Targeting School Threats in the Digital Space

Group 36: Chanse Bhakta, Itbaan Nafi, Modupe Akintan, Raul Ruiz-Solis, Febie Lin, William Wang

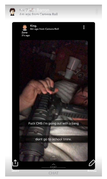
CS152: Trust and Safety Engineering / POLISCI 243C: The Politics of Internet Abuse

Problem Description

Social media use has contributed to the rise of a number of threats in digital platforms that can have negative consequences for communities (public threats) in the physical world. These digital spaces offer abusers a sense of anonymity, as well as a far-reaching audience to which they can invoke fear and panic.

An example of this took place last year when the gunman who killed 19 children and 2 teachers in Uvalde, Texas, posted a message online warning he would shoot up an elementary school minutes before the attack began.

Another example includes a threat against the high school of one of group members...



Fortunately, the situation was resolved by the local authorities and it is a testament of how timely detection and liaison with authorities could help prevent similar incidents and expedite law enforcement responses.

Our team decided to focus on the detection of imminent threats to public safety. We looked at the specific case of school threats.

Policy Language

Policy Rationale Overview

We offer users an intuitive reporting mechanism with the capacity to expedite connection to local law enforcement authorities. Precedents set by previous school shootings (where online expression preceded offline harm) point to the need of implementing this mechanism. We recognize that not all online threats are credible, therefore, our model envisions a 2-tier credibility verification measure. For public safety threats, we remove posts and disable accounts, when we believe the content could result in offline harm. We coordinate with law enforcement authorities and dispatch users to reporting routes when we believe there is a genuine threat to public safety. For general types of abuse (e.g. spam, harassment) we remove posts, disable accounts, and verify previous user behavior before making decisions.

Do not post:

- Threats to commit violent acts against schools and other public venues.
- Threats of violence that could generate a sense of collective panic among the public in certain locations and result in serious harm or accidents.
- For content that is references and/or expresses opinions about school shootings, we require additional context to proceed with enforcement

Enforcement

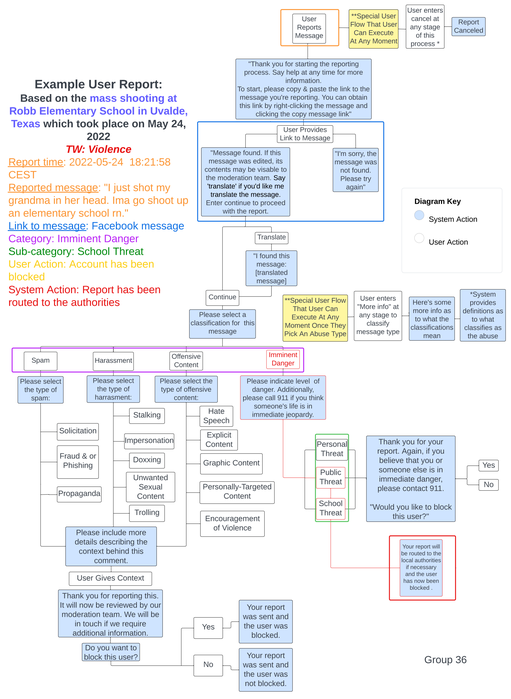
1. The detection technology developed by our team actively identifies threats.
2. Our content review team works diligently to analyze the credibility of threats and avoid false reports that could overwhelm authorities.
3. Once online abuse has been detected and evaluated, we remove content and accounts that fail to threaten public safety or fall in the category of “general types of abuse”. In the case of imminent danger threats, our teams communicate promptly with law enforcement to avert violent incidents.

User Reporting Flow

Example User Report:
Based on the mass shooting at Robb Elementary School in Uvalde, Texas which took place on May 24, 2022

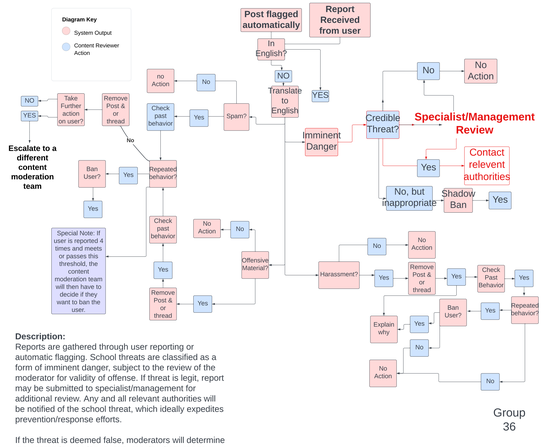
TW: Violence

Report time: 2022-05-24 18:21:58 CEST
Reported message: "I just shot my grandma in her head. Ima go shoot up an elementary school rn."
Link to message: Facebook message
Category: Imminent Danger
Sub-category: School Threat
User Action: Account has been blocked
System Action: Report has been routed to the authorities



