Heliyon 6 (2020) e05066

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Screening for potential school shooters through the weight of evidence

Yair Neuman [a,*], Yoav Lev-Ran [a], Eden Shalom Erez [b]

[a] The Department of Cognitive and Brain Sciences and the Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[b] Independent Researcher, Israel

ARTICLE INFO

Keywords:
Psychology
School shooters
Lone wolf perpetrator
Screening
Weight of evidence
Methodology

ABSTRACT

The challenge of automatically screening for potential school shooters involves several difficulties. In this paper, we present a simple and interpretable methodology for screening for potential school shooters through (1) the psychological textual signature of the shooter and (2) Jaynes approach for measuring the weight of evidence. We have tested our proposed approach on a dataset of texts written by shooters and non-shooters alike (N = 5047). Our major finding is that the methodology can successfully support the screening for potential shooters in interpretable way. The major implication for stakeholders is that there is great potential in developing screening systems for improving the safely of schools. However, developing such a system is a project that must be actualized within an integrated system of "command and control".

1. Introduction

There are certain practical contexts where we may be interested in screening for an event whose incidence in the population is extremely low. In the United States, the incidence of school shooting is extremely low, but the salience of these rare events in the media, and the anxiety they evoke, is relatively high. For example, in 2019 the total injuries and fatalities resulted from school shootings was 128.[1] In comparison, statistics indicate that 310 people are shot in the United States every *day*.[2] Since schools are supposed to provide students with a safe environment, it is clear why school shooting attracts disproportionate attention and why despite the low *actuarial* prevalence of school shooting, there is a place to offer new directions for preventing this form of violence.

The default response to school shooting may be to deploy more security personnel and improve surveillance methods within schools (e.g. cameras). However, preventive measures seem to be the most *proactive* means of improving safety of schools because "targeted prevention" of a potential shooter might save the efforts of neutralizing him in real-time without taking the risk involved in the unexpected attack.

Since solo perpetrators sometimes produce a text (i.e., a *manifesto*) that accompanies or precedes the violent act (e.g. Knoll, 2010), it may be worthwhile to examine whether such texts bear a tell-tale signature, or early warning signs, that may be identified automatically and used to screen for the shooter in advance. Indeed, it is reasonable to hypothesize that such warning signs may appear long before the violent act is perpetrated (e.g. Simons and Meloy, 2017), and may therefore be identified in advance. We use the term "texts" in the widest possible sense to include visual images uploaded to Instagram but here we focus on written texts only. Since the texts produced by the perpetrator may be accessed through social media, analyzing them for early-warning signs and the adoption of preventive steps appears to be a reasonable approach. For instance, Eric Harris—one of the perpetrators of the famous Columbine High School Massacre—kept a journal that included some disturbing signals (Neuman, 2016a). In hindsight, such warning signs could have been used, if not for prediction, then at least for (1) risk analysis; (2) putting in place preventive legal measures; and (3) the application of psycho-therapeutic tools, if possible. Therefore, identifying a potential "signature" of a shooter, may be an important step for identifying potential shooters in advance.

Manual screening of a massive number of subjects in a bid to spot early warning signs is practically unfeasible, and therefore automatic approaches must be used (Neuman, 2016b). However, with a few

* Corresponding author.
E-mail address: yneuman@bgu.ac.il (Y. Neuman).

[1] https://www.chds.us/ssdb/incidents-by-injuries-and-fatalities-annually-2010-present/.

[2] https://www.bradyunited.org/key-statistics?gclid=Cj0KCQiApt_xBRDxARIsAAMUMu-alTQ8oHpbKPq8HSskroN4LBdEqF4oDMxpLwLo3Dt9f2JuzLhQ8uIaAr2dEALw_wcB.

https://doi.org/10.1016/j.heliyon.2020.e05066
Received 8 May 2020; Received in revised form 24 August 2020; Accepted 22 September 2020

exceptions (e.g. Neuman et al., 2015), the challenge of *automatically* identifying the signature of "lone-wolf" perpetrators, such as school shooters, has rarely been addressed even by using tools of Machine Learning (ML) and Natural Language Processing (NLP). This challenge, and its attendant difficulties, may be better clarified within the general context of *diagnosis* and *screening* (Streiner, 2003), and by highlighting what appears to be the proverbial problem of finding a needle in a haystack, and the price of false positives (Neuman et al., 2019).

The aim of *diagnosis* is to confirm, or rule out, the hypothesis that a given individual has a certain attribute—such as posing a risk to others. In contrast, *screening* is broadly used to determine which member of a large group of individuals has the attribute in question. In the case of solo perpetrators, given the low prevalence of mass-shootings, automated diagnosis is highly problematic, and is almost inevitably accompanied by a high rate of false positives (e.g. Neuman et al., 2019). Given the low prevalence of the event, it is difficult to identify an informative signature that can be *efficiently* used for *diagnosis*. This point is critically important: while some psychological characteristics may be attributed to shooters (e.g. Knoll, 2010; Neuman et al., 2015), they are not necessarily informative for real-world and practical interventions. For example, a recent report by the National Council for Behavioral Health, titled *Mass Violence in America*,[3] suggests that while mass violence is a rare event, perpetrators share certain characteristics—such as a feeling of hopelessness. However, feeling hopeless in and of itself cannot serve as an informative marker for diagnostic purposes, since the vast majority of people who feel this way do not commit acts of violence against others. Thus, although the perpetrators probably *share* the attribute of feeling hopeless, its presence does not contribute to the diagnosis of a potential perpetrator. In other words, the issue is not whether or not the perpetrators share a certain attribute, or whether there is a difference between perpetrators and non-perpetrators in that regard, but what is the probability that *given* that attribute (such as a feeling of hopelessness), a given individual is a perpetrator, and whether that probability enhances the detection of such perpetrators with an acceptably low rate of "false alarms."

