Christopher Tuckwood , Drew Boyd

Christopher Tuckwood , Drew Boyd

IDRC GRANT / SUBVENTION DU CRDI : - HATEBASE - GLOBAL NETWORK BUILDING AND ARTIFICIAL INTELLIGENCE FOR MULTILINGUAL HATE SPEECH MONITORING

Project title: Hatebase - Global Network Building and Artificial Intelligence for Multilingual Hate Speech Monitoring

IDRC project number and component number: 109128-001

By: Christopher Tuckwood (project leader) and Drew Boyd (technical team member)

Report type and #: Final Technical Report (Technical Report 3)

Period covered by the report: 1 October 2021 to 30 September 2022

Date: 31 October 2022

Country / region: Global

Full name of research institution: The Sentinel Project for Genocide Prevention

Address of research institution:

1400-18 King Street East
Toronto, Ontario
M5C 1C4
Canada

Name of project leader: Christopher Tuckwood

Contact information of project leader:

Phone: +1 (647) 222-8821

Email: chris@thesentinelproject.org

Table of Contents

1.0 - Synthesis

This final technical report for the “Hatebase - Global Network Building and Artificial Intelligence for Multilingual Hate Speech Monitoring” initiative aims to summarize the activities undertaken, tools and processes developed, and lessons learned over the course of the entire project. The project team face unique and interesting technical challenges in addition to disruptive world events but the findings and outputs produced through their efforts represent a rigorous and dedicated commitment to research on the subject of artificial intelligence (AI), hate speech, and the global networks required to meaningfully engage in confronting the proliferation of online hate.

Significant effort went into developing the tools and methodologies used to demonstrate how AI might be applied in both international development and humanitarian aid fields, as well as in other contexts, to monitor and assess online hate speech. The open-source code base produced by the project is available for other researchers and practitioners who may possess only limited technical expertise. The full findings of this effort can be found in the additional document submitted with this report. Due to COVID-19 restrictions, the travel and in-person collaboration of research nodes was converted to an online collaboration approach. Despite this setback their value in global hate speech monitoring was beneficial and substantial two-way value was generated by the project. Participants contributed to the refinement of the dataset while also benefiting from its outputs. Research into the relationship between online hate speech and offline violence did not yield meaningful statistical correlations and the small body of research on the subject reflects this limitation. Nonetheless, hate speech monitoring logically has value in supporting offline violence prevention efforts because it can inform the sentiment and intentions of groups that traffic in the hate speech that acts as an incendiary element within broader societies.

2.0 - Research Problem

It is still widely accepted that hate speech is a precipitator of violence, particularly when it is directed against vulnerable minorities. This relationship was a major aspect of the research problem that the Hatebase project was intended to address and it remained so throughout the project. While the project did not identify a strong correlational relationship in this regard, this simply highlights the need for further research since there is more to understand about how the use of hateful language, especially by people in positions of influence, can contribute to societal polarization which can lead to instability, persecution of vulnerable groups, and the escalation of violent conflict. This risk dynamic is present in all human societies but the impact of hate speech is particularly pronounced in fragile states, where it has grown in recent years as a result of increased internet connectivity. Greater digital access enables malicious actors to disseminate hate speech through social media and contributes on a large scale to violent persecution, armed conflict, mass atrocities, and genocide in countries as varied as Myanmar, Syria, and South Sudan. Developed countries are not immune to this issue either, as demonstrated by growing polarization and xenophobia in the United States, several European countries, and even Canada. Alongside other critical phenomena such as misinformation, hate speech threatens to undermine

trust and social cohesion in communities all over the world, thereby threatening democracy, peace, human rights, and development.

This situation, which is unlikely to improve on its own, has continued contributing to greater public and governmental pressure on technology companies, such as Facebook and Twitter, to find ways of reducing the proliferation of hate speech on their platforms, though attention on this issue has faded significantly over the course of this project. While reducing the prevalence and impact of hate speech requires a multifaceted approach both online and offline, it was clear at the outset of this project that technology could play a role in addressing the problem. This potential remained clear throughout the project, though the project team's experience has also provided insights on the limitations of technology in this regard, such as how it can support but not replace overburdened human efforts to moderate online content. For example, social media platforms currently focus significant human resources on the moderation of content which users publish on their platforms, such as determining whether to remove posts that other users have flagged as offensive or dangerous. Unfortunately, even if social media companies were to invest in significantly expanding their content moderation teams, the sheer volume of hate speech constantly circulating online exceeds the capabilities of human moderators, who themselves often suffer negative effects from excessive exposure to such harmful content. If major social media companies are not able to properly meet the need for more effective hate speech monitoring and moderation then government agencies and civil society organizations are in even weaker positions to do so. This situation highlights the need for increasingly effective automation which can assist all stakeholders with recognizing hate speech online, maintaining situational awareness about changing trends, and informing effective interventions to counter hate speech and its negative impacts.

