Neuman *et al. J Big Data* (2019) 6:9
https://doi.org/10.1186/s40537-019-0172-9

Journal of Big Data

METHODOLOGY Open Access

How to (better) find a perpetrator in a haystack

Yair Neuman[1*], Yochai Cohen[2] and Yiftach Neuman[3]

*Correspondence:
yneuman@bgu.ac.il
[1] The Department of Cognitive and Brain Sciences and Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel
Full list of author information is available at the end of the article

Abstract

In many real-world contexts, there is a pressing need to automatically screen for potential perpetrators, such as school shooters, whose prevalence in the population is extremely low. We first explain one possible obstacle in addressing this challenge, which is the confusion between "recognition" and "localization" during a search process. Next, we present a pragmatic screening methodology to the problem along Jaynes Bayesian hypothesis testing procedure. According to this approach, we should first focus our efforts on reducing the size of the haystack rather than on the identification of the needle. The third and major methodological contribution of the paper is in proposing that we may reduce the size of the haystack through the identification and use of unique data cues we describe as "impostors' cues". An experiment performed on an artificial data set of 7000 texts, shows that when incorporating these cues in the hypothesis testing procedure, they significantly improve the automatic screening of objects characterized by an attribute of a low prevalence (i.e. a psychopathic signature). The relevance of the proposed approach for Big Data and Homeland security is explained and discussed.

Keywords: Homeland security, Lone wolf perpetrators, Terrorism, Screening, Needle in a haystack, Bayes Factor, Jaynes

Introduction: the challenge

One of the major challenges facing democracies, is the screening of perpetrators before they have launched their targeted violence. From school shooters to lone-wolf terrorists, this is a pressing challenge with no simple, trivial, or ready-made solutions. In contrast with diagnosis, where the aim is to confirm or rule out the hypothesis that a specific individual has a certain attribute, screening is broadly used to determine which member of a large group of individuals has the attribute in question [21]. In the real world these processes must be practically combined, as first a large group of individuals is screened and then an in-depth diagnosis (or inspection) is applied. The screening of potential perpetrators is probably done using very large, unstructured and multidimensional data sets that update in real time. For example, Eric Harris, one of the two perpetrators who conducted the Columbine High School massacre on 1999, wrote a blog where clear warning signals appear years before the actual attack took place [16]. At that time there was no awareness neither analytic tools for screening such a perpetrator through the analysis of public digital signatures. Today however, the availability of Big Data, analytic tools and

law enforcement awareness turn the screening of perpetrators through Big Data analysis into an almost inevitable challenge. Therefore, the relevance of the challenge we discuss in our paper and the solution we present for Big Data and Homeland security researchers and practitioners, is almost self-evident.

Screening for perpetrators, from school shooters to lone wolf terrorists, pose a difficult methodological problem for professionals dealing with profiling, whether expert manual profiling or the automatic profiling of individuals through Big Data analysis. The problem results from the low prevalence of the perpetrators in the population, a problem summarized under the title of "finding a needle in a haystack". This problem results in a very high-rate of false positives. These false alarms are usually accompanied by a high price that might dismiss the benefits of the search process [14, 23].

Technological solutions

Common and technologically informed approaches for addressing the challenge of finding a needle in a haystack, usually involve Machine Learning (ML) and/or Anomaly Detection methodologies where a large amount of data is analyzed for identifying individuals that (1) match the perpetrator's hypothesized profile and/or (2) individuals that do not conform to a "normal" expected pattern.

With regard to the general and powerful methods for anomaly detection (e.g. [22], it is clear that they are limited in identifying lone wolfs as there are so many ways of deviating from an allegedly expected normal pattern; As written by Tolstoy in his intensively cited statement: "All happy families are alike; each unhappy family is unhappy in its own way". This lesson has been learned long time ago by the Israelis who have painfully realized that there is no unique psychological profile of a suicide bomber [10].

A different albeit a related kind of problem is evident when trying to apply ML algorithms. The extremely low prevalence of the perpetrators turns their identification into a search for a needle in a haystack which is accompanied by a high rate of false positives and false alarms [14, 23]. This problem is evident even if some kind of the target's profile is identified and even if the diagnostic test we apply is characterized by high sensitivity/specificity as clearly illustrated by Wainer and Savage [23]. In this context, technology may come to our help but still, common machine learning procedures are facing the problem of class imbalance or more accurately extreme or high-class imbalance (e.g. [1, 6, 8, 11]) and fall back into the same needle in the haystack situation. This criticism should be taken with a grain of salt, as advances in ML (e.g. the "Deep Learning" approach) have made a substantial contribution to the field, and produced results far beyond those who were familiar at the time Wainer and Savage [23] published their paper. Nevertheless, even the powerful Deep Learning approach cannot resolve the needle in the haystack problem and it is also accompanied by some bothering difficulties, from the need to gain a substantial amount of tagged data [4], which is usually unavailable in the case of lone wolf perpetrators, to the impossibility of understanding the Black Box of the neural network. While understanding the black box of a neural network may be irrelevant for practical needs, it is nevertheless highly important in the context of lone wolf perpetrators given the high political sensitivity of some features used by the machine, and given our need to understand the contributing features and their validity in screening the subjects. Quite recently, some innovative and insightful

methodologies for explaining the Black Box of Machine Learning have been proposed (e.g. [5, 18, 19]). However, it must be emphasized that given the extreme class imbalance evident in the case of solo perpetrators, Machine Learning tools cannot be trivially applied and therefore powerful Black Box methodologies such as LIME [18] cannot be applied either. Nevertheless, we currently consider the possibility of combining LIME with the methodology proposed in the current paper. Given the abovementioned context, and similarly to other situations of searching for an object in a complicated and difficult situation (e.g. [20]) it seems that a Bayesian oriented approach, combining experts' knowledge with some simple and meaningful hypothesis testing procedure, may be of high relevance.