Another recent report—prepared by the United States Secret Service, and titled *Protecting America's Schools*[4]—argues that while there is *no profile* of a student attacker, a grievance with classmates is reportedly the most common motivation. Again, since the vast majority of individuals who experience such grievances do not resort to extreme violence, this data is of no informative value in the diagnosis or screening of potential offenders. This important methodological point, which is grounded in a long tradition of Bayesian reasoning (e.g., Fenton and Neil, 2012) and methodologies for computing the weight of evidence (e.g., Good, 1985), has been reiterated both in the scientific literature (e.g. Neuman et al., 2019) and in the popular media,[5] but ignored by many studies and practical applications.

If there is an informative signature of a perpetrator, it is probably a high-dimensional one, whose exact pattern cannot be easily identified by human experts and manual analysis. Therefore, the screening of school shooters can benefit from the development of automatic screening methodologies leaning on ML and NLP. However, given the low prevalence of the event, even the most powerful tools of ML, such as Deep Neural Networks (DNN), and methodologies for handling imbalanced sets such as the Synthetic Minority Oversampling Technique (SMOTE) cannot easily be used to address this challenge, as there are (1) not enough cases to train the model and attune the parameters, and (2) to validate the model through n-fold cross-validation.

Moreover, DNN cannot provide yet easily *interpretable* results, which are critical to justifying the use of preventive measures. While the "black box" approach of DNN may be adequate for the classification of visual images, it may not be acceptable when applied to the screening of individuals, or for taking preventive measures that require easily interpretable justification. When choosing a subject for in-depth inspection, we must judge the *weight of evidence* in favor of the hypothesis that the subject is a potential risk. In such instances, *the evidence must be clearly interpretable*, and must not hide within the "black box" of the neural network. This argument is the *main justification* for the measurement approach which is proposed in this paper.

Given this challenge and its attendant constraints and difficulties, as we have described, the question is whether there is a simple, practical and interpretable methodology that may be used for the specific task of screening a population and *ranking* potential suspects by their respective "risk factor," in order to establish priorities for an in-depth inspection. Positively answering this question is the major aim of the current study. For example, in instances such as the Eric Harris case, local FBI officers might use OSINT (i.e. Open Source Intelligence) tools to screen social media texts for potential offenders. A helpful automated screening methodology might "flag" texts of potential perpetrators as top-priority candidates for in-depth inspection. The basic working assumption of such a procedure is that by analyzing and ranking the texts (and images) published by individuals in social media, the analyst, as an expert, can validly and reliably identify those who should be priority subjects of in-depth inspection, as they present certain early-warning signs. It is very important to emphasize that by moving from the task of classification to the task of ranking, we may bypass the problem of false positives. This is a context, where a sharp criterion doesn't exist and the problem of false positives is replaced by a softer version of decision making and its benefits.

An automated screening methodology may therefore save the screening efforts of the human analyst, and allow for analysis of massive data sets in a short time. In this real-world context, the success of such a methodology should be measured *mainly by its practical success in saving screening efforts* (Neuman et al., 2015), rather than by conventional measures of ML performance—such as AUC, precision, or recall. Therefore, such a screening procedure should involve the ranking of individuals according to their potential risk, as identified through their psychological-textual signature, the identification of the top-k subjects, and their selection for secondary, in-depth inspection. Top-k ranking methodologies have been intensively studied in Information Retrieval (e.g. Niu et al., 2012; Zehlike et al., 2017), and appear to be particularly relevant to the challenge of screening for shooters. Such a top-k learning-to-rank methodology should be judged by its ability to identify top-k-ranking potential shooters, and by the screening efforts saved for the human analyst. Moreover, it should be judged by the *interpretability* of its results, and its ability to provide the human analyst with simple *evidence-based justification* for marking a given subject for in-depth inspection. In sum, to gauge the usefulness of such a screening methodology, we may measure the extent in which texts written by a school shooter are prioritized by the system, in a manner that saves such efforts by a human analyst and by providing him with a simple methodology for measuring the weight of evidence that the person is a shooter. The aim of the present paper is to introduce such a methodology.

A search on Google Scholar for papers at the intersection of "school shooters" and "machine learning" and/or "natural language processing" produces very few studies that used indirect measures of risk, rather than the texts of actual shooters, did not focus on school shooters, used a limited number of comparative texts, or a limited number of texts produced by a few shooters (Neuman et al., 2015). With very few exceptions (Neuman et al., 2019), the overwhelming majority of the papers on lone-wolf perpetrators ignore the needle-in-the-haystack problem, or the real-world constraints facing an analyst.

[3] https://www.thenationalcouncil.org/wp-content/uploads/2019/08/Mass-Violence-in-America_8-6-19.pdf.

[4] https://www.secretservice.gov/data/protection/ntac/usss-analysis-of-targeted-school-violence.pdf.

[5] https://www.vox.com/science-and-health/2018/2/22/17041080/predict-mass-shooting-warning-sign-research.

2. Materials and methods

2.1. Data

We downloaded all of the available *personal* written texts of school shooters (N = 18) from a site devoted to studying this subject.[6] The reason for using personal texts rather than other forms of texts, is that personal texts expose the inner world of the individual and may be used for psychological profiling (Neuman, 2016a). Following Neuman et al. (2015), and for comparative analysis, we also used the texts from the Blog Authorship Corpus (Schler et al., 2006). For maximizing the overlap with the shooters' set, and as women are not shooters, we used the texts of male bloggers only aged between 15 and 25. The texts written by each subject were merged into a single file; subjects who produced fewer than 100 words were removed from the analysis. Overall, the final dataset we used for the analysis featured the texts written by 5047 subjects, 18 of whom (i.e. 0.003%) are shooters.

