2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Predicting Targeted Violence from Social Media Communication

Lisa Kaati
Stockholm University
Stockholm, Sweden
Email: lisa.kaati@dsv.su.se

Amendra Shrestha
Mind Intelligence Lab
Uppsala, Sweden
Email: amendra@mindintelligencelab.com

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

***Abstract*—For decades, threat assessment professionals have used structured professional judgment instruments to make decisions about, for example, the likelihood of violent behavior of an individual. However, with the increased use of social media, most people use online digital platforms to communicate, which is also the case for potential violent offenders. For example, many mass shootings in recent years have been preceded by communication in online forums. In this paper, we introduce methods to identify markers of the warning behaviors *Leakage*, *Fixation*, *Identification*, and *Affiliation* and examine their discriminant validity. Our results show that violent offenders score higher on these markers and that these markers were present among a significantly higher proportion of violent offenders as compared to the normal population. We argue that our method can be used to predict potential planned, purposeful, or instrumental targeted violence in written communication. Automated methods for detecting warning behavior from written communication can serve as a complement to traditional threat assessment and provides unique opportunities for threat assessment beyond traditional methods.**

I. Introduction

Risk assessment of individuals involves collecting data about an individual to assess the probability of certain behavior. Analysts who are conducting risk assessments commonly use various structured professional judgment instruments. There are a large number of different instruments that are used to assess various forms of risks. Most risk assessment instruments are designed for specific uses, and there is a number of different instruments that are constructed to assess, for example, risk of workplace violence, risk of sexual violence, or risk of abuse of children.

Several instruments are designed to measure the risk of violence in general. One such instrument is HCR-20 [5] a structured professional judgment instrument with guidelines for violence risk assessment and management. Another instrument that is built on HCR-20 but that also considers ideological motivation is the Violent Extremism Risk Assessment (VERA-2) [6]. VERA-2 is specifically developed to assess the likelihood of future violence by an identified offender who has been convicted of unlawful ideologically motivated violence [21]. The Terrorist Radicalization Assessment Protocol (TRAP-18) [17] is another structured professional judgment instrument that is used in risk assessments of persons of concern for acts of terrorist violence [8]. TRAP-18 is specifically used to assess the risk of individuals engaging in lone-actor terrorism to assist threat assessors with prioritizing cases for risk management [21].

Most existing risk assessment instruments have been developed for assessments in hospitals, prisons, and other institutions and are often only applicable in that particular environment. That is, though there is scientific support for the operation of a particular risk assessment instrument in psychiatry, the ability to accurately forecast the patient's risk of violence in society is limited. Most research concerning threat assessment of individuals is focused on offline settings with accessible information about an individual or where the individual is present and can answer questions. However, there have been some attempts to detect individuals that pose a risk of committing violent attacks. One example where several online behaviors of individuals are combined is presented in [2] and in [9], [12]. Other ways to detect violent lone offenders combining online behavior and other traits are presented in [10].

A. Research question

Our research is focused on detecting markers of four patterns of behavior associated with targeted violence: ***Leakage***, ***Fixation***, ***Identification***, and ***Affiliation*** in social media communication. Hence, we first present methods for automatic detection of these markers from written communication. Next, we examine the discriminant validity of the markers by comparing the presence of these markers in texts written by violent offenders and texts written by a normal population. While the texts from violent offenders consist of writings that have been communicated prior to their violent attack, the texts from the normal population consist of samples of writings from 14 different online platforms. Compared to the normal population, we expect the violent offenders to score significantly higher on all four markers. Specifically, we examine the discriminant validity of our markers, one by one and jointly, by calculating

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

the area under a receiver operating characteristic (ROC) curve. We expect the discriminative ability (ROC) of our markers to be significantly higher than the chance level (0.50).

II. Threat Assessment and Warning Behaviors

Warning behaviors for targeted or intended violence play a central role in threat assessment. Warning behaviors are described by Meloy et al. [15] as any behavior that "precedes an act of targeted violence, is related to it, and may, in certain cases, predict it." The risk assessment protocol TRAP-18 includes eight proximal warning behaviors, and ten distal characteristics [17]. While the presence of warning behaviors is commonly analyzed in the behavior of an individual, Cohen et al. [3] argue that some warning behaviors are detectable in written online communication. In this study, we focus on the warning behaviors Leakage, Fixation, and Identification (from TRAP-18). We also include behavior that we denote as Affiliation.

