Eur J Secur Res
https://doi.org/10.1007/s41125-018-0029-9

ORIGINAL ARTICLE

Policing of Terrorism Using Data from Social Media

Robert Pelzer[1]

Received: 14 August 2017 / Accepted: 29 January 2018

Abstract This article considers the challenges of policing terrorism in social media and investigates whether and how these challenges are being addressed in the research and development of tools to detect radicalisation in social media. The availability of big data tools for the analysis of social media has raised concerns about the onset of algorithm-driven profiling techniques leading to social control that extends to large populations. However, it has not been clarified whether such tools are used by the police and whether they fit the requirements of policing in the field of terrorism. The article provides an overview on research and development of big data tools using data from social media in the field of terrorism and extremism. The tools presented predominantly use supervised machine learning classifiers to discriminate between two classes of content, e.g. radical and non-radical. Most of the tools follow a technology-driven approach aiming to optimise the precision of algorithms. They lack both a conceptual model of (violent) radicalisation as well as a clear focus on providing decision support in terms of helping human analysts to filter data. The article argues that technology-driven approaches do not fit with current practices of policing in the field of terrorism and extremism, which build on professional judgement rather than algorithm-driven pattern identification in big data. The paper concludes with the hypothesis that the application of machine learning algorithms can support analysts in generating predictive knowledge, but will not lead to an algorithm-driven security production.

Keywords Predictive policing · Terrorism · Radicalisation · Social media

✉ Robert Pelzer
pelzer@ztg.tu-berlin.de

[1] Technische Universität Berlin, Zentrum Technik Und Gesellschaft, Hardenbergstrasse 16-18, 10623 Berlin, Germany

Published online: 07 February 2018

1 Introduction

Law enforcement agencies increasingly draw on data from social media for the monitoring of domestic extremism and terrorism (Dencik et al. 2017). At the same time, there is a growing field of research and development of predictive tools that are designed to help security analysts identify (violent) radicalisation in social media. This has provoked controversies over a tendency towards *big data-driven profiling* in counter-terrorism policing, involving the deployment of predictive analytics for a large-*N* population (Heath-Kelly 2017: 37). In the envisioned scenario of big data-driven counter-terrorism, threats are not being defined by humans, but by algorithms that "authorize what or whom is surfaced for the attention of a security analyst who, in turn, cannot meaningfully access this process of authorizing and surfacing" (Amoore and Raley 2017: 4). It is assumed that surveillance practices are then no longer based on pre-existing profiles; the profiles are inductively generated from large amounts of surveillance data (Leese 2014; Heath-Kelly 2017).

However, it has not been clarified how available tools to detect or predict (violent) radicalisation in social media actually work, whether they are relevant for the practice of policing in the field of counter-terrorism and how they might affect the current practice. The aim of this article is to compare the current practice of policing in the field of counter-terrorism with the research and development of big data tools for this practice. Hence, it compares actual practices with possible future scenarios of this practice. The article investigates tools and models that are intended to help security analysts detect (violent) 'radicals' or 'extremists' in social media or to predict involvement in 'terrorism' but do not take into account tools and models addressing the prediction of protests or social unrest. Thus, regarding policing, the main focus is on counter-terrorism policing, not taking into account the broader field of protest policing. Even so, the use of the categories 'radical', 'extremist' or 'terrorist' by researchers and developers as well as by police to classify individuals or groups remains highly problematic due to the potentially negative personal consequences of these labels. It becomes even more problematic when researchers or practitioners are trying to predict behaviour or to identify dispositions of individuals along these categories and in doing so produce high numbers of "false positives". Considering this problem, it appears to be important to critically reflect on the concepts and criteria that are used in big data tools to predict or to identify (violent) radicalisation.

The article is structured in three sections. First, the article introduces important concepts for the analysis of policing in the field of counter-terrorism and presents preliminary considerations about the use of big data in policing terrorism. Second, based on these concepts the challenges of counter-terrorism policing are described and implications for the application of big data-driven approaches are discussed. Third, the article investigates current models and tools that make use of algorithms to detect radicalisation in social media. It concludes with the analysis of which kind of predictive knowledge is being produced in these tools and how this reflects on current practices of policing in the field of terrorism and extremism.

2 Pre-emptive Policing and Big Data

As Zedner (2007) points out, the main characteristic of pre-crime oriented control is a shift of the temporal perspective to anticipating and predicting future criminal behaviour and incidents that have not yet occurred. With regard to policing, pre-crime orientation builds upon the concept of *pre-emptive policing*, which is primarily focussed on preventing crime (Van Brakel and de Hert 2011) but need not be based on data-driven forecasting (Moses and Chan 2016). The goal of *predictive policing* is to "forecast where and when crimes will take place in the future" and to make decisions based on these forecasts (Moses and Chan 2016). Predictive policing can thus be understood as pre-emptive policing based on or informed by the predictive analysis of data on past crimes.

However, pre-crime orientation goes beyond a mere shift of the temporal perspective. Take the German legal context for example where the pre-crime shift is discussed under the heading *forward displacement* ("Vorverlagerung"). In criminal law, this involves the forward displacement of criminal liability by the definition of pre-crime offences, e.g. the preparation of a terrorist offence. In police law ("Gefahrenabwehrrecht", danger aversion law),[1] forward displacement refers to *pre-danger territory* ("Gefahrenvorfeld"), enabling police action without preconditions of 'concrete danger' (probable occurrence of a damaging event) being present. As Bäcker (2015) demonstrates, this does not necessarily imply a shift of the temporal perspective. The pre-danger territory (with regard to individual behaviour) is rather defined as a probable disposition of the individual to commit crimes. Policing in this case is based on—as Bäcker calls it—"dispositional facts" ("Dispositionstatbestände") (Bäcker 2015: 228ff.). Police need to make a judgement call about the probability of the existence of a criminal disposition but do not need (and are not able) to specify the probability of its realisation in criminal behaviour (Bäcker 2015: 206ff.). Starting point here is that the criminal behaviour cannot be predicted due to either unknown or in the individual case unobservable conditions of the occurrence of future criminal behaviour. In other words: it is not possible to specify the individual's risk of becoming involved in criminal behaviour. Therefore, it is proposed here to differentiate between two modes of pre-emptive/predictive policing: on the one hand policing that is based on prognoses of *future criminal behaviour* which implies the ability to specify an individual's risks of becoming involved in criminal behaviour; on the other hand policing that is based on assessments of a *present disposition* for future criminal behaviour.

