The Hague, 07/09/2018

Assessment of Risk in Written Communication

Introducing the Profile Risk Assessment Tool (PRAT)

This paper was presented at the 2nd European Counter-Terrorism Centre (ECTC) Advisory Group conference, 17-18 April 2018, at Europol Headquarters, The Hague.

Author: Nazar Akrami (Uppsala University), Amendra Shrestha (Uppsala University), Mathias Berggren (Uppsala University), Lisa Kaati (Swedish Defense Research Agency), Milan Obaidi (Uppsala University), and Katie Cohen (Swedish Defense Research Agency)

1 Abstract

In this report, we introduce a tool for risk assessment in written communication – the profile risk assessment tool (PRAT). PRAT extracts a profile consisting of 30 personality and risk-behavior-related variables from any given text. PRAT includes a theoretically generated threat profile to be used as a comparison norm. To assess threat potential, the extracted profile is compared to a theoretical profile as well as 27 834 profiles including known cases of violent lone offenders, school shooters, and social media users from various sources (Google blogs, Stormfront, Reddit, Islamic Awakening, and Boards). A preliminary case study showed that the theoretical profile was highly similar (interclass correlation) to profiles of lone offenders and school shooters but not profiles of individuals from the normal population extracted from text communicated on social media. Another case study examined the extent to which the PRAT-extracted profiles from various sources (e.g., Google blogs vs. violent lone offenders) could be identified using quadratic discriminant analyses. The results of this study showed a promising outcome in some cases but limited in others due to the small number of included cases (e.g., violent lone offenders). Some limitations and challenges of PRAT are identified and will be subject to development and further elaboration. PRAT is aimed to be open for authorized researchers and law enforcement professionals.

2 Background

In the past couple of decades, there has been a significant development in efforts to counter terrorism. Much of these efforts have been directed toward assessment of risk from individuals that are considered to be a major source of threat, or so-called lone offenders. These efforts have resulted in the development of several risk assessment tools (see Meloy & Gill, 2016). While research on counter-terrorism continues, terrorism and terrorist activities have in recent years structurally changed to, in most cases, include a digital component.

Thus, counter-terrorism research needs to meet this structural change by focusing on risk assessment beyond traditional tools, for example checklists. The Profile Risk Assessment Tool (PRAT) described in this report is aimed to contribute in this development. The main aim of PRAT is to assist counter-terrorism specialists, researchers as well as law enforcement professionals, to assess risk based on written communication.

3 Issues in Risk Assessment

Considering the significant decrease in number of terror attacks in western countries in the past two years (globalterrorismindex.org, see also Europol, 2018), counter-terrorism research and practice can, with no doubt, be considers successful. Nevertheless, law enforcement professionals need to be steps ahead in the process of detection of potential threats. This necessitates further development of risk assessment and detection capacities. Enabling this development, some critical issues in contemporary counter-terrorism research that need to be addressed (for a review, see Monahan, 2016). For example, previous research tends to leave out core psychological variables, particularly nonclinical personality that constitutes a fingerprint-like signature for every individual. Previous research shows that these personality variables provide a strong predictive power in distinguishing between individuals with and without intentions to use violence (Obaidi et al., 2018a). Another critical issue that needs to be addressed is the tendency to treat significant variables and approaches independently. Thus, research needs to systematically integrate variables and approaches by providing useful measures. More importantly, counter-terrorism research needs to engage methodological developments as well as studies examining the validity and the reliability of contemporary methods. We believe that dealing with these issues can

be facilitated by providing a common flexible platform that can be used by researchers and law enforcement professionals. We hope that initiating the Profile Risk Assessment Tool (PRAT) project is a step in this direction.

4 The Profile Risk Assessment Tool (PRAT)

As mentioned above, while there are a variety of risk assessment instruments based on structured professional judgment, none is directly aimed to assess risk in written communication using established linguistic technologies. PRAT was developed using open source components that are freely available for academic purposes. It is a web-based tool where authorized users can upload any text or write paragraphs and receive a threat assessment of the uploaded text file or written text. Depending upon their need, there is a possibility to download the result either in text or pdf format. The goal was to build a simple, convenient and visually informative system for law enforcement agencies and researchers, which can be used without vast knowledge in computers. The PRAT is aimed to combine recent technologies with flexibility of development, modification potentials and accessibility for researchers as well as law enforcement professionals. The idea is to successively integrate the best components of contemporary risk assessment instruments and make an integrated system available for research and practice. The initial version of the tool includes a number of variables that are automatically extracted from written communication by target individuals. Each variable is based on occurrence of information related to that specific variable in the written text. The choice of variables is based on previous research and theory regarding risk behavior but also core personality information that constitutes the individual´s psychological fingerprint. The extracted variables are aimed to provide a profile that can be compared to a theoretical risk profile extracted from a wide range of targets in a variety of populations. Profiles will be compared in terms of absolute similarity. Each profile will also be subjected to normative comparison, that is, the target individual´s position on each variable will be compared to the entire sample population.

4.1 Psychological Variables

In this section we will describe the set of psychological variables we have included in the first version of our tool. These variables include core personality variables, emotions and a set of variables related to the social part of the self.

Personality

Generally, personality refers to the psychological makeup of an individual. A more specific definition is "the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought" (Allport, 1961, p. 28). Personality is also assumed to comprise a set of psychological traits that are relatively enduring (e.g., McCrae & Costa, 1997). A significant number of personality psychologists, not to say the majority, consider personality to be expressed in traits that are related to different aspects of human behaviors (e.g., interpersonal relations). While there is some agreement about central elements that define personality, this is not the case with regard to the structure of it, especially when it comes to the organization of personality traits. There are, thus, a variety of models promoting various structures for and numbers of traits. In this regard, we focus on the Five Factor Model (McCrae & John, 1992) to be integrated in the profile risk assessment tool. This focus is made for various reasons, for example, a) extensive previous research documenting the validity and reliability of the factors, b) extensive availability of various measurement methods and measures both within and across various languages, and c) the Big Five factors being linked to a long list of human behavior (e.g., Roberts et al., 2007) including political extremism (see Thomsen et al., 2014). The basic assumption in the Five Factor Model is that human personality can be captured by a set of factors; Neuroticism, Extraversion, Openness to experience (henceforth Openness), Agreeableness, and Conscientiousness. In our threat assessment tool, we assess indicators of the positive and negative aspects of each factor in a given text, resulting in ten different personality variables, two for each factor. We refer to these positive and negative aspects of each factor by adding a plus (+) or minus (–) sign after the factor name. High scores on a specific factor indicate that it is descriptive of the writer. We now describe these factors.

