Richtlinie zum „Software-Sprint“

Hassreden-Tracker

Schlussbericht

Zuwendungsempfänger:

Travis Robert Brown

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S09 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Wie im Vorschlag angegeben, besteht das Ziel des Projekts darin, Benutzer in die Lage zu versetzen, Muster von Belästigung und Hassreden zu erkennen, um ihre Communities zu verteidigen und die Täter zu entlarven. Das Projekt konzentrierte sich speziell auf die Erweiterung eines bestehenden Satzes von Softwaretools, die vom Autor entwickelt wurden, und darauf, diese Tools skalierbarer, robuster und über sichere Webschnittstellen für eine breitere Gruppe potenzieller Benutzer verfügbar zu machen.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Welche Bezüge gibt es zu den Themenfeldern „Civic Tech“ und „Data Literacy“ des Software Sprints oder zu weiteren gesellschaftlich relevanten Zielen bzw. Lösungsansätzen?

Viele aktuelle extremistische Bewegungen haben sich im Kontext der sozialen Medien entwickelt. Die Teilnehmer löschen oft Inhalte oder ändern Benutzernamen, was es den Communities erschweren kann, sich gegen deren Aktionen zu wehren. Beispielsweise haben prominente rechtsextreme Accounts auf Twitter und Facebook ein gut dokumentiertes Muster entwickelt, um kontroverse und extremistische Inhalte an ihre Follower zu verteilen und diese dann zu löschen, bevor Moderatoren die Möglichkeit haben, darauf zu reagieren. So erreichen diese Accounts ihr Ziel, ihr Publikum zu radikalisieren und gleichzeitig die Gefahr zu minimieren, dass sie ihre Plattformen, ihre Finanzierung oder gemäßigtere Unterstützer verlieren.

Die Anonymität und Kurzlebigkeit von Social-Media-Plattformen kann auch dazu genutzt werden, um Desinformationen zu verbreiten, schutzbedürftige Personen und Gruppen gezielt zu belästigen und sogar Gewalttaten zu provozieren.

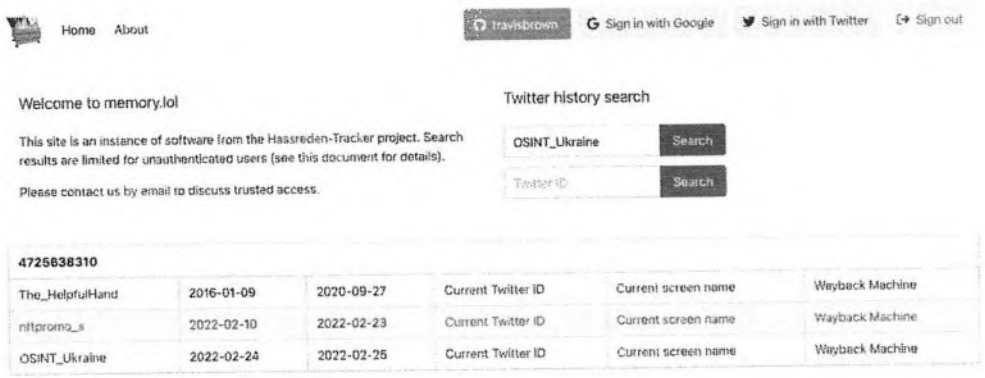
Archivierungsprojekte von Organisationen wie das Internet Archive spielen eine zentrale Rolle bei der Bekämpfung dieses Verhaltens, aber diese Archive sind oft nicht so indiziert, dass Zielgruppen oder ihre Verbündeten die benötigten Informationen finden können. Das Hassreden-Tracker-Projekt soll diese Lücke schließen, indem es Indizes und Suchwerkzeuge für Datensätze bereitstellt, die sonst schwer zu navigieren wären.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Die Projektentwicklung hatte mehrere große Themen, die vom ursprünglichen Vorschlag geleitet wurden, aber auch von den Bedürfnissen der Nutzer, die in jeder Phase Feedback gaben. Das prominenteste Produkt des Projekts ist ein Webdienst und eine Webanwendung, die einen Index historischer Twitter-Profilinformationen bereitstellen. Im Laufe des Projekts wurde dieser Webdienst von Journalisten genutzt, um wichtige Geschichten zu veröffentlichen, von anderen Plattformen (wie CounterSocial) integriert und von prominenten OSINT-Experten (Open Source Intelligence) wie Michael Bazzell vorgestellt.