2.2. Pre-processing

The texts produced by each individual were automatically analyzed by means of Natural Language Processing tools. For all NLP tasks, we used the Natural Language Toolkit—NLTK 3.4.5[7]. Lemmatization and POS tagging were carried out, leaving only nouns, verbs, adjectives, and adverbs for analysis. Next, the *tf-idf* score was calculated for the words used by each subject. Each subject was represented by the 100 highest-scoring words. Thus, and *irrespective of the number of texts* produced by the individual and their length, each individual was represented by a vector of 100 words that best represented his text. The data used in this study (plus the code used for the analysis) is available in the following link: https://github.com/YoavLevR/ScreeningForShooters.

Next, a topic analysis tool was applied. To this end, we used the *Empath* tool (Fast et al., 2016), which is a free topic analysis tool, and translated each of the 100 words into a normalized (i.e. percentage-based) distribution of topics. This way, each individual was ultimately represented by a normalized vector of length 194, reflecting the distribution of the 194 topics evident in the 100 selected words. Among the topics that can be identified by Empath are: *violence*, *shame*, *pain*, *love* etc. Our procedure is based on the idea that shooters can be screened through the signature of their topics, as evident in the topics' distribution.

2.3. Measuring the weight of evidence

We adopted the measure developed by the physicist E.T. Jaynes (1996) to compute the weight of evidence in favor of hypothesis *H*, as opposed to the alternative hypothesis, *-H*. In our case, we compared the weight of evidence that a subject is a shooter against the evidence for a non-shooter. The weighted-evidence approach is interpretable, and appears to relate naturally to how sensory information is integrated in the nervous system (e.g. Gold and Shadlen, 2001). As such, it provides a simple and interpretable approach for integrating evidence that exists in the text, and directs a learning process that may be used for the ranking and screening of individuals.

It is important to emphasize that our use of Jayne's idea does not purport to compete with common ML approaches to classification and ranking, but rather to present a *different* approach that may have some benefits. Further below, we explain the justification for using Jaynes' proposal, and compare our results to those gained by ML algorithms for classification and anomaly detection.

[6] https://schoolshooters.info/original-documents.
[7] https://www.nltk.org.

Jaynes proposed to compute the *posterior odds* in favor of hypothesis *H* (e.g. being a shooter) given certain evidence *D* (such as writing about revenge). The posterior odds are:

$$\text{Posterior Odds} = \frac{P(H/D)}{P(-H/D)} \tag{1}$$

The posterior odds are equal to the *prior odds* in favor of *H*:

$$\text{Prior Odds} = \frac{P(H)}{P(-H)} \tag{2}$$

multiplied by the Likelihood Ratio:

$$\text{Likelihood Ratio} = \frac{P(D/H)}{P(D/-H)} \tag{3}$$

which is called the *Bayes Factor* (BF) (Goodman, 1999; Kass and Raftery, 1995). In the context of hypothesis testing, the Bayes Factor is the ratio between the likelihood of the data given *H*, and the likelihood of the data given the alternative hypothesis *-H*. Therefore, the *odds* in favor of hypothesis *H* given evidence *D* is calculated as follows:

$$\frac{P(H/D)}{P(-H/D)} = \frac{P(H)}{P(-H)} * \frac{P(D/H)}{P(D/-H)} \tag{4}$$

Jaynes (1996) proposed measuring the weight of evidence in favor of hypothesis *H* by translating this equation into a decibel system, where the prior evidence for hypothesis *H* is:

$$e(H) = 10\log_{10}\left[\frac{P(H)}{P(-H)}\right] \tag{5}$$

the evidence for *H*, given *D* is:

$$e(H|D) = e(H) + 10*\log_{10}\left[\frac{P(D|H)}{P(D|-H)}\right] \tag{6}$$

and when the process involves several pieces of evidence:

$$e(H|D) = e(H) + 10\sum_{i=1}^{N}\log_{10}\left[\frac{P(Di|H)}{P(Di|-H)}\right] \tag{7}$$

We describe e (H|D) as the *Jaynes score* (i.e. Jaynes(e)). Following Neuman et al. (2019), our basic idea is that by using Jaynes(e), we may screen for potential shooters using the topical evidence that exists in their texts, where each piece of evidence corresponds to the existence of a specific Empath category in the text. Integrating this weight of evidence across all the topics mentioned by the subject may give us a tool for screening for the shooters.

What are the justifications for using Jaynes' measure, rather than common ML algorithms for classification? Jaynes(e) has several appealing properties. First, we are not simply measuring the probability of a given individual being a shooter or non-shooter given the evidence *D*. As explained earlier in the case of shooters, P(H/D) is known to be extremely low. What we are actually measuring are the *odds*—i.e., the ratio of favoring one hypothesis over the other—and how it improves our inference beyond the prior odds given by P(H)/P (-H). While, for practical reasons, this may not provide an advantage over other ML algorithms (such as Naïve Bayes) that use the same underlying logic of Bayesian inference, it may provide analysts with a more intuitive and easily interpretable way to work with the results—because when they must decide whether or not to send a subject for in-depth inspection, it provides a score that can be interpreted much more easily. In other words, the Jaynes(e) seem to be much more appealing to human judgment, because it provides results whose meaning can be easily interpreted both in terms of odds and the ten-based decibel system (Jaynes, 1996), and where a *threshold* for decision can be identified more effectively: a score above *1* provides some odds in favor of the hypothesis. In

contrast, a ML classifier may produce predicted probabilities that a person is a shooter, but determining the cut-point of decision is less intuitive than the odds-based and decimal score proposed by Jaynes. Moreover, Jaynes' measure has been proposed in the context of hypothesis-testing and inference, while the ultimate context of ML is one of *optimization* (Battiti and Brunato, 2014). While this subtle difference may be practically irrelevant in terms of improved screening for shooters, it is relevant to achieving interpretable results, since by definition optimization focuses more on minimizing a cost function, while inference through weight-of-evidence is more attuned to revealing a valid chain of reasoning in support of a given decision-making process.