Fortunately, the point about informing effective interventions highlights the fact that the pervasiveness of online hate speech presents not only a threat but also an opportunity since these large volumes of data are, when combined with other data sources, potentially useful as indicators of spiraling instability and enable the possibility of early intervention. Despite the cautionary note above regarding the apparently weak statistical relationship between online hate speech and offline violence, there remains significant potential in this regard. Much as was stated at the beginning of this project, it is still true that being able to effectively address hate speech requires the enhancement of several aspects of current monitoring efforts. Existing technologies still need to be improved in order to be able to effectively recognize hate speech across multiple languages, cultural contexts, and data sources. Another closely related aspect continues to be that effective multilingual and multi-regional monitoring also requires significant non-technical effort since language is constantly evolving and software can only be effectively trained to monitor hate speech through regular engagement with the native speakers of various languages who also contribute critical contextualizing factors. Lastly, if software and data are to be turned into useful warning tools then it is necessary to develop an understanding of how online hate speech relates to offline "real world" events, such as determining whether there truly is a relationship between changes in the incidence of online hate speech and negative offline outcomes. A new addition to this aspect of the research problem concerns not only the relationship between online hate speech and offline violence but also the relationship between both of those phenomena and other

types of harmful content and potential warning indicators, such as rumours and misinformation. It appears that various individuals and institutions working as researchers and practitioners are increasingly interested in this multifaceted relationship and what it means for efforts to predict and prevent violence and other forms of instability.

3.0 - Research Findings

The project team has made the following progress in relation to the research questions for this project.

(1) Question - How can automated hate speech monitoring be improved to reduce the need for human moderation and, specifically, in what ways can AI and natural language processing be most efficiently employed for this purpose?

Progress - Due to the complexity of this issue, the project team has submitted a separate document detailing the research findings and other outcomes related to this particular research question along with this report.

(2) Question - How can global networks of research nodes collaborate sustainably to keep pace with the constantly evolving nature of online hate speech?

Progress - Human moderation thresholds for recognizing hate speech with a significant degree of reliability are still surprisingly high even when considering the automated process of analysis that can be accomplished by machine learning algorithms. Global networks of hate speech research nodes can ease this requirement by distributing the workload across organizations, individuals, language groups, and cultural contexts.

These nodes also serve the purposes of keeping pace with the evolution of different languages, including the introduction of new terminology, especially ever-changing colloquial usage which has a close relationship with hate speech. Most importantly, human moderation and data submission ensures that the highest-priority vocabulary and most relevant terms are acquired since human review is able to distinguish importance in a way that machine learning still struggles with. Rather than scooping up vast amounts of potentially low-priority terminology (e.g. rarely used and increasingly obscure terms) it takes resources away from addressing the data that is more relevant to the discussion of hate speech.

Hate speech research nodes can attain sustainable collaboration by acting as a component of the research community rather than a director of it, where all participants see value in not merely extracting data but also providing it to a centralized system for analysis. This broader analysis - conducted through the Hatebase algorithm - then provides a basis for additional insights from participants

without necessitating the development of their own software tools, which can be a specialist-intensive process.

The value and sustainability of a global network largely depend upon (1) whether the benefits of such a network compel a sufficient number of participants, (2) whether the data collected is relevant to all participants, and (3) the inputs and outputs, whether they be new vocabulary or analysis provided by the Hatebase algorithm, and whether they provide meaningful insights to the discussion surrounding online hate speech.

(3) Question - What is the relationship between online hate speech and societal instability (including violence and state fragility)?

Progress - The relationship between online hate speech and societal instability is complicated and difficult to accurately assess. For this research project the project team divided these relationships into online hate speech linked to real-world violence and hate crimes, and online hate speech linked to state fragility. Though there are linkages between violence and state fragility in general, we have divided these interrelated subjects to better delineate their roots in order to assess the influence of online hate speech.

State fragility and the larger socio-political influences of hate speech appear to have a much smaller affinity in the broader sense due to the fact that these challenges are exacerbated by the adoption behaviours of a sizable and meaningful number of individuals within a given state. A larger number of individuals involved makes it less likely that one single variable (such as hate speech) would be a significant motivator for all or even most of them. For that reason the statistical connection between online hate speech and state fragility may exist but the evaluation of it as a variable is incredibly difficult to ascertain against the background noise of many other varied influences. For this reason, the research team was not able to observe meaningful statistical relationships but believe that further research may shed light on these dynamics.

