Supplementing Police Intelligence: A Research Design Plan

Karoline Pershell, Ph.D.

Note: *This report details a research design plan for building a constellation of classifiers to identify and monitor individuals displaying language/behaviors in public social media that align with behavioral indicators of active shooter planning. This document is intended to illustrate how the Zenti system could be used.*

Zenti proposes to create a constellation of classifiers around school shooting (active shooter) phenomena, aiming to identify commonalities between people who commit these acts. *Post facto* research in this field is based on individual profiles, which provides rich data on specific characteristics (constructs) and language used to discuss such constructs. Unfortunately, such studies fail to provide a mechanism or context for gathering and analyzing data in real time. The intention of the proposed program is to allow for interventions with individuals displaying a high number or high severity of specific classifiers that have been shown to correlate with persons who have committed school shootings.

Research Question*.*

How can police identify and monitor communication leading up to an event like a school shooting?

Recommended Solution for Client.

The goal is to identify potential perpetrators of these events based on the pattern of signals they are communicating. This will be achieved by:

1) Monitoring the live Twitter stream for *Red Flags*, meaning single Tweets that score highly on one of several classifiers.
2) Once a Red Flag has been identified, the user's Twitter history is scored against a constellation of classifiers, trained to identify specific constructs.
3) If the Twitter profile can be matched with a facebook profile, the Facebook profile will also be scored against the same constellation of classifiers.
4) Potential expansions of this work would be investigation of additional social media accounts: YouTube, Vine, Instagram, etc.

Overview

Designing the Constellation of Classifiers.

Each classifier (or class) will be trained on language used to indicate a certain construct, meaning one component of a psychological profile. The constellation (i.e., collection) of classifiers will cover all psychological, social, and behavioral constructs. For example, classifier topics that may be used in identifying potential persons of interest include: Profanity, Racism, Action Planning, Anti-school sentiment, Disdain for peers, and Violence adoration.

Clearly Defined Constructs.
The real-world applicability of this system depends on the constructs, and the authenticity of the language used to train those constructs. Classifiers are built to identify these constructs. As such, classifiers need to be created by, or with, Subject Matter Experts (SMEs).

SMEs should be able to say, “We know what it should sound like when they talk about this.” Zenti’s team can work with the SMEs to map the language, identifying the latent variables and building observable characteristics of similar verbiage around several examples. (For illustrative purposes, a first pass at identifying constructs for classifier training is below. Note that it is critical to still work with SMEs to assess the gaps and identify authentic language for these constructs.)

Red Flags vs. User History.
A subset of the constellation will be used to identify red flags from the live Twitter Stream. The reason we would use a subset is because we expect that the constructs are necessary, but not sufficient, to indicate a potential person of interest. For example, it may be the case that the majority of previous school shooters used high levels of profanity in their social media, however, profanity alone is not an indicator, since many Americans use high levels of profanity in their social media.

When a red flag is indicated, the user’s Twitter history is pulled and each Tweet is scored against every classifier in the constellation, providing a readout of the intensity, frequency, and duration of these constructs for a single Twitter user. To make this data actionable, thresholds for individual scoring will need to be established, based on historical data and SME assessment for risk.

Methodology for Verifying Classifiers

The goal is to create a constellation of hyper-narrow classifier that will identify one particular construct (may include content and intent), by training the classifier on all possible language that may be used for that single construct. The following methods vary in how they (1) define the classifiers that we should use, (2) define the corpus of text that we should use to train the classifier, and (3) establish how we test the quality of the classifiers.

<u>Method 1</u>: *Use current research to <u>TRAIN</u> the classifiers. Use Historical Social Media Accounts of Known Perpetrators to <u>TEST</u> classifiers.*

(1) Build classifiers in partnership with SMEs. A first pass at identifying constructs for classifier training is below.
(2) Identify social media accounts of known perpetrators.

(3) Score the account against the classifiers. This will provide intensity, frequency, and duration of tweets across the constellation.
(4) As these were known perpetrators, we can set a baseline threshold for risk below the scoring of these individuals' Tweets.
(5) The result will be the ability to identify a subset of high-scoring variables, for which we can then monitor the Twitter stream.

In addition to real time review of the Twitter stream, Zenti maintains a historical database of [include measure of time or volume of Twitter feeds that Zenti has], which can be scored against these classifiers.