Practices of pre-emptive/predictive policing also differ in how risks of future criminal behaviour are assessed, calculated and managed. Risks can be assessed in an *individualised or de-individualised* manner. De-individualised risk assessment is based on actuarial approaches that abstract from the individual subjectivity. This means that the risk is not specified with regard to the particularities of an individual case—it is the objective probability that an individual may

[1] In German law, police powers to maintain public order and security are regulated in the federal states' police laws whereas criminal procedural law regulates powers while investigating crimes.

be or become an offender that matters (Norris and McCahill 2006). Actuarial approaches to risk assessment are based on statistical calculations of probability, correlating specific risk factors with data on known characteristics of a criminal population. Although actuarial approaches are gaining increasing importance in crime control (Feeley and Simon 1994), in many cases risk assessments in the context of prevention including terrorism prevention are based on "professional judgments" (Sarma 2017). These judgements can solely depend on the professional's experience and knowledge of the individual being assessed ("unaided professional judgement"), or elements of both actuarial and unaided approaches can be combined in a so-called "structured professional judgment" (SPJ). In an SPJ, risk factors are not being combined by an algorithmic formula, but rather "serve to guide the assessor through a process of systematically identifying and interpreting risks, with the overall risk evaluation being based on a broader review of the individual in context" (Sarma 2017: 280). This distinction between actuarial approaches and approaches that support structured, evidence-based judgements is important when it comes to the question of how researchers and developers in the field of security informatics are designing analytical or predictive tools. They can be designed as expert systems, providing analyses that can be accepted or declined by the human operator, or as decision support systems designed to help analysts in gathering, processing and analysing data.

The types of analytics used in predictive policing so far are largely based on traditional empirical approaches and methods used in quantitative criminology (Chan and Moses 2016: 29; Perry et al. 2013: 10–12). However, with the developments in applied statistics, mathematics and computational science in areas such as machine learning and big data analytics, new opportunities to generate prognostic knowledge are emerging. The European Union-funded project PROFILING defines *big data analytics* as the "process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information" (Ferraris et al. n.d., 11). Apart from the application of machine learning techniques, the main difference between big data analytics and earlier data-driven tools from a technical point of view is the size of the data sets being analysed. Another difference is their ability to process and combine data sets with different structures, enabling the identification of correlations and patterns that otherwise might have been missed or could not be statistically justified if based on a small or moderate sample (Moses and Chan 2014: 664). The capability to make unexpected predictions, however, reduces transparency compared to statistical conclusions drawn from smaller and homogenous data sets (Moses and Chan 2014).

As it is the nature of big data analytics to be deployed on data with large sample sizes, using big data analytics for security purposes implies a tendency not only to combine existing data sets on criminal offenders, but also to create larger data sets that include large-N populations by trend, allowing for analysis and identification of new potential risk patterns that were previously unseen (Joh 2016). In this big data scenario of predictive policing, the boundary between predictive analysis and surveillance practice would be blurred because the generation of predictive knowledge such as criminal profiles would go hand in hand with the screening of large data sets in order to hunt for these profiles (see also Leese 2014; Heath-Kelly 2017).

3 Predictive Policing and Prevention in the Field of Terrorism

3.1 Problems Identifying a Pre-terrorist Profile

The main objective of counter-terrorism policing is to prevent terrorist attacks. Thus, the most important criterion of success is the ability to uncover a terrorist attack at the planning stage. As the planning of a terrorist attack or even the intention to do so is usually punishable by law as a criminal offence, the shift of the temporal perspective is already inscribed in the definition of terrorism-related crimes.[2] In any case, as terrorist planning activities remain unknown and unreported, they need to be exposed in pro-active investigations. In order to direct and focus these investigations, counter-terrorism policing is dependent on intelligence providing predictive knowledge about individuals and groups that are 'at risk' of being or becoming involved in the preparation of terrorist attacks. From this point of view, policing terrorism can be understood as predictive or prediction-led policing as it builds on implicit or explicit prognoses about an individual's future involvement in the preparation of terrorist attacks.

More than two decades of research in both academia and security authorities, however, have merely revealed that there is no single profile of the 'pre-terrorist'. This applies to socio-economic as well as psychological and biographical attributes that would allow a distinction between individuals who are on violent and non-violent trajectories (Taylor and Horgan 2006; Horgan 2008; Corner et al. 2016; Silke 2008). State-of-the-art research on radicalisation emphasises a dynamic understanding of violent radicalisation. Becoming involved in terrorism is a complex, dynamic process that varies among individuals and is not predictable from initial individual-biographical and social circumstances (Borum 2010; Horgan 2008). Furthermore, trajectories from non-violence to violence may not only vary between different types of individuals, but also across different forms of terrorism (in terms of ideologies, geographic contexts, organisational linkages, etc.) (Sarma 2017).

