GIFCT WORKING GROUPS OUTPUT 2022

Research Call for Proposals: Machine Translation

Technical Approaches Working Group

GIFCT
Global Internet Forum to Counter Terrorism

For development of a multilingual machine translation model capable of recognizing the nuanced use of language specific to a violent extremist context.

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 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. Since 2017, GIFCT's membership has expanded 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. 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 moderation is a challenging and resource-intensive task. The subject matter expertise required to accurately assess if a piece of content is terrorist or violent extremist material is rare and to achieve this level of understanding in multiple languages even more so. As critiques by groups such as EFF point out, "Automated technology doesn't work at scale; it can't read nuance in speech the way humans can, and for some languages it barely works at all. Over the years, we've seen the use of automation result in numerous wrongful takedowns. In short: automation is not a sufficient replacement for having a human in the loop."[1] As we design systems to support content moderation by skilled human reviewers, we should aim to ensure that they are provided the nuanced information they need in as accessible a format as possible.

Existing machine translation models can help to a certain extent. However, current multilingual

1 Jillian York and Corynne McSherry, "Automated Moderation Must be Temporary, Transparent and Easily Appealable," Electronic Frontier Foundation, April 2, 2020, https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable.

models may not deeply model the subtleties of languages and language varieties to their full extent.[2] More concerning is how 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.[3] While inequity is not an inevitable consequence of these models, it is essential to identify such potential effects through proactive and reactive means.

Aims

GIFCT seeks the development of a multilingual machine translation system that is capable of recognizing the nuanced use of language specific to a violent extremist context, enabling subject matter experts to apply it toward moderating content efficiently in multiple languages. We are also conscious of the human rights and ethical implications of applying such technologies and seek to apply a human rights-based approach to the development, evaluation, and application of any solutions developed.

Requirements:

- A solution that is capable of recognizing the nuanced use of language specific to a violent extremist context.
- The model will be capable of translating text from multiple different languages into a target language (English).
- The model can be executed using a standard machine learning framework such as PyTorch or TensorFlow.
- The model and the process of building and training the model 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 model and in the process of building and training the model.
- To the extent the model includes or is integrated with any third party intellectual property, such IP will preferably be licensed as open-source (e.g., https://opensource.org/licenses/), or alternatively (and only after consultation with GIFCT), can be made available in perpetuity with a fully paid license in favor of GIFCT and associated entities and persons for use in preventing terrorist or violent extremist content.
- Vendors must not have any conflict of interest with GIFCT or GIFCT staff.

Evaluation

- Performance of the model across a broad range of languages in line with the XTREME benchmark

2 Zihan Wang et al., "Extending Multilingual BERT to Low-Resource Languages," Findings of the Association for Computational Linguistics: EMNLP 2020, (November, 2020): 2649–2656, https://aclanthology.org/2020.findings-emnlp.240/.

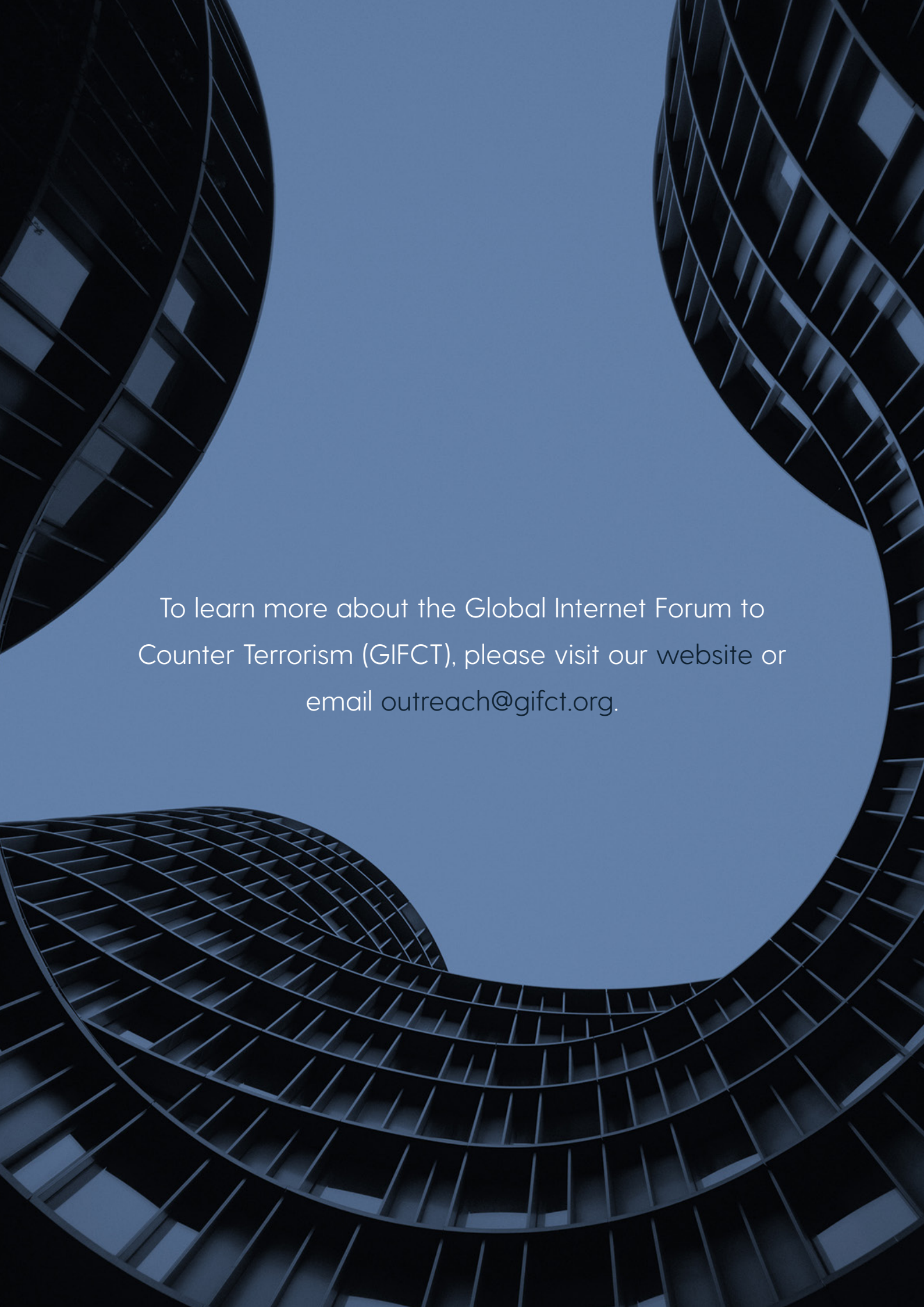
3 Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," Center for Research on Foundation Models (CRFM), August 16, 2021, https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models.

- Performance of the model to capture the cultural specificity of different violent extremist groups' use of language
- Alignment with GIFCT's Mission and Values
- Sensitivity to Human Rights and Ethical issues that may arise

Deadlines and Format

GIFCT aims to begin the research project in July 2022 for completion before the end of the calendar year. Proposals or other inquiries should be submitted to tech@gifct.org with the email subject line "Machine Translation Proposal" by Friday, June 24, 2022. Proposals should include the following:

- 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.