Neuroticism+ indicates that the person expresses negative emotions such as anxiety, sadness, and insecurity, as well as anger, and could mean that the person is more impulsive. Thus, high scores on Neuroticism+ indicate that the person is more easily affected emotionally and, therefore, probably also more "explosive". Individuals low on Neuroticism+ express less anxiety and negative emotions. Neuroticism+ is of special interest in this context, as low levels of negative emotions mean fearlessness and being unfazed when dealing with more extreme/risky situations. **Neuroticism**–, on the other hand, means stability, even temperedness, calmness and an immunity to experience negative feelings. A high score on Neuroticism– is usually a protective factor against falling into a negative spiral of destructive behavior.

Extraversion+ is indicative of a talkative and active person who prefers social gatherings and activities. Individuals scoring high on Extraversion+ are signified by energy, positive emotions and are attention-seeking and relatively dominant in social settings. Low levels of extraversion, or high on Extraversion–, indicates that the person talks less about friends and social experiences, and more about other things such as what they have achieved, about home and work. Individuals with high scores on Extraversion– are not driven by social factors in their life and find meaning in other domains. They are shy and reserved. High scores on **Extraversion**+ could function as a protective factor when it comes to extremism – as low degrees of social connection means lower social costs and lack of social support.

Openness+ deals with a willingness to test new things, in particularly intellectually, such as taking in different perspectives and elaborating on different ideas. High Openness+ is coupled with a tendency to be creative, imaginative and curious. High Openness+ is probably a protective factor against things like conspiracy theories and radicalization, as it allows people to think more critically about their own ideas. High scores on **Openness**– on the other hand, indicate that the person is relatively closed to new experiences, and is overly sure that their view is the "right" view without taking in different perspectives – that is: a tendency to be dogmatic. Individuals high in Openness– also tend to be more traditional and conventional and cite numbers and "old wisdoms" that support one's view. High Openness– also often means that the writer is rather conservative politically and religiously. Political or religious extremists are therefore more likely to show high Openness–.

Agreeableness+ indicates more concern for others, including thinking about feelings of people, as well as generally more altruistic behavior. Individual high on Agreeableness+ tend to trust others and are soft-hearted and generally good-natured. **Agreeableness–** indicates that an individual is competitive rather than cooperative with others. A high Agreeableness– individual tend to be relatively selfish, lacks empathy or sympathy to the sufferings of others, and views the world in a more mechanical "cause" and "effect"" way. Aggressive behavior such as threats and insults are also strong indicators of Agreeableness–. An individual who shows clear signs of antisocial behavior, and for, in layman's terms, being a "psychopath", typically scores high on Agreeableness–. It is important to note, however, that not everyone with this pattern needs to be a psychopath, and psychopaths might fake agreeableness.

Conscientiousness+ is mostly related to work-life and is considered to be a sign of motivation, ambition, hard-working, punctuality, and organization.

Conscientiousness+ is usually something valuable in society, as it indicates determination to follow through on work and projects. However, this also means that people high in Conscientiousness+ are likely to follow through on any plan they have, including plans that might to be unfavourable for others or society, for example a criminal act. **Conscientiousness**–, on the other hand, is indicative of being lazy, disorganized and having a general tendency to be late. Individuals high on Conscientiousness– tend to talk more about failure, negative emotions, and what they want (but cannot have).

Emotions

Emotions are important for predicting our behavior and actions and they play a significant role in our everyday life. Also, emotions are prime indicators of the interaction between the individual's way of thinking and the social world surrounding him/her. Emotions have been also emphasized by various scholars in predicting violent extremism (e.g., Atran, 2003; Gerges, 2005; Khosrokhavar, 2005; Richardson, 2006; Ricolfi, 2005; Stern, 2003) and many have documented the effect of emotions on various forms of extremist violence. For example, Obaidi et al. (2018b) showed that anger was a reliable predictor of violent intentions in defense of one's group, and Tausch et al. (2011) demonstrated that contempt was associated with support for violent extremism.

Indeed, some theorize that emotions are associated with violations of cherished values such as ethics and morality (Haidt, 2012), and hence it is believed that emotions are important drivers of group-based aggression and violence (Molho et al., 2017). Similarly, others have hypothesized that emotions are particularly important for understanding violence and aggression because they play a unique role in our moral and cultural convictions (Molho et al., 2017; Rozin et al., 1999). Hence, emotions may play a role in providing the motivation for aggression and violence given their social and relational functions (Fiske, 2002) and relatedness to our ethics and morality (Haidt, 2012).

Sternberg (2003) has, for instance, proposed emotions as one of the main causes of aggressive behavior toward the outgroup members. For example, anger, contempt, and disgust have been proposed to constitute the three components of hate, which in turn may be the main driver of violence and acts of aggression, such as genocide and terrorism. In sum, emotions are instrumental in inciting groups to commit violence against each other (Mackie et al., 2000). More importantly, emotions can be a reliable predictor of extremist behavior. For instance, analyzing emotional content of written documents can provide indications of future potential acts of aggression toward certain groups and individuals in society. For example, the

emotions expressed in speeches by leaders of ideologically motivated groups showed an increase in expression of anger, contempt, and disgust immediately before acts of violence toward the out-group (Matsumoto et al., 2012). Similarly, the verbal expressions of anger, contempt, and disgust toward outgroups across time have been shown to be associated with violence and hostility against that particular out-group members (Matsumoto et al., 2016). More importantly, an overt expression of anger has also been found to be highly correlated with physical aggression (see e.g., Buss & Perry, 1992).