3. The experiment

To test the proposal procedure of using the Jaynes(e) for screening subjects, we ran an experiment on the dataset. First, however, we informally presented and detailed the procedure. The pseudo-code appears in Appendix 1. To test our methodology, we first divided our experiment into two phases: *learning* and *testing*.

3.1. The learning-phase

In the learning phase, we randomly sampled 70% of the shooters, and 70% of the comparison group (non-shooters). Next, we measured the median score (M_i) for each of the Empath categories (i.e., D_i) produced by a subject: if the subject scored above the median in a given Empath category (i.e. $D_i > M_i$), their score was *1*— in all other instances, it was *0*. The median score was chosen as an arbitrary cut-point. (Optimizing the cut-points for each Empath category is possible, but beyond the scope of this paper.)

Next to be calculated was the Bayes Factor for D_i, given the hypothesis that the individual is a shooter (i.e. H), versus a non-shooter (i.e. -H):

$$BF_{Di} = \frac{P(Di/H)}{P(Di/-H)} \tag{8}$$

In line with common norms (Kass and Raftery, 1995), we considered a BF_{Di} to be significant only when it was equal to, or higher than, 3 ($BF_{Di} \geq 3$): if so, we selected that particular Empath category for the next phase of the analysis. The topical categories that scored 3 or above on average over 50 runs (i.e. folds) of the learning phase were:

(1) Prison (including words such as *torture*, *kill*, *escape*)
(2) Government (e.g. *poverty*, *tyranny*, *unjust*)
(3) Kill (e.g. *exterminate, massacre, hunt*)
(4) Legend (e.g. *occult*, *magic*, *hero*)
(5) Neglect (e.g. *fear*, *suffer*, *humiliation*)
(6) War (e.g. *destroyer, assassination, enemy*)
(7) Weapon (e.g. *gun*, *pistol*, *sniper*), and
(8) Exasperation (e.g. *outraged*, *disgust*, *loathing*)

These content categories have face validity and are clearly interpretable in the context of shooters. To test the extent to which this procedure identifies the informative attributes (i.e., Empath categories), we tested several other procedures of attributes selection, using a 10-fold cross-validation procedure. Using the Weka platform,[8] we tested two methods of attributes selection: *InfoGain* and *CfSubset*. The top eight attributes identified by our procedure, and the two attributes selection procedures, are presented in Table 1.

This shows that among the top-eight attributes identified by our procedure, two of them—*Kill* and *Weapon*—correspond to those identified by the InfoGain and by the CfSubset. However, the simple procedure we used seems to have a clear face validity.

[8] https://www.cs.waikato.ac.nz/ml/weka/.

Table 1. Attributes selected by the three procedures.

Jaynes	InfoGain	CfSubset
Prison	Negative emotion	Negative emotion
Government	Death	Death
Kill	Kill	Kill
Legend	Sadness	Neglect
Neglect	Traveling	Traveling
War	Weather	Weather
Weapon	Weapon	Weapon
Exasperation	Pet	Hate

After identifying the Empath categories that can serve as evidence in favor of the hypothesis that the subject is a shooter, and computing the BF associated with each of these pieces of evidence, we calculated a "noise factor" associated with each given BF_{Di}.

3.2. Calculating the noise factor η

We assumed that calculating the BF of each piece of evidence is not free of noise, and therefore calibrated the Jaynes score by adding a noise factor η to it. The noisier is the feature the less we trust it and the less weight we give it in calculating the final Jaynes score. Inspired by the models of *multisensory cue integration* (Rohde et al., 2016), the procedure for measuring the noise factor η was as follows: we performed 50 runs, in each of which we randomly sampled 50% of the learning set, and calculated the BF of each Empath category. At the end of this procedure, we had a distribution of the BF scores for each of the Empath categories. Next, we calculated the variance σ2 of each BF_{Di}. The noise factor of BF_{Di} was then calculated as:

$$\eta_i = \frac{1}{variance\ of\ BFDi} \tag{9}$$

3.3. The test-phase

We completed the learning phase by identifying the BF_{Di} for all Empath categories that might signal a shooter, and the noise factor associated with each BF_{Di}. We then moved to the test set, which comprised 30% of the shooters and 30% of the non-shooters. For each subject, we measured the score of each Empath category (i.e. D_i) that had been identified in the learning phase: if the score was higher than the median score of the test set it was replaced by:

$$BF_{Di} = \left[\frac{P(Di|H)}{P(Di|-H)} * \eta i\right] \tag{10}$$

which had been calculated at the learning phase. The final Jaynes score for each subject in the test set was then computed as follows:

$$J(e) = e(H) + 10\sum_{i=1}^{|E|} Log10(BFDi * \eta i) \tag{11}$$

The final test file comprised a list of subjects tagged as "shooter" or "non-shooter," and their respective Jaynes scores. The J(e) measures the weight of evidence for the hypothesis that the subject is a shooter. To validate the screening performance of the Jaynes score, we ran the above procedure (from the learning to the test phase) 50 times with random sampling, according to the pre-defined split preferences (70/30). At the end of each run, we ranked the subjects in descending order of their respective Jaynes scores (i.e. high-to-low), and analyzed the percentage of shooters in the top N% of the ranked texts. For example, we examined how many shooters were found in the top 1% of the ranked texts; in the top 2% of ranked texts; in the top 3% of ranked texts; and so on. The final results were averaged over the 50 runs.

4. Analysis and results

4.1. Analysis 1

The aim of analysis 1 was to gain a better understanding of the difficulty in screening for shooters. To address this challenge, we trained several ML models on the complete dataset of all 5047 subjects. All models were trained through a 10-fold cross validation procedure, with the aim of correctly identifying the shooters.