3.2 Current Tools for Risk Assessment

Given this lack of an empirical evidence base for specific risk factors for the involvement in terrorist behaviour and taking further into account the potentially devastating consequences of missing "true positive" cases, it is not surprising that existing risk assessment models and guidelines are mainly built on unspecific risk criteria and subsequently prioritise sensitivity at the expense of specificity[3] and thus at the

[2] However, the construction of terrorism-related crimes is not only driven by the objective of terrorism prevention but may also involve the criminalisation of sympathy for terrorists' ideologies and so forth.

[3] Sensitivity refers to the ability of a tool to accurately identify those who are at high risk of transitioning into terrorism (true positives), whereas specificity refers to the accurate identification of those who are not at risk.

expense of high numbers of "false positives" (Sarma 2017: 282).[4] The application of these criteria may allow identifying an individual's disposition or rather 'vulnerability' to becoming involved in terrorism but do not tell us anything about the conditions under which this disposition would actually be realised in terrorist behaviour.

For example, the UK government has developed the "vulnerability assessment framework" that focuses on aspects of 'vulnerability' which are considered on three levels (HM Government 2012): (1) factors that promote "engagement with a group, cause or ideology"; (2) intent factors that indicate readiness to use violence, including dehumanization of those targeted by terrorists; and (3) the capability to cause harm, referring to individual skills, competencies and access to networks and equipment (see also Sarma 2017). Prevention and de-radicalisation programmes in Germany do not yet use a common guideline for risk assessment (Köhler 2017). In the policing domain, the Federal Criminal Police Office (BKA) has introduced the risk assessment system RADARite to be used by the police investigators in the federal states (BKA 2017). Building on a standardised questionnaire the system allows a rule-based classification of an individual's risk of getting involved in acts of 'destructive' violence in three categories: 'high', 'noticeable' or 'moderate'. As the system is designed to be applied to individuals already classified as 'dangerous' and to help counter-terrorism units with prioritising resources, it appears that the criteria used to assess an individual's risk of becoming involved in terrorism are more specific than in the case of the aforementioned "vulnerability assessment framework".[5] In both examples for risk assessment tools presented, the criteria being used are not drawn from big data. In RADARite, the criteria appear to be deduced from analyses of past cases of violent extremism whereas the UK government's "vulnerability assessment framework" seems to be deduced from the general knowledge base in radicalisation and terrorism research.

Apart from that, the NSA Skynet program has demonstrated that in the intelligence domain big data, for example telecommunication metadata, are used or could be used to identify suspicious patterns of activity or communication. However, it is argued here that such approaches, apart from the highly problematic tapping of private data on a mass scale, would not solve the problems of counter-terrorism policing in identifying pre-terrorist profiles. First, as mentioned above, terrorism is extremely rare behaviour compared to the kind of behaviour usually predicted when using big data analytics such as consumer or voting behaviour. This means that data sets on terrorist behaviour are small by nature (Leman-Langlois 2016). Second, terrorist behaviour is clandestine behaviour that does not always become visible in the data traces produced during daily activities of individuals. So in many cases it does not seem to be possible to identify suspicious patterns of pre-terrorist behaviour in big data. Rather, information about patterns of pre-terrorist behaviour needs to be reconstructed through police investigations and thus—at a systematic level—only

[4] Due to the negligible low base rate for involvement in terrorism in Western societies (Sarma 2017; Roberts and Horgan 2008), which means due to the very small number of terrorists within the population, risk assessment instruments with low specificity generate high numbers of "false positives".

[5] The detailed criteria are not publicly available.

exist in specific data sets, such as police files. As a result, data on terrorist behaviour are pre-structured data on offenders and offences that are categorised according to the criteria relevant for law enforcement and counter-terrorism purposes. Typically, these are data about the offender's socio-demographic background, such as gender, nationality, age, family status and health-related issues like drug abuse and mental health problems. Data sets usually also contain information on prior offences as well as on the modus operandi of the terrorist offence itself. Profiles of pre-terrorist behaviour can be generated using traditional statistical methods of empirical criminological research by identifying correlations between the aforementioned variables within data sets on terrorists or in comparison with other population samples (criminal and non-criminal). From this point of view, apart from the stated general problems identifying pre-terrorist profiles, these profiles can only be identified with 'labelled' data which only exist in 'small data' sets.

By contrast, in her work on the new UK Prevent strategy, Heath-Kelly (2017) comes to the conclusion that the identification of 'vulnerable' individuals in the context of the National Health Service (NHS) follows the logic of big data methodologies. The hypothesis is based on the argument that NHS staff is trained for 'inductive profiling' instead of following static checklists of vulnerability criteria. In fact, within the Prevent framework, the NHS staff is obliged to recognise signs of vulnerability to violent radicalisation. The "Prevent Training and Competencies Framework" (NHS England 2015) encourages NHS staff to use their professional judgement in assessing behaviours and risks, as there are no static risk factors for the vulnerability to violent radicalisation. Heath-Kelly's conclusion that "the training attempts to operationalise inductive profiling, whereby NHS staff constitute the extremist profile as a kind of 'data derivative'" (Heath-Kelly 2017: 38) is not convincing. It would only be a case of inductive profiling if staff judgements were treated as a valid judgement about vulnerability to extremism and, furthermore, if data on these 'positive' cases were used in the follow-up process to search for new patterns of vulnerability—but this is precisely not the case. The role of the NHS staff is to generate suspicion against individuals and, if needed, to convene a "Prevent Management Group" (PMG) that can "escalate concerns" and refer the individual to the Channel programme, where the risk of the referral is finally assessed in a panel discussion (Home Office 2015).[6] This indicates that the concerns raised by NHS staff are part of a selection process for 'vulnerability'. Out of the total figure of 3934 Channel referrals[7] between 2007 and 2014, some 777 (20% of referrals) were assessed to be vulnerable to violent radicalisation by a multi-agency panel (NPCC). The NHS staff nominates candidates for assessment. Thus, they take on an important role in the selection process, comparable to the selection process in the criminal

[6] The Channel programme aims to provide support and redirection for individuals who were assessed to be 'vulnerable' for transitioning into violent extremism. Risk assessment and intervention planning are completed by panels of practitioners from different domains including police, community, health, education. The police practitioner responsible for coordinating the Channel process in their area pre-selects whether a referral is appropriate for the Channel programme or not (Home Office 2015).

