2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Introducing Digital-7

Threat Assessment of Individuals in Digital Environments

Amendra Shrestha
Uppsala University
Uppsala, Sweden
Email: amendra.shrestha@psyk.uu.se

Nazar Akrami
Uppsala University
Uppsala, Sweden
Email: nazar.akrami@psyk.uu.se

Lisa Kaati
Uppsala University
Uppsala, Sweden
Email: lisa.kaati@it.uu.se

***Abstract*—One of the most challenging threats towards the security of the society is attacks from violent lone offenders, individuals that act alone or with minimal help from others without any economic gains or direct orders from organizations. Over the past few years, several terror attacks have been accompanied by manifestos published on social media platforms that outline ideology, motivation, and in some cases tactical choices. The trend in publishing manifestos and other communication on social media sites before committing an attack has increased the need for threat assessment in digital environments. Most existing methods for threat assessment are developed to be used in offline settings where information about an individual is accessible and cases where the individual is present and can answer questions. In this paper, we present seven indicators that can be used to assess the potential threat of violence based on digital communication only. The seven indicators are designed to be used when analyzing texts and can be seen as a complement to other risk assessment protocols.**

I. Introduction

One of the most challenging threats towards the security of the society is attacks from violent lone offenders - individuals that act alone or with minimal help from others, without economic gains or direct orders from organizations. In this paper, we use the term *lone offender* when referring to violent lone offenders - individuals that commit targeted violent attacks alone or with minimal help from others. This includes mass murderers, solo-terrorists, single-issue offenders, and school shooters. However, being able to identify violent lone offenders before they commit an act of violence is a constant struggle for security services and law enforcement authorities. The problem is that detection is very difficult - partly because it is hard to infiltrate a solo actor and partly because there are so many different underlying reasons why a person chooses to commit an act of violence. Threat assessment of individuals is one approach that could help in the process of detection and to reduce violent attacks. For this purpose, a number of risk assessment protocols have been developed. Some of these protocols focus on assessing the risks that someone who has already committed acts of violence will do it again while others assess the risk that someone will commit a violent attack for the first time [19].

In recent years, digital environments have played an important role in many attacks conducted by lone offenders: the mass shootings that took place during 2018 and 2019 in the United States (the San Diego-area synagogue shooting and the El Paso shooting), New Zealand (the Christchurch mosque shooting), Norway (the Baerum mosque shooting), and Germany (the Halle mosque shooting) were all preceded by communication in digital environments.

While most research about threat assessment of individuals have focused on offline settings where there is accessible information about an individual or where the individual is present and can answer questions, some researchers have focused threat assessment on digital environments. The question is, how we can identify certain indicators that are present in digital environments that can be used for threat assessment. For example, the expressions of emotions in extremist environments were studied in [9], and techniques for measuring the level of hate in digital environments are described in [4]. The ability to identify individuals with a radicalized mindset (a style of understanding and relating to the world that has often been observed among violent extremists) in digital environments is described in [28] and the possibilities to identify certain warning behaviors in social media have been examined by [15]. In threat assessment, warning behaviors for targeted or intended violence play an important role since they can be viewed as indicators of increasing or accelerating risk. Warning behaviors are described by [21] as any behavior that "precedes an act of targeted violence, is related to it, and may, in certain cases, predict it." While the presence of warning behaviors is commonly analyzed in the behavior of an individual, Cohen et al. argue that the warning behaviors that are likely to be the most easily detectable in written online communication are leakage, fixation, and identification. These ideas were later implemented by [13], [16] and by [11]. Another approach towards analyzing written communication is a tool called PRAT (Profile Risk Assessment Tool). PRAT can be used for risk assessment of digital communication and is further described in [27] and [1]. PRAT uses a number of variables to create a risk score of a given text. The score is computed automatically and leaves little or no room for an analyst to interpret the results.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

In this paper, we present a new threat assessment method, the digital-7. The digital-7 consists of seven indicators that can be used for threat assessment of written communication. These indicators can be assessed either manually, with computer support, or by a combination of both. The benefit of using the seven variables in threat assessment is that the assessment can be done in a structured way and that the analyst can always verify computer generated assessments manually or receive a second opinion by completing the manual assessment with these made by computer. Importantly, the combination of the characteristics of the digital seven makes this assessment method unique and opens for assessment of potential risk individuals by using digital communication and without direct in-person contact/interviews.