Among four tested classifiers (i.e., Naïve Bayes (NB), k-NN, Random Forest and Logistic Regression) only the Logistic Regression and the Naïve Bayes gained some classification results: The NB gained 1% Precision and 28% Recall—i.e., that the classifier was able to identify 28% of shooters, but when it classified a subject as a shooter, it was correct *only* 1% of the time. The Logistic Regression classifier gained 10% Precision and 22% Recall—namely it, too, offered low precision and recall rates. Examining the Threshold Curve produced by the NB classifier, and recall as a function of sample size, we found that in order to identify half of the shooters by the NB algorithm, we needed 37% of our subjects' population, which is quite unhelpful. The major implication of these results is that they increase our awareness to the difficulty of automatic diagnosis through ML and point to the need to consider an alternative screening procedure, such as the one we propose. We then returned to the results gained by our procedure.

4.2. Analysis 2

Our test set comprised 1515 subjects, only five of which were tagged as shooters ($p = 0.003$)—i.e., an extremely low prevalence of shooters in each test set. We ranked the subjects according to their average J(e) across the 50 runs of the experiment, and analyzed the percentage of shooters of different percentages up to the top 20% of the ranked texts. In other words, *the main results of this paper are the results gained over the 50 runs of the full experiment.*

The major aim of analysis 2 was to test the performance of our procedure in ranking potential shooters. If the procedure is effective, then we should find at the top-ranked subjects, a higher percentage of shooters than the one predicted by chance.

The percentages of shooters identified within the top-ranked subjects are presented in Figure 1.

Result 1. *The majority of the shooters (i.e., 53.2%) were identified within the top 8% percent of the highest-ranking cases.*

The major implication of this finding is that given this result, an analyst conducting a top-down screening of the subjects based purely on their Jaynes scores would be able to identify the majority of shooters within the top 8% of the dataset—a significant reduction in the analyst's workload. Similarly, if we round up the averages, 60% of the shooters may be identified within the top 10% of the cases, and 80% of them may be identified within the top 20% of the cases.

As we have previously argued, the performance of our procedure should be basically compared to the one gained by chance. For this comparison, we have conducted the third analysis.

4.3. Analysis 3

If our methodology has a practical benefit, then in each of the N% of the top-ranked cases we should expect to find more shooters, on average, than expected by chance. The aim of analysis 3 was to test this hypothesis. Figure 2 presents the expected number of shooters within a given percentage of the ranked cases (e.g., the top 1%), and the actual average number of shooters identified by our procedure.

The ratio between the expected number of shooters and the actual number of shooters is indicative of the procedure's utility. For example, the ratio between the observed number and the expected number of shooters in the top 1 percent of ranked cases, is 16—which means that, by using our methodology, the analyst's performance is increased by a factor of 16. This is the major implication of our analysis. Figure 3 presents the ratio for all percentiles.

We can see that the ratio ranges between 16 and 4.36—namely, that in the worst-case scenario, we can improve our identification of the shooter by a factor of ~4.

4.4. Analysis 4

Following the previous results, we may examine the performance of our procedure when compared to the performance of various ML classifiers. It must be remembered that our procedure is a procedure of hypothesis testing and not a ML classifier. Therefore, we didn't expect it to outperformed ML classifiers but just examined how far is the performance from the one gained by ML classifiers.

For performing this comparative analysis, we trained several ML classifiers on 2/3 of the data, tested them on the rest, and produced the predicted probability score that a given subject is a shooter. For each classifier, we ran the procedure three times, by using three different folds, ranking the subjects from high to low according to the predicted

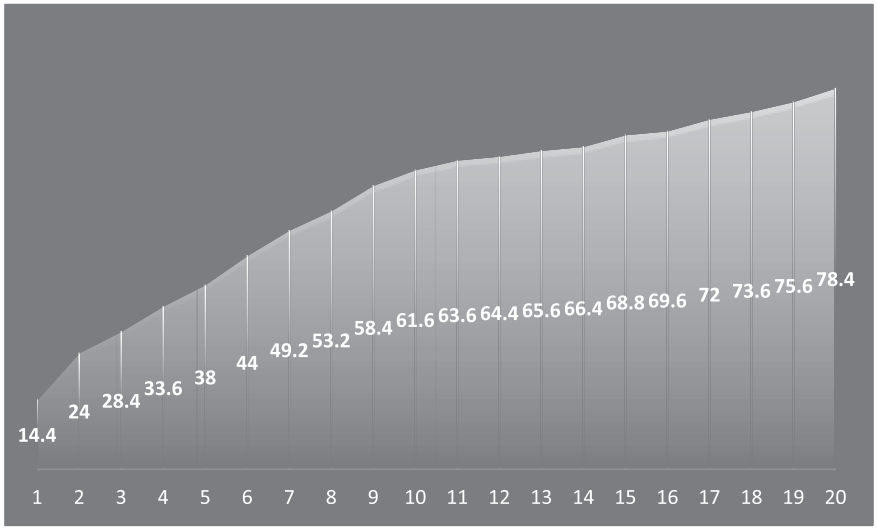

Figure 1. The percentage of shooters (Y-axis) identified within each of the top percentages (1–20) of ranked cases (X-axis).

