, Arntz, A., & Verschuere, B. (2019). *Detecting deceptive communication through linguistic concreteness*. https://doi.org/10.31234/osf.io/p3qjh

Koopmans, R., & Muis, J. (2009). The rise of right-wing populist Pim Fortuyn in the Netherlands: A discursive opportunity approach. *European Journal of Political Research*, *48*(5), 642–664. https://doi.org/10.1111/j.1475-6765.2009.00846.x

Koopmans, R., & Olzak, S. (2004). Discursive Opportunities and the Evolution of Right-Wing Violence in Germany. *American Journal of Sociology*, *110*(1), 198–230. https://doi.org/10.1086/386271

Kop, M., Read, P., & Walker, B. R. (2019). Pseudocommando mass murderers: A big five personality profile using psycholinguistics. *Current Psychology*. https://doi.org/10.1007/s12144-019-00230-z

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M. (2010). *Variable Selection Using The caret Package*.

Lahiri, S. (2014). Complexity of Word Collocation Networks: A Preliminary Structural Analysis. *ArXiv:1310.5111 [Physics]*. http://arxiv.org/abs/1310.5111

Langman, L. (2005). From Virtual Public Spheres to Global Justice: A Critical Theory of Internetworked Social Movements*. *Sociological Theory*, *23*(1), 42–74. https://doi.org/10.1111/j.0735-2751.2005.00242.x

Leschke, J., & Schwemmer, C. (2019). *Media Bias Towards African-americans Before and After the Charlottesville Rally*. Proceedings Of The Weizenbaum Conference 2019 Challenges Of Digital Inequality. https://www.ssoar.info/ssoar/handle/document/62623

Lewis, H. (2019). *How Anti-feminism Is the Gateway to the Far Right—The Atlantic*. https://www.theatlantic.com/international/archive/2019/08/anti-feminism-gateway-far-right/595642/

Lewis, R. (2018). *Broadcasting the Reactionary Right on YouTube*. 61.

Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest* [R].

Ma, A. (2019, March 25). *New Zealand made it illegal for anyone to download or share the Christchurch shooter's manifesto*. Business Insider Nederland. https://www.businessinsider.nl/new-zealand-bans-christchurch-shooter-manifesto-livestream-2019-3/

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning Word Vectors for Sentiment Analysis*. 9.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press. https://cds.cern.ch/record/2135372

Marcusson-Clavertz, D., Kjell, O. N. E., Persson, S. D., & Cardeña, E. (2019). Online validation of combined mood induction procedures. *PLOS ONE*, *14*(6), e0217848. https://doi.org/10.1371/journal.pone.0217848

Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J. L., & Stringhini, G. (2018). "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *ArXiv:1805.08168 [Cs]*. http://arxiv.org/abs/1805.08168

Masuzzo, P., & Martens, L. (2017). *Do you speak open science? Resources and tips to learn the language* (e2689v1). PeerJ Inc. https://doi.org/10.7287/peerj.preprints.2689v1

Mayer, H. (2018). *The alt-right manipulates free-speech rights. We should defend those rights anyway. - The Washington Post*. https://www.washingtonpost.com/news/made-by-history/wp/2018/08/21/the-alt-right-manipulates-free-speech-rights-we-should-defend-those-rights-anyway/

McCausland, P. (2017). *White Nationalist Leads Torch-Bearing Protesters Against Removal of Confederate Statue*. NBC News. https://www.nbcnews.com/news/us-news/white-nationalist-leads-torch-bearing-protesters-against-removal-confederate-statue-n759266

McGarty, C., Thomas, E. F., Lala, G., Smith, L. G. E., & Bliuc, A.-M. (2014). New Technologies, New Identities, and the Growth of Mass Opposition in the Arab

Spring. *Political Psychology*, *35*(6), 725–740. https://doi.org/10.1111/pops.12060

Meloy, J. R., & Amman, M. (2016). Public Figure Attacks in the United States, 1995–2015. *Behavioral Sciences & the Law*, *34*(5), 622–644. https://doi.org/10.1002/bsl.2253

Meloy, J. R., & Hoffmann, J. (2014). *International Handbook of Threat Assessment*. OUP USA.

Mihalcea, R., & Strapparava, C. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, *August*, 309–312.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing Semantic Coherence in Topic Models*. 11.

Mohammad, S. M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In H. L. Meiselman (Ed.), *Emotion Measurement* (pp. 201–237). Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Roth, L. H., Grisso, T., & Banks, S. (2001). *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. Oxford University Press.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94. https://doi.org/10.1145/3078714.3078723

Munger, K., & Phillips, J. (2019). *A Supply and Demand Framework for YouTube Politics*. 38.

Nagle, A. (2017). *Kill all normies: The online culture wars from Tumblr and 4chan to the alt-right and Trump*. Zero Books.