Twenty questions for a lone wolf

There is nothing new about the false positives problem in profiling as presented by Wainer and Savage [23], but its persistence and the persistence of the common approaches to deal with it, even in the context of big data and powerful ML algorithms, may suggest that there might be some deep conceptual and methodological flaws hindering our progress. The current sections present such a mental bug. The reader interested in our own constructive approach may skip these sections with no loss of information whatsoever, and go directly to the main part.

Our first claim is that one possible reason for the difficulties we are experiencing in identifying the needle, may be a kind of "mental bug". This mental bug involves the confusion between recognition and localization during a search task. The argument is that some major theoretical approaches in the field of perpetrators' profiling and screening, such as the protocol proposed by Meloy and Gill [13], represent this logic and that when such approaches are translated into screening procedures, they almost inevitably fall into the needle in the haystack pitfall.

Let us start with a common solution to the problem of identifying a lone wolf. While a unique signature/profile of a lone wolf hasn't been identified yet, there are some features and warning behaviors that may be used as guidelines. For example, two leading experts in the field of forensic psychology and solo terrorists [13] present a protocol that includes a few warning behaviors theorized to be associated with targeted violence. For example, a "pathway warning behavior" is any behavior that is a part of research, plans, preparation or implementation of an attack. Given a set of such warning behaviors (i.e. features), it seems that the simplest approach is to automatically apply a search procedure guided by the logic of the famous twenty questions game.

Let's assume that we have ~ one million potential suspects, each represented through a vector of twenty warning behaviors in their binary form. That is, we score each individual on each of the twenty warning behaviors to produce his unique profile as represented by a vector of numbers and we may use some cut-points to dissect each score into "0" and "1" in a way that may use us for a binary decision tree (i.e. a search procedure).

To sort out who a potential perpetrator is, we may start asking questions such as "Is this person exhibits a pathway warning behavior?" (YES/NO), "Is this person presents sympathy with a radical and violent ideology?" (YES/NO), and so on. In theory, a phase space of relevant features and well-chosen questions should halve our search space at each step, providing us with a one bit of information at each crossroad, and finally with

the ability to identify our object among 1,048,576 objects in twenty steps only. This search procedure is precisely the one defining the meaning of Shannon's entropy. Even a softer version of the above procedure seems to be adequate for our search task, as implemented for instance by the decision-tree machine learning classifier used by Neuman et al. [15] for identifying school shooters. If this approach is so simple and appealing why doesn't it work so well in practice? The argument that this simple and appealing approach doesn't work so well in practice is probably a consensus among practitioners in the field who realize that there is no magic bullet for targeting the needle. One may argue that it is our shortcomings in asking the right questions, or in the context of ML, of gaining access to the right features, that hinder our progress. According to this explanation, the increase of data available to the ML algorithms would finally lead to the solution. This is a reasonable argument. However, an alternative explanation for our extreme difficulties in finding the needle is presented below.

Why can't we identify a lone wolf in twenty steps?

In the previous section, we have described a search procedure that actually involves the *recognition* of an object (i.e. the perpetrator) in a prescribed phase space (i.e. a high dimensional space of features). However, two physicists trying to explain the meaning of an entropy known as Tsallis Entropy, have insightfully pointed to an important point that may help us to better understand what is wrong with our search mission.

As argued by Wilk and Włodarczyk [25]: "the way in which one collects information about an object decides the form of the corresponding information entropy". This is a highly important point. The kind of questions we ask invites the kind of answers we get, and the form of information entropy we use, should be adapted to the specific question we ask. We will follow and explain this argument through the lone wolf example in order to explain the shortcoming on the Shannon's information model and why we should think differently about the problem.

Let's assume that a terrorist is an object hidden in a system of M cells (i.e. M partitions of the features space) where the probability of identifying it in a given cell is 1/M. That is, we search for an object defined as a point in a high dimensional space where each point is a microstate of the system. Wilk and Włodarczyk [25] explain that in the Shannon's information model locating an object in a simple space is equivalent with finding the respective cell containing this object. *Localization* is therefore equivalent with *recognition* as can be illustrated in Fig. 1.

In this context, the Shannon entropy of the system is defined as:

$$H(X) = \sum p(xi) log \frac{1}{p(xi)} \tag{1}$$

and the information required for identifying the object in the above phase space is indeed the number of YES/NO questions needed to locate the object. Therefore, if we adopt the Shannon's idea of entropy then the most important challenge facing us in identifying/locating a lone wolf, is the formation of the appropriate questions/features that will lead us to the localization/recognition of the perpetrator. In such a simple and "perfect" world asking the right questions (i.e. identifying the relevant features, psychological or others) is the only ultimate key for both locating and recognizing the required