A. Leakage

The warning behavior leakage is the communication of intent to harm a specific target. Leakage can be done using written statements, verbal statements to the public, and statements to family and friends [16]. 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 regard to the act. Studies on public figure attacks and assassinations have, according to Meloy and O'Toole [16] 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 [16]. However, according to the same study, threats are typically not posed directly at the target. In a study of 198 lone actors [22], 86% of the sample communicated their radical or extremist convictions to others, and 58% of the sample gave others the idea that they were involved in suspicious and potentially violent activities. In the cases of school shootings, the numbers are even higher [23].

B. Fixation

The warning behavior fixation is defined by Meloy et al. [15] as "any behavior indicating an increasingly pathological preoccupation with a person or a cause, for instance increasing perseveration on the object of fixation, increasingly strident opinion, or increasingly negative characterization of the object of fixation". This definition indicates that fixation might be rather challenging to extract from social media communication unless we have data that supports analysis over time (as in [7]). However, fixation on a specific object might be possible to detect even when using static data.

C. Identification

Identification is a warning behavior defined as a behavior indicating a desire to be a "pseudo-commando", have a warrior mentality, closely associate 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 by Cohen et al. [3] the warning behavior identification can be divided into two subcategories: identification with radical action and identification with a role model. Offenders often identify themselves as a warrior, a person 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. When the identification is with a role model, it is common that the role models are school shooters, mass murderers, and solo-terrorists.

D. Affiliation

According to Rahman, Zheng, Meloy [20] acts of terrorism, mass murders, and hate crimes are often motivated by Extreme Overvalued Beliefs that are shared by others in a person's cultural, religious, or subcultural group. In the digital space, online forums can be seen as subcultural groups or a subculture. This is, for example, the case for the online subculture Incel (involuntary celibacy), where men blame women and society in general for their lack of romantic success [18]. The presence of extreme overvalued beliefs is complex to measure from written communication. However, a possible approach is to measure the extent to which an individual (their text) is affiliated with a specific group (cultural, religious, or subcultural) – for details, see below. Therefore, we name the indicator Affiliation rather than extreme overvalued beliefs.

III. Approaches to Detect Warning Behaviours in Written Communication

Detecting psychological constructs using text analysis has been done using several different approaches. The most common approach is to use text analysis tools such as LIWC [19], or various machine learning approaches [14], [26] or combinations of both [25], [27]. In this work, we have used a combination of different text analysis methods and machine learning.

Psychological constructs are latent and cannot be directly observed. Another characteristic of psychological constructs is that they have no absolute values and are meaningful only in relative terms. Thus, it is necessary to use norm data (a normal population) for comparison. Previous research has compared writings by lone offenders with different populations, such as non-violent activists [1], standard control writings and emotional writings [13], and white supremacy discussion forums [14], [28]

Our approach toward detecting warning behaviors that may indicate a risk of targeted violence is built on linguistic markers. Each warning behavior is represented by a linguistic marker. Each marker is assigned a score, and each such score is compared to a normal population consisting of a sample of texts from a wide range of forums and social media posts. For example, a common way forward is to create a dictionary where that represents a theme or a psychological variable.

For each variable, the relative frequencies of words from the dictionary that are present in the text material that is analyzed are computed. The frequencies of the dictionary words in the text are standardized (divided by total word counts for that specific text), 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 in the text. The scores for each variable for a specific individual/text/ can be compared to the scores of other individuals/texts or the norm data, that is, the normal population. Before describing how we detect the different warning behaviors, we will provide a description of the data set, the normal population, and the violent offenders.

IV. Data

To create a normal population, we have used data from 15 different digital environments and a sample of violent offenders (see Table IV). Some of the environments included are environments with content moderation and user rules that forbid some content, free speech environments where almost all kinds of content are allowed, and environments that allow and, in some cases, promote extremist ideologies. The reason for selecting such a wide range of environments is that we want the data to represent a normal population and to make a more robust evaluation of our method to detect valid risk indicators. The number of users included from each environment is limited to a representative sample (95% confidence interval and margin of error equal to 5%) from available users in each environment. Each environment is briefly described below, and some statistics about each environment can be found in Table I.