Nach Rückmeldungen von Community-Mitgliedern im Juli 2022 wurde der Zugriff auf den Webdienst und die Webanwendung auf vertrauenswürdige Nutzerbeschränkt, und die Authentifizierung wird über GitHub, Google und Twitter unterstützt. Als Beispiel zeigt der folgende Screenshot, wie der vertrauenswürdige Zugriff auf diese Webanwendung verwendet werden kann, um den Profilverlauf von *@OSINT_Ukraine* zu identifizieren, einem Spam-Konto, das im Februar 2022 eine große Fangemeinde auf Twitter aufgebaut hat, indem es behauptete, über die russische Invasion der Ukraine zu berichten.



Das Projekt entwickelte auch eine Browsererweiterung, die es Nutzern ermöglicht, Profilverläufe direkt in Twitter-Profilen anzuzeigen. Hier als Beispiel das Profil für *@libsoftiktok*, ein bekanntes Hasskonto, mit aktivierter Erweiterung, das die früheren Identitäten des Kontos zeigt.



Die Browser-Erweiterung liefert zusätzlich Informationen über gesperrte und deaktivierte Konten, in diesem Beispiel ein wegen rassistischer Belästigung gesperrtes Konto:



Diese Erweiterung funktioniert derzeit in den Browsern Firefox und Chrome und unterstützt nur Twitter, könnte aber für andere Plattformen erweitert werden.

Das Projekt hat auch eine Reihe von Berichten, Datensätzen und Schnittstellen entwickelt, die auf derselben zugrunde liegenden Indizierungs- und Datenerfassungs-Engine wie der Webdienst basieren. Ein Beispiel ist eine Reihe von Berichten, die Kontosperrungen nachverfolgen und seit ihrer Einführung durchschnittlich über tausend gesperrte Twitter-Konten pro Tag identifiziert haben. Diese Berichte bieten einen interessanten Einblick in die Entwicklung der Moderationsrichtlinien von Twitter, aber was noch wichtiger ist: sie unterstützen die Arbeit von Aktivisten, die gegen koordinierte Belästigung auf Twitter kämpfen, indem sie Accounts zur Umgehung von Verboten identifizieren und melden. Neben der Identifizierung von Sperren sammelt das System, das die Berichte erstellt, auch Informationen über neu erstellte Konten mit ähnlichem Verhalten und ermöglicht es in vielen Fällen, neue Konten, die von missbräuchlichen Personen und Gruppen erstellt wurden, mit gesperrten Konten zu verknüpfen.

Ein wichtiges Ziel, das in den Projektmeilensteinen skizziert wurde, war, dass die Software sowohl skalierbar als auch effizient ist. Es sollte möglich sein, große Datenmengen zu indizieren (in der Größenordnung von beispielsweise Milliarden von Profil-Schnappschüssen), aber die Kosten für die Erstellung und Bereitstellung dieser Indizes sollten minimiert werden, da das Ziel darin besteht, dass die Software von Aktivistengruppen oder anderen mit begrenztem Budget betrieben wird. Aus diesem Grund haben wir uns entschieden, den größten Teil der Datenplattform in Rust zu implementieren, einer Programmiersprache, die Wert auf Leistung legt. Insofern war das Projekt erfolgreich. Unsere Beispielinstanz baut auf einem Index von rund 34 Milliarden Profil-Snapshots auf, der auf einem einzigen Desktop-Computer erstellt wurde und seit seiner Einführung auf einer einzigen Amazon Web Services-Instanz des Typs *t4g.small* läuft, die pro Stück weniger als zwanzig Euro pro Monat kostet, um betrieben zu werden.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weiter-gehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Bisher waren die meisten Nutzer der Software Aktivisten, Journalisten und Forscher, die die Dienste und Berichte des Projekts genutzt haben, um Themen wie Verbindungen zwischen extremistischen Gruppen oder Social-Media-Aktivitäten rechtsextremer politischer Kandidaten zu untersuchen (siehe Links unten für einige Beispiele).