Outline This paper is outlined as follows. In Section II, we describe the seven variables that we propose for digital risk assessment of individuals. Section III described how the seven variables can be assessed. Section IV describes how we can identify the set of variables using text analysis, and then we analyze the presence of our seven variables on a set of recently published manifestos. A discussion of the results is presented in Section V, and finally, some conclusions and directions for future research are presented in Section VI.

II. Seven Variables for Threat Assessment of Written Communication

Existing research on lone offenders and threat assessment largely build on information from medical or prison journals and interviews with people who have lived close to them [14]. For an overview of risk assessment tools see [31]. However, several of the common psychological factors that are considered in threat assessment are expressed in communication that several lone offenders have left behind on social media. Specifically, these factors can be extracted from text using various text mining techniques.

Based on research in psychology, linguistics, and computer science, we describe seven variables that can be used for threat assessment of written communication. The seven variables for threat assessment of written communication can be used to assist professionals in risk assessment of potential lone offenders. Each of the seven variables is described below.

A. Anger

Emotions are important for predicting behavior and actions since 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 world surrounding them. Emotions have also been emphasized by various scholars as strong predictors of violent extremism.

For example, Pennebaker and Chung [24] studied texts written by al Qaida and found them to be relatively high in emotion compared to other texts. They also noticed that the relationship between positive and negative emotions differed from what is usually found in natural conversation. The natural conversation contains almost twice as many positive words than negative emotion words, but the al Qaida-texts expressed more negative emotions, mostly anger words. We argue that anger is one of the key emotions in threat assessment and thus one of our seven variables.

B. Grievance

Grievance is often based on a perception of having been wronged or treated unfairly or inappropriately. This may result in a desire or even a mission to right the wrong and reach deserved justice/status. Research suggests that lone-actor terrorists and mass murderers may be better conceptualized as lone-actor grievance-fueled violence [7] due to the presence of grievance that could be personal grievances or political grievances.

In [6], a model of the pathway to violence is suggested. The model proposes six different milestones on the path to violence and one of these is a grievance. Grievance among a set of 155 mass murderers was studied in [12]. The study showed the presence of some of the following grievances:

- Sense of being treated unfairly at work
- Desire for revenge for being bullied at school
- Belief in being unjustly denied fame
- Lack of success in forming romantic relationships
- Belief that a failing grade in high school resulted in adult unemployment
- Sense of being targeted for harassment by acquaintances and family

While the source of grievance may differ among individuals who commit targeted violence, the presence of grievance is something that can be found among many of them. Some research [7] even suggests that the term lone-actor grievance-fueled violence (LAGFV) offenders should be used instead of lone-actor terrorists and/or mass murderers. Thus, propose that the expressions of grievance as one of the seven variables for threat assessment.

C. Othering

The use of pronouns in natural language has been examined in various studies. Pronouns have been linked to different aspects of personality and emotion [23]. Frequent use of third-person plural (they, them, etc.) in a group suggests that the group is defining itself to a high degree by the existence of an oppositional group [24]. The use of pronouns is a reliable indicator of negative identification with an out-group, something that MacCauley and Moskalenko consider [5] a precursor of terrorism. As Pennebaker and Chung point out: wars, prejudice, and discrimination are based on the psychological distinction between "us" and "them" [25]. The use of third-person plural pronouns in the analysis of online groups such as American Nazis and animal rights groups has been proven to be the best single predictor of extremism according to [24]. In previous work, Kaati et al. [15] analyzing lone offender manifestos showed that the use of third-person plural was significantly higher in texts written by violent lone offenders compared to a normal group. Therefore, we propose othering as one of the seven variables for threat assessment.

D. Leakage

Leakage is the communication of intent to harm a specific target, and it can be done using written statements, verbal statements to the public, verbal statements to family/friends. Data suggest that leakage commonly occurs in cases of targeted violence, ranging from school shootings to attacks on public figures. Leakage can be intentional or unintentional, and more or less specific with regards to the act.

Studies on public figure attacks and assassinations have, according to Meloy and O'Toole, found a suggestive pattern of leakage, in which an attack has often been preceded by indirect, conditional, or direct threats aimed at people associated with the target, or bizarre or threatening communication to politicians, public figures, or police forces [21]. However, according to the same study, threats are typically not posed directly at the target. In different studies, the occurrence of pre-attack leakage ranges from 46% to 67% (and even higher for school shootings [26]). Thus, we propose leakage of attack plans as one of the seven variables for threat assessment.