When assessing the potential for causal influences of online hate speech and offline violence outside of the context of larger state fragility, we can glean a great deal from the small but growing body of research conducted by other academics and to which this research project hopes to contribute. In the article “Race, Ethnicity and National Origin-based Discrimination in Social Media and Hate Crimes Across 100 U.S. Cities” [link] the authors make no claims of a causal relationship between online hate speech and offline violence, though noted that in the regions studied which correlated to high hate crime rates the nature of online hate speech geolocated to those same cities had unique qualitative differences. Most notably, the online hate speech for these regions was targeted rather than self-narrated. Targeted hate speech is speech made with discriminatory intent and

self-narrated is merely the sharing of exposure to discrimination. This would appear to indicate some manner of relationship between explicit, pointed hate speech online and real-world violence but the direction and magnitude of this relationship is not clear given the dataset being used. It is also worth noting that the researchers selected Hatebase's hate speech terminology dataset as the basis for its study.

In "Hate In The Machine: Anti-Black And Anti-Muslim Social Media Posts As Predictors Of Offline Racially And Religiously Aggravated Crime" [link] the authors argue that online hate speech may influence offline hate crimes but that it is "unlikely that online hate speech is directly causal of offline hate crime in isolation." Hate speech is not a single act but part of a larger process. This perspective offers a more realistic - if unsatisfying - hypothesis on the role of hate speech in offline violence. When a subject as complex and multi-variable as violence is analyzed, it is common to find that the causes which shaped the event are a confluence of elements rather than a single reason and this complexity increases as one delves further into the motives of multiple unconnected perpetrators. For this reason, a definitive causal relationship between online hate speech and offline violence remains elusive even if the relationship appears to have logical linkages.

These findings are echoed in *Understanding Online Hate Speech as a Motivator and Predictor of Hate Crime* [link] in which the authors state that their research failed to show strong links between online hate speech and offline violence and therefore no significant conclusions can be drawn. It also notes that despite social media platforms implementing policies which meaningfully cut down on the occurrence and viewership of hate speech, most statistics show a steady upward trend in hate crimes. This likely hints at a far more complicated relationship than common sense might indicate. Further research is required to better understand these dynamics and to investigate the factors of online hate speech which may or may not influence offline violence.

Ultimately, a concept which appears reasonable upon general reflection is a notion that is difficult to accurately identify with data due to the intensely complex relationships between an enormous number of variables. State fragility remains far more complex so meaningful causality is not possible within the scope of this research project. However, even where quantitative and qualitative data is more available such as with social media platforms, the causal link between online hate speech and offline violence is not readily apparent. This is not to suggest that no link is present, nor that any initiative premised on a link between online hate speech and offline violence is inherently flawed. Instead it indicates that such links cannot simply be assumed and their effects known. Proceeding in the belief that online hate speech *may* influence or incite offline violence is a sensible and cautious approach that can balance the understanding of more scientific support being

required for the theory while not removing potentially impactful action because of this dearth of data.

(4) Question - How can online hate speech monitoring meaningfully support offline efforts to prevent violence and build societal cohesion?

Progress - It is entirely possible to use correlations between online hate speech and real-world violence to inform offline initiatives to prevent violence and build social cohesion.

The mechanism for this action is to flag terminology, intent, and thresholds for these elements and track their occurrence. When these flagged terms and intent indicators rise beyond a specific threshold it can trigger an alert which notifies on-the-ground actors to take additional steps to ease tensions in their communities. Beyond that, the long-term ebb and flow of these indicators may provide insights into what sort of events are most likely to prompt alerts and community actors can work to address those factors before they become prevalent concerns.

Importantly, future efforts will need to consider the relationship between hate speech, rumours, and misinformation that circulate both online and offline, including their collective relationship with physical violence. It is increasingly recognized in the field that these phenomena cannot be considered in isolation from each other.

Lastly, though the project team recognizes that robust causative links could not reasonably be established by this limited research project, it is still useful to inform further research efforts.

4.0 - Project Implementation and Management

The project largely progressed as planned, despite significant pandemic-related disruptions, and included the following notable milestones and activities.

- Deployment of application program interface (API) module for interacting with vocabulary and sightings database
- Development of general data cleaning tools for end users
- Initial development of natural language processing (NLP) algorithm using sample dataset
- Language processing, entity recognition, and sentiment analysis elements mostly complete
- Implementation of k-means clustering algorithm and accompanying visualization framework

- Onboarding of citizen linguists and new hate speech vocabulary added to database
- Planning and implementation for regional node development completed
- Online panel discussion held in August 2021 with 55 attendees from North America, South America, Europe, Asia, and Africa
- Refinement of initial training models for machine learning algorithms
- Initial structure for final software package deployment
- Invited to present at the following conferences:
    - Mozfest - March 2021
    - Davos Lab Dialogue on Hate Speech - May 2021
    - RightsCon - June 2021
    - University of Cambridge Centre for Geopolitics Symposium - September 2021
    - Bread and Net Digital Rights Conference - November 2021
    - GAAMAC Conference - November 2021
    - United Nations Internet Governance Forum - December 2021
    - MIGS AI and Human Rights Forum - April 2022
    - Code For All Conference - September 2022

4.1 - COVID-19 Impact

The COVID-19 pandemic struck as this project was being implemented and subsequently altered many aspects of the initiative. Measures were taken to circumvent any challenges faced with respect to pandemic-related obstacles and, as such, the project team feels there was no significant detriment to the project overall as a result.