Based on the above, we regard emotions as an important component in predicting violent extremism and an important component of the threat assessment tool. In this regard, we mainly focus on anger and the expression of positive and negative emotions. We include anger because, in the literature on extremism, anger is often associated with non-normative collective actions (e.g., Livingstone et al., 2009; Obaidi et al., 2018b). We include positive and negative emotions as people's thoughts and feelings about themselves and others influence their negative and positive emotions reactions (Stucke & Sporer, 2002). Hence, we propose that the difference and intensity of positive and negative emotions expressed in written communication are valuable sources of information in assessing potential threat. Research has shown that intense positive and negative emotions can interfere and undermine rational, advantageous decision making (Bechara, 2004, 2005; Dolan, 2007; Dreisbach, 2006; Shiv et al., 2005) and can, hence, lead to ill-considered or rash actions (Cyders & Smith, 2008). To this end, we integrate assessments of **positive emotions**, **negative emotions** and **anger** in our tool.

The Social Self

Humans are social animals and an important part of people's self-concept related to our relations with other. We belong to various social groups and categories that define our social self. Thus, our personality and self-concept contains information on how we deal with social groups and individuals belonging to our group or groups that we do not belong to (e.g., out-groups). Being a part of a group is a basic human need, fundamental for survival and, therefore, group identification and the concept of social identity are two important components in explaining what motivates a person's actions (Shrestha et al., 2017).

Studying the use of pronouns can inform us how people consider themselves in relations to others, groups as well as individuals (Pennebaker, 2011). Also, previous research has documented that, for example, radicalized individuals tend to have an inflated self, expressed by frequent use of "I" and by a tendency to see boundaries between groups, expressed by frequent use of "we", "them" "they".

Therefore, our threat assessment tool includes a social-self index comprising of lists assessing the frequency of words related to I**,** we, **they** and **friends** in any given text.

4.2 Warning Behaviors

The idea that there are identifiable behaviors that precede targeted violence is a core concept in the threat assessment literature (Meloy et al., 2012). Such behaviors are sometimes referred to as warning behaviors. Leakage, along with fixation, is the most frequently studied warning behavior in the literature. A brief description of these warning behaviors is presented below.

Leakage

Leakage is a warning behavior where an intent to do harm is communicated to a third party. Leakage is a warning behavior that has been noticed among school shooters in a study done by O'Toole (2000). O'Toole defined leakage as the intentionally or unintentionally revelation of feelings, thoughts, fantasies, attitudes, or intentions that may signal an impending violent act. Leakage can be intentional or unintentional and it can be done for a number of different reasons, such as the need for excitement, a desire to create panic, attention-seeking, or fear and anxiety about the impending act. Sometimes, leakage could be a result of the subject's desire to memorialize their action after their death or incarceration. Hempel et al. (1999), studied adult mass murders (who killed three or more) and found that 67% of the subjects did express some kind of threat before committing their attacks. In another study of mass murders (who killed more than three), it was found that subjects discussed the act of murder with at least one person prior to the event in 44% of the cases (Meloy et al., 2001). In our tool we assess leakage by including three different themes related to leakage: intent, killing, and power.

The assessment of intent is aimed to capture words signalling that the writer aims to act or do something. Specifically, high scores on intent would suggest that the writer could be likely to be planning an action that needs a closer look. For example, if the communicated text is also high on words relating to killing (see below), this should be a warning signal.

The assessment of **killing** is aimed at capturing any communication related to death, killing, and hurting others. A text with high scores on killing could therefore be cause for alarm. In this regard, leakage is the most important aspect of warning behaviors. The assessment of **power** is aimed to capture various aspects of status,

control, strength, and domineering. A high use of power words is indicative of a tendency to glorify the self, the groups to which the individual belongs, or a cause or ideology. It could also express a devaluing and dehumanization of opponents or enemies, and in the case of a planned attack: the intended victims. Previous research has shown that the use of power words tends to increase before an attack (Kaati et al., 2016).

Fixation

The warning behavior fixation is defined as any behavior indicating a nearly pathological preoccupation with a person or a cause (Meloy & O´Tool, 2011). This category of warning behavior is directly related to or indicative of the possible target of an action. Thus, fixation is rather specific and deals with, for example, a social category, ideology or an individual. The threat assessment tool assesses fixation by measuring how much a person communicates about specific themes or individuals. We acknowledge that the choice of themes is directly related to the detection ability of our tool and, thus, we are open to extend the measurement of fixation beyond the current version. The current set of themes were chosen to illustrate some of the common ideologies or ideas that have been present among previous lone offenders. The present version of the tool includes the assessment of communication related to **weapons**, **military terms**, **well-known lone offenders** or **school shooters**. A frequent use of military terminology or mentions of weapons would, in this case, indicate that the subject identifies with a warrior (having a warrior mentality). Frequent mentioning of well-known murderers would indicate the possibility of being influenced by them or a sense of identification with them. The tool also assesses fixation related to **Jews, migration, Islamisation, Islamic state terminology**, and **incel terminology** ("involuntary celibacy" community terminology).

5 Method

In this section we present the methodological features of the PRAT and the rationale for using some methods before others. PRAT was developed using Python, and the web interface was developed using HTML, CSS and JavaScript. The backend of the web interface was developed using Python's microframework Flask based on Werkzeug and Jinja 2. Beyond these technical aspects other methodological features are outlined below.

5.1 Psychological Variables and Linguistic Indicators

The aim of PRAT is to automatically identify psychological (and other) variables in written communication based on linguistic indicators. Previous research has linked the use of various linguistic indicators to, for example, self-rated personality (e.g., Park et al., 2015; Yarkoni, 2010). In this regard, the Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker, 2010) is used to count the relative frequencies of words in a text and divide words into categories that are psychologically meaningful. Today, linguistic indicators are used in a variety of branches to extract constructs like cognition and emotion, drives, and personality in addition to grammatical and other writing style markers from any text (Tausczik & Pennebaker, 2010).