All text/data used in our analysis were collected from English-speaking forums and are in English. The Python library langdetect[1] was used to ensure that the language for each text is English.

- Boards (Boards.ie) is a large Internet forum in Ireland. Users on the forum discuss a wide variety of topics such as entertainment, jobs and work life, politics, and personal relationships. The forum is moderated, and offensive posts are removed.
- Blogger (or Google Blogs) is a blog publishing service. The blogs are hosted by Google and accessed from a subdomain of blogspot.com. Most of the blogs in our dataset are concerned with personal interests, news, fashion, and photography.
- Reddit is a discussion forum that calls itself ”the front page of the internet”. Reddit has more than 500 million visitors every month and is one of the most visited sites on the internet. Reddit is open to all discussion topics. Since 2008, users have been given the opportunity to start subdivisions themselves where specific topics are discussed, so-called subreddits. Currently, there are over one million subreddits. Discussions are moderated, which means that many of the most offensive or provocative posts are removed.
- The Daily Stormer is an American white power and anti-Semitic news site founded in 2013. On The Daily Stormer users can both comment and read the content anonymously. The majority of the posts consist of racist and anti-Semitic films and images with accompanying comments. The Daily Stormer has been shut down several times for violating different Internet service provider's terms of use, for example after publishing insulting comments about the woman who was killed in connection with the Unite the Right gathering in Charlottesville in 2017.
- Gab is a social network where users can write messages of up to 300 characters called gabs. Gab, which was created in 2016 as a freedom of expression-friendly alternative to Twitter, welcomes users who have been banned from other social networks. Several radical-nationalist organizations are represented on Gab. There is virtually no moderation of posts on Gab, which gives room for radical and pro-violence voices.
- Gates of Vienna is a digital meeting place for the European counter-jihad movement. Gates of Vienna publishes posts from a variety of writers and contains descriptions of the historical development of the counter-jihad movement and information about European counter-jihad gatherings.
- Stormfront is one of the most well-known white supremacy discussion forums. The forum was founded in 1995 and the forum describes its members as follows: “We are a community of racial realists and idealists. We are White Nationalists who support true diversity and a homeland for all peoples. Thousands of organizations promote the interests, values and heritage of non-White minorities. We promote ours.” In addition to an active discussion forum, Stormfront provides daily radio broadcasts, blogs, and chat opportunities.
- The Vanguard News Network Forum (VNN Forum) was founded in 2000 by a former National Alliance member as an uncensored forum for “whites”. VNN Forum is sometimes referred to as the place for users who are banned from Stormfront [11].
- Ni**ermania is a website with an associated discussion forum where condescending jokes and racist comments about Black people are published. Ni**ermania started in 2003 and is divided into sub-forums with headings such as “nigger crime”, “coontacts” and “uppity niggers”. Most posts on Ni**ermania talk about Black people as underdeveloped and innately inferior to whites. Social problems such as crime and unemployment are described as an issue of race. The side is dominated by aggressive racist jokes and images.
- Incels The forum incels is the largest active digital environment for incels (involuntary celibates), which according to its own statement has about 11 000 registered members and 3.3 million posts. The forum was founded in November 2017 as an alternative to the then just closed sub-forum /r/incels on Reddit, which had been one of the

