The Black Box

Authors: Arijit Douglas Sen and Derêka Bennett

Introduction

When Social Sentinel representatives pitched their service to Florida's Gulf Coast State College in 2018, they billed it as an innovative way to find threats of suicides and shootings posted online.

But for the next two years, the service found nothing dangerous.

One tweet notified the school about a nearby fishing tournament: "Check out the picture of some of the prizes you can win - like the spear fishing gun."

Another quoted the lyrics from a hit pop song from 2010: "Can we pretend that airplanes in the night sky are like shooting stars? I could really use a wish right now."

As police and administrators fielded a flood of alerts about posts that seemed to pose no threat, the company told the school in emails that it had eliminated more than half of all irrelevant alerts. Months later, they said the number had decreased by 80%. By January 2019, the company told schools its service flagged 90% fewer irrelevant posts.

But at Gulf Coast, the problem continued.

One alert from March 2019 read, "Hamburger Helper only works if the hamburger is ready to accept that it needs help."

"Nothing ever came up there that was actionable on our end," David Thomasee, the executive director of operations at Gulf Coast, said in an interview earlier this year. The college stopped using the service in April 2021.

Gulf Coast was not the only college inundated with irrelevant alerts. Officials from 12 other colleges raised concerns about the performance of Social Sentinel in interviews and emails obtained by *The Dallas Morning News* and the Investigative Reporting Program at UC Berkeley's Graduate School of Journalism. Only two of the 13, North Central Texas College and the University of Connecticut, still use the service.

As schools and universities confront a worsening mental health crisis and an epidemic of mass shootings, Social Sentinel offers an attractive and low-cost way to keep students safe. But experts say the service also raises questions about whether the potential benefits are worth the tradeoffs on privacy.

Records show Social Sentinel has been used by at least 38 colleges in the past seven years, including four in North Texas. The total number is likely far higher — The company's co-founder wrote in an email that hundreds of colleges in 36 states used Social Sentinel.

The News also analyzed more than 4,200 posts flagged by the service to four colleges from November 2015 to March 2019. None seem to contain any imminent, serious threat of violence or self-harm, according to a *News* analysis, which included all of the posts obtained through public records requests.

Some schools contacted by *The News* said the service alerted them to students struggling with mental health issues. Those potential success stories were outweighed by complaints that the service flagged too many irrelevant tweets, interviews and emails between officials show. None of the schools could point to a student whose life was saved because of the service.

Launched in 2015 by two former university police chiefs, Social Sentinel told colleges and K-12 schools around the country that its service scanned more than a billion social media posts across multiple platforms each day by comparing them to its "language of harm," allowing officials to become aware of threats in near real-time.

In October 2020, the company was acquired by the private Ohio-based school safety firm Navigate360 for an undisclosed sum. Earlier this year the company changed the name of the service to Navigate360 Detect.

Though many of its college clients seem to have canceled their use of the service, it remains popular among K-12 schools. A News investigation last year revealed that at least 52 school districts in Texas have adopted Social Sentinel as an additional security measure since 2015, including Uvalde CISD where a gunman killed 19 children and two teachers in May. In an interview with *The News* in February, Navigate360 CEO J.P. Guilbault said the company's services were used by one in four districts in the country.

A Navigate360 spokesperson called *The News* investigations' findings "inaccurate, speculative or by opinion in many instances, and significantly outdated." Social Sentinel co-founder Gary Margolis declined to comment.

Privacy and legal experts say the service may give schools the false impression that technology can help avert tragedies, while potentially exposing them to even greater liability.

“People know that it is not going to work,” said Andrew Guthrie Ferguson, a law professor at American University’s Washington College of Law. “And yet they still will spend the money because they need to have an answer to these really sad, unanswerable tragic responses in our community.”

The Black Box *Section 1*

Authors: Arijit Douglas Sen and Derêka Bennett

“Useless alerts”

For one former Social Sentinel employee, it only took three days before they had serious doubts about the effectiveness of the service.

The worker estimated that 99.9% of the flagged posts sent to clients were not threatening. The service often crashed because it flagged too many posts. At least 40% of clients dropped the service every year, the employee said.

Over the course of several months, the employee repeatedly raised concerns with supervisors and fellow employees about flaws in the system, but those complaints were often ignored, the worker said.