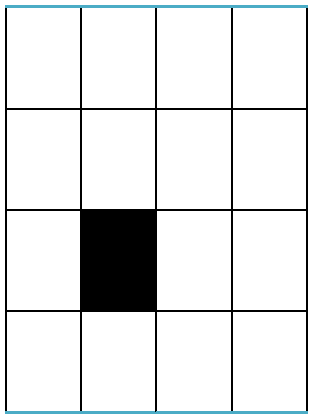

Fig. 1 An object (the black square) located in a phase space

object. To emphasize this important point once more: asking the right questions is our Holy Grail for resolving the uncertainty as represented by Shannon's entropy. All other aspects of the challenge are technical only (e.g. the collection of the data, the choice of the appropriate classifier etc.). In fact, the protocol quite recently proposed by Meloy and Gill [13] epitomizes this approach by seeking for the behaviors that are most indicative of targeted violence and hence of the location and the recognition of the perpetrator who might pose a threat. This approach, which is highly intuitive and common in addressing our challenge, is imbued with difficulties not only as it ignores some basic ideas of Bayesian inference but mainly as it doesn't address the problem of finding the right features and reducing the false positives evident in the case of extreme class imbalance. For example, Corner and Gill [2] have analyzed the prevalence of mental illness among lone-actor and group-based terrorists. By comparing the two groups, they have found that mental illness was much more prevalent among lone-actor terrorists (32% vs. 3%). They conclude by saying (ibid., p. 32):

> *"If mental health professionals were aware of these findings then screening processes can be carried out by security agencies on patients that present similar antecedents and behaviors in medical evaluations".*

However, the issue is not about finding attributes differentiating between the two groups but about computing the probability that an individual is a lone-actor terrorist GIVEN the evidence that he suffers from a mental illness. In this context, the probability that a person is a lone-actor terrorist given the attribute of mental illness is so low, that one may wonder what is the practical aspect if any, in presenting such a feature and attributing to it a high informative value while ignoring the most basic ideas of Bayesian inference.

Let's further elaborate the argument presented by Wilk and Włodarczyk [25]. They point out that there are two alternative forms of search in which one of them is clearly relevant for our challenge. They suggest that the object we seek may have (1) additional features one should account for and (2) that there may be more than one object in a cell, which is a situation in which "one finds the cell with a particle (i.e., object or perpetrator) in it but one is still not sure that this is the right ("true") particle" (ibid. p. 4810). In this case, localization

is not equivalent with recognition. In the best-case scenario, you may locate the right cell where your target resides but may still have difficulties in recognizing it among many other objects (i.e., "impostors"). This is probably the problem we encounter in screening for a potential perpetrator, as despite its possible localization in a certain cell (i.e. a subspace of the features' space), it is accompanied by many other objects/people (i.e. false positives) located within the same "cell".

Hence, the failure of common approaches for the identification of lone wolfs might first result from the inability to conceptually differentiate between localization and recognition and from the difficulties in recognizing the "true" object among several other objects sharing highly similar characteristics.

To illustrate this point let me use a simple example. Let assume that the prevalence of pedophiles in a given population is approximately 1%. Let's further assume that the results of a diagnostic test for a pedophile are normally distributed for the general population and that the higher the score the more likely it is that the person is a pedophile. Imagine a situation in which a person's test score is located slightly higher than one standard deviations above the mean. It means that roughly 84% of the population got a lower score and one may intuitively bet that the probability that this person is a pedophile is higher than the probability that he is a normative citizen. However, in a population of one million subjects, we should expect that 160,000 people would gain such a score or higher. As the prevalence of pedophiles in the general population may be around 1% (i.e. 10,000 people), and even if all of them scored one standard deviations above the mean or higher, then a person who gained the score of one standard deviation above the mean is more likely to be a normative citizen then a pedophile … That is even if we identify the "cell" in which a pedophile is located then there are far more objects (i.e. individuals) in these cell that are not pedophiles than pedophiles.

Wilk and Włodarczyk [25] further develop their argument that may be summarized as follows:

1. An object can be located in a cell with probability p or not to be located at the cell with probability $1-\mathrm{p}$.
2. In a cell there can be maximally $\mathrm{k}-1$ other (i.e. false) objects.
3. The probability of the occurrence of r such particles is p^r.
4. The average probability to register a false object per one false object is therefore:

$$Pk-1=\frac{1-p}{k-1}\sum_{r=1}^{k-1}p^r=-\frac{p}{k-1}\left(p^{k-1}-1\right) \tag{2}$$

and in this case, the entropy for a system with M cells is:

$$H=-\sum_{i=1}^{M}p\frac{p^{k-1}-1}{k-1} \tag{3}$$

which is actually the Tsallis entropy (Tsallis, 2009/2014):

$$Sq(pi)=\frac{1}{q-1}\left(1-\sum pi^q\right) \tag{4}$$

in which the entropy index q is equivalent to the number of maximal objects per cell (i.e. k).

Wilk and Włodarczyk [25] argue that in the context of the more challenging search process, the analogy to the YES/NO questions is the sum over all cells of the probability to not register the object but a false object. This is a highly important point as if we are seeking to identify an object in a context where it may be accompanied by some false objects (i.e. "impostors") then we should take into account the probability of registering a false object. The most important lesson we may learn from Wilk and Włodarczyk [25] analysis is that in a case where our search task is such that localization is not identical with recognition, then the entropy of the system cannot be validly estimated through Shannon's entropy and that a binary search procedure working along this line of reasoning might be sub-optimal. A possible implication resulting from this theorization is that when there is more than one object in a cell then we trivially have to ask more questions.[1] However, we would like to argue that at some point the nature of the questions we ask should start changing and that our screening procedure should be framed in a different manner given the existence of "impostors".