[1] https://pypi.org/project/langdetect/

TABLE I
NORMAL POPULATION

Source	*n*	Number of Words				
		M	*SD*	Min	Max	*Mdn*
Boards	25 587	2 435	1 271	83	4 774	2 610
Reddit	9 874	2 073	1 199	51	4 874	1 794
Google blogs	3 391	3 051	1 118	248	13 317	3 528
Stormfront	2 206	3 487	181	2 558	4 163	3 496
Gab	2 179	1 751	1 071	67	4 167	1 544
Incels	1 512	2 831	1 202	529	4 072	3 621
Daily Stormer	1 383	2 636	1 157	471	4 097	3 347
Turn to Islam	1 333	2 387	1 217	462	4 299	2 693
Gates of Vienna	1 327	1 888	1 185	379	4 103	1 416
VNN Forum	1 177	2 683	1 106	186	4 056	3 303
Islamic awakening	1 044	2 539	1 167	384	3 932	3 215
Looksmax	986	2 930	1 155	542	4 111	3 664
Neogaf	2 187	2 067	1 050	39	3 986	1 899
Ni**ermania	455	2 499	1 252	506	4 086	2 837
Lookism	44	1 811	1 150	567	3 949	1 342
Lone Offenders	68	18 947	99 278	97	81 1967	1 590

n = number of users selected from each environment (representative sample)
M= mean number of words
SD = standard deviation
Min= minimal number of words
Max = maximal number of words
Mdn= median
Total number of users in the population (*N*) = 52 542

most popular incel environments with about 1.2 million posts.

- Lookism was founded in 2015 and is one of the oldest active incel environments on the web with over 10 000 members and 3.8 million posts. Lookism has been an important part for the development of the incel culture. Lookism is the place where the world view and jargon typical of incels have been developed. Today, lookism includes general discussions, appearance advice and methods to improve one's relationship status. Unlike incels the forum has no special requirements on who may become a member.
- Looksmax is a sister forum to Incels and run by the same owner. It is intended as a forum for men who want to discuss options to improve their appearance, with the goal of increasing their success with women. The forum does not allow any female members. Officially the members are not required to be incels, and anyone who is interested in appearance improvements is allowed to post. Nevertheless, the discussions are characterized by incel jargon and incel-inspired theories of appearance. Looksmax was founded in 2018 and according to its own statement it has almost 1.4 million posts and 3,400 members.
- Turn to Islam is an English language forum with the goal of "correcting the common misconceptions about Islam". Turn to Islam is considered as a lifestyle network for Muslims.
- Islamic Awakening is an English language Islamic forum with members from the UK and some other countries. The forum identifies itself as "dedicated to the blessed global Islamic awakening".
- NeoGAF is a video game message board. The forum was launched on April 4, 2006. Discussions are related to gaming, the video game industry, and gaming communities.
- The violent offenders our data set included a sample of 68 violent offenders who have committed acts of targeted violence.

V. METHOD

A. Detecting leakage

To detect leakage, we have used the same approach as in [24] and developed a dictionary that contains words related to weapons and violence. Examples of words are *killing*, *glock, gun, ammo, attack*, and *firearm*.

B. Detecting identification

To detect identification, we created two different dictionaries: one that can detect the presence of military terminology and one that can detect mentions of previous offenders. Examples of military terminology are *solidier*, *combat*, *invasion*,

general, *warfare*, *squadron*, and, *uniform*. Previous offenders are individuals that have committed targeted violence. Examples of what is included in our dictionary are the Columbine shooters Eric Harris and Dylan Klebold, the Unabomber Ted Kaczynski, and active shooters such as Brenton Tarrant.

C. Detecting fixation

Grover and Mark [7] present an approach to detect three sub-components for fixation on a group level. The three sub-components that they focus on are:

1) an increasing perseveration on a person or cause,
2) an increasingly negative account of the object of fixation, and
3) an increasingly strident opinion and angry emotional undertone

To detect the three different sub-components, Grover and Mark use a combination of term Frequency, TF-IDF (term frequency-inverse document frequency), LIWC, and HateSonar (detection of hate speech mentioned in [4]).

When detecting fixation behavior, we focus only on a pathological preoccupation with a person or a cause. This means that we do not know the topic or subject of fixation. Instead, our way of detecting fixations considers the number of mentions of a person, a topic, or a cause.

To detect fixation, we first clean the data by removing all personal pronouns, stop words, and words less than two characters. We also perform text lemmatization to capture the root of a word. Lemmatization removes inflectional endings, returns the base or dictionary form of a word, and can also extract the meaning of words. To detect fixation, we extract the 15 most frequently used (lemmatized) words and count the frequency of how often these words are used.