Neter, E., & Ben-Shakhar, G. (1989). The predictive validity of graphological inferences: A meta-analytic approach. *Personality and Individual Differences*, *10*(7), 737–745. https://doi.org/10.1016/0191-8869(89)90120-7

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

Neuman, Y., & Cohen, Y. (2014). A Vectorial Semantics Approach to Personality Assessment. *Scientific Reports*, *4*(1), 4761. https://doi.org/10.1038/srep04761

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, *45*(3), 211–236. https://doi.org/10.1080/01638530802073712

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). *How Old Do You Think I Am?: A Study of Language and Age in Twitter*. 10.

Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2011, July 5). Towards Discovery of Influence and Personality Traits through Social Link Prediction. *Fifth International AAAI Conference on Weblogs and Social Media*. Fifth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2772

Nickerson, R. S. (1998). *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*. https://journals.sagepub.com/doi/abs/10.1037/1089-2680.2.2.175

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 145–153. https://doi.org/10.1145/2872427.2883062

Nyberg, P. (2010). *Explosions in Stockholm believed to be failed terrorist attack—CNN.com*. CNN. http://edition.cnn.com/2010/WORLD/europe/12/11/sweden.explosion/index.html?hpt=T1

Oberlander, J., & Nowson, S. (2006). Whose thumb is it anyway?: Classifying author personality from weblog text. *Proceedings of the COLING/ACL on Main …, July*, 627–634. https://doi.org/10.1177/0266382105060607

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2013). *The Extreme Right Filter Bubble*.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018, June 15). The Effect of Extremist Violence on Hateful Speech Online. *Twelfth International AAAI Conference on Web and Social Media*. Twelfth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17908

Olver, M. E., & Wong, S. C. P. (2006). Psychopathy, Sexual Deviance, and Recidivism Among Sex Offenders. *Sexual Abuse*, *18*(1), 65–82. https://doi.org/10.1177/107906320601800105

Ortega, A., & Navarrete, G. (2017). Bayesian Hypothesis Testing: An Alternative to Null Hypothesis Significance Testing (NHST) in Psychology and Social Sciences. *Bayesian Inference*. https://doi.org/10.5772/intechopen.70230

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. https://doi.org/10.1080/13600834.2018.1458455

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 497–501. https://www.aclweb.org/anthology/N13-1053

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 309–319.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Peltz, J. (2017, August 14). *Protests, vigils around US decry white supremacist rally—NY Daily News*. https://web.archive.org/web/20170814021328/http://www.nydailynews.com/newswires/new-york/protests-vigils-decry-white-supremacist-rally-article-1.3408420

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42–45. https://doi.org/10.1016/S0262-4079(11)62167-2

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.1068/d010163

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLoS ONE*, *9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pennebaker, J. W., & King, L. A. (1999). Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291–301. https://doi.org/10.1037/0022-3514.85.2.291

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perraudin, F., & Murphy, S. (2019, October 31). Alarm over number of female MPs stepping down after abuse. *The Guardian*. https://www.theguardian.com/politics/2019/oct/31/alarm-over-number-female-mps-stepping-down-after-abuse

Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. *Synthesis Lectures on Human Language Technologies*, *13*(4), 1–175. https://doi.org/10.2200/S01057ED1V01Y202009HLT047

Polletta, F., & Jasper, J. M. (2001). Collective Identity and Social Movements. *Annual Review of Sociology*, *27*(1), 283–305. https://doi.org/10.1146/annurev.soc.27.1.283

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, *50*(3), 1198–1216. https://doi.org/10.3758/s13428-017-0938-y

Porter, M. (2001). *Snowball: A language for stemming algorithms*. /paper/Snowball%3A-A-language-for-stemming-algorithms-Porter/0d8f907bb0180912d1e1df279739e45dff6853ee

Postmes, T., & Brunsting, S. (2002). Collective Action in the Age of the Internet: Mass Communication and Online Mobilization. *Social Science Computer Review*, *20*(3), 290–301. https://doi.org/10.1177/089443930202000306

Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., & Ungar, L. (2016). Studying the Dark Triad of Personality through Twitter Behavior. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, 761–770. https://doi.org/10.1145/2983323.2983822

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

QSR Internation Pty Ltd. (2014). *NViVo (Version 10)*. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). *Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations*. 35.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*(1), 31. https://doi.org/10.1140/epjds/s13688-016-0093-1

Regner, T. (2020). Crowdfunding a monthly income: An analysis of the membership platform Patreon. *Journal of Cultural Economics*. https://doi.org/10.1007/s10824-020-09381-5

Reuters. (2019, June 28). Charlottesville: White supremacist gets life sentence for fatal car attack. *The Guardian*. https://www.theguardian.com/us-news/2019/jun/28/charlottesville-james-fields-life-sentence-heather-heyer-car-attack

Rinker, T. W. (2019). *Quantitative Discourse Analysis Package.* https://cran.r-project.org/web/packages/qdap/citation.html

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R Package for Structural Topic Models. *Journal of Statistical Software*, 41.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Sahlgren, M., Isbister, T., & Olsson, F. (2018). *Learning Representations for Detecting Abusive Language*.