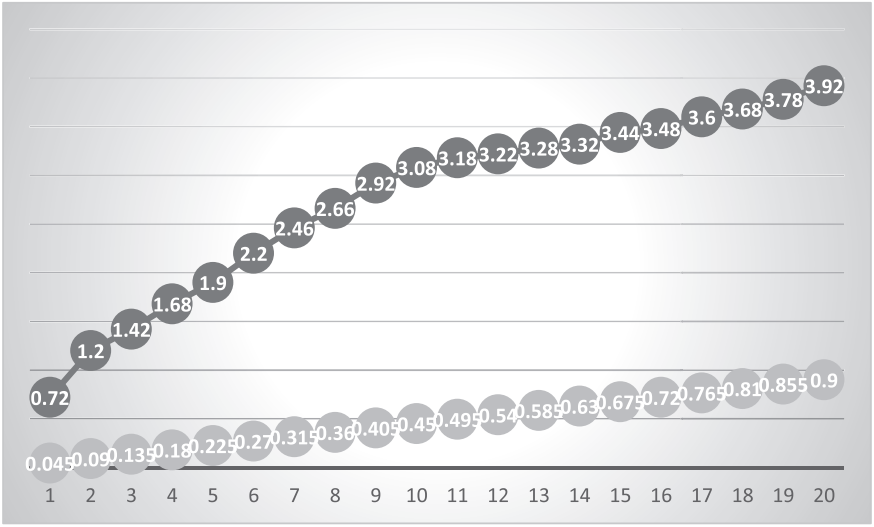

Figure 2. Actual (black) vs. expected (gray) number of shooters for each of the top 20 percentiles of ranked cases (X-axis).

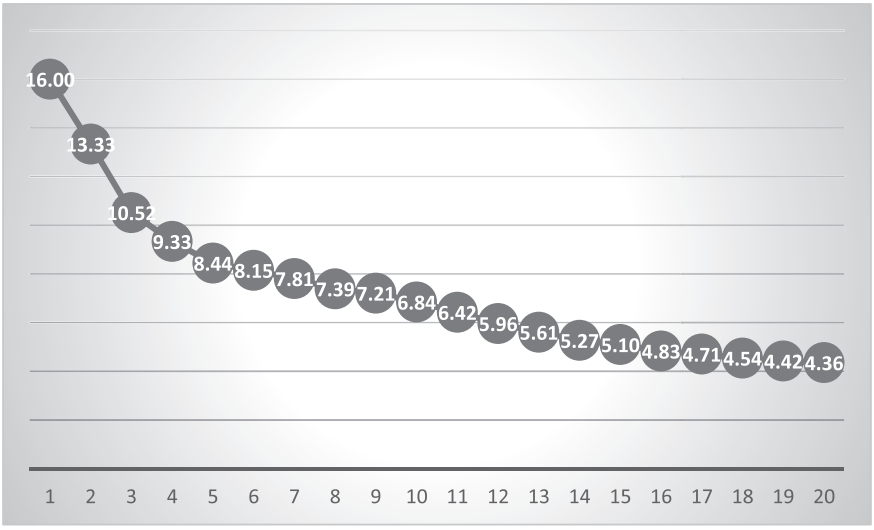

Figure 3. The ratio between the actual and expected number of shooters (Y-axis) for the top 20 percentiles of top-ranked cases (X-axis).

probability score that they are shooters, and averaged the identified percentage of shooters of the k-top cases.

To this end, we used the Scikit-Learn platform (https://scikit-learn.org/stable/) and the following classifiers: Random Forest, Gradient Boosting, Gaussian Naïve Bayes, and AdaBoost. The results for the top ranked percentages (5, 10, 15, and 20 percent) appear in Figure 4.

As evident from this chart, the best results were gained by the Gradient Boosting classifier—although our very simple hypothesis testing procedure fares well compared with other, much more powerful, classifiers. For example, for the top 5% of the ranked texts it performed better than the AdaBoost, the NB and the Random Forest. The second main result of the paper is that:

Result 2. The Jaynes hypothesis testing procedure performs well even when compared with some common ML algorithms (e.g. AdaBoost).

However, here the issue of interpretability, that we mentioned earlier as a justification for using the Jaynes approach, comes into the picture. The next analysis aims to test the interpretability of the scores produced by our procedure as compared to those scores gained through the use of the ML classifiers.

4.5. Analysis 5

Table 2 presents the ranking of the three top-ranked shooters in the three test-folds, and the predicted probability (calculated by the Gradient Boosting that gained the best performance) that a given subject is a shooter.

This reveals that in test fold 1, the first identified shooter ranks No. 8, but his predicted probability of being a shooter according to the Gradient Boosting classifier is only 0.04! With the exception of one case, where the predicted probability is high ($p = 0.96$), in all other cases the probabilities are *extremely* low.

If we consider the probability produced by the classifier as the degree of belief that the subject is a shooter, this is a problem, regardless of the subjects' rankings, as it means that *possibly none of the ranked subjects is a shooter*, and yet the procedure would still rank them among the top scorers, along with the shooters! In other words, the relative ranking score may not be enough and a low probability might produce enormous difficulties in interpreting the results and making a decision. In contrast, the Jaynes score provides us with more easily interpretable results. Table 3 presents the rankings and Jaynes scores of the top-ranked shooters, in three randomly selected test folds.

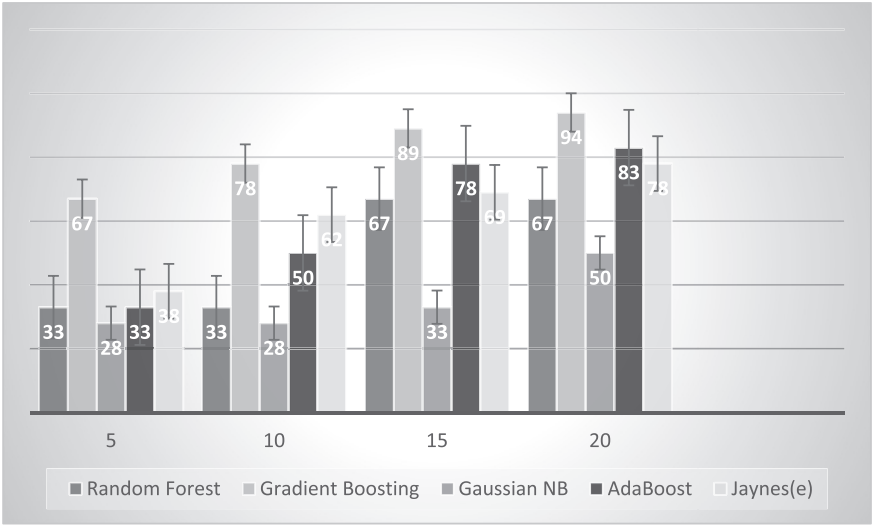

Figure 4. Percent of identified shooters (Y-axis) within the top-ranked cases (top 5%, 10%, 15%, and 20%) (X-axis).