Therefore, the less trivial implication of Wilk and Włodarczyk's [25] is that if we are searching for a true object among several false objects within the same cell then the number and nature of the "impostors" (i.e. false objects) in the cell is crucial for identifying the needle in the haystack. Searching for a potential perpetrator through a search procedure motivated by Shannon's entropy assumes an incremental and "positive" process of getting closer and closer to the target trough a binary procedure as epitomized by the protocol proposed by Meloy and Gill. It is a "more of the same" approach, where the more and more questions we ask (i.e. the relevant features we collect about the individuals) the closer we get to the answer. In other words, the mental bug, we have discussed before is that we believe that just adding more features will solve our problem. In contrast, it seems that our main non-trivial challenge is to differentiate between the "impostors" and the real object rather than between the real object and the whole haystack. In other words, our main pragmatic challenge may be in identifying a specific type of cues sorting out the impostors from the true objects. In this context, it seems that a Bayesian approach may come to our help specifically through a "Jaynesian" hypothesis testing procedure [7]. This argument actually calls for developing and using methodologies that differentiate between the true objects and "impostors". Here, we would like to propose a simple form of Bayesian Analysis that may help us to better address the challenge of finding a perpetrator in a haystack. The procedure is introduced in the next sections through a worked-out example. However, first we present our second claim that a screening procedure may first focus on reducing the size of the haystack.

The proposed methodology

The information-oriented approach described in the previous section involves a positive process of getting closer and closer to the object by asking the right questions. Here we would like to present a reverse approach proposing that we "*negatively*" progress to our

[1] We thank Grzegorz Wilks for clarifying this point.

target by reducing the size of the haystack. It must be emphasized that negatively reducing the size of the haystack is not the same as positively getting closer to the needle, as these two processes are not necessarily symmetric. This point will be better understood when presenting our methodology.

The challenge of finding a perpetrator in a haystack may be better approached if we consider it in Bayesian terms. Given a piece of data D we may want to measure the evidence in support of the hypothesis HP that a subject is a potential perpetrator. This hypothesis—HP—is meaningful though only when compared to the alternative hypothesis that our subject is not a perpetrator but a normal person—HN. For this process of hypotheses testing we may first compute the Bayes Factor:

$$\mathrm{K_{HP/HN}} = \frac{Pr(D|\mathrm{HP})}{Pr(D|HN)} \tag{5}$$

But more specifically and following Jaynes [7], we may quantify the evidence strength for our first hypothesis that the subject is a perpetrator using the following equation:

$$e(HP|D) = e(HP) + 10\log_{10}\left[\frac{\Pr(D|HP)}{\Pr(D|HN)}\right] \tag{6}$$

where the prior support for hypothesis HP [i.e. $e(HP)$] is being updated through the addition of the log transformed Bayes Factor multiplied by ten. Summing the evidence across various D's (i.e. pieces of evidence or "cues") we get:

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

This process of hypothesis testing is explained and illustrated in the next section with emphasis on the "negative" process toward which we have pointed before.

The Jihadist and the cat

ISIS has been one of the most notorious terrorist movement in history, gaining its reputation through orchestrated and well-documented atrocities such as public beheading. As its publicity increased, it has gained more and more popularity and support mainly from young male Muslims who have been recruited through the social media and have joined the jihad. Now, let's assume that we would like to proactively screen for UK Jihadists before they are sneaking the border to war zones. According to some sources[2] the estimated number of UK male citizens who have joined the Jihad is 700 and given the size of the population in the UK, their prevalence is extremely low ($\mathrm{p} = 0.00001$) and the Jaynes score for the a priori evidence that one is a Jihadist (i.e. H) is:

$$e(H) = 10\log_{10}\left[\frac{\Pr(H)}{\Pr(-H)}\right] = -50 \tag{8}$$

a number that represents a strong a priori evidence against our hypothesis that one is a jihadist. However, we also know that the Jihadists are all Muslims and therefore may

[2] https://www.brookings.edu/wp-content/uploads/2016/06/En-Fighters-Web.pdf.

ask for the extent in which the data that one is a Muslim (DM) is somehow telling about being a Jihadist. Let's answer this question by screening a population of 1,000,000 subjects. The Muslims are approximately 5% of the UK population[3] and therefore we should have 50,000 Muslims in our sample. The approximated number of Jihadists consists of 0.0002 of the Muslim population and therefore we should expect to find 10 Jihadists and 49,990 non-Jihadists among the Muslims. The hypothesis testing is therefore:

$$\mathrm{e(H/DM)} \;=\; \mathrm{e(H)} + 10\log_{10}\left[\frac{\Pr(DM|H)}{\Pr(DM|-H)}\right] \;=\; -37. \tag{9}$$

We can see that in itself being a Muslim is clearly a very poor evidence for supporting the hypothesis that a subject is a Jihadist. This is precisely the source of the criticism powerfully raised by Wainer and Savage [23] against ethnic profiling. However, the question we ask is different. The question is not whether one is a Jihadist given that he is a Muslim, but whether being a Muslim should be taken as a screening cue for reducing the size of the haystack. A hypothetical team of engineers working for the MI5 may built an automatic system for screening potential jihadists through the analysis of Facebook pages. Should they dismiss the evidence that one is a Muslim when building the system?

Jaynes hypothesis testing procedure proves that when taken into account the evidence that one is a Muslim moves our scale of confidence that one is a Jihadist from -50 to -37, which mean that when designing a screening procedure, we should take this piece of evidence into account, regardless of the fact that in itself and as an isolated piece of evidence it cannot validly screen for the Jihadists as the false positives are still such that we are left with the needle in the haystack problem. This is a very important point.