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., & Alani, H. (2017). *A Semantic Graph-Based Approach for Radicalisation Detection on Social Media* (pp. 571–587). Springer, Cham. https://doi.org/10.1007/978-3-319-58068-5_35

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 6.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sammut, C., & Webb, G. (2011). *Encyclopedia of Machine Learning—Google Books*. Springer Science & Business Media.

Scanlon, J. R., & Gerber, M. S. (2014). Automatic detection of cyber-recruitment by violent extremists. *Security Informatics*, *3*(1), 5. https://doi.org/10.1186/s13388-014-0005-5

Scanlon, J. R., & Gerber, M. S. (2015). Forecasting Violent Extremist Cyber Recruitment. *IEEE Transactions on Information Forensics and Security*, *10*(11), 2461–2470. https://doi.org/10.1109/TIFS.2015.2464775

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. http://www.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings Ofthe Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Schumann, S., Vegt, I. van der, Gill, P., & Schuurman, B. (2019). Towards Open and Reproducible Terrorism Studies: Current Trends and Next Steps. *Perspectives on Terrorism*, *13*(5), 13.

Schuurman, B. (2020). Research on Terrorism, 2007–2016: A Review of Data, Methods, and Authorship. *Terrorism and Political Violence*, *32*(5), 1011–1026. https://doi.org/10.1080/09546553.2018.1439023

Scrivens, R. (2020). Exploring Radical Right-Wing Posting Behaviors Online. *Deviant Behavior*, *0*(0), 1–15. https://doi.org/10.1080/01639625.2020.1756391

Scrivens, R., Burruss, G. W., Holt, T. J., Chermak, S. M., Freilich, J. D., & Frank, R. (2020). Triggered by Defeat or Victory? Assessing the Impact of Presidential Election Results on Extreme Right-Wing Mobilization Online. *Deviant Behavior*, *0*(0), 1–16. https://doi.org/10.1080/01639625.2020.1807298

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. https://doi.org/10.1080/19434472.2016.1276612

Scrivens, R., Davies, G., & Frank, R. (2020). Measuring the Evolution of Radical Right-Wing Posting Behaviors Online. *Deviant Behavior*, *41*(2), 216–232. https://doi.org/10.1080/01639625.2018.1556994

Şenel, L. K., Utlu, İ., Yücesoy, V., Koç, A., & Çukur, T. (2018). Semantic Structure and Interpretability of Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769–1779. https://doi.org/10.1109/TASLP.2018.2837384

Serwer, A. (2018). The White Nationalists Are Winning. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2018/08/the-battle-that-erupted-in-charlottesville-is-far-from-over/567167/

Shear, M. D., & Haberman, M. (2018). Trump Defends Initial Remarks on Charlottesville; Again Blames 'Both Sides.' *The New York Times*. https://www.nytimes.com/2017/08/15/us/politics/trump-press-conference-charlottesville.html

Shrestha, A., Kaati, L., & Cohen, K. (2017). A Machine Learning Approach towards Detecting Extreme Adopters in Digital Communities. *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, 1–5. https://doi.org/10.1109/DEXA.2017.17

Siegel, A. A., Nikitin, E., Barbera, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2018). *Measuring the prevalence of online hate speech, with an application to the 2016 U.S. election*. 22.

Silge, J., & Robinson, D. (2019). *Text Mining with R*. https://www.tidytextmining.com/

Silver, J., Horgan, J., & Gill, P. (2019). Shared Struggles? Cumulative Strain Theory and Public Mass Murderers From 1990 to 2014. *Homicide Studies*, *23*(1), 64–84. https://doi.org/10.1177/1088767918802881

Simon, B., Trötschel, R., & Dähne, D. (2008). Identity affirmation and social movement support. *European Journal of Social Psychology*, *38*(6), 935–946. https://doi.org/10.1002/ejsp.473

Simons, A., & Tunkel, R. (2013). The Assessment of Anonymous Threatening Communications. In *International Handbook of Threat Assessment*. Oxford University Press.

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

Smith, L. G. E., Wakeford, L., Cribbin, T. F., Barnett, J., & Hou, W. K. (2020). Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior*, 106298. https://doi.org/10.1016/j.chb.2020.106298

Smith, S. S., Woyach, R. B., & O'toole, M. E. (2013). Threat Triage: Recognizing the Needle in the Haystack. In *International Handbook of Threat Assessment*.