More specifically, PRAT uses a dictionary-based approach to identify various linguistic indicators. For each dictionary, the relative frequency of words in the text material is counted. The relative frequencies of the dictionary words are then averaged, and this produces an aggregated score for each linguistic indicator that represents its relative frequency of occurrence.

There are many drawbacks of using a dictionary-based approach when detecting linguistic markers. One is that the meaning of words can be context-dependent, which means that words may have several different meanings depending on the context. To develop dictionaries that capture the different themes, we have included experts with domain knowledge of the environment that we study. The dictionaries were augmented with words using a distributional semantic model that was pre-trained on relevant data. Each of the words suggested by the distributional semantic models was manually verified by experts before inclusion in the dictionary. Acknowledging the limitation, we intend to improve the ability to detect some of the linguistic indicators using technologies such as machine learning. However, since machine learning requires training data – in this case, training data annotated by experts – it requires significantly more time and resources.

5.2 Normative Data

Psychological (and some other) characteristics are latent constructs that cannot be directly observed. These latent constructs have no absolute values and are meaningful only in relative terms. It is therefore more informative (not to say necessary) to provide the score of a specific individual on a specific characteristic in relation to the population, or a subpopulation. Therefore, PRAT, in connection to

any extracted score, provides the normative standing of that score. PRAT indicates not only the specific score of anger in a text but also that specific score´s position among the scores of a population.

5.3 Current Sample

In its current version, PRAT includes comparison samples from a variety of sources including major general population blogs and discussion forums, discussion forums related to Islam and extreme right-wing discussion forums. These sources and the number of posts for each individual and the final number of included individuals are shown in Table 1.

Forum/Source	Criteria	Number of users
Boards[a]	30 > = posts < = 400	8 612
Google blogs[a]	10 > = posts < = 200	2 355
Islamic Awakening[b]	-- posts > = 5	2 094
Turn to Islam	5 > = posts < = 1500	3 674
Reddit	50 > = posts < = 250	3 514
Stormfront	30 > = posts < = 400	7 477
School Shooters	all	50
Lone offenders	all	11
IS supporter blogs	all	47
[a]Normal population, [b]Muslim normal population		

Table 1: Datasets and criteria for selecting users from each forum

5.4 Profile and Profile Comparison

When extracting the necessary variables from the text, PRAT automatically creates a profile for each text/target person. In principle, any profile, regardless of whether it comprises personality data or any other indicators, contains three basic characteristics: *shape*, *scatter* (variability), and *elevation* (see Figure 1). Shape refers to the pattern of high and low scores across a profile; scatter refers to the degree of variation around average within a profile; and finally, elevation refers to the overall level or mean across elements of a profile. A profile can be compared with any other profile containing the same number of elements/variables. Profile similarity can be identified using various measures (e.g., Cronbach & Gleser, 1953; Livingston et al., 2003). PRAT uses the *Intraclass Correlation Coefficient* (ICC,

absolute agreement) as the main measure of comparison. The ICC is primarily sensitive to differences in profile shape, but also to differences in elevation and scatter (e.g., Edelbrock & McLaughlin, 1980). ICC values vary between 1.0, indicating perfect similarity between two profiles (profiles are literally overlapping) and 0.0 indicating no similarity at all. Notably, ICC can be negative and thus interpreted as high profile dissimilarity. However, in this case a negative correlation can be taken as zero similarity (Bratko, 1976).

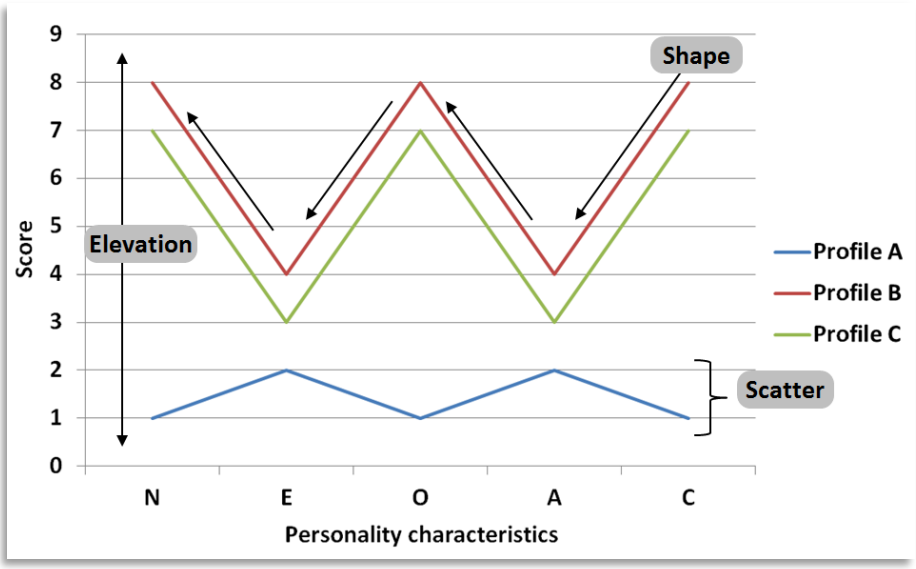

Figure 1: An illustration of the basic elements of a profile together with three profiles of two which rather similar. The intraclass correlations between A with B and C are zero and that between Profile B and C is .90 (B and C have the same shape and scatter but differ in elevation).

5.5 Threat Indicator/Case Classification

The ultimate goal of PRAT is to assist law enforcement and researchers in their threat assessment of written communication and for each analyzed text provide an indication of threat potential – that is, the probability that the individual who wrote the text is a potential security risk or is having a profile similar to a violent lone offender or a school shooter.