D. Detecting Affiliation

To detect affiliation, a set of machine learning models that can recognize different subcultures and ideologies were created. The affiliations we consider are:

- ***Incel*** (involuntary celibate) - an online subculture consisting of heterosexual men who are active on incel forums and blame women and society for their lack of romantic success.
- ***Jihadist ideology*** - ideology promoted by terrorist groups such as the so-called Islamic state.
- ***Counter jihad*** - a movement that considers Muslims, in particular Muslims living within Western boundaries, a potential threat to Western society and culture.
- ***White supremacy*** - a belief that white people are superior to those of other races and thus should dominate them.
- ***Alt-right*** (alternative right) - a online phenomenon that can be described as a loosely connected far-right white nationalist movement.

For each subculture/ideology, we collected online data that represents the subculture/ideology. For Incel, we used the forums Incel, Lookism, and Looksmax, for Jihadist ideology we used text from the Islamic state produced magazines Dabique and Rumiyah, for white supremacy we used a set of selected white supremacy comments from Stormfront and VNN forum, for counter jihad we used the website Gates of Vienna, and finally, for Alt-right we used the Daily Stormer as the positive class.

Before training our models, the data was cleaned. Each character was converted to lowercase, and English stop words were removed. We use a bag of words model with term frequency-inverse document frequency (TF-IDF) features and a linear support vector machine (SVM) to build a set of models (one for each subculture/ideology). For each subculture/ideology, we create two different classes: a positive class and a negative class. The positive class contains data that is specific to the subculture/ideology, i.e. data from digital environments that are specific to the subculture/ideology that we want to recognize.

The negative class for each subculture/ideology consists of all data in Table I except the positive class. When training the SVM model, hyperparameter tuning was done using grid search to estimate the optimal parameters of the classifier. TF-IDF was used as a feature, and English stop words were removed from the text. While building the TF-IDF vocabulary features, words that appear in more than 20% of the documents and words that appear in less than 0.1% of the documents were removed. This process eliminates the most common words and words that seldom appear in the corpus.

Importantly, as a measure of Affiliation, we calculated the average affiliation score across the five different subcultures/ideologies for each text/user. The minimum score for the average would be 0 (zero) and the maximum 100.

VI. Results

To examine whether the behavior markers discriminate between violent offenders and the normal population we have implemented our approaches of detecting markers. For each text in our dataset, we have applied our approach. Next, we calculated the mean scores on each of the warning behaviors for the offenders and the normal population and conducted a *t*-test to examine whether these means were significantly different (See Table II). As can be seen in Table II, the results are inline with our predictions showing that the violent offenders, on average, score significantly higher on all warning behaviors than the normal population.

More importantly, we conducted receiver operating characteristic (ROC) analyses to test whether the indicators discriminated between the 68 violent offenders and the normal population. These analyses showed that the area under the curve (AUC), varied between 0.628 and 0.839 and departed significantly from chance level (0.50) for all four markers, see Table III. While all indicators show significant discriminant validity, Leakage and Affiliation were more powerful to distinguish between the offenders and the normal population.

Subsequently, using the coordinates of the curve, from the receiver operating characteristic analyses, we identified the optimal discrimination threshold for each of the markers for our binary variable (offenders vs. normal population). Based on the threshold we generated four new variables where each

TABLE II
MEANS AND T-TEST RESULTS FOR WARNING BEHAVIORS

	Violent Offender		Normal Population			
Warning behavior	*M*	*SD*	*M*	*SD*	*t*	*p*
Leakage	**0.00430**	0.00441	0.00100	0.00153	18.141	<0.001
Fixation	**0.00510**	0.00212	0.00450	0.00255	2.145	<0.02
Identification	**0.00150**	0.00201	0.00040	0.00095	9.587	<0.001
Affiliation	**14.57240**	12.71105	4.32210	7.67401	10.996	<0.001

Means (*M*) in boldface denote the higher scores for violent offenders
t = independent sample t-test
p = two-tailed p-value

TABLE III
RESULTS OF RECEIVER OPERATING CHARACTERISTIC

		95% Confidence Interval	
Warning behavior	Area Under Curve	Lower Bound	Upper Bound
Leakage	0.839	0.786	0.893
Fixation	0.628	0.561	0.695
Identification	0.689	0.614	0.763
Affiliation	0.826	0.793	0.859