[7] Becoming a Channel referral is already the outcome of a selection process, since Prevent Management Groups inside the NHS organisations decide on escalating concerns.

justice system, starting with citizens who can raise suspicion against potentially criminal behaviour. But there is no indication that NHS staff take part in the definition process of 'vulnerability'. One could say that it is in their sphere of influence whether a person remains in the 'dark field of vulnerability', but not whether the person is actually participating in the Channel process. In the end, as a social phenomenon, vulnerability is constituted in the Channel-statistic. This statistic reports individuals that have been registered as either positive or negative referrals (with regard to vulnerability). Hence, the statistic does not appear to be the outcome of a big data methodology but of a multi-stage selection and definition process of vulnerability. Such statistics can be interpreted and criticised in the same way as crime statistics.

4 Generating Predictive Knowledge Using Social Media Data

4.1 Tools and Models

Although there is no indication for big data-driven 'profiling' of terrorists in current counter-terrorism practices, there is a growing number of tools and models that have been developed in the field of Intelligence and Security Informatics aiming to detect online radicalisation and warning signs for pre-terrorist behaviour on social networking sites, micro-blogging sites such as Twitter, or on blogs and forums. This raises questions as to how these tools actually work, whether and how the aforementioned problems to predict terrorism are addressed and whether the tools fit with current practices of predictive policing in the field of terrorism and extremism.

There are many studies from the computational sciences dealing with the analysis, modelling or detection of radicalisation in social media. My discussion will draw on a comprehensive meta-analysis of previous research conducted by Agarwal (2015) as well as an in-depth investigation of eight selected papers that present solutions to *detect radicalisation* (see Table 1).[8] Agarwal reviewed more than 40 publications on the topic of radicalisation detection between 2001 and 2015. Her results match with the finding of my review perfectly.

The problem of predicting radical content is usually conceptualised as a *binary text classification* problem for a categorical output variable, for example "promoting hate" and "extremism" (Agarwal and Sureka 2015a, b), "supporting jihadist groups" (Ashcroft et al. 2015; Kaati et al. 2015; Ferrara et al. 2016) or "cyber-recruitment" (Scanlon and Gerber 2014). A distinct approach is taken by Brynielsson et al. (2013, see also Cohen et al. 2014), who develop a theoretically oriented "threat model" for violent intent (outcome variable) consisting of five indicators ("active on radical

[8] A keyword-based search (radicalisation, social media) has been conducted in Google Scholar (since 2014) to identify relevant pieces of research. We have selected open papers or papers accessible via German libraries dealing with predictive modelling of radicalisation. Hence, we excluded papers that do not explicitly present solutions for the detection of radicalisation (e.g. Rowe and Saif 2016).

Table 1 Example of current approaches for detecting radical content in social media data

Authors	Aim and scope	Techniques and features	Training data
Agarwal and Sureka (2015a, b)	Identifying "hate"- and "extremism"-promoting content on Twitter	Topical crawler; semi-supervised machine learning classifier for binary text classification; multiple features	Annotation of "hate-promoting" content by students
Ashcroft et al. (2015) and Kaati et al. (2015)	Detecting users supporting jihadist groups and disseminating propaganda ("media mujahedeen") on Twitter	Supervised machine learning classifier for binary text classification; stylometric, time-based and sentiment-based features	Known jihadists' accounts on Twitter as well as manually identified "pro ISIS" accounts
Azizan and Izzatdin (2017)	Detecting "terrorism" in tweet patterns of Twitter users	Supervised machine learning classifier for sentiment classification, comparison of sentiment score of a user's latest and previous tweet	Terrorism keywords such as "ISIS", "Bomb", "Muslim"
Brynielsson et al. (2013) and Cohen et al. (2014)	Identifying signs of "warning behaviours" for lone wolf terrorists on websites	Webcrawler; different indicators of "warning behaviour" are considered in a "threat model", such as leakage, identification and radical expression. Both supervised machine learning and discriminant word lexicon used for text classification	Manually created word lexicon and other
Ferrara et al. (2016)	Detecting ISIS supporters on Twitter and predicting adoption of content and interaction with extremist	Supervised machine learning classifier based on metadata only, including user metadata, network statistics and temporal patterns of activity	List of over 25 thousand Twitter accounts labelled as supportive to ISIS by the crowd-sourcing initiative "Lucky Troll Club"
Nouh et al. (2017)	Detecting "extremist content" on Twitter	Unsupervised machine learning for clustering; linguistic (e.g. frequent words), psychological and behaviour-based (likes, reply, social network) features	None, but results of the clustering algorithm are presented to the analyst (user) who tags "radical" content to train the algorithm
Scanlon and Gerber (2014)	Detection of cyber recruitment in Islamist/jihadist web-forums	Supervised machine learning classifier for binary text classification, *n*-grams	Two "independent judges" annotate content as either containing or not containing "violent extremist recruitment"

Table 1 (continued)

Authors	Aim and scope	Techniques and features	Training data
Wadhwa and Bhaita (2014, 2015)	Discovering "hidden radical communities"	Rule-based message classification; topic-based social network analysis; detection of changes in the graph over time ("rate of overlap") to identify radicalisation	"Security dictionaries" created with domain experts

websites", "leakage", "radical expression", "fixation", "identification").[9] However, the majority of approaches do not explicate their conceptual model of 'radicalisation' or 'violent extremism', but simply report results of machine learning-driven models to the Intelligence and Security Informatics community (see also Etudo 2017: 57).