There are different ways to arrive at an indicator or a probability. One of these is using artificial intelligence techniques where a given text is compared with either neutral texts or texts with known signals of threat (i.e., a text written by someone who actually went through with an attack). These kinds of analyses demand a large amount of data, neutral as well as texts containing signals of threats. Thus, with

very few cases available, we are currently not able to conduct these analyses reliably. Another way of doing an assessment of threat potential, however, is to analyze profile similarity by comparing the extracted profile for a given text to all available cases in the sample (the datasets included in PRAT are listed in Table 1). Taking the average similarity between the profile of the extracted text and individual cases within each dataset would be informative of the writer. By doing so, we are able to see whether an individual behind a given text is, on average, similar to, for example, Google bloggers, lone offenders, and school shooters. We acknowledge the limitation of such a comparison but like any other method, this is to be subjected to a validity test, which will be presented in the case study below.

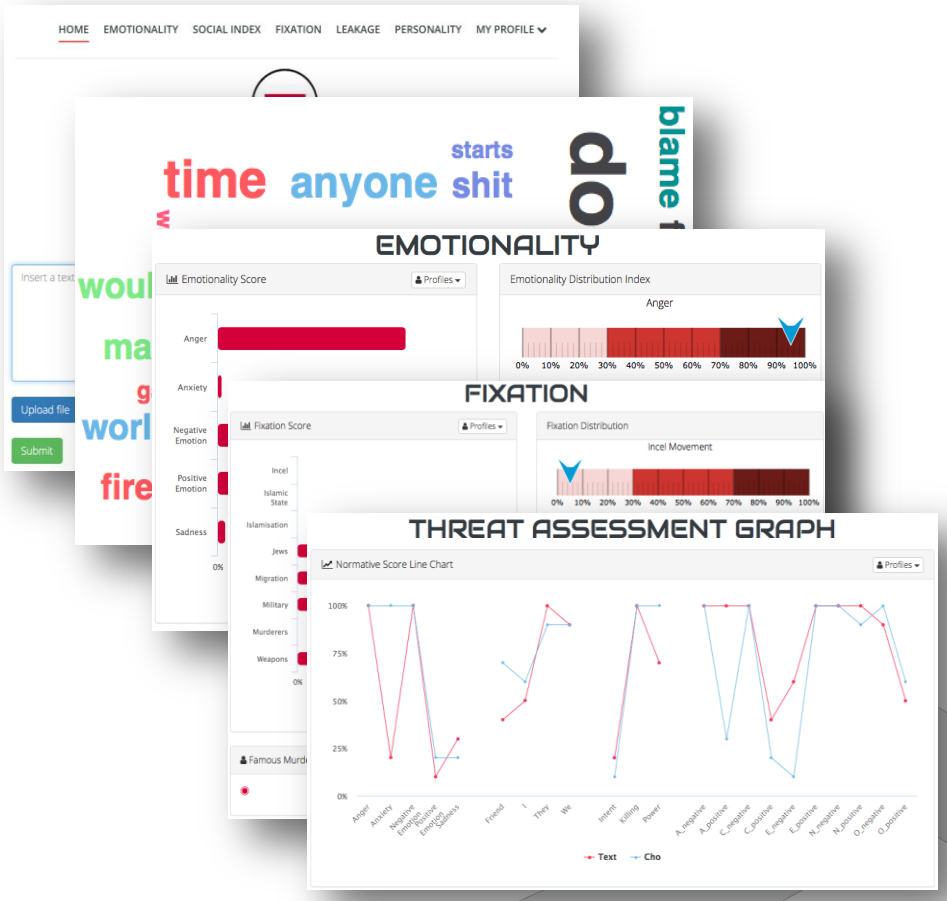

Figure 2: Screenshots of the profile risk assessment tool (PRAT)

5.6 PRAT Illustrated

In this section, we briefly present the user-end features of PRAT. Initially, users need to be authorized (contact the first author). When logged into the main page, users can either paste their target text or upload a file containing the text. As illustrated by Figure 2, PRAT provides a word cloud summarizing the text content. The word cloud is followed by a presentation of the values for each of the variables (e.g., the emotionality variable anger) followed by percentile score for anger in the text. The percentile in the case in Figure 2 means that the expression of anger in the text (or for that individual) is higher than 95% of all other texts (individuals) in the entire sample (see Table 1).

The analysis in PRAT ends by presenting the full profile extracted from the text. At this stage, the user is able to compare the extracted profile with a series of profiles stored in the PRAT database. The stored profiles include well-known cases of lone offenders and school shooters. More importantly, the user is also able to explore an index of similarity (intraclass correlation) between the extracted profile and profiles extracted from a variety of forums (see Table 1). A case study illustrating this is presented in the next section.

Case study 1 – Similarity assessment

In this section, we briefly illustrate some possibilities to examine the threat potential of a profile. Initially, we created a theoretical threat profile. We started by averaging the scores of the texts written by the 11 lone offenders that we had access to, arriving at an average profile of a lone offender. Next, based on theoretical reasoning, we set the scores on some variables in the theoretical profile to either maximum (e.g., anger, negative emotions, power) or minimum (e.g., positive emotions, friends). This step was done to make the theoretical profile more extreme. After establishing the theoretical profile, we calculated the intraclass correlation (using absolute agreement method) between the theoretical profile and each of the 27 834 individual profiles in our database.

As shown in Table 2, the average similarity with the theoretical profile is highest for lone offender, followed by that of the school shooter. These results provide some support for the validity of our theoretical profile. The correlations were averaged after Fisher's z transformation.

Forum/Source	Average Similarity	Number of profiles
Violent lone offender	.356	11
School Shooter	.203	50
Reddit	.134	3 514
IS supporter blogs	.092	47
Islamic Awakening	.067	2 094
Stormfront	.049	7 477
Google Blog	-.026	2 355
Boards	-.058	8 612
Turn to Islam	-.090	3 674
Total sample	.004	27 834

Table 2: Average similarity (intraclass correlation) between a theoretical profile with threat potential and profiles from various forums/sources

The distribution of similarity scores (intraclass correlation) between the theoretical profile and individuals within each group is depicted in Figure 3. As indicated by the Box plots in the figure, compared to other groups, the median score of the lone offenders indicated the highest similarity with the theoretical profile while all other groups are close to zero.

