tech against terrorism

Case study: Using the GIFCT hash-sharing database on small tech platforms

This case study is intended to provide insight into how a Public-Private Partnership has contributed to the fight against terrorist use of the internet by using the Global Internet Forum to Counter Terrorism's (GIFCT) hash-sharing database to support small platforms. The study was completed with the assistance of JustPaste.it, one of the leading small platforms that has joined the hash-sharing consortium established by the GIFCT to identify terrorist imagery. This guide provides insight into the benefits and challenges of deploying this technology.

What is the Global Internet Forum to Counter Terrorism (GIFCT)?

The GIFCT is an industry-led coalition formed by Facebook, Twitter, Google, and Microsoft which collaborates with a wide range of NGOs, academic experts, and governments. The objective of the initiative is to substantially disrupt terrorists' ability to promote terrorism, disseminate violent extremist propaganda, and exploit or glorify real-world acts of violence on the internet. The GIFCT focusses on: conducting and funding research; sharing knowledge, information and best practices; and employing and leveraging technology. The GIFCT is working in close partnership with Tech Against Terrorism initiative to share knowledge and expertise with smaller tech platforms.

What is JustPaste.it?

JustPaste.it is a "pasting" site that allows users to share content anonymously. The site is run by one person, who works on the site in his spare-time. Unfortunately, groups like ISIS have exploited the site to re-share propaganda material. Despite these challenges, JustPaste.it has made significant strides in removing terrorist content shortly after it is posted.

The GIFCT's hash-sharing consortium

Terrorist organisations use the internet to recruit, propagandise, and connect. Technology companies, both independently and collectively, are pushing back. This is not easy: smaller platforms in particular often do not have the resources to build large operations teams or develop sophisticated algorithmic policy enforcement tools. That is why the GIFCT has developed a shared database of hashes of known terrorist content. Updated continuously by participating companies, the hash-sharing consortium empowers smaller platforms to find and act on terrorist content on their platforms. Currently, there are more than 80,000 visually distinct images hashed in the database and more than 8,000 visually distinct videos in the database.

How does it work?

The hash-sharing consortium is designed to be flexible for participating companies. They sign an NDA, MOU, and, if necessary, no-cost licenses to use specific content-hashing techniques. Companies are NOT required to agree on shared content policies in order to participate, which means that if a company finds a match to content in the database they are not obligated to report on that match to the consortium or anyone else. They are not even required to take action. The purpose is simply to empower companies to enforce their own standards more effectively by enabling companies to tip one another off to potentially dangerous material at scale.

The database itself is hosted on Facebook's ThreatExchange platform which was originally developed to allow companies to share hashes of malware with one another. ThreatExchange is secure, flexible, and robust; data shared via the hash-sharing consortium is available only to other members of the consortium, not all users of ThreatExchange. ThreatExchange also has a mature API, and many consortium members choose to integrate with that platform in that way. Others use a graphical interface. The consortium uses multiple hashing techniques to facilitate easy engagement by companies that may have familiarity with one technique over another.

The consortium has developed guidelines for the type of content to be uploaded into the database. When a hash is uploaded it includes metadata that corresponds to the company that uploaded it and a code that corresponds roughly to the type of content reflected in the hash. The hash-sharing consortium does not share personally identifiable information.

How Do Small Companies Get Involved?

The first step to joining the hash-sharing consortium is signing an NDA, which facilitates more detailed discussion about ThreatExchange, the coding scheme, and the hashing techniques we use. This also gives existing members of the consortium an opportunity to vet interested companies. Companies then sign an MOU, register for ThreatExchange, and get set up with one or more of the supported hashing techniques. The Facebook team provides support and guidance for all of those technical aspects, and the hash-sharing companies more broadly sometimes collaborate to write and share useful scripts and tools. After that, implementation is a matter of engaging the ThreatExchange API and developing tools and processes for managing matches to hashes that are found on the platform. Many of the GIFCT member companies are happy to provide support and guidance for developing those protocols, but they are ultimately the purview of each platform.

TECH AGAINST TERRORISM

Tech Against Terrorism is an initiative launched and supported by the United Nations Counter Terrorism Executive Directorate (UN CTED) that supports the tech sector in tackling terrorist exploitation of the internet whilst respecting human rights.