Method 2: *Use Social Media Accounts of Known Perpetrators to TRAIN Classifiers.*
Key user profiles provide a wealth of training events. When training classifiers to identify a personality feature, it is likely the case that a person who has the desired characteristic has expressed it more than once. We recommend viewing this data through several lenses to identify meaningful correlations between perpetrators.

Train on Phases of Life. Working under the assumption that reaching proclivity for extreme violence is a process, it would be meaningful to segment a known perpetrator's life into phases of extremism, and then train classifiers to identify language used in these phases. For example, phases may include:

(1) Unaffected
(2) Susceptibility to ideas of extreme violence
(3) Presented with ideas of extreme violence
(4) Accepted ideas of extreme violence
(5) Fantasized about extreme violence
(6) Planned for extreme violence
(7) Prepared to carry out plan

Use a Grounded Theory Approach. Grounded Theory is a method in psychology where the structure is determined from the information, rather than the structure being applied to the information. This process would require review of a corpus of text associated with the known perpetrators and development of the constructs (and supporting language) based on the user accounts. Potential sources for text include:

(1) Suicide/farewell notes
(2) Social Media profiles (Twitter, Facebook, etc.) of known perpetrators
(3) The last set of tweets/posts of known perpetrators

The result of the approaches of Method 2 is that we will be able to identify people who talk like known perpetrators.

Recommendation. Pursue all of the methods listed above. By developing this multi-pronged approach, at best it is developing multiple tools for identifying different classes of potential perpetrators, and at worst it is incorporating a redundancy check.

Knowledge Mapping Example

Zenti will work with SMEs to carefully map the behavioral, social, and psychological components of known perpetrators, and to extrapolate that to social media language used. As mentioned earlier, potential classifiers may be Profanity, Racism, Action Planning, Anti-school sentiment, Disdain for peers, or Violence adoration. Our goal is hyper-narrow classifiers so that we can better establish intent, and as such, these suggested classes are only a starting point for a more in-depth knowledge mapping.

For example, Action Planning (making plans for an actual attack), is a dense construct that should be further unpacked. We may eventually decide that specific items are not relevant and can be omitted, or that some elements would have similar outcome and can be combined, but this exercise is to make sure that (1) we break a compound idea up into singular ideas, (2) each of our ideas provides actionable data, and (3) we have a holistic (as opposed to piecemeal) approach so that we can identify gaps.

<u>Example:</u> Mapping out *Action Planning* with regards to an active shooter/school shooting scenario:

(1) Planners who
- Intend to escape
    - Stage vehicles in close proximity
- Do not intend to escape
    - Differ from our typical criminal in that their "mission" is to exact the greatest amount of carnage until they are killed.
    - Have a vision of themselves going out in a blaze of glory, and would be willing to commit suicide when confronted.

(2) Planning for Equipment
- Researching equipment
- Purchasing equipment
    - Handguns
    - Automatic Weapons
    - Communications devices, and hence is this intended to be a lone gunman or coordinated attack?
- Identifying sources for borrowing equipment
    - Parents, friends, other groups
- Purchasing materials

    - IED materials
    - Ammunition
- Purchasing tools or materials to build the equipments
    - 3D Printers
    - Machining tools
    - Soldering equipment

(3) Training in
- Weapons use
- Hand to hand combat
- Building clearing (room-to-room)
- Communication techniques (with coordinated team, with media)

(4) Site selection
- Time
- Locations

This non-exhaustive deep dive into *Action Planning* is just a demonstration of the type of knowledge mapping that would need to happen for all constructs which would need classifiers.

Additional Considerations

Population Identification.
It is not required that Twitter users divulge identifying information, like their name or geographic location. What we can know:

- Self-reported geographic location, when it exists.
- By filtering time zone data, “likely US” and “less likely US” Twitter users can be identified.
- Some Twitter users have geographic and time zone data left blank.

Because of the US’s unique geographic location, only the time zone on Eastern Standard Time has overlap with South America. (That is, New York, Peru and Columbia are all on the same time zone.) Additionally, Central, Mountain and Pacific time zones overlap with Mexico, the US and Canada.

Recommendation: Because of the restrictions on knowing geographic location, this method of identification and monitoring cannot be reasonably done by local law enforcement officers. Clear lines of communication must be outlined between the centralized data analysis and local law enforcement officers, as the initial suspect identification cannot be done by geographic region using only Twitter-mining.

Persons of interest could be more identified by the proper legal authorities (ie. twitter could provide IP address, cell phone number from mobile app, GEO location tags, etc. with the appropriate warrants/police involvement).