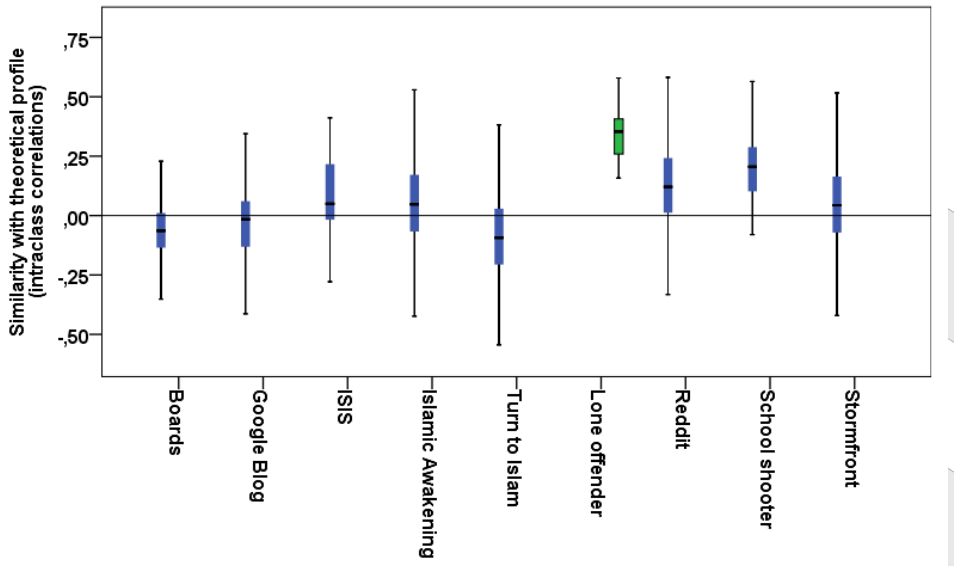

Figure 3: The distributions of similarity scores between the theoretical profile and profiles from various populations.

Having established some validity support for the theoretical profile on group level (average correlations), we aimed to do this at the individual level. Thus, first, we calculated the intraclass correlation between the theoretical profile and each of the 11 lone offenders. More importantly, for each lone offender, we randomly selected a profile from the remaining 27 823 profiles in the database (these profiles were from the following forums/sources: 2 from Google blogs, 3 from Stormfront, 2 from Boards, 2 from Turn to Islam, 1 from Islamic Awakening, and 1 from Reddit). Now, we calculated the intraclass correlation between the theoretical profile and each of the 11 randomly chosen profiles. As can be seen in Figure 4, similarity between the theoretical profile and the profiles of lone offenders was rather high ranging between .16 and .66 but low for the randomly selected profiles, ranging between .03 and -.18 (meaning no similarity at all).

The analyses above show that the theoretical profile is capable of fairly accurately distinguish between profiles of lone offenders and randomly selected profiles from what can be considered a general population. However, these analyses are not comparable with a classification task where profiles are classified according to their threat potential.

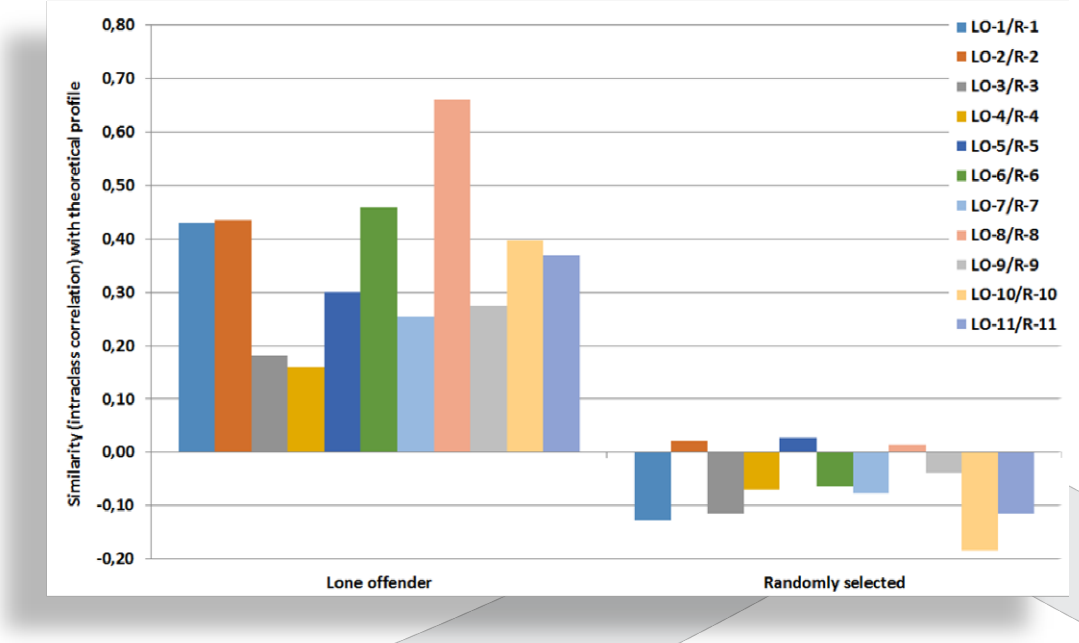

Figure 4: Similarity (intraclass correlation) between the theoretical profile and the profiles of 11 lone offenders (LO) and 11 randomly selected profiles.

Case study 2 – Classifying profiles

In this section, we briefly examine the possibility to classify profiles. To find a rule for classifying an observation as belonging to one of two different groups (e.g.,

normal population vs. lone offenders) we used sparse Poisson linear discriminant analysis classifier, which is a Bayes optimal classifier. Poisson classification (Witten, 2011), is especially suitable for variables consisting of annotated text data. The method assumes that the variables follow a Poisson distribution in both groups, but with different means (making discrimination between the two groups possible). This is a better approximation for the data than, for example, the normal distribution assumption in Quadratic Discriminant Analysis, due to, among other things, the probability of many low and null-observations leading to highly skewed data. When an observation is put into the formula, the likelihood of those values is then computed and the group with the highest likelihood is then chosen as the most likely for the observation to be from. This method, like all classification methods, will not always succeed. However, if the Poisson distribution assumption is justified, this method is optimal meaning that it classifies cases into the right group as often as possible. In addition to the Poisson distribution assumption being justified, it is necessary that there is sufficient number of observations.