When screening for the perpetrators through Jaynes Bayesian hypothesis testing procedure, and given the low prevalence of the target group, *it is almost inevitable that the data we collect work "negatively" from a strong negative e score to a less negative e score*. As such, this process might be totally dismissed by the argument presented by Wainer and Savage [23] if considered in positive terms. However, if we consider it in negative terms as a process of reducing the size of the haystack, then it may be perfectly legitimate as a phase in the screening process.

While being a Muslim is a significant screening cue, it cannot differentiate between Muslims who are Jihadists and Muslims who are not Jihadists. Our second claim is therefore that while Jaynes hypothesis testing procedure cannot provide us with the smoking gun for identifying the perpetrator, it clearly provides us with a way of reducing the size of the haystack in which we are searching. Using Jaynes Bayesian form of hypothesis testing is therefore a kind of reasoning via *Negativa*. In the above example it reduces our certainty that the individual is a normal citizen while it cannot provide us with the positive evidence that one is a Jihadist.

Identifying and using "impostors' cues"

Here comes our third claim which is that one possible approach for improving the screening process is as follows. Search for a piece of evidence Dc such that:

$$\frac{\Pr(Dc|HJ)}{\Pr(Dc|HI)} > \tau \quad \text{and} \quad \mathrm{Dc} \notin \mathrm{D} \tag{10}$$

[3] https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/muslimpopulationintheuk/.

where HJ is the hypothesis that one is a Jihadist, HI is the hypothesis that one is an impostor (i.e. a non-Jihadi Muslim) and τ is a criterion whose value will be elaborated later. Dc's are the cues used to test the hypothesis of HJ over HI, and D is the set of cues used to identify HP over HN. Identifying these Dc cues—The Impostors Cues—is at the heart of our proposed screening procedure. Next, we perform Jaynes hypothesis testing procedure by incorporating this piece of evidence. Let us illustrate and explain this idea.

Let's assume that through the analysis of Facebook images taken by the Jihadist, experts have noticed that the Jihadists are taking pictures with … cats (Mohammad's favorite pet) before traveling to the war zones. We may examine this piece of evidence and may found out that among the ten Jihadists in our sample, six have taken a picture with a cat. In contrast, among those who are non-Jihadist Muslims, 2500 people have taken a picture with a cat. Among the Muslims who didn't take a picture with the cat, only four are Jihadists and 47,490 non-Jihadists. Using $\tau = 3$ [9], the relevant Bayes Factor is:

$$\mathrm{K}_{\mathrm{HJ/HI}} = \frac{\Pr(Dc|HJ)}{\Pr(Dc|HI)} = \frac{0.6}{0.05} > 3 \tag{11}$$

And in contrast:

$$\mathrm{K}_{\mathrm{HJ/HN}} = \frac{\Pr(Dc|HJ)}{\Pr(Dc|HN)} = \frac{0.6}{0.2375} < 3 \tag{12}$$

We can see that the D that one is a Muslim is "telling" about being a Jihadist in a certain sense but it is not telling whether one is a Jihadist or a false positive (i.e. impostor). In contrast, the Dc that one is taking a picture with a cat is not telling whether one is a Jihadist or a non-Jihadist in general but may screen the Jihadists from the impostors. Now merging our two different pieces of evidence, we may first screen for Jihadists first by reducing the size of the haystack by taking the Muslim cue into account and then by moving from a strong evidence against the hypothesis that one is a Jihadist (-50) to a much more skeptical stance (-37). When examining whether the object is taking a picture with a cat, we may move further to -26. That is by combining a piece of evidence supporting HJ over HN with a piece of evidence supporting HJ over HI, we were able to improve our screening procedure by reducing the size of the haystack. One may even extend this procedure by looking for the Bayes Factor of:

$$\mathrm{K}_{\mathrm{HI/HJ}} = \frac{\Pr(Dc|HI)}{\Pr(Dc|HJ)} > 3 \tag{13}$$

in the case where more supporting evidence exist for this BF, and after screening for the Jihadists, simply screen out as imposters cases where the above Bayes Factor holds. This is actually the heuristic we have applied as detailed in the next sections.

The methodology and the experiment

For explaining and testing our approach, we have designed an experiment that aims to screen for texts with a psychopathic signature. We illustrate our approach through the experiment. The general pseudo-code of our procedure appears in Appendix 1 and accompanied by a link to the Python code used in this study.

Table 1 Round percentages of the topics' distribution

Topic	%
Food	18
Music	20
Politics	22
Sports	17
Travel	24

The pre-processing phase

The estimated prevalence of psychopathy in the population is according to some researchers up to 4% [23]. Therefore, we have designed the following experiment. First, we have built a corpus of texts by retrieving from Reddit (https://www.reddit.com/) 7000 texts. The average length of a text was 632 words (Sd = 453) and they ranged from a minimum number of 173–2685 words. The distribution of the texts by topic is presented in Table 1.

Second, as texts written by psychopaths are extremely rare, we have built an artificial data set of psychopathic sentences. Drawing on the Meanness dimension of the psychopathic personality [17], we have used the psychopathy questionnaire of Patrick et al. [17] and other sources to form a set of 100 items expressing Meanness. For example,

"I enjoy humiliating others"
"I'm always looking for a good fight"
"Revenge is sweet"

Third, as the prevalence of psychopaths in the population is up to 4%, we have randomly sampled 300 texts and added to each text six items/sentences from the set of psychopathic items we have formed. This procedure, has formed as set of texts (N = 300) with a clear textual signature of psychopathy/meanness.

Next, we have used Empath [3].) which is an automatic content analysis tool, and for each text in our data set produced a vector of 194 features normalized to represent the percentage of each content category in the text (e.g. torment, violence, vacation etc.).