The employee, who asked not to be named for fear of retribution, said problems with the service were an open secret at the company, and described it as “snake oil” and “smoke and mirrors.” *The News* also contacted more than two dozen other former company employees, who either did not respond or said they had signed nondisclosure agreements preventing them from speaking publicly about their time at the company.

At the University of Texas at Dallas, which started using the service in 2018, campus police officers in charge of the service also grew increasingly skeptical of its performance, emails obtained through a records request show.

“Does the company have any data (not anecdotal) to show its success rate in mitigating harm or disaster through its alert system?” UT Dallas Police Lieutenant Adam Perry asked his chief in an email obtained by *The News*. The chief forwarded the email to a company employee who didn’t answer the question.

Perry said that while the school used the service, the technology never alerted police to legitimate threats of suicide or shootings.

“I think in concept, it’s not a bad program,” Perry said. “I just think they need to work on distinguishing what a real threat is.” UT Dallas ended its use of the service last year.

Ed Reynolds, police chief at the University of North Texas, defended the system, but also estimated that “99.9 percent (of the alerts) were messages we didn't need to do anything with.” After using the service for about three years, UNT ended its contract with the company in November 2018.

Todd Simmons, an associate vice chancellor at North Carolina A&T, told the school’s procurement director in a 2019 email that he found the service to be “of little value” and asked that it be canceled.

“Over the course of the past year, we can only point to a couple of situations that Social Sentinel reported to us on social media that we were not already aware of,” Simmons wrote. “On the flip side, the platform generated scores of useless alerts over posts that were not problematic and had no potential of becoming problematic. We estimate that 90 percent of what it alerted us to were false positives.”

Guilbault, the Navigate360 CEO, said earlier this year that the company errs on the side of caution when sending alerts.

Guilbault also said the company has seen multiple instances of the service working to prevent students from harming themselves or others, but declined to discuss specific clients.

“We've seen school staff and law enforcement liaison show up at a house and stop the kids from killing themselves,” Guilbault said.

When asked to identify a specific instance of the service preventing a student suicide or shooting on a college campus, he did not provide examples.

The Black Box *Section 2*

Authors: Arijit Douglas Sen and Derêka Bennett

Vague examples

The News did not find a student whose life was saved because of an alert from the service.

In two cases, alerts from Social Sentinel may have helped people who were not students.

The News identified 38 schools that used Social Sentinel using purchasing data, news reports and records requests, and asked all of them to cite any instances when the service helped prevent a suicide or shooting. At Gulf Coast State in Florida, Thomasee, the school's executive director, wrote in an email to another employee that the college had prevented a suicide in another state.

Thomasee told *The News* the school had received information and passed it along to a local sheriff's department, but never learned the outcome.

At Kennesaw State, emails show the service flagged a post about a student's brother, who may have been in crisis. Emails show police tried to communicate with the brother but the outcome was unclear. A Kennesaw State spokesperson denied requests for comment.

Kennesaw State initially used the service often. But documents show the university gradually stopped checking its alerts. From April to September 2020, it viewed only 34% of the alerts it was sent, records show. The former Social Sentinel employee said it was so common for schools to not check their alerts that company workers deleted particularly irrelevant ones every day without schools ever noticing.

David Perry, the former police chief at Florida State University and UNC-Chapel Hill, said police at FSU once used the service to help a student in crisis.

"We attribute that system to saving a student's life who posted on social media a cry for help," Perry said at a campus event in 2020. "Ultimately [it] concluded in a very happy ending, but there are numerous situations where this technology has saved lives, has detected weapons on school campuses as well as colleges and universities."

Perry did not respond to multiple requests via email, phone and direct message for comment from *The News*. Florida State University confirmed that it used Social Sentinel, but did not answer questions about the incident or say how many times the school used the service to prevent students from harming themselves or others.

The former Florida State chief appeared to share the same story in a testimonial in Social Sentinel marketing material.

“Our team engaged quickly and thoughtfully to check on the welfare of a student who appeared to be in crisis,” Perry’s testimonial said. “As we worked to make contact with the student it became more obvious that our intervention was necessary.”

Documents obtained by *The News* show Perry remained a strong supporter of the service when UNC was considering its contract renewal in November 2020.

“This just came in and coincidentally is where I received the tweet about the gentleman carrying his firearm,” Perry wrote in an email to administrators, referring to an invoice for the service. “I consider this a mission critical item for our intelligence as well as health and safety efforts on and round campus.”