A text classifier for discriminating radical content can be built in two ways. One way is to manually create a *discriminant word lexicon*. Another way that is used more frequently is the application of Natural Language Processing (NLP)[10] techniques where linguistic features (characteristics) of the content are used to train a *machine learning classifier*. Machine learning involves algorithms that learn (i.e. build a model) from sample data to make data-driven predictions. Predominantly supervised machine learning classifiers are used where the classifier is taught from labelled data. This means that an expert defines the output variable, classifying the content accordingly. The algorithm draws on pre-defined rules and concepts and automates the application of these rules, enabling the analysis of large data sets. In semi-supervised learning, the algorithm learns from small samples of labelled data or from a manually created list of seed terms and then starts to develop new classification rules, e.g. based on terms that are synonyms or co-occurring with already labelled terms. In this case, the algorithm starts to innovate the existing concept and rules. In unsupervised machine learning, there are no corresponding output variables for the input data. The algorithm learns and autonomously establishes patterns in the data in order to detect meaningful abnormalities. Here, the rules of defining deviant behaviour are autonomously invented by the algorithm.

However, the machine learning classifiers used in the reviewed papers, both supervised and unsupervised, draw on pre-defined linguistic "features" for text classification. The approaches mainly differ in feature section and number. The simplest features used are *n*-grams of words. Scanlon and Gerber (2014), for example, just use unigrams (single words) as the discriminatory feature. Others use sentiment-oriented lexicon features, such as SentiWordNet, that enable polarisation between positive or negative emotions in postings to discriminate between radical and non-radical content (e.g. Azizan and Izzatdin 2017; Bermingham et al. 2009). However, the trend is using multiple features from different feature classes (Yang et al. 2011; Kaati et al. 2015; Agarwal and Sureka 2015a; Nouh et al. 2017): stylistic features (e.g. punctuation, digits, and word length), content-specific features (e.g. hashtags, "religious terms", "presence of war") and semantic-oriented lexicon features (e.g. emotion words). Multi-modal approaches can also include non-text-based features like metrics of Social Network Analysis (see Wadhwa and Bhaita 2015; Nouh et al. 2017; Bermingham et al. 2009) or metadata from postings and user profiles. Ferrara et al. (2016) do not use any linguistic features to build their classifiers for ISIS

[9] A further theoretically driven approach for a (semi-)automated analysis of radicalisation in texts (not focussing on social media) has been developed by Sanfilippo et al. (2011). The approach draws on social-scientific theories on radicalisation, particularly social movement theory.

[10] NLP is a sub-field of artificial intelligence research involving the computational analysis of natural language, enabling e.g. the automated classification of postings across different categories such as sentiments, attitudes, etc.

supporters on Twitter but only user metadata, network statistics and temporal patterns of activity. One example of a multi-modal approach including different linguistic features complemented with social network-based features is the model developed by Nouh et al. (2017). In this approach, an unsupervised machine learning classifier is first used to cluster content on Twitter into groups based on three classes of features: semantic and syntactic features (bag of words, *n*-grams, most frequent words, ratio of bad words, number of emoticons), lexical features to identify psychological characteristics (e.g. openness, conscientiousness, extroversion, thinking style, positive or negative tone) and behavioural-based features including social relationships of a user, as well as Like and Reply actions. The results of the clustering algorithm will be presented to the user analyst who is then asked to tag "radical" content in order to train the algorithm.

The solutions presented use different approaches to create data sets and to label the data. Ferrara et al. (2016) draw on a list of ISIS-related Twitter accounts created by a crowd-sourcing initiative. Kaati et al. (2015) identified jihadists' Twitter accounts based on primary sources (a list provided in a jihadist forum) and expert knowledge. Other approaches use a topical crawler to generate the data set for further analyses (e.g. Agarwal and Sureka 2015b; Brynielsson et al. 2013). A topical crawler starts with a seed node, crawls through navigation links and websites, and returns relevant nodes (users, websites) to a given topic based on pre-defined rules—for instance, a keyword list. Topical crawlers are not only used to identify relevant data, but also to identify "hidden communities" and to perform Social Network Analyses (Wadhwa and Bhaita 2014, 2015).[11]

All papers claim to aid or envisage to help practitioners to detect online (violent) radicalisation. The problem statement usually includes a large amount of social media data, making it practically impossible for analysts to manually identify more than small selections of radical content or signs of violent radicalisation. However, most papers finish with presenting their technical solution for the classification of online content, but do not discuss how practitioners could or should use the technology to solve their problems. The main focus is on the optimisation of the machine learning classifier in terms of precision. Correspondingly, the role assigned to the human analysts is to feed the algorithm with expert knowledge to improve the precision of the predictive models. With this focus, it appears the vision is to design expert systems rather than decision support systems that explore and filter web content while maintaining space for the interpretation and assessment of data by the human analysts.