In the final pre-processing phase, we have randomly split the original 7000 texts file into three different files each having an equal proportion of texts with a psychopathic signature (N = 100) and non-psychopathic signature (N = 2333). The first file we have used is titled L1. Our procedure is built around three phases as described below.

Learning phase 1

At learning phase 1, we first measured the Median score for each Empath category of each text in L1, and turned the score into a binary score. If the category scored above the Median then it was tagged "1″ otherwise "0". We consider each binary score of an Empath category as D and compute the Bayes Factor where H is the hypothesis that that the text is a one with a psychopathic signature and the alternative hypothesis is that it is a normative text.

$$K_{H/-H} = \left[\frac{\Pr(D|H)}{\Pr(D| - H)}\right] \tag{14}$$

The first output file is a list of the Empath categories (i.e. the Ds) and for each category, its BF score as we have computed above. Following some common norms [9], we selected for the next phase of the analysis only Empath categories that have produced a Bayes Factor > 3 for Eq. 14. In our case, we have identified 42 relevant empath categories where the seven top ranked categories were:

1. Torment.
2. Monster.
3. Ridicule.
4. Exasperation.
5. Weakness.
6. Irritability.
7. Kill.

As can be seen, the existence of these content categories in a text clearly provides evidence for the existence of a psychopathic signature in the text.

Learning phase 2

In the second learning phase, we have used the evidence gathered using L1 and tested it on L2. Again, we have measured the percent of each Empath category in each text and turned the Empath scores into binary values using the procedure described above. We have used the Bayes Factor identified for each Empath category and applied the following procedure:

For text = 1 to 2333
and for D_1 to D_{42}
If Di = 1 THEN compute:

$$\text{JaynesH} = \text{e(H)} + 10\sum \log_{10}\left[\frac{P(D|H)}{P(D|-H)}\right]$$

where e(H) is defined as:

$$\text{e(H)} = 10\log_{10}\left[\frac{P(H)}{P(-H)}\right] \tag{15}$$

which in our case is "− 13.97".

For testing the screening procedure of the above measure, we have ranked the texts in a descending order and analyzed the top 100 texts. The probability of finding a psychopathic text among the top 100 is 0.04 (i.e. 4 texts). In practice, we have found among the top 100 texts, 52 texts tagged as psychopaths. We have selected the top 100 texts for further analysis.

Learning phase 2.1

The aim of this phase was to use the data in order to build a measure for differentiating between true and false positives. Texts that were included among the top 100 cases but were lacking a psychopathic signature have been considered as "impostors" (i.e., HI). All

other texts have been considered as true positives (i.e., HP). We have re-used all of the Emapth categories, scored the 100 texts and turned the scores into binary scores. Next, we have measured the BF for HI over HP and selected the categories with BF > 3 that were not included in the set of Empath categories identified previously for computing:

$$\text{JaynesI} = \text{e(HI)} + 10\sum \log_{10}\left[\frac{P(D|HI)}{P(D|HP)}\right] \tag{16}$$

Experiment and results

For the concluding experiment, we have used the third dataset TEST. We have built for each text in this dataset a vector of Empath binary values as done before. Using the 42 Empath categories identified during Learning phase 1, we have computed for each text its JaynesH score. Using the results of Learning phase 2, we have computed for each text its JaynesI score by using the 22 Empath categories that have been identified in the previous phase.

Through the Min–Max transformation, we have standardized our two features (i.e. JaynesH and JaynesI) and scored them on a scale ranging from 0 to 1. If the impostors' cues identified through JaynesI are important, then we should expect to find that JaynesI may significantly contribute to the classification of the cases as psychopaths beyond what can be gained through JaynesH. It is important to emphasize, that given our toy example and artificial dataset with quite a strong psychopathic signature among the psychopathic texts, we don't aim to compete with the performance of ML classifiers but just to prove the idea that using specific kind of cues may improve the screening of the low prevalence objects in a negative way.

To test the hypothesis that the impostors' cues are significant, we have used a Classification and Regression Decision Tree Classifier (CRT) [12] with the text's tag (psychopath vs. non-psychopath) as the dependent variable, and with JaynesH and JaynesI as the independent features. The CRT has been applied by using a tenfold cross-validation procedure. Using JaynesH as a single feature in the CRT didn't yield any identification of the psychopaths. However, including JaynesI has resulted in 67% precision and 64% recall. Pruning the decision tree to avoid overfitting we can examine the decision-tree produced by the classifier (Fig. 2).

We can see that the decision-tree clearly works along the lines provided by our "via Negativa" approach of reducing the size of the haystack. First, it classifies cases as non-psychopaths through the first left branch. If the evidence in support of HP over HN is lower than 0.4 the decision is that most chances the text is normal. If the evidence in support of HP over HN is higher than a certain value then still most chances are that the text is normal (i.e. 63%) but in this case, the classifier is taking the impostors cues into account and suggests that if the evidence in support of the hypothesis that a text belongs to an impostor is higher than a certain value, *then the text should be dismissed as a false alarm*. This decision tree clearly illustrates the logic presented in the paper, where first the size of the haystack is reduced and then the impostors' cues are used to reduce it further.