Planned in-person activities were converted to remote sessions with no discernable impact on deliverables. Travel was limited due to public health restrictions the world over but due to a 12-month no-cost extension being granted with this in mind, it was possible to reallocate funds with IDRC permission to make better use of the original travel budget, which could not feasibly be spent as planned. Network building deliverables, communication and engagement with various stakeholders, and team cohesion were not impacted negatively despite the unique challenges faced.

5.0 - Project Outputs and Dissemination

All project outputs and dissemination as outlined in the grant agreement have been completed. The relevant output targets are listed below along with descriptions of the progress that has been made to date.

Outputs	Status
Open dataset	The Hatebase dataset has been expanded to nearly 4,000 terms across 98 languages and with more than 1,000,000 sightings in 175 countries since the start of the project. The dataset has been accessed by 149 academic institutions since the start of the project. The dataset is openly and freely available to academic, governmental, and non-profit researchers through a dedicated API which enables the integration of third-party tools.
Software	Data cleaning functions provide users with the ability to refine data outputs into workable formats for analysis. Language processing capabilities, including Naive Bayes classification, named entity recognition, natural language sentiment analysis, and k-means clustering have been implemented with plans for additional improvements. The final software package structure has been organized and is awaiting finalization of each component part. Machine learning algorithms have been generated and training models have been produced. The software code is now available under an open-source license and can be found here.
Regional node capacity building	Hatebase has gained 215 citizen linguists since the start of this project. These contributors represent 50 countries and have made nearly 800 contributions across 46 languages.

	Working relationships with various organizations who have expressed an interest in our work have been established in South America, Europe, Asia, and Africa.

6.0 - Impact

The project has had the most visible impact in two central ways.

First and foremost, the software component of Hatebase provides not only the basis for additional analysis of online hate speech but also the building blocks for other researchers to form new algorithms, training models, and information theories. Substantial effort was put into generating both a technical and approachable research document (see additional document submitted with this report) to demonstrate the cohesive framework of AI processes necessary for sensible online hate speech monitoring.

Secondly, the project has resulted in the development of a community of practice and stakeholders to advance both Hatebase's multilingual lexicon and the research network required to better tailor its language processing rules. Non-English terminology collection expanded significantly during the project period, which addresses a noted shortcoming in the dataset at the outset of the project. Furthermore, the research and its underlying theory was demonstrated at nine international conferences which engaged hundreds of academics, researchers, and practitioners.

The project team is confident that the project has contributed positively to the field of research on hate speech and smarter monitoring methods in addition to setting new standards for the practical implementation of AI deployments for international development, humanitarian aid, and other social benefit purposes.

7.0 - Recommendations

While concerns about bias in machine learning algorithms have been raised frequently among researchers and commentators, there is a dearth of research on other shortcomings of the tools themselves. New technologies are often treated as a panacea for various societal problems, including applications for which they were not designed or suited. Recognizing these limitations is an important outcome for this research project but further efforts should be made to investigate this component of machine learning and AI usage in international development, humanitarian aid, and other social benefit fields.

The project team believes that efforts by IDRC to facilitate connections between this project and others within IDRC's funding portfolio, or external contacts, whose project work overlaps on the subjects of AI, hate speech, and online monitoring, would be tremendously beneficial in expanding the reach and impact of the project and its outputs.

The final recommendation is for funding organizations to consider the natural progression of research on the issue of hate speech monitoring. As mentioned above, this issue needs to be considered in conjunction with the related phenomena of rumours and misinformation (including disinformation) and has important potential early warning applications. The project team has previously raised this topic with IDRC staff members under the broader topic of weaponized information, which draws together hate speech, rumours, and misinformation to look at how these issues relate to each other and physical violence. Conducting expanded action research on weaponized information is a natural progression of the project described in this report and has important implications for predicting, preventing, and mitigating violence and instability. Such work is globally relevant and has the potential to contribute significantly to more peaceful and equitable societies in which democracy, human rights, and development can flourish. Most importantly, it allows the lessons learned from this project to be applied in a meaningful and pragmatic way to solve real-world problems.