E. Military terminology

The use of military terminology may be related to the warning behavior identification that is defined by Meloy et al. [20] as a behavior indicating a desire to be a "pseudo-commando", have a warrior mentality, closely associated with weapons or other military or law enforcement paraphernalia, identify with previous attackers or assassins, or identify oneself as an agent to advance a particular cause.

As described in [8], the warning behavior identification can be divided into two subcategories: identification with radical action and identification with a role model. Offenders often tend to identify themselves as a kind of warrior, a person who is prone to use structured violence for a "higher cause". In these cases, the use of military terminology and a strong interest in weapons and military strategies can be observed. We suggest the use of military terminology as one of the seven variables for threat assessment.

F. Influence

The Internet and social media have made it easier than ever to find like-minded people to communicate with and to find spaces for subcultures and groups that glorify previous offenders. There are digital environments where school shooters and mass murderers are seen as heroes and sources of inspiration.

The warning behavior identification [20] can, as previously mentioned, also be identification with a role model. In these cases, it is not uncommon that school shooters, mass murders, and solo-terrorists are mentioned in social media communication. We have identified mentions of previously known lone offenders as one of the seven variables for threat assessment.

G. Personality - the dark triad

Generally, personality refers to the psychological makeup of an individual. A more specific definition refers to "the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought" [2]. Personality is also assumed to comprise a set of psychological traits that are relatively enduring [17]. 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 (e.g., stability, change, and importance for everyday life behavior)) 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 (e.g., number of trait and hierarchical order of traits). There are, thus, a variety of models promoting various structures for and numbers of traits.

Also, there are a number of personality traits that fall outside the traditional models of personality. A set of three among these traits is what is known as the dark triad [22]. The dark triad includes Machiavellianism, narcissism, and psychopathy. These traits encompass a wide range of characteristics, such as manipulative behavior (Machiavellianism), grandiosity, dominance, and superiority (narcissism), and impulsivity, and thrill-seeking along with lack of empathy (psychopathy) [22]. Thus, the dark triad encompasses an index of negative/dark and subclinical aspects of personality. These traits are overlapping, but conceptually distinct [10]. Most importantly, research has shown that the traits within the dark triad are predictors of a wide range of antisocial behavior [10]. We consider the dark triad as one of the seven variables for threat assessment.

III. Assessing the Variables

When making a manual assessment of the seven variables, the analyst needs to assess the occurrence of each variable in the text. This is a demanding task since psychological characteristics are latent constructs and have no absolute values, and are meaningful only in relative terms. This means that when making a manual assessment, the analyst needs to make a judgment based on his/her previous knowledge and experience. To overcome this challenge, text analysis technologies can be used to aid analysts in their assessments.

Detecting psychological constructs using text analysis has been done using several different approaches. The most common approach is to use existing psycho-linguistic tools such as LIWC [30] or various machine learning approaches [16].

The most common way forward is to create a set of dictionaries where each dictionary represents a theme or a psychological variable. For each variable, we use a dictionary containing words that represent the psychological variable, and then we count the relative frequencies of such words in the text material that is analyzed. The frequencies of the dictionary words in the text are standardized (divided by total word counts), producing a score for each variable that represents its relative frequency of occurrences in the text. This gives an indication of the presence of each variable. The scores for each variable can be compared to the scores of the normal population.

A. Creating the dictionaries

When working with dictionary-based analysis, it is important to consult experts with significant domain knowledge of the studied environment. We have consulted a number of experts to create a set of dictionaries for the variables that we present here. The experts create dictionaries according to the steps outlined below:

1) Create a clear definition of what is intended to measure/examine.
2) Specify the environment the dictionary should be applied to - dictionaries are domain-dependent.
3) Examine the environment by spending time reading and analyzing the language and expressions that are used.
4) Create a preliminary list of keywords that are considered to be important for what you want to measure.
5) Create a word embedding of the environment that you are analyzing. A word embedding is a learned representation for text where words that have the same meaning have a similar representation. The word embedding provides words that are semantically similar to the preliminary list of keywords created in the previous step.
6) Exclude irrelevant words that do not relate to what should be measured as well as words that have different meanings.
7) Find out if certain words are missing by going back to the definition.
8) Expand the words in the word space again.
9) Examine the list and exclude irrelevant words, but try to be a liberal/generous and also include words that you are less sure of (see it as a way to have control words that you want to investigate further).
10) Invite at least two reviewers and ask the reviewers to answer the following question about each word: is the word relevant to the definition? and If any words are missing.
11) Examine the reviewer's answers and revise the list - include any suggestions from the examiner if appropriate.

