GIFCT WORKING GROUPS OUTPUT 2022

Research Call for Proposals: Multimedia Content Classifiers

Technical Approaches Working Group

GIFCT

Global Internet Forum
to Counter Terrorism

For development of a system to classify multimedia content as terrorist or violent extremist content.

Background

The Global Internet Forum to Counter Terrorism (GIFCT) is a non-profit organization with a mission to prevent terrorists and violent extremists from exploiting digital platforms. Our vision is to build a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.

Founded by Facebook, Microsoft, Twitter, and YouTube in 2017, GIFCT was established to foster technical collaboration among member technology companies, advance relevant research, and share knowledge with all our member companies. Since 2017, GIFCT's membership has expanded beyond the founding companies to include eighteen diverse digital platforms committed to cross-industry efforts to counter the spread of terrorist and violent extremist content online.

Three strategic objectives provide the focus for GIFCT to realize its vision:

1. Be a leading organization to convene, engage, and provide thought leadership on the most important and complex issues at the intersection of terrorism and technology, demonstrating with concrete output that multistakeholderism can deliver genuine progress.
2. Create a global, diverse, and expansive community of GIFCT member companies reflective of the ever-evolving threat landscape.
3. Build the collective capacity and capability of the industry by offering cross-platform technology solutions, information sharing, and practical research for GIFCT members.

Content moderators need to make decisions about whether specific content violates the content policies of social media platforms. Given the large volume and breadth of content, it is important to be able to prioritize specific content for human moderation.

As we design systems to support content moderation by skilled human reviewers, we should aim to ensure that they are provided the nuanced information that they need in as accessible a format as possible.

In many contexts, machine learning has been shown to contribute to, and potentially amplify, societal inequity, furthering the unjust treatment of people who have been historically discriminated against[1]. While inequity is not an inevitable consequence of these models, it is essential to identify such potential effects through proactive and reactive means.

[1] Bommasani, R. (2021, August 16). On the Opportunities and Risks of Foundation Models. THE FREEMAN SPOGLI INSTITUTE FOR INTERNATIONAL STUDIES. https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models.

Aims

GIFCT is seeking proposals for the development of a system to classify multimedia content as conforming to some definition of terrorist and violent extremist content in a way that is contextualized and explainable, and provides some degree of confidence or probability to the user (hereinafter, the "Solution"). The Solution is intended to be used to inform human content moderators decisions about terrorist and violent extremist content and help prioritize their reviews.

Requirements

- Given some definition of violent extremism, the Solution should be able to classify content as belonging to that definition or not, and to what probability or confidence that it is part of that definition.
- The Solution classifies content as being violent extremist or terrorist content based on a broad understanding from experts in the field.
- The Solution will provide sufficient context and explanation to the user that can help determine why the classification decision was made.
- The Solution will be able to classify at least one content type but may also be multi-modal and consider content types such as audio, video, images, text.
- The Solution can be executed using a standard machine learning framework such as PyTorch or TensorFlow.
- The Solution can be further fine-tuned by users through training it on additional content.
- The Solution and the process of building and training the Solution will be shown to provide sufficient data protection to protect user privacy in line with GDPR and other regulations.
- The vendor will be shown to have taken reasonable steps to identify and address potential issues of bias in the Solution and in the process of building and training the Solution.
- To the extent the Solution includes or is integrated with any third-party intellectual property, such IP will preferably be licensed under an open source license (such as those listed here: https://opensource.org/licenses/), or alternatively, and only after consultation with GIFCT, can be made available with a perpetual, fully paid-up license in favor of GIFCT and associated entities and persons for use in preventing terrorist or violent extremist content.