It’s unclear what incident Perry is referring to. Local news reports from the days before the email show that UNC Police investigated a report of an armed person at the campus hospital in response to a 911 call, but found no evidence of any threat.

UNC said they did not have any further details to provide about Perry’s message.

Documents obtained by *The News* show Perry produced marketing material for Social Sentinel while he was chief at Florida State. Social Sentinel representatives also suggested him as a point of contact for other schools interested in the service. It is unclear whether he was paid for that role. Perry’s financial disclosures from FSU show he worked as a security consultant, but do not name his clients.

The Black Box *Section 3*

Authors: Arijit Douglas Sen and Derêka Bennett

Measuring effectiveness

The posts flagged by Social Sentinel frequently contained words like “shooting,” “kill” or “bomb,” which could indicate violence.

More often, the service sent schools posts related to sports, politics and jokes.

The News requested alerts from dozens of public colleges across the country. Only four — the University of North Texas, Gulf Coast, Indian River and Palm Beach state colleges — provided the text of flagged posts. Combined they include more than 4200 social media posts flagged to the colleges between 2015 and 2019.

Many referenced major political figures, including former Presidents Donald Trump and Barack Obama.

“Dear Mr. Obama: the majority of Americans are thrilled that you’re no longer in the oval office and would love to see you go away,” a tweet sent to the Palm Beach school reads. “Every time you crawl out from under the rock that you live and open your mouth, we thank God that you are no longer calling the shots.”

The word “trump” showed up at least 75 times, making it one of the 100 most common words in the posts, after removing common terms like “the” “it” “and” or “but.”

The service even picked up on Bible verses.

“How long, O Lord, wilt thou look on?” one tweet sent to Palm Beach State college read. “Rescue me from their ravages, my life from the lions!”

The flood of irrelevant alerts may have something to do with how the service was designed. In its marketing material, the company says it uses sophisticated natural language processing, an area of machine learning which focuses on text, to understand and classify social media posts as threatening or not.

The company's website also says the service evaluates the context of the posts and distinguishes between the different meanings of the same words.

Experts say that the technology may not reliably distinguish threats from jokes.

“This software is new and experimental,” said Elizabeth Laird, the director of equity in civic technology at Center for Democracy and Technology. “There's no evidence that it is effective or does what it says it will do.”

Beyond its marketing materials, Navigate360 has offered few details about the kinds of models they use to evaluate whether a tweet is threatening. Experts say that without an understanding of the type of model the company uses and the data it was trained on, the system may perform poorly or perpetuate biases.

Studies show even the most sophisticated AI language models can perpetuate racial and gender biases if they were trained on text that contained such prejudices.

Models trained on short social media posts may be particularly susceptible to the problem. A 2015 study found some popular models frequently misclassified tweets written in a dialect of English commonly spoken by Black people as another language altogether.

Nigam Shah, a professor of biomedical informatics at Stanford’s School of Medicine, said it’s also unclear how technology might predict suicidal thoughts because there is no agreement on which behaviors computers could use to reliably predict that risk.

“Suicides are such that you only find out [whether someone is suicidal] if someone succeeds,” Shah said.

Experts also raised concerns that misunderstood posts could lead to unfair punishment.

For one student, the consequences of a flagged tweet were life-changing.

The Black Box

Section 4

Authors: Arijit Douglas Sen and Derêka Bennett

The football player

As Robbie Robinson prepared to graduate from high school in 2016, he looked like he was on his way to the top of the football world.

Less than two years later he sat behind bars in a Phoenix jail cell, charged with a making a terrorist threat.

In 2016, Robinson accepted an offer from Arizona State, forgoing 26 other full scholarships from elite private schools and football powerhouses alike. The school's coaches praised both his football prowess and high grades, which landed him in the university's honors college.

As the season progressed, Robinson started to retreat from the game. He soon spiraled into depression and ultimately decided to quit the team. He took to Twitter and expressed his frustrations.

In early 2018, an athletics department staffer spotted Robinson's tweets and sent them to the school's police department. He soon received a message from someone who said they were an ASU police detective, but ignored it because he thought it was a scam.

Around the same time, ASU's police department plugged his information into Social Sentinel, a university spokesman confirmed. (ASU did not comment further on the Robinson case.)

Soon the service alerted the department to a potentially concerning message.