TABLE IV
CLASSIFICATION RESULTS BASED ON THE OPTIMAL DISCRIMINATION THRESHOLD FROM THE RECEIVER OPERATING CHARACTERISTIC ANALYSES

	Proportion (%) above the threshold			
Warning behavior	Offender	Normal Population	*z*	*p*
Leakage	80.9	20.1	228.7	<.0001
Fixation	50.0	25.5	162.6	<.0001
Identification	42.6	7.5	170.8	<.0001
Affiliation	86.8	31.8	226.3	<.0001

z = z-test for two independent proportions
p = two-tailed p-value

user was assigned 0 (zero if they were below the threshold and 1 (one) if they were above the threshold. Thus, this is an implementation of the classification implied by the receiver operating characteristic analyses and we use this to test our prediction on the proportion of cases being positive (above the threshold) and negative (below the threshold) within each group (offenders and normal population). We also conducted a series of z-score tests for two population proportions to see whether the proportions of cases classified as "above the threshold" among the offenders and the normal population were significantly different. The classification results, presented in Table IV showed that the proportion classified as "above the threshold" in the offender sample was significantly higher than those in the normal population.

Finally, we examine the discriminate validity of the four markers jointly. We created a composite score by taking the average of the four classification variables and submitted the composite score to the receiver operating characteristic analyses. The analysis showed a ROC score of 0.885 (95% confidence interval, 0.849-0.921), which is higher than the ROC scores for all individual indicators. Again, also for the composite measure, we identified the optimal discrimination threshold for our binary variable (offenders vs. normal population) and generated a new variable where each user was assigned zero (0) if they were below the threshold and 1 (one) if they were above the threshold. The results of these analyses showed that 85 percent of the offenders scored higher than the threshold, compared with the normal population where only 22 percent scored higher.

VII. CONCLUSIONS

In this work, we aimed to introduce methods to identify markers of the warning behaviors Leakage, Fixation, Identification, and Affiliation and to test their discriminant validity, one by one and jointly. We used a combination of machine learning and linguistic methods to construct the markers. The statistical analyses revealed that the markers had

good discriminant validity, especially the markers for Leakage and Affiliation. The composite measure combining all four markers was even more powerful and could correctly classify 85 percent of the offenders in our dataset.

We argue that this outcome is promising, and digital threat assessment could already by these variables contribute to risk assessment in the digital era. Despite our optimism, one could ask why we did not do better. A critical reader could say that we, after all, had 15 percent false negative and 22 percent false positive. Here, we can point out several reasons, for example, the small number of offender cases in the offender sample and the size of the dataset in general. However, the most important reason is the diversity in our dataset, both when it comes to the offender cases and the normal population. For example, we decided not to leave out any cases when it comes to offenders, despite poor data quality and short text. Also, the offender cases had very diverse backgrounds and included school shooters (with varying motives), Jihadists, racially or ethnically motivated offenders, Incels, and some also some cases that cannot be classified. To match this diversity, we decided to include a very diverse normal population. One could, for example, argue that the data texts/users from Ni**ermania are not really "normal." However, we decided to introduce the diversity in the normal population – both to match the diversity in the cases of but also to enable a model that can classify beyond the very normal sample. To give an example, ROC analyses base on the data from the offenders (n = 68, positive class) and the internet forum Boards (n = 25 587, negative class) only revealed ROC scores of 0.892, 0.652, 0.756 and 0.955 for Leakage, Fixation, Identification, and Affiliation respectively. All these figures are higher than those for the model with a diverse normal population reported in the results section (0.839, 0.628, 0.689, and 0.826, respectively). We argue that a model based on diverse data would work better in applied settings – that is, for prediction in real life. Also, there is not sufficient amount of data within each category of offenders (e.g., School shooters, Jihadists, racially or ethnically motivated offenders, Incel) to create a model for a specific type of ideologically motivated offender. Thus, we set to create robust indicators and a model for use in real life rather than a model that shows a good fit but has little implication with reality. Another issue that can be raised is the classification method we used for detecting affiliation. One way to improve the results for detecting affiliation could be the use deep-learning-based methods.