Using the Poisson classification, we tested how well we could discriminate between pairs of samples/groups (e.g., Stormfront vs. Lone offenders, see Table 3) by estimating which group each individual originated from and compared this to the true group belonging. Table 3 shows the results of these analyses. Overall, the percentage of correctly grouped individuals was quite high, between 67.6 and 94.2%.

Model	Group 1	% Correctly classified	Group 2	% Correctly classified
1	Normal population[a]	93.4	Lone offender (LO)	90.9
2	Normal population[a]	94.2	School Shooter (SS)	68.0
3	Normal population[a]	93.7	IS supporter blogs (ISSB)	70.2
4	Normal population[a]	91.5	LO, SS, & ISSB	67.6
5	Normal population[a]	79.3	Stormfront	83.4
6	Muslim normal population[b]	81.7	IS supporter blogs	68.1
7	Stormfront	81.8	Lone offender	81.8
8	Stormfront	77.6	Reddit	80.2

[a]Google Blog & Boards, [b]Turn to Islam & Islamic Awakening

Table 3: Results of Poisson Linear Discriminant Analysis Comparing Two Groups in Each of the Eight Models

I think we should write that the same data is used for training and testing while classifying Group 1 vs Group 2. It is not common to use same data for training and test.

However, while the classification results above provide valuable information about the validity of our method, it does not generalize to new individuals/cases. We also wanted to know the sufficiency of our method to classify a new case. For this reason, we conducted the same analyses above but with one essential exception – we used the so-called leave-one-out method on the extreme/critical groups (e.g., Lone offenders). The basic idea of the leave-one-out method is to leave out one observation from a specific group and test whether that observation is classified into the right group using the classification rule acquired from the remaining observations. The method is repeated for all observations in the target group. This method is optimal when it comes to examining generalizability of the classification procedure. The results of our analyses are shown in Table 4. As can be seen in the table, the method fairly accurately classified Lone offender (81.8%) and Stormfront (83.4%) members while being only decent (66%) in classifying new school shooters and IS supporters, when compared with the normal population.

Should we write what does target group means?

Model	Group 1	Group 2 (Leave-One-Out)	% Correctly classified
1	Normal population[a]	Lone offender (LO)	81.8
2	Normal population[a]	School Shooter (SS)	66.0
3	Normal population[a]	IS supporter blogs (ISSB)	66.0
4	Normal population[a]	LO, SS, & ISSB	65.7
5	Normal population[a]	Stormfront	83.4
6	Muslim normal population[b]	IS supporter blogs	59.6
7	Stormfront	Lone offender	54.5
8	Reddit	Stormfront	77.5

[a]Google Blog & Boards, [b]Turn to Islam & Islamic Awakening

Table 4: Results of Leave-One-Out Comparing Two Groups in Each of the Eight Models with One Group Being the Target Leave-One-Out Group

6 Caveats and Future directions

We initiated PRAT as an expandable research tool allowing to integrated past and future advances in risk profile identification. Acknowledging that PRAT is at this initial stage, we still find the outcome of the studies presented above is promising. Case study 1 showed that a theoretical threat profile is possible to implement and can provide a fairly accurate indication of threat in a given text. Case study 2,

showing fairly high classification accuracy provide further support for the potential of PRAT as toll for threat assessment.

While we are optimistic about the potential of the tool, we still predict that much work is needed before considering the instrument fully reliable. In this regard, we argue that process of developing PRAT has just started, and we will be able to introduce new functions in the near future. For example, we will be able to introduce new statistical approaches to improve the precision of identifying risk profiles. We are also in the process of introducing tailored artificial intelligence solutions. Also, we believe that the development of a tool like PRAT will benefit from being used by both researchers and law enforcement professionals. For example, we regard the filed knowledge of law enforcement professionals to be crucial for such an instrument. It is therefore our aim to open PRAT for authorized users and allowing them to (noncompulsory) contribute with data (profiles/text) to be used to improve the tool.

While optimistic about the current and future effectiveness of PRAT, we still see some challenges that need to be highlighted. Specifically, there is no perfect threat assessment or classification, and the risk of false positives (innocent classified as offender) and false negatives (offenders classified as innocent) is always evident and needs to be considered. This risk is especially evident for tools under development, like PRAT. That being said, we still hope that PART can be developed to assist law enforcement professionals in their efforts to making the world a safer place.

References

Allport, G. W. (1961). *Pattern and growth in personality*. New York: Holt, Rinehart, and Winston.

Atran, S. (2003). Genesis of suicide terrorism. *Science, 299*, 1534-1539.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*, 762-765.

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition, 55*, 30 - 40.

Bechara, A. (2005). Decision making, impulse control and loss of willpower to resist drugs: A neurocognitive perspective. Nature Neuroscience, 8, 1458 -1463.

Buss, A. H. & Perry, M. (1992). The aggression questionnaire. *Journal of personality and social psychology, 63*, 452.

Cronbach, L. J. & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological bulletin, 50*, 456-473.

Cyders, M. A. & Smith, G. T. (2008). Emotion-based dispositions to rash action: positive and negative urgency. *Psychological bulletin, 134*, 807-828.

Dolan, R. J. (2007). The human amygdala and orbital prefrontal cortex in behavioural regulation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*, 787-799.

Dreisbach, G. (2006). How positive affect modulates cognitive control: The costs and benefits of reduced maintenance capability. *Brain and Cognition, 60*, 11-19.

Edelbrock, C. & McLaughlin, B. (1980). Hierarchical cluster analysis using intraclass correlations: A mixture model study. *Multivariate Behavioral Research, 15*, 299-318.