B. Dictionaries for the seven variables

We have created our own dictionaries for six variables: anger, grievance, othering, leakage, military terminology, and influence. The dictionaries and some sample words are presented in Table I. Each dictionary is created using the approach described above. The domain we consider is a contemporary internet jargon in line with the meme and image culture that characterizes the internet in late 2010. Therefore, we have created a set of word embeddings trained on some of the environments that we are interested in analyzing. This includes embeddings for the incel community (that engages people that live in involuntary celibacy), radical nationalistic environments, and large discussion forums including Reddit and 4chan. For the dark triad, we used the same approach as [29].

There are several possible critiques of using a dictionary-based approach when analyzing textual data. One issue is that the meaning of words can be context-dependent, which means that words may have several different meanings depending on the context, something that was noticed in [18]. For example, the word "execute" can be used in the meaning "to kill", but also in the meaning "to carry out a task". By only considering the frequency of occurrence of the word "execute" with no regard to the context will lead to inaccurate analysis.

TABLE I
THE DICTIONARIES WE USE AND SOME SAMPLE WORDS

Variable	Example words
Anger	war, attack, bastards, destroy, fucking
Grievance	failed, destroy, lost, despair, dying
Othering	they, them, themselves, their
Leakage (violence)	killing, gun, attack, firearms
Military terminology	solider, invasion, warfare
Influence	Breivik, Columbine, Tarrant, Oswald

Another issue of dictionary-based analysis is that the dictionaries are often defined as a priori, without any consideration of the domain that they are supposed to analyze. This introduces bias, and the analysis may be sensitive to vocabulary variation that is introduced by slang words, different spellings, and domain-specific terminology. Inability to handle vocabulary variation increases the risk of under-estimating, which will lead to inaccurate analysis.

To avoid misinterpretations, we strongly suggest extracting sentences that can be manually analyzed. A manual inspection of sentences containing words from the dictionaries is important since there are some issues that need to be considered when using dictionary based text analysis.

The score for each variable is created by counting the occurrences of words from the corresponding dictionary. For the normal group (see below), we have limited the analysis to 20,000 characters. Before the analysis, all texts are preprocessed by removing punctuations and html-links and converting all letters to lower case. The score for each variable is created by dividing the frequency of words from each dictionary by the total amount of words in the text.

C. Normal group

Many psychological characteristics are latent constructs that cannot be directly observed. These latent constructs have no absolute values and are meaningful only in relative terms. When applying automatic technologies for threat assessment, it is necessary to add a normal group for comparison. Previous research has compared writings by lone offenders with different populations, such as non-violent activists [3] and standard control writings and emotional writings [15].

When examining the presence of the seven variables, we will analyze the score of an individual in relation to a population or a sub-population. We have created a set of comparison samples from a variety of sources consisting of blogs and discussion forums. The comparison samples are selected to

provide a snapshot of the Internet and communication that takes place on the Internet. The comparison samples are from a wide range of sources; some samples are from digital milieus where known lone offenders have been active, and some are from more mainstream environments. The samples are from discussion forums with a focus on Islam (Turn to Islam and Islamic Awakening), incel forums (Incel, Lookism, Lookmax), Counter jihad webpages (Gates of Vienna), white supremacy forums, and webpages (Stormfront, VNN Forum, Daily Stormer), blogs on different topics (Google blogs), racist forums (Niggermania), and forums with a wide range of discussions (Reddit and Boards). The wide range of sources from forums that can be considered as extreme or deviant due to their expression of, for example, hate speech and hate propaganda, were selected to challenge our analysis.

Table II shows the sub-population we use as a normal group: the different sources from where we have collected our samples and the number of samples from each source. Our comparison sample consists of writings from a total of 52,498 individuals. We refer to the comparison group using the terminology normal group.

TABLE II
DATA USED FOR THE NORMAL GROUP.