Group 36

Manual Review Flow



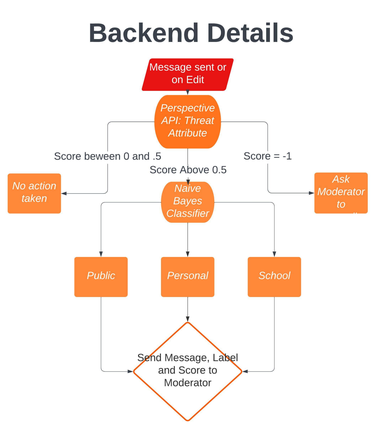
Description:
Reports are gathered through user reporting or automatic flagging. School threats are classified as a form of imminent danger, subject to the review of the moderator for validity of offense. If threat is legit, report may be submitted to specialist/management for additional review. Any and all relevant authorities will be notified of the school threat, which ideally expedites prevention/response efforts.

If the threat is deemed false, moderators will determine the appropriate action to take, ranging from shadow ban to no action.

Group 36

Technical Back-end

Our group employs a two stage classification system. The first stage uses Google's Perspective API as the backbone of our threat moderation system. The messages are fed into the 'threat' label of the API and assigned a rating between 0 and 1 based on how likely the message is to be a threat. If the score is above 0.5, we employ the Naive Bayes classifier to distinguish the message as a personal, public, or school threat. A threshold of 0.5 was shown to give the best balance in distinguishing between threats and non threats in our tests. The full message along with the label is then sent to the mod channel where a moderator can make the final decision.



If for whatever reason the Perspective API is unable to handle the message, the message is assigned a rating of -1. A message pops up to the moderator indicating that the message from a user could not be parsed and suggests reading that user's message manually.

Adversarial Cases

Our system also handles the following adversarial cases:

1. Messages that are sent in foreign (non-English) languages or non-ASCII characters (e.g. $\delta\ f \cdot \neg\ \emptyset\ \Delta$)
2. Messages that are initially sent and later edited to an abusive message

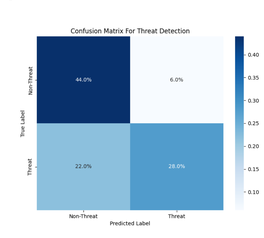
Evaluation

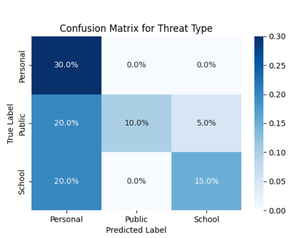
Our group's back-end technology use of Google's Perspective API in combination with our Naive Bayes classifier is a powerful tandem of tools to help detect and classify threatening messages in the context of DMs and group chats as one might see on Facebook or WhatsApp. The implementation accomplishes our original goal of detecting threats and giving the moderator power of whether a threat should be escalated to the authorities or if other outcomes should be taken. It can successfully detect messages that appear in unicode, foreign languages, or if a user edits a message. However, the detection and classifier possesses issues when context is missing. A message such as "don't come to school tomorrow" implies a threat, but is not caught our classifier. If we make the threat assessment more sensitive, we risk overwhelming the authorities with false reports. By the same accord, we need to detect any plausable threat that may arise, which our bot accomplishes with moderate effectiveness.

Qualitative

1. The perspective classifier performs relatively well in practice. However, using our data of synthetic threats and real threats, we found that the model is not as robust as we had expected. The data we tested took into account threats that did not have sufficient context, which is likely why has a high false negative rate.
2. While the performance of the Naive Bayes classifier is currently limited, there is potential for improvement with a larger data, and more robust data set.

Quantitative





The recall of the perspective model was 0.56 which means the positive threats were only flagged correctly a little over half the time; Precision was also 0.82 which means that there are less false alarms that may overwhelm authorities. Finally, the accuracy was 0.72 which means the model makes the correct prediction well.

Looking Forward

Our team believes that our design would fare well on digital platforms for the purpose of moderating threats. Our system provides both automatic detection and manual reporting of threats, the result of which is expedited reporting to necessary authorities. Additionally, the reporting flows are easy to follow such that reporters are not discouraged from acting on abusive content.

Future directions for this project include:

1. Expand training data (possibly with the use of a generative model) to improve the accuracy of our classifier
2. Allowing reporters to link multiple messages within a single reporting
3. For the "see user history" functionality, provide a robust description of the previous violations committed by the user