Table 2. Three top-ranked shooters in the three test-folds, and their probability of being a shooter, according to the Gradient Boosting classifier.

Test Fold 1		Test Fold 2		Test Fold 3	
Rank	Prob.	Rank	Prob.	Rank	Prob.
8	0.04	2	0.96	19	0.05
43	0.002	32	0.025	37	0.02
64	0.001	37	0.012	46	0.01

Table 3. Ranking of the top three identified shooters, and their respective Jaynes scores.

Test Fold 1		Test Fold 2		Test Fold 3	
Rank	J(e)	Rank	J(e)	Rank	J(e)
3	43	5	33	9	26
26	26	39	17	27	16
106	8	40	17	141	-0.7

Here, we can see that for the majority of the top-ranked shooters, the Jaynes score is considerably higher than 1—meaning that they are more likely to be shooters than non-shooters, which makes it much easier to interpret the results, irrespective of their *relative* ranking. Therefore, the third main result of the paper is that:

Result 3. When compared with the ML classifiers, the Jaynes procedure produces better interpretable results.

5. The utility of the proposed procedure

We now return to the results gained by our proposed procedure. A simple and practical illustration of the practical benefits of the methodology is to plot the percent of identified shooters against the work saved by the analyst—as presented in Figure 5.

With our dataset, the analyst can avoid engaging in in-depth examination of 80% of the cases: by focusing on only the top 20% of the rankings, approximately 80% of the shooters can be identified. Similarly, reducing the workload by 90% by focusing on the top 10% of the rankings would lead to the identification of 60% of the shooters. The trade-off is clear—and so, too, are the benefits. Using Linear Regression analysis, we may try to measure the model fit of predicting the percentage of identified shooters by means of the independent variable of work saved.

Among several regression models, the Linear Regression model exhibited the best fit with the data ($R^2 = 0.92$, $p < 0.001$): the fewer cases that are screened and the more work one wants to save, proportionally the fewer shooters can be identified. The linear model fit is important, as it shows that the percentage of identified shooters is proportionate to the work saved—an understanding that may be useful for practical real-world applications of our proposed screening methodology.

6. Discussion and conclusions

The challenge of automatically identifying solo perpetrators is extremely difficult, and for good reasons. However, people's loss of privacy online, and their voluntary self-exposure on social media, can provide a rich source of information when designing a screening process based on the subjects' own personal "texts" from his written documents to images and music files. The problem is that manual screening of such massive amounts of data is practically unfeasible. A practical combination of automated analysis and human expertise is therefore required, but there is, as yet, no silver bullet that may fully address this challenge. Increasing the safety of school students is a challenging task that requires a system with multiple levels of protection. In this study, we have highlighted the importance of taking preventive steps, and proposed that

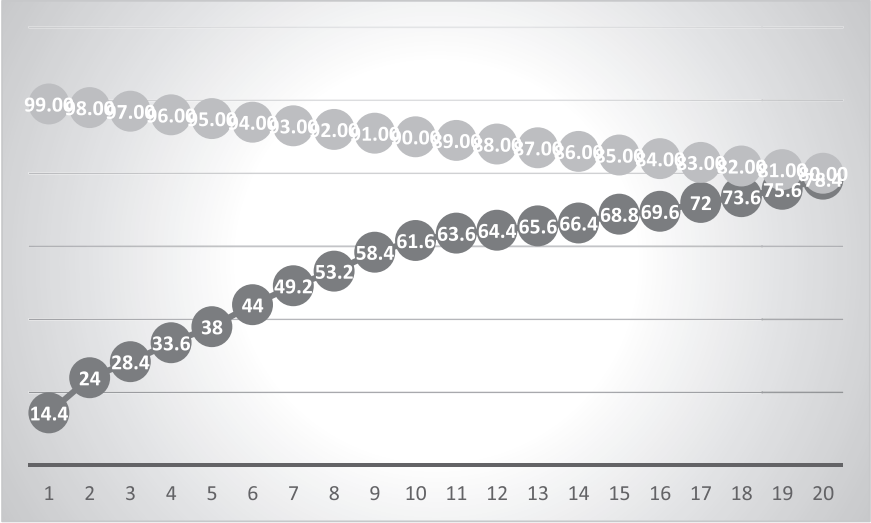

Figure 5. Percent of shooters identified (Black) against the percent of work saved (Gray) for top-ranked cases (percentages 1–20) (X-axis).

by screening various sources from OSINT, and by identifying a signature of early warning signs, we may improve the process of screening for school shooters. It is difficult to precisely measure the expected utility of our proposed approach, as there no ready consensus on the cost of human lives and the benefits of targeting a shooter before an attack. Nevertheless, at least in terms of human labor, the proposed approach appears to offer added value to the human analyst, both in terms of saved screening efforts, and in terms of easily interpretable results. In practice, the methodology can probably be significantly improved by applying better tools to the identification of content categories in a given text; by optimizing the cut-points to include a content category that is indicative of a shooter; by experimenting with various Top-k approaches and ML algorithms; by incorporating the "impostors' cues" (Neuman et al., 2019) and combining it with other ML approaches; and by merging together various sources of information (such as criminal records).