Source	Number of users
Boards	25,587
Daily Stormer	1,383
Gab	2,179
Gates of Vienna	1,327
Google blogs	3,391
Incel	1,512
Islamic Awakening	1,044
Lookism	44
Looksmax	986
Niggermania	455
Reddit	9,874
Stormfront	2,206
Turn to Islam	1,333
VNN Forum	1,177

IV. Testing the Seven Variables

A. Subjects

To test our seven variables will assess a set of lone offenders that have committed violent attacks within the last six years and communicated their intentions or believes in written text. The subjects and the texts we are assessing are described briefly below.

- **Brenton Tarrant** Conducted two consecutive mass shootings that occurred at mosques in a terrorist attack in Christchurch, New Zealand, on 15 March, 2019. A manifesto was posted online before the attack.

- **Patrick Crusius** Conducted a mass shooting at a Walmart store in El Paso, Texas, United States on 3 August, 2019. A manifesto was posted on 8chan shortly before the attack.

- **Stephan Balliet** Charged for the Halle synagogue shooting that occurred on 9 October, 2019 in Halle, Germany. Before the attack, a manifesto was posted online.

- **John Earnest** The Poway synagogue shooting occurred on April 27, 2019. Before the attack, a text was published on the repository Pastebin and posted on the message board 8chan.

- **Dylan Roof** Conducted the Charleston church shooting on June 17, 2015 where nine people were killed. Before the attack, a manifesto was published on a website.

- **Elliot Rodger** Conducted the Isla Vista killings on May 23, 2014. Before the attack, Rodger posted a video on Youtube. We have used a transcript of the video.

B. Experiments

We started by calculating mean scores and standard deviations for the normal population (N = 52,489). Thus, we arrived at a mean and standard deviation for each of the variables in digital-7. Next, we calculated the mean scores on each variable for each of the six lone offenders. The results are presented in Table III.

Finally, for each of the seven variables means, we conducted a single/one-sample t-test with the population means (and standard deviation) as the test variable (e.g., mean anger of the normal group) and the related score (e.g., anger score for one lone offender) for each of the lone offenders as the test value (/reference/expected value). Table IV shows the t-values for each lone offender compared with the normal population. A negative t-value indicates that the lone offenders scored higher than the normal population and a positive value indicates that the lone offenders scored lower than the normal population. Table V shows the result of the t-test where a + sign means that the lone offender had a higher score than the population while a – sign means that the lone offender scored lower than the normal group. All comparisons were significant.

As can be seen in Table V these analyses showed that the lone offenders had significantly ($p < 0.001$, at least, df = 52,497) higher scores than on the digital seven on 39 of the total 42 comparisons. Thus, three of the comparisons showed that the lone offenders had lower scores as compared to the normal population. T-values varied between 7 and 1,983 (mean t-value = 323).

We also tested whether the mean scores of all included lone offenders differed from that of the normal population on each of the digital seven variables. The results of these analyses showed that lone offenders scored higher on all digital seven variables, as compared to the normal population.

TABLE III
MEAN OF THE DIGITAL-7 VARIABLES FOR THE NORMAL POPULATION AND THE LONE OFFENDERS.

	Anger	Greivance	Influence	Military	Othering	Dark triad	Leakage
Normal population							
Mean	0.004777	0.001786	0.000168	0.000658	0.010143	0.027086	0.000953
Standard deviations	0.004982	0.002685	0.000741	0.001670	0.006093	0.010356	0.002007
Lone Offenders							
Brenton Tarrant	0.016164	0.004712	0.000596	0.003102	0.020518	0.037735	0.002237
Dylan Roof	0.009331	0.005274	0.000406	0.000000	0.019878	0.036646	0.001014
Elliot Rodger	0.009589	0.006849	0.001370	0.000000	0.010959	0.041096	0.002055
John Earnest	0.015426	0.003403	0.006579	0.000907	0.015653	0.041289	0.004310
Patrick Crusius	0.012315	0.007800	0.001642	0.002874	0.017241	0.029146	0.007184
Stephan Balliet	0.019608	0.002179	0.002179	0.001089	0.005447	0.039216	0.007081
Mean - Lone Offenders	0.013739	0.005036	0.002129	0.001329	0.014949	0.037521	0.003980

TABLE IV
T-VALUES FROM SINGLE/ONE-SAMPLE T-TESTS COMPARING THE MEAN OF EACH OF THE DIGITAL SEVEN FOR THE NORMAL POPULATION WITH RESPECTIVE SCORES FOR LONE OFFENDERS. NEGATIVE T-VALUES INDICATE THAT THE LINE OFFENDERS SCORED HIGHER AND POSITIVE VALUES (**BOLD**) INDICATE THAT THE LONE OFFENDERS SCORED LOWER THAN THE NORMAL POPULATION. ALL VALUES ARE SIGNIFICANT AT $p < 0.001$, AT LEAST, $df = 52{,}497$).