The aim of automatic threat assessment of written communication is to assist threat assessment analysts and law enforcement in the assessment process. While there is still much work to do in the development of methods for digital threat assessment, our results show that digital threat assessment can be a powerful tool. However, while we are optimistic regarding the technical aspects of digital threat assessment, considering the number of false positives, we think that this is one of several tools in the toolbox of threat assessment. In the process of constructing the markers, we were repeatedly reminded that the interpretation of the results when using machine learning models remains a challenge. Insight like these shows that researchers need to examine the optimal position of digital methods in the process of threat assessment. Regardless, it is important to stress that automatic text analysis cannot entirely replace a human analyst, but it can assist in making threat assessment in the digital space faster, more reliable, and in some cases, also provide information beyond human perception.

References

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

[2] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson. Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, 2(11), 2013.

[3] 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.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May 2017.

[5] K. S. Douglas, S. D. Hart, C. D. Webster, H. Belfrage, L. S. Guy, and C. M. Wilson. Historical-clinical-risk management-20, version 3 (hcr-20v3): Development and overview. *The International Journal of Forensic Mental Health*, 13(2):93–108, 2014.

[6] D. Elaine Pressman and J. Flockton. Calibrating risk for violent political extremists and terrorists: the vera 2 structured assessment. *The British Journal of Forensic Practice*, 14(4):237–251, 2012.

[7] 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.

[8] A. Guldimann and J. R. Meloy. Assessing the threat of lone-actor terrorism: the reliability and validity of the trap-18. *[Einschätzung der Bedrohung durch Einzeltäter-Terrorismus: Reliabilität und Validität des TRAP-18]. Forensische Psychiatrie, Psychologie, Kriminologie*, 14(2):158–166, 2020.

[9] 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.

[10] L. Kaati, K. Cohen, and N. Akrami. Lone offenders. profiles, risk assessment and digital traces (report is on swedish). In *Swedish Defence Research Agency (FOI)*, FOI-R–4736–SE, 2019.

[11] L. Kaati, K. Cohen, and B. Pelzer. *Heroes and scapegoats : right-wing extremism in digital environments*. European Commission and Directorate-General for Justice and Consumers. Publications Office, 2021.

[12] L. Kaati and F. Johansson. Countering lone actor terrorism : Weak signals and online activities. In *Understanding Lone Actor Terrorism : Past experience, future outlook, and response strategies*, pages 266–279. 2016.

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

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

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

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

[17] J. R. Meloy. *TRAP-18. Terrorist radicalization assessment protocol. User Manual*. North Tonawanda NY: Global Institut of Forensic Research, 2017.

[18] B. Pelzer, L. Kaati, K. Cohen, and J. Fernquist. Toxic language in online incel communities. *SN Social Sciences*, 2021.

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

[20] T. Rahman, L. Zheng, and J. Meloy. Dsm-5 cultural and personality assessment of extreme overvalued beliefs. *Aggression and Violent Behavior*, 60, Sept. 2021.

[21] RTI International. Countering violent extremism: The application of risk assessment tools in the criminal justice and rehabilitation process. literature review. 2018.

[22] B. Schuurman, E. Bakker, P. Gill, and N. Bouhana. Lone actor terrorist attack planning and preparation: A data-driven analysis,. *Journal of Forensic Sciences*, 63(4):1191–1200, 2018.

[23] 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.

[24] A. Shrestha, N. Akrami, and L. Kaati. Introducing digital-7 threat assessment of individuals in digital environments. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 720–726, 2020.

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

[26] A. Shrestha, L. Kaati, and K. Cohen. A machine learning approach towards detecting extreme adopters in digital communities. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 1–5, 2017.

[27] A. Shrestha, L. Kaati, and K. Cohen. Extreme adopters in digital communities. *Journal of Threat Assessment and Management*, 7(1-2):72–84, 2020.

[28] I. van der Vegt, M. Mozes, B. Kleinberg, and P. Gill. The grievance dictionary: Understanding threatening language use. *Behavior Research Methods*, (53):2105–2119, 2021.