In the tweet, Robinson tagged another user and wrote that he had once tried to get a gun from him. "I'm tryna buy his gun because I was about to clap [the n-word] Spray the stadium up. Give me 25. Give me life. Give me liberty or give me death," Robinson wrote. "I'm not staying quiet. I'm hurt."

Shortly after sending the tweets, Robinson said a friend told him the police were outside of the apartment where he was staying. He peered out the window and saw men dressed in black retrieving AR-15 rifles from their cars. As more than 60 officers from ASU and Tempe's police departments converged on the building, Robinson called his uncle on FaceTime. He expected it to be the last friendly voice he ever heard.

After two hours of negotiation, he finally stepped out of the apartment to face the SWAT team and surrendered.

No gun was ever found.

In an interview with *The News*, Robinson said he was trying to express his past experience with mental health and suicidal thoughts and that police had misunderstood his tweets, in part because he is Black. According to court documents, he was later diagnosed with a bipolar type schizoaffective disorder, though he was deemed competent to stand trial.

In September 2018, after spending nearly seven months in jail, he pleaded guilty to a lesser felony, disruption of an educational institution, and misdemeanor marijuana possession. He was sentenced to three years of probation and released about a year and a half later.

Until *The News* contacted him, Robinson said he had never heard of the social media monitoring service that led to his arrest.

The Black Box *Section 5*

Authors: Arijit Douglas Sen and Derêka Bennett

Deadly errors

Robinson is far from the only college student to struggle with mental health issues.

Between 2017 and 2019, more than one in 10 of college-aged people reported having serious thoughts of killing themselves, according to the Center for Behavioral Health Statistics and Quality. For 15 to 24-year-olds, suicide was estimated to be the third-leading cause of death in 2020, according to the Centers for Disease Control and Prevention. Shootings on school grounds, while comparatively rare, have killed at least 50 people this year, according to Everytown for Gun Safety.

Glen Coppersmith, chief data officer of an online counseling service called SonderMind, said attempting to accurately assess the risk of a suicide from a single piece of information at a single point in time is extremely difficult.

“Any time you try to turn this into a small data problem, it’s a problem,” said Coppersmith, who also co-authored several papers on using machine learning models for suicide prevention. “A single message, a single user, a single type, a single whatever is more prone to error, especially if you're trying to look at the things that are highest risk.”

Because language changes over time and varies between groups of people, Coppersmith said it is difficult to predict whether someone poses a danger based on a tweet. His research sought to find whether there were certain shared characteristics of suicidal social media users, such as how often they repeat certain words or phrases. Coppersmith is optimistic that artificial intelligence systems using this method may allow users who opt-in to generate more insights about their mental health, before they reach the point of crisis.

Despite the potential risk, some police departments also have started to use big data tools more frequently, Ferguson, the law professor at American, said. In his book The Rise of Big-Data Policing, Ferguson said there were three main problems with how police use big data tools — racial bias, conflicts with Constitutional privacy law and a lack of transparency about how the systems work.

“We are looking at black box technologies that are promising something without the ability to look under the hood,” Ferguson said.

SOI: The presentation of commonality, diversity and interconnection are communicated with a specific purpose and impact audience reactions.

“It is distorting everything we have built a legal system around.”

Experts contacted by *The News* also raised concerns that police officers, not mental health professionals, were often the ones tasked with evaluating alerts from Social Sentinel.

Studies show police incidents involving mentally ill people are far more likely to turn deadly. A 2015 Washington Post analysis of police shootings found more than a quarter of those killed were suffering from a mental health crisis. That same year, a study by the Treatment Advocacy Center found mentally ill people were 16 times more likely to be killed by police.

“When we say things we don't mean, or we talk about things in different ways or we maybe use colloquialisms or other terms you are going to have error,” Ferguson said. “And when you're talking about error and police use of SWAT teams, that error can really turn deadly.”

In June 2002, Margolis, the Social Sentinel co-founder, testified before the U.S. Senate about how police could better respond to mental health incidents, recommending that officers and emergency dispatchers receive more training and that they work with teams of trained mental health professionals.

“Traditional criminal justice interventions for people with mental illness who commit minor crimes decimate lives,” Margolis testified. “These lives are difficult, if not impossible, to rebuild — for them and their loved ones.”

If you or someone you know is having suicidal thoughts, you can call 988 for help.