Europol, *EU Terrorism Situation and Trend Report* (TE-SAT) (2018).

Fiske, A., P. (2002). Socio-Moral Emotions Motivate Action to Sustain Relationships. *Self and Identity, 1*, 169-175.

Gerges, F. A. (2005). *The Far Enemy: Why Jihad went global* (Second ed.): Cambridge University Press.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York, NY: Pantheon.

Hempel, A. G., Meloy, J. R. & Richards, T. C. (1999). Offender and offense characteristics of a nonrandom sample of mass murderers. *Journal of the American Academy of Psychiatry and the Law Online, 27*, 213-225.

Kaati, L., Shrestha, A. & Cohen K. (2016). Linguistic Analysis of Lone Offenders Manifestos. In proceedings of Cybercrime and Computer Forensic (ICCCF).

Khosrokhavar, F. (2005). *Suicide Bombers*. London: Pluto Press.

Livingston, R. B., Jennings, E., Reynolds, C. R. & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18*, 487-507.

Livingstone, A.G., Spears, R., Manstead, A.S.R., & Bruder, M. (2009). Illegitimacy and identity threat in (inter)action: Predicting intergroup orientations among minority group members. *British Journal of Social Psychology, 48*, 755-775.

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology, 79*, 602-616.

Matsumoto, D., Hwang, H. C. & Frank, M. G. (2016). The effects of incidental anger, contempt, and disgust on hostile language and implicit behaviors. *Journal of Applied Social Psychology, 46*, 437-452.

Matsumoto, D., Hwang, H. S., & Frank, M. G. (2012). The role of emotion in predicting violence. *FBI Law Enforcement Bulletin, 81*, 1–11. Retrieved from https://leb.fbi.gov/2012/january/the-role-of-emotion-in-predicting-violenc

McCrae, R. R. & Costa Jr., P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516.

McCrae, R. R. & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

Meloy, J. R. & O'Toole, M. E. (2011). The concept of leakage in threat assessment. *Behavioral Sciences & the Law, 29*, 513-527.

Meloy, J. R., Hempel, A. G., Mohandie, K., Shiva, A. A., & Gray, B. T. (2001). Offender and offense characteristics of a nonrandom sample of adolescent mass murderers. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*, 719-728.

Meloy, J.R. & Gill, P. (2016). The lone-actor terrorist and the TRAP-18. *Journal of Threat Assessment and Management, 3*, 37-52.

Meloy, J.R., Hoffmann, J., Guldimann, A. & James, D. (2012). The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences and the Law, 30*, 256-279.

Molho, C., Tybur, J.M., Güler, E., Balliet. D., & Hofmann, W. (2017). Disgust and anger relate to different aggressive responses to moral violations. *Psychological Science, 28*, 609-619.

Monahan, J. (2016). The individual risk assessment of terrorism: Recent developments. In G. LaFree & J. Freilich, *The handbook of criminology of terrorism* (eds.). Hoboken, NJ: Wiley.

Obaidi, M., Bergh, R. & Akrami, N. (2018). *The normal personalities of extremists - examining violent and non-violent defense of Islam*. Manuscript in preparation.

Obaidi, M., Bergh, R., Sidanius, J. & Thomsen, L. (2018b). The Mistreatment of My People: Victimization by Proxy and Behavioral Intentions to Commit Violence Among Muslims in Denmark. *Political Psychology, 39*, 577-593.

O'Toole, M. E. (2000). The School Shooter: A Threat Assessment Perspective. Critical Incident Response Group (CIRG), National Center for the Analysis of Violent Crimes (NCAVC), FBI Academy, Quantico, VA 22135.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*, 934-952.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. CT New York: Bloomsbury Press.

Pennebaker, J. W., & Chung, C.K. (2008). Computerized text analysis of Al-Qaeda transcripts. In K. Krippendorff & M. Bock (Eds.), A content analysis reader. Thousand Oaks, CA: Sage.

Richardson, L. (2006). *What terrorists want. Understanding the terrorist threat*. London: John Murray.

Ricolfi, L. (2005). Palestinians, 1981-2003. In D. Gambetta (Ed.), *Making sense of suicide missions* (pp. 77-129). New York: Oxford University Press.

Roberts, B. W., Kuncel, N., Shiner, R., N., Caspi, A. & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socio-economic status, and cognitive ability for predicting important life outcomes. *Perspectives in Psychological Science, 2*, 313-345.

Rozin, P., Haidt, J., McCauley, C., Dunlop, L. & Ashmore, M. (1999). Individual differences in disgust sensitivity: Comparisons and evaluations of paper-and-pencil versus behavioral measures. *Journal of Research in Personality, 33*, 330-351.

Shiv, B., Loewenstein, G., & Bechara, A. (2005). The dark side of emotion in decision-making: When individuals with decreased emotional reactions make more advantageous decisions. *Cognitive Brain Research, 23*, 85-92.

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), Lyon, 2017, pp. 1-5.

Stern, J. (2003). *Terror in the name of God*. New York: HarperCollins.

Sternberg, R. J. (2003). A duplex theory of hate: Development and application to terrorism, massacres, and genocide. *Review of General Psychology, 7*, 299-328.

Stucke, T. S. & Sporer, S. L. (2002). When a grandiose self-image is threatened: Narcissism and self-concept clarity as predictors of negative emotions and aggression following ego-threat. *Journal of personality, 70*, 509-532.

Tausch, N., Becker, J. C., Spears, R., Christ, O., Saab, R., Singh, P., & Siddiqui, R. N. (2011). Explaining radical group behavior: Developing emotion and efficacy routes to normative and nonnormative collective action. *Journal of Personality and Social Psychology, 101*, 129-148.

Tausczik, Y. & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24-54.

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*, 377-378.

Witten, D. (2011). Classification and clustering of sequencing data using a Poisson model. *The Annals of Applied Statistics, 5*, 2493-2518.

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*, 363-373.