Evaluation

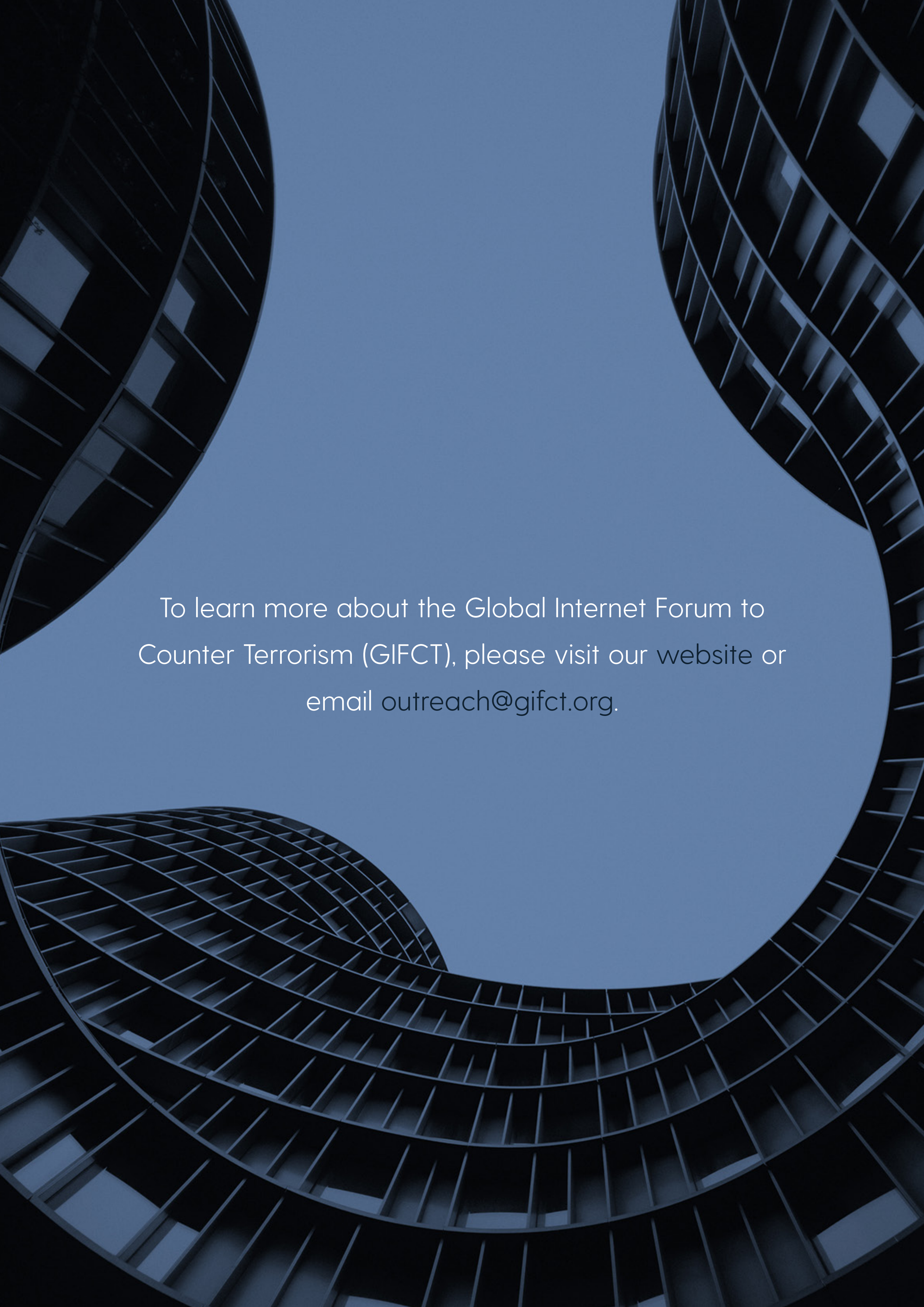
- Performance of the Solution across a broad range of content types using metrics such as precision and recall.
- Performance of the Solution to capture a broad range of violent extremist groups and ideologies.
- Alignment with GIFCT's Mission.
- Alignment with GIFCT's Values.
- Sensitivity to Human Rights and Ethical issues that may arise.

Deadlines and Format

GIFCT aims to begin the research project in September 2022 for completion by July 2023.

Proposals or other inquiries should be submitted to tech@gifct.org with the email subject “Classifier Proposal” by Friday, June 24, 2022 and proposals should include:

- A brief overview of your proposed approach to this project
- An estimated timeline for delivery
- An estimated budget or costs
- A brief overview of the organization or team that would deliver the project



To learn more about the Global Internet Forum to Counter Terrorism (GIFCT), please visit our website or email outreach@gifct.org.