Subject	Anger	Grievance	Influence	Military	Othering	Dark triad	Leakage
Brenton Tarrant	-524	-250	-132	-335	-390	-236	-147
Dylan Roof	-209	-298	-73	**90**	-366	-212	-7
Elliot Rodger	-221	-432	-372	**90**	-31	-310	-126
John Earnest	-490	-138	-1983	-34	-207	-314	-383
Patrick Crusius	-347	-513	-456	-304	-267	-46	-711
Stephan Balliet	-682	-33	-622	-59	**177**	-268	-700
All	-412	-277	-606	-92	-181	-231	-346

TABLE V
RESULTS OF THE COMPARISONS BETWEEN SCORES OF VARIOUS LONE OFFENDERS AND THE POPULATION MEAN ON DIGITAL SEVEN VARIABLES. A + SIGN MEANS THAT THE LONE OFFENDER HAD A HIGHER SCORE THAN THE POPULATION WHILE A – SIGN MEANS THAT THE LONE OFFENDER SCORED LOWER THAN THE NORMAL GROUP. ALL COMPARISONS WERE SIGNIFICANT.

Subject	Anger	Grievance	Influence	Military	Othering	Dark triad	Leakage
Brenton Tarrant	+	+	+	+	+	+	+
Dylan Roof	+	+	+	-	+	+	+
Elliot Rodger	+	+	+	-	+	+	+
John Earnest	+	+	+	+	+	+	+
Patrick Wood Crusius	+	+	+	+	+	+	+
Stephan Balliet	+	+	+	+	-	+	+
All	+	+	+	+	+	+	+

V. DISCUSSION

In this paper, we have introduced digital-7: seven variables that can be used for threat assessment of online communication. The seven variables can either be analyzed manually or with text analysis. We have used text analysis to test the variables, and we assess a set of six lone offenders that committed violent attacks within a six-year period. All lone offenders that we assess were active on social media and posted their intentions and/or believes on social media before their attacks.

We examined to what extent the seven variables are present in the communication from our group of lone offenders. We compare the results with a normal group consisting of communication from 52,489 individuals from a variety of places on the internet.

The results show that three of our subjects had higher scores on all variables compared to the normal group. The other three lone offenders had higher scores of six of the variables compared to the normal group. The group as a whole (all six subjects together) had a higher presence of all variables compared to the normal group.

The results indicate that digital-7 is a possible tool for

threat assessment of written communication, although it is important to stress that a digital risk assessment should only be seen as one component in threat assessment when digital communication is available.

VI. Conclusion and Directions for Future Work

The goal of digital risk assessment is to assist law enforcement and analysts in their threat assessment of written communication in digital environments. While there is still much work to do in the area of digital threat assessment, the variables we have introduced here have support in previous research and can be checked either manually or by using automated text analysis.

It is important to stress that automatic text analysis can not entirely replace a human analyst and should only be used in combination with human analysis. When using machine learning models and classification such as in [16], the analyst needs to depend entirely on computerized methods, and the result is difficult to interpret. Such methods can be used in the first step for detecting potential individuals at risk. They should not be used without human analysis.

There is still a lot of work to do when it comes to digital threat assessment. An important consideration when using linguistic analysis for threat assessment is that the results of the analysis should be transparent and understandable for the analyst. The seven variables that we present here can either be analyzed manually, by using automatic text analysis, or with a combination of both.

For future work, we will apply our seven variables for digital threat assessment on more cases to further examine the reliability and the predictive power of these variables.

References

[1] N. Akrami, A. Shrestha, M. Berggren, L. Kaati, M. Obaidi, and K. Cohen. Assessment of risk in written communication: Introducing the profile risk assessment tool (PRAT). https://www.europol.europa.eu/publications-documents/assessment-of-risk-in-written-communication, 2018.

[2] G. W. Allport. *Pattern and growth in personality*. Holt, Reinhart and Winston, 1961.

[3] S. J. Baele. Lone-actor terrorists' emotions and cognition: An evaluation beyond stereotypes. *Political Psychology*, 38(3):449–468, 2017.