One may question the above results as they have been gained through a specific sample. To address this concern, we have run the experiment ten times with full random

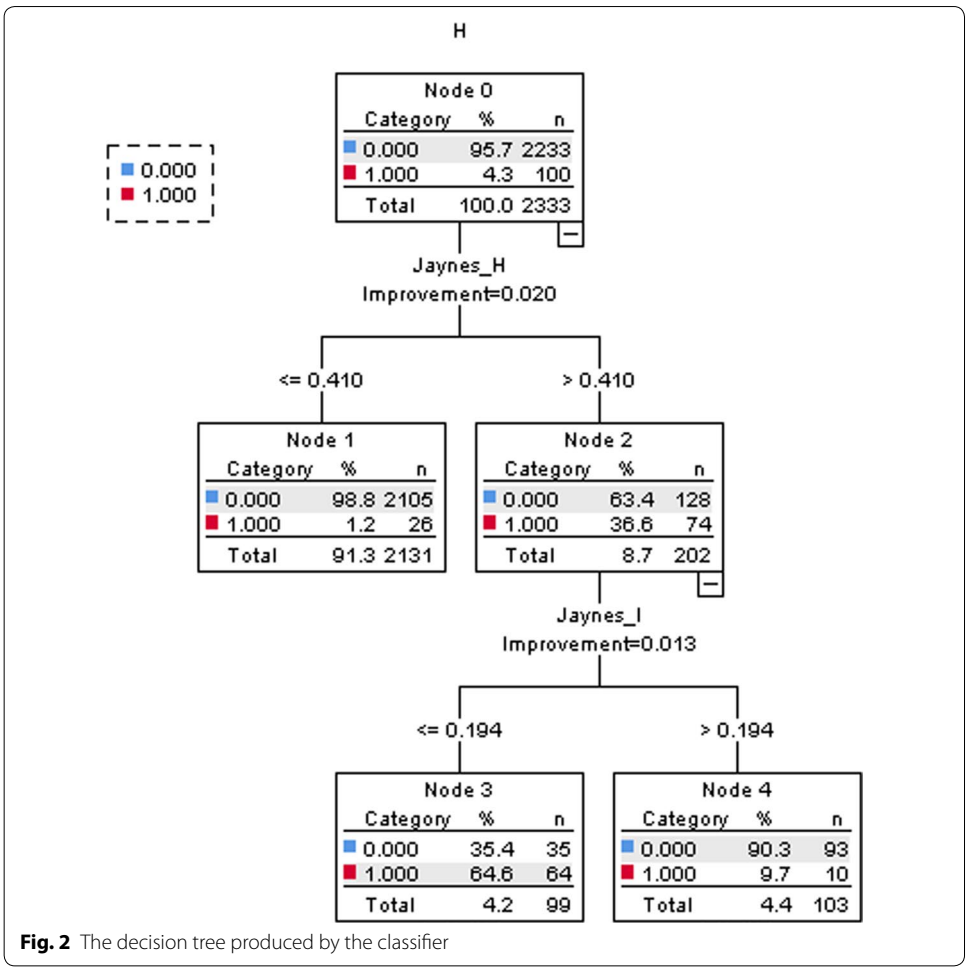

Fig. 2 The decision tree produced by the classifier

sampling of the data, from re-sampling the Reddit texts to the assignment of the psychopathic items to the texts. Using the same CRT classifier with a tenfold cross validation procedure for each run, we have on average gained 67% Precision using both JaynesH and JaynesI and only 35% Precision using JaynesH only.

Conclusions

The bread and butter of screening tests is the ability to diagnose in binary terms the existence of a "disease" based on positive versus negative test results. Such an approach, common and grounded in psychometric measurements, is limited when searching for a needle in haystack. In this paper, we have discussed the difficulty of finding a needle in a haystack, specifically in the context of screening for solo perpetrators, and made three main arguments.

The first argument, is that one possible reason for the difficulties we experience in addressing the challenge is that we adopt, whether consciously or unconsciously, the Shannon's form of information entropy that mislead us in identifying the location of the object (i.e. the perpetrator) with its recognition (i.e. identification).

Our second argument is that using Jaynes Bayesian form of hypothesis testing procedure, we may advance our screening procedure by working "negatively" and by reducing the size of the haystack.

Our third argument was that in this context the real challenge is to identify non-trivial data cues that may help us to differentiate between the real "object" that we are seeking and other objects. We've described these data cues as "impostors' cues" and further proposed that a natural way of looking for the needle is to extend Jaynes Bayesian form of hypothesis testing by including in it two different types of hypothesis testing. Using the two features produced through Jaynes elementary hypothesis testing procedure, we have shown the contribution of the "impostors' cues" for better screening the perpetrators in our dataset.

The limitations of the current study are clear. We have tested our proposed approach and hypothesis in the context of an artificial dataset where the perpetrator's signature is very strong. It is still an open question whether the proposed approach may work with regard to real and messy data-sets. However, there seems to be several benefits in applying the proposed approach. First, it urges us to seek for specific data cues that discriminate real perpetrators from impostors. Second, the cues we have used, as well as the hypothesis testing procedure we conducted, are comprehensible and can be well-understood by experts and non-experts in order to validate the process, examine it without the shadow of the "black box" threat, and improve it by incorporating experts' knowledge. Third, the screening process we have presented along the line of reasoning via *Negativa* is a highly pragmatic solution for the engineering of screening processes that do not have the pretention of identifying the needle but of reducing the size of the haystack in which it resides. The particularities of the approach we have presented must be further developed but this alternative seems promising at least as a mind challenging alternative for the practitioners working in the field. The current paper is just a first step in a possible direction and has no pretensions beyond presenting a possible way of thinking about a current and extremely bothering challenge.

Authors' contributions

Proposed the main thesis and developed the methodology: YN and YN. Wrote the code: YC. Wrote the paper: YN, YN and YC. All authors read and approved the final manuscript.