Further, it can be noted that the majority of approaches is based on simplified conceptualisations of 'radicalisation' or 'violent extremism' which, above all, do not distinguish between radical opinions and violent behaviour and thus do not reflect on

[11] Social Network Analysis (SNA) applies different mathematical methods of graph analysis to identify the structure of networks generated by the users of various online platforms such as friendships and joint membership of groups. It can differentiate between explicit communities (friendships) and implicit communities, indicated through other forms of interaction such as likes, links and comments (Bartlett et al. 2013).

the related problem to predict involvement in violent behaviour, as discussed before. A positive exemption from technology-driven approaches is the work done by a research group in the Swedish Research Defence Agency (Brynielsson et al. 2013; Cohen et al. 2014; Johansson et al. 2016). Their approach is based on a multidimensional threat model consisting of theoretically relevant indicators of violent radicalisation. The aim of the data analysis is to filter out users with a high score in one or more of the threat indicators and to provide "a list of potential lone wolf terrorists that need further investigation" (Brynielsson et al. 2013: 8). The analyst can adapt the sensitivity of the model and investigate/reconstruct the assessment score of each alias on the list by analysing the scores for every single indicator. This allows (critical) reflection on the theoretical assumptions behind the classification algorithm. At the same time, the classifiers are designed as supervised machine learning classifiers, allowing the user to adapt the markers for the different indicators. To cope with the problem of "human biases", Brynielsson et al. (2013) plan to "include checks and balances in the system, both in the technology and in the form of peer reviews of both analysis models (including markers) and the results of analyses".

4.2 The Use of Social Media Intelligence in Policing Terrorism and Extremism

Although social media intelligence (SOCMINT) is gaining increasing relevance for police work,[12] little information is available regarding the actual use of social media data and aforementioned tools and techniques in the field of counter-terrorism/extremism policing. In the European context, only the study by Dencik et al. (2015, 2017) investigates the use of social media by British police officers that are involved in policing of domestic extremism and disorder. The results are as follows: The monitoring of social media for policing extremism is still an emerging practice within the British police. The tools used are available on the market, predominantly marketing-driven tools. At the time of completing the study in 2014, the British police—according to Dencik et al.—had neither developed its own software, nor was it involved in the design and development of predictive tools. Furthermore, the authors note that monitoring has a clear focus on Open Source Intelligence (OSINT) and is used for both pre-emptive policing, as well as real-time police tactics, and responses in the context of events. Keyword-based searching is reported to be the most dominant practice used to filter data relating to particular events and potential threats, and to make a decision whether further action is needed. In this application, algorithm-based tools are used to 'filter the noise' in order to raise awareness on upcoming threats, but not to define the threats or to suggest any actions. A second important use is reported to carry out risk assessment before and during events, e.g. assessing the expected number of militants. This may also include the identification of "influencers".

[12] See Bartlett et al. (2013), for Germany, e.g. the report of the Berlin Police social media project group (Der Polizeipräsident in Berlin 2013). In the US, the LexisNexis company conducted an online survey of more than 1200 law enforcement professionals at the local and national level in 2014, finding that approximately 80% used social media platforms as intelligence gathering tools (LexisNexis 2014).

The authors highlight that social media intelligence is no isolated practice in policing, but integrated in other forms of intelligence (Dencik et al. 2015: 29). Furthermore, automated processes are not used as expert systems but rather to 'filter the noise', allowing analysts to focus on relevant information. The authors conclude that "big data is predominantly used to identify patterns that are subjectively (humanly) interpreted and assessed" (Dencik et al. 2015). Thus, "pre-existing knowledge, intelligence and broader societal understandings of events continue to shape and determine big data analyses" (Dencik et al. 2015).

If the key findings of Dencik et al.'s study are representative for policing of extremism (at least across Western Europe), the following conclusions can be drawn: Big data from social media are used in the policing of extremism/terrorism to generate new knowledge. However, this knowledge seems to be generated as an outcome of *human–machine–human interactions* where a human is defining patterns of interest (e.g. by generating keyword lists) which are used by the machine to filter content and identify patterns that need to be interpreted by the human analyst. Knowledge production and decision making thus appear to be *supported by algorithms* but *not driven by algorithms.* This demonstrates that the human element of the analysis as well as accountability in decision making is key to predictive policing (see also Perry et al. 2013: 123, Moses and Chan 2014). Given this, we can assume that technology-driven approaches seeking to develop expert systems are not appropriate to the practice of (predictive) policing.

5 Conclusions

Big data tools enable the identification of hidden patterns in large amounts of data independent from pre-defined criteria. The research and development of big data tools for the purpose of helping security authorities to detect radicalisation and extremism in social media have raised concerns about the onset of algorithm-driven profiling techniques leading to social control that extends to large populations. In big data scenarios of predictive policing, the boundary between predictive analysis and surveillance practice is blurred. In the case of counter-terrorism policing, police are drawing on and generating predictive knowledge about (pre-)terrorist behaviour. The argument put forward in this article is that the scope for big data-driven approaches that are seeking to identify unknown patterns of pre-terrorist behaviour is limited because pre-terrorist behaviour rarely becomes visible in big data. Given this, pre-terrorist profiles still need to be deduced from 'labelled' data on pre-terrorist behaviour which only exist in the confines of law enforcement databases of terrorist offenders and suspects.

Due to these structural problems of identifying pre-terrorist behaviour, current practices of risk assessment in the field of terrorism and radicalisation prevention predominantly appear to build on structured professional judgements of an individual's risk of getting involved in terrorism rather than on actuarial or algorithmic approaches. This also applies to the policing of terrorism and extremism using data from social media. Although big (social media) data seem to be used to generate predictive knowledge, the interpretation of data and the analyst's discretion

in judging risks remain key components of policing in the field of terrorism and extremism. Contrasting this, tools and models that have been or are being developed in Intelligence and Security Informatics mainly follow a technology-driven approach, seeking to optimise the precision of machine learning classifiers. They lack both a conceptual model of (violent) radicalisation as well as a clear focus on decision support by providing analyses that explicitly leave room for interpretation and discretion. This means placing the analyst's professional judgement at the centre of the approach, instead of addressing them as 'algorithm-trainers' who only provide their expert knowledge to enhance the predictive value of the algorithm.