As recognized by various agencies, from the NSF to DARPA, one of the current computing challenges is the development of real-time learning, prediction, and decision-making by ML. The real-time aspect of ML may significantly improve the identification of potential shooters but such a real-time system must have a layer of incoming data-stream arriving from heterogenous OSINT sources. Gaining access to high-quality OSINT sources in real-time is not a trivial task as it involves both legal, ethical and technical considerations. A near future challenge may therefore be to identify alternative data sources that are more easily available and can be fed into the system in real-time. In addition, such as system must have an architecture that not only integrates real-time data from heterogenous sources, but a decision-making component designed as a part of an integrated Command and Control (C2) center which is in charge of operating in a case where a strong warning signal appears. In sum, building a real-time ML system nurturing from incoming high-quality OSINT signals and integrated with a decision-making component of a C2 center, may significantly advance our ability to protect schools. This is a complex engineering project while our paper only modestly points to one possible academic direction that may be used in such a project.

In sum, the present paper should be considered more as a proof-of-concept only, which is based on a single selected approach that may be useful in the unique and sensitive context of screening for shooters. Since the modest aim of this paper is to incrementally enhance the safety of schools by harnessing the power of ML for screening for shooters, we leave all other improvements to real-world applications.

Declarations

Author contribution statement

Y. Neuman, Y. Lev-Ran, E. S. Erez: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to thank Dr. D. Vilenchik and the anonymous reviewers for a critical and constructive reading of the paper.

Appendix 1. The pseudo-code

The learning-phase

Randomly sample 70% of the shooters and 70% of the non-shooters group.
Input: The set L includes subject's texts.
For i = 1 to 194 (number of Empath categories).

For j = 1 to |L|
D [j][i] = Empath category$_i$ (of text$_j$).
Find the Median score M_i of Empath category i.
For j = 1 to |L|
IF D [j][i] > M_i
THEN D [j][i] = 1.
ELSE D [j][i] = 0.
Calculate the Bayes Factor (BF) for each D_i (noted as BF_{Di}).
IF $BF_{Di} \geq 3$.
THEN add i to E.

Outputs:

E is the set of evidence indexes to be used in the test phase.
BF_{Di} is the Bayes Factor for Empath category i.

Sub-procedure: Computing the noise factor.
For each $i \in E$.

For k = 1 to 50.
Randomly sample 50% of L
calculate the corresponding BF_{Di}
Set [i][k] = BF_{Di}
calculate the variance σ2 of each BF_{Di} from Set [i].
The noise factor of BF_{Di} is:

$$\eta_i = \frac{1}{variance\ of\ BFD_i}$$

Output: η_i is the noise factor of BF_{Di}

The test-phase

Use the test set T composed of 30% of the shooters and 30% of the non-shooters.
The set T includes subject's texts (|T| = 1509).
For i = 1 to 194 (number of Empath categories).

For j = 1 to |T|
D [j][i] = Empath category$_i$ (of text$_j$).
Find the Median score M_i of Empath category D_i
For j = 1 to |T|
IF D [j][i] > M_i
THEN D [j][i] = BF_{Di}
ELSE D [j][i] = 0.

Calculate the Jaynes score:

$$J(e) = e(H) + 10\sum_{i=1}^{|E|} Log10(BFDi*\eta i)$$

END.

References

Battiti, R., Brunato, M., 2014. The LION Way. Machine Learning Plus Intelligent Optimization. LIONlab, University of Trento, Italy, 978-14-960340-2-1.

Fast, E., Chen, B., Bernstein, M.S., 2016. Empath: understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, USA, pp. 4647–4657.

Fenton, N., Neil, M., 2012. Risk Assessment and Decision Analysis with Bayesian Networks. CRC Press, .

Gold, J.I., Shadlen, M.N., 2001. Neural computations that underlie decisions about sensory stimuli. Trends Cognit. Sci. 10–16.

Good, J., 1985. Weight of evidence: a brief survey. In: Bernardo, J.M., DeGrot, M.H. (Eds.), Bayesian Statistics. Elsevier, North Holland, pp. 249–270.

Goodman, S., 1999. Toward evidence-based medical statistics. 2: the Bayes factor. Ann. Intern. Med. 130, 1005–1013.

Jaynes, E.T., 1996. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Knoll, J.L., 2010. The "pseudocommando" mass murderer: Part I, the psychology of revenge and obliteration. J. Am. Acad. Psychiatr. Law 38, 87–94.
Neuman, Y., 2016a. Computational Personality Analysis: Introduction, Practical Applications and Novel Directions. Springer. N.Y.
Neuman, Y., 2016b. Artificial intelligence in public health surveillance and research. In: Luxton, D. (Ed.), Artificial Intelligence in Behavioral and Mental Health Care. Academic Press, N.Y., pp. 231–254
Neuman, Y., Assaf, D., Cohen, Y., Knoll, J.L., 2015. Profiling school shooters: automatic text-based analysis. Front. Psychiatr. 6, 1–5.
Neuman, Y., Cohen, Y., Neuman, Y., 2019. How to (better) find a perpetrator in a haystack. J. Big Data 6, 9.
Niu, S., et al., 2012. Top-k learning to rank: labeling, ranking and evaluation. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 751–760.
Rohde, M., et al., 2016. Statistically optimal multisensory cue integration: a practical tutorial. Multisensory Res. 29, 279–317.
Schler, J., Koppel, M., Argamon, S., Pennebaker, J., 2006. Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. AAAI, Palo Alto, CA, pp. 191–197.
Simons, A., Meloy, J.R., 2017. Foundations of threat assessment and management. In: Handbook of Behavioral Criminology. Springer, Cham, pp. 627–644.
Streiner, D.L., 2003. Diagnosing tests: using and misusing diagnostic and screening tests. J. Pers. Assess. 81, 209–219.
Zehlike, M., et al., 2017. Fa* ir: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1569–1578.