Snow, D. (2001). *Collective Identity and Expressive Forms*. https://escholarship.org/uc/item/2zn1t7bj;jsessionid&

Soldner, F., Ho, J. C., Makhortykh, M., van der Vegt, I. W. J., Mozes, M., & Kleinberg, B. (2019). Uphill from here: Sentiment patterns in videos from left- and right-wing YouTube news channels. *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 84–93. https://www.aclweb.org/anthology/W19-2110

Spitzberg, B. H., & Gawron, J. M. (2016). Toward Online Linguistic Surveillance of Threatening Messages. *Journal of Digital Forensics, Security and Law*, *11*(3), 43–78.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. *2012 11th International Conference on Machine Learning and Applications*, *2*, 386–393. https://doi.org/10.1109/ICMLA.2012.218

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tanveer, M. I., Samrose, S., Baten, R. A., & Hoque, M. E. (2018). Awe the Audience: How the Narrative Trajectories Affect Audience Perception in Public Speaking. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems  - CHI '18*, 1–12. https://doi.org/10.1145/3173574.3173598

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thomsen, L., Obaidi, M., Sheehy-Skeffington, J., Kteily, N., & Sidanius, J. (2014). Individual differences in relational motives interact with the political context to produce terrorism and terrorism-support. *Behavioral and Brain Sciences*, *37*(04), 377–378. https://doi.org/10.1017/S0140525X13003579

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., & Liu, T.-Y. (2014). A Probabilistic Model for Learning Multi-Prototype Word Embeddings. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 151–160. https://www.aclweb.org/anthology/C14-1016

Tilly, C. (1993). Contentious Repertoires in Great Britain, 1758–1834. *Social Science History*, *17*(2), 253–280. https://doi.org/10.1017/S0145553200016849

Townsend, M., & Traynor, I. (2011). *Norway attacks: How far right views created Anders Behring Breivik*. The Guardian. https://www.theguardian.com/world/2011/jul/30/norway-attacks-anders-behring-breivik

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0018963

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, *29*(4), 402–418. https://doi.org/10.1177/0894439310386557

Turney, P. D. (2012). Domain and Function: A Dual-Space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research*, *44*, 533–585. https://doi.org/10.1613/jair.3640

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

Vine, V., Boyd, R. L., & Pennebaker, J. W. (2020). Natural emotion vocabularies as windows on distress and well-being. *Nature Communications*, *11*(1), 4525. https://doi.org/10.1038/s41467-020-18349-0

Vossekuil, B. (2004). *The final report and findings of the safe school initiative: Implications for the prevention of school attacks in the United States*. DIANE Publishing.

Voué, P., De Smedt, T., & De Pauw, G. (2020). 4chan & 8chan embeddings. *ArXiv:2005.06946 [Cs]*. http://arxiv.org/abs/2005.06946

Vrij, A. (2015). Verbal Lie Detection Tools: Statement Validity Analysis, Reality Monitoring and Scientific Content Analysis. In *Detecting Deception: Current Challenges and Cognitive Approaches*.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wallace, P. (2015). *The Psychology of the Internet*. Cambridge University Press.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning*, 977–984. https://doi.org/10.1145/1143844.1143967

Walter Lübcke: Man on trial admits to killing German politician. (2020). *BBC News*. https://www.bbc.com/news/world-europe-53662899

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a Lexicon of Abusive Words – a Feature-Based Approach. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1046–1056. https://doi.org/10.18653/v1/N18-1095

Wiegand, M., Schulder, M., & Ruppenhofer, J. (2016). Separating Actor-View from Speaker-View Opinion Expressions using Linguistic Features. *Proceedings of NAACL-HLT*, 778–788.

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 347–354. https://www.aclweb.org/anthology/H05-1044

Windsor, L. (2018). The Language of Radicalization: Female Internet Recruitment to Participation in ISIS Activities. *Terrorism and Political Violence*, 1–33. https://doi.org/10.1080/09546553.2017.1385457

Yan, H., & Sayers. (2017). *Virginia governor on white nationalists: They should leave America.* CNN. https://www.cnn.com/2017/08/13/us/charlottesville-white-nationalist-rally-car-crash/index.html

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., & Blackburn, J. (2018). What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. *Companion Proceedings of the The Web Conference 2018*, 1007–1014. https://doi.org/10.1145/3184558.3191531

Zannettou, S., Finkelstein, J., Bradlyn, B., & Blackburn, J. (2019). A Quantitative Approach to Understanding Online Antisemitism. *ArXiv:1809.01644 [Cs]*. http://arxiv.org/abs/1809.01644

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An *R* Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, *7*(2). https://doi.org/10.18637/jss.v007.i02