Since the policing of terrorism is (irreducibly) dependent on professional judgement, this article concludes that machine learning algorithms can support analysts in generating predictive knowledge, but will not lead to an automation of predictive analysis through algorithms that invent new rules of defining risks as envisioned in scenarios of algorithm-driven security production. Still, the technology allows training the algorithm with a practitioner's definitions of deviance or risks. The following scenarios may arise and should be investigated in further research: The use of trained algorithms creates new opportunities to expand surveillance on larger parts of the population in social media. As a consequence, it is very likely that the visibility of risks will increase with further efforts needed to assess these risks. Analysts would need to develop strategies to assess these risks efficiently based on the fragmented information available, which could make it necessary to adapt criteria. This could lead to either scaling down (a smaller scope of behaviour targeted as risk) or scaling up (a broader scope of behaviour targeted as risk) the definitions of deviance and risk. Further, it is not 'objective' knowledge involving formalised criteria and definitions of deviance or risk that is objectified in the machine learning classifier when the algorithm is trained by a practitioner, but rather the practitioner's subjective knowledge. Given this, it could make sense to develop a methodology to peer-review the results of content classification, as suggested by Brynielsson et al. (2013). This could also open up the opportunity to involve external actors into the peer-review and thus to increase not only predictive, but also reflective policing.

Acknowledgements Funding was provided by Bundesministerium für Bildung und Forschung (Grant no. 13N14340).

References

Agarwal S (2015) Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. https://pdfs.semanticscholar.org/fe12/d70b0d849fb4d80153bc2acd11a6d4125ecc.pdf. Accessed 12 Jul 2017

Agarwal S, Sureka A (2015a) Using kNN and SVM based one-class classifier for detecting online radicalization on Twitter. In: Natarajan R, Barua G, Patra MP (eds) Distributed computing and internet technology. 11th International conference on (ICDCIT). Springer, Cham, pp 431–442

Agarwal S, Sureka A (2015b) A topical crawler for uncovering hidden communities of extremist microbloggers on Tumblr. In: Rowe M, Stankovic M, Dadzie A (eds) Proceedings of the 5th workshop on making sense of microposts, vol 1395. CEUR, pp 26–27. http://ceur-ws.org/Vol-1395. Accessed 15 Jun 2017

Amoore L, Raley R (2017) Securing with algorithms: knowledge, decision, sovereignty. Secur Dialogue 48(1):1–8

Ashcroft M, Fisher A, Kaati L, Omer E, Prucha N (2015) Detecting jihadist messages on twitter. In: Proceedings of the intelligence and security informatics conference (EISIC), September 2015, pp 161–164

Azizan SA, Izzatdin AA (2017) Terrorism detection based on sentiment analysis using machine learning. J Eng Appl Sci 12(3):691–698

Bäcker M (2015) Kriminalpräventionsrecht: Eine rechtsetzungsorientierte Studie zum Polizeirecht, zum Strafrecht und zum Strafverfahrensrecht. Mohr-Siebeck, Tübingen

Bartlett J, Miller C, Crump J, Middleton L (2013) Policing in an information age. DEMOS, London. https://www.demos.co.uk/files/DEMOS_Policing_in_an_Information_Age_v1.pdf?1364295365. Accessed 20 Jul 2017

Bermingham A, Conway M, McInerney Lisa, O'Hare N, Smeaton AF (2009) Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In: ASONAM 2009—advances in social networks analysis and mining, 20–22 July 2009, Athens, Greece. http://doras.dcu.ie/4554/3/DCU_asonam09.pdf. Accessed 20 Jun 2017

BKA (2017): New risk instrument for the risk assessment of violent offenders. Press release from February 2017. https://www.bka.de/DE/Presse/Listenseite_Pressemitteilungen/2017/Presse2017/170202_Radar.html. Accessed 15 Jul 2017

Borum R (2010) Understanding terrorist psychology. Mental Health Law & Policy Faculty Publications. University of South Florida, Scholar Commons. http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1575&context=mhlp_facpub. Accessed 18 Jul 2017

Brynielsson J, Horndahl A, Johansson F, Kaati L, Martenson C, Svenson P (2013) Harvesting and analysis of weak signals for detecting lone wolf terrorists. Secur Inform 2(11):11–15. https://doi.org/10.1186/2190-8532-2-11

Chan J, Moses LB (2016) Is big data challenging criminology? Theor Criminol 20(1):21–39

Cohen K, Johansson F, Kaati L, Clausen MJ (2014) Detecting linguistic markers for radical violence in social media. Terror Political Violence 26(1):246–256

Corner E, Gilla P, Mason O (2016) Mental health disorders and the terrorist: a research note probing selection effects and disorder prevalence. Stud Confl Terror 39(6):560–568

Dencik L, Hintz A, Carey Z, Pandya H (2015) Managing 'threats': uses of social media for policing domestic extremism and disorder in the UK. Project report. Cardiff School of Journalism, Media and Cultural Studies. http://orca.cf.ac.uk/85618/1/Managing-Threats-Project-Report.pdf. Accessed Jul 20 2017

Dencik L, Hintz A, Carey Z (2017) Prediction, pre-emption and limits to dissent: social media and big data uses for policing protests in the United Kingdom. New Media Soc. https://doi.org/10.1177/1461444817697722

Der Polizeipräsident in Berlin (2013) Abschlussbericht Projektgruppe Neue Medien Version 1.1.2. https://netzpolitik.org/wp-upload/Abschlussbericht_Projektgruppe_Neue_Medien.pdf. Accessed 17 Jul 2017

Etudo U (2017) Automatically detecting the resonance of terrorist movement frames on the web. Dissertation, Virginia Commonwealth University. http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=5941&context=etd. Accessed 20 Jun 2017

Feeley M, Simon J (1994) Actuarial justice: the emerging new criminal law. In: Nelken D (ed) The futures of criminology. Sage, London, pp 173–201