[4] T. Berglind, B. Pelzer, and L. Kaati. Levels of hate in online environments. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 842–847, 2019.

[5] M. C and M. S. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, 20(3), 2008.

[6] S. Calhoun, F.S. ans Weston. Contemporary threat management: A guide for identifying, assessing, and managing individuals of violent intent. San Diego, CA:Specialized Training Services, 2003.

[7] C. Clemmow, P. Gill, N. Bouhana, J. Silver, and J. Horgan. Disaggregating lone-actor grievance-fuelled violence: Comparing lone-actor terrorists and mass murderers. *Terrorism and Political Violence*, 0(0):1–26, 2020.

[8] K. Cohen, F. Johansson, L. Kaati, and J. C. Mork. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, 26:246–256, 2014.

[9] L. Figea, L. Kaati, and R. Scrivens. Measuring online affects in a white supremacy forum. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 85–90, 2016.

[10] A. Furnham, S. C. Richards, and D. L. Paulhus. The dark triad of personality: A 10year review. *Social and Personality Psychology Compass*, 7(3):199–216, 2013.

[11] T. Grover and G. Mark. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the Thirteenth International Conference on Web and Social Media, ICWSM 2019, Munich, Germany, June 11-14, 2019*, pages 193–204, 2019.

[12] J. G. Horgan, P. Gill, N. Bouhana, J. Silver, and E. Corner. Across the universe? a comparative analysis of violent radicalization across three offender types with implications for criminal justice training and education. National Criminal Justice Reference System, 2016.

[13] F. Johansson, L. Kaati, and M. Sahlgren. *Detecting linguistic markers of violent extremism in online environments*. Artificial Intelligence: Concepts, Methodologies, Tools, and Applications. IGI Global, 2017.

[14] L. Kaati, K. Cohen, and N. Akrami. Ensamagerande våldsverkare. profiler, riskbedömningar och digitala spår. In *Swedish Defence Research Agency FOI-R–4736–SE*, 2019.

[15] L. Kaati, A. Shrestha, and K. Cohen. Linguistic analysis of lone offender manifestos. In *International Conference on CyberCrime and Computer Forensics (ICCCF)*, 2016.

[16] L. Kaati, A. Shrestha, and T. Sardella. Identifying warning behaviors of violent lone offenders in written communication. In *ICDM workshop SoMeRis*, 2016.

[17] R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60 2:175–215, 1992.

[18] M. Mehl, M. Robbins, and S. Holleran. How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences*, 3(2), 2013.

[19] J. Meloy and P. Gill. The lone-actor terrorist and the TRAP-18. *Journal of Threat Assessment and Management*, 1(3):37–52, 2016.

[20] J. Meloy, J. Hoffmann, A. Guldimann, and D. James. The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, 30(3):256–279, 2012.

[21] J. Meloy and M. E. O'Toole. The concept of leakage in threat assessment. *Behavioral Sciences and the Law*, 29:513–527, 2011.

[22] D. L. Paulhus and K. M. Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563, 2002.

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

[24] J. W. Pennebaker and C. K. Chung. Computerized text analysis of al-Qaeda transcripts. In K. Krippendorf and M. A. Bock, editors, *The Content Analysis Reader*. Sage, London, UK, 2008.

[25] J. W. Pennebaker and C. K. Chung. Language and social dynamics. In *Technical Report 1318*. University of Texas at Austin, Texas, USA, 2012.

[26] A. Semenov, J. Veijalainen, and J. Kyppö. Analysing the presence of school-shooting related communities at social media sites. *Int. J. of Multimedia Intelligence and Security*, 1:232 – 268, 01 2010.

[27] A. Shrestha, L. Kaati, and N. Akrami. PRAT - a tool for assessing risk in written communication. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4755–4762, 2019.

[28] A. Shrestha, L. Kaati, and K. Cohen. Extreme adopters in digital communities. *Journal of Threat Assessment and Management*, 2020.

[29] C. Sumner, A. Byers, R. Boochever, and G. J. Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *ICMLA (2)*, pages 386–393. IEEE, 2012.

[30] Y. Tausczik and J. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), March 2010.

[31] L. van der Heide, M. van der Zwan, and M. van Leyenhorst. The practitioner's guide to the galaxy-a comparison of risk assessment tools for violent extremism. In *ICCT Research paper*, 2019.