Our work consists of three pillars:

1. **Outreach:** Convening tech companies of all sizes, counter-terrorism experts, government officials, and civil society at in-person workshops and seminars, as well as video conferences, to learn more about the nature of the threat and how smaller companies can play their part to help address terrorist exploitation

2. **Emerging Best Practice and Knowledge-Sharing:** Facilitating knowledge-sharing amongst tech companies through in-person workshops and our online Knowledge Sharing Platform.[1] We work with companies to improve their Terms of Service / Community Guidelines, develop balanced approaches to content moderation, and produce transparency reports to increase accountability. We also collaborate with the GIFCT to facilitate knowledge-sharing with smaller tech companies

3. **Capacity Building:** Working with companies to provide individual training workshops on Terms of Service, Content Moderation and Transparency Reporting. Building online tools to help small tech companies tackle terrorist exploitation of their platforms

USING THE HASH-SHARING DATABASE

Getting access to the hash-sharing database

JustPaste.it was given access to the GIFCT's hash-sharing database by Facebook. JustPaste.it saw this as a means of getting "the biggest impact possible" in terms of tackling exploitation of the platform and explained that they wanted to examine "all images hosted on the site". Before granting access, JustPaste.it signed a non-disclosure agreement and an agreement "licensing" JustPaste.it to use the GIFCT hash-sharing database.

The hash-sharing database in practice

Initially JustPaste.it deployed the hash-sharing database as a one-off procedure but then implemented an almost real-time version of the platform that enables continuous and automated consultation of the database. Given that JustPaste.it does not host video, only photos are matched. In total, this process took around a month.

JustPaste.it received support from Facebook at all stages of the technical integration. For example, Facebook provided JustPaste.it with a handbook containing a detailed description of the technology and on-boarding procedure. Furthermore, a Facebook developer assisted JustPaste.it with building a plug-in for the programming language used by JustPaste.it that was used to generate PDQ hashes[1] from the images on JustPaste.it. Once these hashes were generated, JustPaste.it was able to compare them with the hashes in the GIFCT's hash-sharing database. In order to compare the hashes, Facebook supported JustPaste.it with Hamming distance[2] measuring, a procedure that measures the proximity between hashes and can therefore determine whether two hashes are similar enough to constitute a "match".

JustPaste.it then ran the hashes comparisons between its database and GIFCT's hash-sharing database. As an additional measure, JustPaste.it verified the matches to determine that they had correctly matched identified terrorist content. JustPaste.it told Tech Against Terrorism that all matches were accurately identifying terrorist content, and this allowed for JustPaste.it to remove 10-12,000 entries containing terrorist imagery.

[1] PDQ hashes are hashes developed by Facebook

[2] Oxford Math Center, "Hamming Distance and Error Correcting Codes": http://www.oxfordmathcenter.com/drupal7/node/525 (accessed on 14 May 2018).

CONCLUSIONS

1. Benefits of using the hash-sharing database according to JustPaste.it

- According to JustPaste.it, all image matches made through the database identified terrorist content with “almost 100% accuracy”. As a result, JustPaste.it were able to remove 10-12,000 entries containing terrorist imagery.
- Human verification is still possible. Companies can verify that matches made from the database are correct, thereby ensuring that there is a “human in the loop”.
- The database gives platforms more independence in their moderation efforts - having access to the database allows companies to consult verified terrorist content and decreases dependence on user content reporting and government takedown requests.
- Companies do not have to implement the database by themselves and have access to GIFCT support. JustPaste.it received assistance from Facebook in both hash generation and matching.

2. Challenges in using the hash-sharing database

- The process requires some time commitment. JustPaste.it estimates that it took a month from initiated to content removal.
- Using the database once might not be enough. JustPaste.it initially consulted the hash-sharing database as a one-off procedure however then made some changes to its platform allowing the company to examine uploaded imagery against the database in “almost real time”.
- Since the process can be time-consuming for a small company with limited time and resources, JustPaste.it suggests starting small and implementing as a one-off procedure rather than trying to immediately implement a “perfect” solution.

3. Key learnings

- The procedure is repeatable, and more small tech companies could use the GIFCT hash-sharing database. Given the assistance supplied by Facebook in hash generation and matching, other companies could replicate JustPaste.it’s procedure. Since JustPaste.it does not host video they only used matching for image content, although the database does hold hashes from videos. Platforms hosting video can therefore deploy the database for both types of imagery.
- Examining a platform’s entire repository of imagery over all history is time-consuming given that it entails generating hashes and matching them against all content uploaded on the platform. JustPaste.it therefore recommends initially matching content from a more limited time period, for example one month prior, to more easily get a grasp of how the hash-sharing database works.
- Companies should strive towards deploying the database in tandem with Tech Against Terrorism’s Knowledge Sharing Platform, which provides companies with operational support on how to moderate content once detected by the hash-sharing database.