Author details

[1] The Department of Cognitive and Brain Sciences and Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel. [2] Gilasio Coding, Tel-Aviv, Israel. [3] Independent Researcher, Lehavim, Israel.

Acknowledgements

The authors would like to thank Grzegorz Wilks for clarifying several points of his paper and the anonymous reviewers for their support and constructive comments.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The code used in this study is available at: http://www.gilasio.com/home/academic-publications/JaynesCode.zip.

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

No funding.

Appendix 1. The Pseudocode

Phase 1

After gathering English texts filter them and keep the ones with word count between 40 and 400. Select 7000 texts and add to 300 of them 6 sentences randomly selected from the meanness items sentences list and mark the text to be mean. Now run empath on the 7000 texts and keep the results in a list with all the 194 categories results and the id and mark. Split the 7000 to 3 groups L1, L2, L_test equal size, each with 100 mark texts and the rest unmarked.

Phase 2

Run a median function on the L1 columns (the empath categories) and mark the entries as Boolean one if above the column median or zero if below. Count the marked texts and the unmarked as total_h1 and total_h0 and:

For each column:
For each row:
If is true and If is marked then:
Add one to count_d1h1
Else:
Add one to count_d1h0
If (count_d1h1/total_h1)/ (count_d1h0/total_h0)>3
Take the result for the category for later use

Select all the filtered results and mark them as Jaynes set 1

Phase 3

Take L2 and select only the categories in Jaynes set 1 and run the median function

For each row:

Sum the log10 of the entries where true using the result of the
Corresponding Jaynes set 1 calculation, then multiply by 10 and add − 13.76

Now select the top 100 texts by the results gained and calculate:

Probability of h1 and Probability of h0
(I.e. total h1 from the 100 and total h0 from the 100)
E_h_3 = 10*log10 (prob_h0/prob_h1)

Now run the median function on the filtered 100 texts

As before take where the category is not in Jaynes set 1
(Count_d1h0/total_h0)/(count_d1h1/total_h1) > 3 as Jaynes set 3

Phase 4

Use the L_Test
Use the median function on it and

For each row

Sum the log10 of each category corresponding to Jaynes set 1 or 3

Multiply by 10 and add − 13.76 for Jaynes 1 score or e_h_3 for Jaynes score 3

Publisher's Note

Received: 29 November 2018 Accepted: 18 January 2019
Published online: 01 February 2019

References

1. Attenberg J, Provost F. Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2010. p. 423–32.
2. Corner E, Gill P. A false dichotomy? Mental illness and lone-actor terrorism. Law Hum Behav. 2015;39:23–34.
3. Fast E, Chen B, Bernstein MS. Empath: understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI conference on human factors in computing systems. ACM; 2016. p. 4647–57.
4. Goodfellow I, Bengio Y, Courville A. Deep learning (adaptive computation and machine learning series). Adaptive Computation and Machine Learning series. Cambridge: The MIT Press; 2016. p. 800.
5. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):93.
6. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6:429–49.
7. Jaynes ET. Probability theory: the logic of science. Cambridge: Cambridge University Press; 2003.
8. Kaati L, Shrestha A, Sardella T. Identifying warning behaviors of violent lone offenders in written communication. In: Data Mining Workshops (ICDMW). IEEE; 2016. p. 1053–60.
9. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90:773–95.
10. Kimhi S, Even S. Who are the Palestinian suicide bombers? Terrorism Polit Violence. 2004;16:815–40.
11. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5:42.
12. Loh WY. Classification and regression trees. Wiley Interdiscip Rev. 2011;1:14–23.
13. Meloy JR, Gill P. The lone-actor terrorist and the TRAP-18. J Threat Assess Manag. 2016;3:37–52.
14. Munk TB. 100,000 false positives for every real terrorist: Why anti-terror algorithms don't work. First Monday. 2017. https://doi.org/10.5210/fm.v22i9.7126.
15. Neuman Y, Assaf D, Cohen Y, Knoll JL. Profiling school shooters: automatic text-based analysis. Front Psychiatry. 2015;6:1–5. https://doi.org/10.3389/fpsyt.2015.00086.
16. Neuman Y. Computational personality analysis: Introduction, practical applications and novel directions. New York: Springer; 2016.
17. Patrick CJ, Fowles DC, Krueger RF. Triarchic conceptualization of psychopathy: developmental origins of disinhibition, boldness, and meanness. Dev Psychopathol. 2009;21:913–38.
18. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. pp. 1135–44.
19. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: AAAI conference on artificial intelligence. 2018.
20. Stone LD, Keller CM, Kratzke TM, Strumpfer JP. Search for the wreckage of Air France Flight AF 447. Stat Sci. 2014;29:69–80.
21. Streiner DL. Diagnosing tests: using and misusing diagnostic and screening tests. J Pers Assess. 2003;81:209–19.
22. Tellenbach B, Burkhart M, Sornette D, Maillart T. Beyond Shannon: characterizing internet traffic with generalized entropy metrics. In: International conference on passive and active network measurement. Berlin: Springer; 2009. p. 239–48.
23. Wainer H, Savage S. Visual revelations: until proven guilty: false positives and the war on terror. Chance. 2008;21:55–8.
24. Werner KB, Few LR, Bucholz KK. Epidemiology, comorbidity, and behavioral genetics of antisocial personality disorder and psychopathy. Psychiatric Annals. 2015;45:195–9.
25. Wilk G, Włodarczyk Z. Example of a possible interpretation of Tsallis entropy. Physica A. 2008;387:4809–13.