Ferrara E, Wang E, Varoly O, Flamminiy A, Galstyan A (2016) Predicting online extremism, content adopters, and interaction reciprocity. https://www.researchgate.net/publication/301819535_Predicting_Online_Extremism_Content_Adopters_and_Interaction_Reciprocity. Accessed 17 Jun 2017

Ferraris V, Bosco F, Cafiero G, D'Angelo E, Suloyeva Y (n.d.) Working paper defining profiling http://www.unicri.it/special_topics/citizen_profiling/WP1_final_version_9_gennaio.pdf. Accessed 24 Jul 2017

HM Government (2012) Channel: vulnerability assessment framework. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/118187/vul-assessment.pdf. Accessed 20 Jun 2017

Heath-Kelly C (2017) Algorithmic autoimmunity in the NHS: radicalisation and the clinic. Secur Dialogue 48(1):29–45

Home Office (2015) Channel duty guidance. Protecting vulnerable people from being drawn into terrorism. Statutory guidance for Channel panel members and partners of local panels. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/425189/Channel_Duty_Guidance_April_2015.pdf. Accessed 22 Jul 2017

Horgan J (2008) From profiles to pathways and roots to routes: perspectives from psychology on radicalization into terrorism. Ann Am Acad Political Soc Sci 618(1):80–94

Joh EE (2016) The new surveillance discretion: automated suspicion, big data and policing. Harvard Law Policy Rev 10(2):15–42

Johansson F, Kaati F, Sahlgren M (2016) Detecting linguistic markers of violent extremism in online environments. In: Khader M et al (eds) Combating violent extremism and radicalization in the digital era. Information Science Reference, Hershey, pp 374–389

Kaati L, Omer E, Prucha N, Shrestha A (2015) Detecting multipliers of Jihadism on Twitter. In: ICDM workshops 2015. pp 954–960

Köhler D (2017) Understanding deradicalization. Methods, tools and programs for countering violent extremism. Routledge, London

Leese M (2014) The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. Secur Dialogue 45(5):494–511

Leman-Langlois S (2016) Big data against terrorism. In: Research workshop 12–14 May, Kingston, Ontario, Canada. http://www.sscqueens.org/sites/default/files/6_big-data-against-terrorism-stephane_leman-langlois.pdf. Accessed 16 Jul 2017

LexisNexis (2014) Social media use in law enforcement: crime prevention and investigative activities continue to drive usage. http://www.lexisnexis.com/risk/downloads/whitepaper/2014-social-media-use-in-lawenforcement.pdf. Accessed 16 Jul 2017

Moses LB, Chan J (2014) Using big data for legal and law enforcement decisions: testing the new tools. Law J 37(2):643–678

Moses LB, Chan J (2016) Algorithmic prediction in policing: assumptions, evaluation, and accountability. Polic Soc. https://doi.org/10.1080/10439463.2016.1253695

NHS England (2015) NHS England prevent training and competencies framework. NHS England Nursing Directorate, London. http://www.lewishamccg.nhs.uk/news-publications/Policies/Documents/Prevent%20training%20and%20competencies%20framework.pdf. Accessed 20 Jul 2017

Norris C, McCahill M (2006) CCTV: beyond penal modernism? Br J Criminol 46(1):97–118

Nouh M, Jason RC, Nurse, MG (2017) Detection of online radical content using multimodal approach. In: Poster presented at 2nd IEEE European symposium on security and privacy, 26–28 April 2017 in Paris, France. https://www.ieee-security.org/TC/EuroSP2017/posters/poster10.pdf. Accessed 22 Jul 2017

Perry WL, McInnis B, Price CC et al (2013) Predicting policing: the role of crime forecasting in law enforcement operations. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf. Accessed 15 Jul 2017

Roberts K, Horgan J (2008) Risk assessment and the terrorist. Perspect Terror 2(6). http://www.terrorismanalysts.com/pt/index.php/pot/article/view/38/html. Accessed 14 Jul 2017

Rowe M, Saif H (2016) Mining pro-isis radicalisation signals from social media users. In: Proceedings of the tenth international AAAI conference on web and social media (ICWSM 2016). pp 329–338. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13023/12752. Accessed 18 Jun 2017

Sanfilippo A, McGrath L, Whitney P (2011) Violent frames in action. Dyn Asymmetric Confl 4(2):103–112

Sarma KM (2017) Risk assessment and the prevention of radicalization from nonviolence into terrorism. Am Psychol 72(3):278–288

Scanlon JR, Gerber MS (2014) Automatic detection of cyber-recruitment by violent extremists. Secur Inform 3:5. https://doi.org/10.1186/s13388-014-0005-5

Silke A (2008) Holy warriors: exploring the psychological processes of jihadi radicalization. Eur J Criminol 5:99–123

Taylor M, Horgan J (2006) A conceptual framework for addressing psychological process in the development of the terrorist. Terror Political Violence 18(4):585–601

Van Brakel R, de Hert P (2011) Policing surveillance and law in a pre-crime society: understanding the consequences of technology based strategies. Cahiers Politiestudies 20:163–192

Wadhwa P, Bhaita MPS (2014) Discovering hidden networks in on-line social networks. Int J Intell Syst Appl 05:44–54

Wadhwa P, Bhaita MPS (2015) An approach for dynamic identification of online radicalization in social networks. Cybern Syst 46(8):641–665

Yang M, Kiang M, Ku Y, Chiu C, Li Y (2011) Social media analytics for radical opinion mining in hate group web forums. J Homel Secur Emerg Manag. https://doi.org/10.2202/1547-7355.1801

Zedner L (2007) Pre-crime and post-criminology? Theor Criminol 11:261–281