Diese drei Kategorien von Nutzern waren ein bedeutender Teil der im Vorschlag definierten Zielgruppen, aber das Projekt sollte auch Community-Betreuern und Einzelpersonen dienen, die von Belästigung betroffen sind, und zu diesem Zeitpunkt wurde es von diesen Gruppen weniger genutzt. Eine Ausnahme ist der „Follower Red Flag"-Bericht, der von Personen verwendet wurde, die Ziel von Belästigungen oder Massenmeldungen auf Twitter waren, um potenziell missbräuchliche Follower zu identifizieren (dies ist besonders wertvoll für Nutzer, die ihre Konten gesperrt haben und kontrollieren müssen, wer Zugang zu ihren Informationen hat).

Wir haben die meisten Elemente des Projekts nach dem Förderzeitraum weiter betrieben, einschließlich der öffentlichen Webanwendung und verschiedener Arten von Berichten, und beabsichtigen angesichts der geringen finanziellen Kosten und des minimalen Wartungsaufwands, diese Software auf unbestimmte Zeit weiter zu betreiben, unabhängig von der zukünftigen Finanzierung. Es gibt mehrere Richtungen, in die die zukünftige Entwicklung gehen könnte, einschließlich insbesondere einer zusätzlichen Indizierung anderer Social-Media-Plattformen (in der aktuellen Phase wurde hauptsächlich auf Twitter fokussiert). Angesichts der aktuellen Annahme des Projekts und der Sichtbarkeit einiger seiner Anwendungen denken wir, dass es möglich wäre, in Zukunft weitere Mittel zu akquirieren.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hinter-gründe, und wie bist Du alternativ vorgegangen?

In den frühen Phasen des Projekts begannen wir mit der Arbeit an einer asynchronen Berichterstellungs-Engine, die auf dem Issue-Tracker von GitHub aufbaut. Der Nutzer würde ein Problem öffnen, in dem er anfordert, dass ein Bericht mit einem eingeschränkten Vokabular erstellt wird, das Problem würde von einem „Bot" verarbeitet. Der Bot würde den Bericht generieren und das Problem schließen, indem er einen Link zur Ausgabe bereitstellt.

Dieser Ansatz hat mehrere Vorteile: Er bietet Nutzern Zugriff auf Datenanalysen, die für eine synchrone Schnittstelle (z. B. eine Webanwendung) zu rechenintensiv sind, ohne dass das Projekt eine asynchrone Berichtsplattform erstellen oder betreiben muss.

Wir haben jedoch festgestellt, dass die Anforderung, dass der Nutzer ein GitHub-Konto haben muss, zusammen mit der Schwierigkeit, mit dem Anforderungsvokabular zu arbeiten, zu einer geringeren Verbreitung des Ansatzes führ, als wir erwartet hatten. Wir haben auch festgestellt, dass viele nützliche Möglichkeiten für den Zugriff auf die Daten synchron bereitgestellt werden könnten (in einem Webdienst oder einer Anwendung). Wir haben daher die Arbeit an diesem Berichtsmodul depriorisiert, und obwohl die grundlegende Grundlage abgeschlossen ist, erfordern viele Arten von Berichten, die vom Projekt bereitgestellt werden, immer noch, dass der Nutzer Befehle über die Befehlszeile ausführt.

Wir begannen auch mit der Arbeit an einer Engine zum Sammeln von Benutzerprofilbildern von Plattformen wie Twitter, Gab und Telegram, wobei wir uns hauptsächlich auf die Wayback-Maschine als Datenquelle stützten. Diese Art der Datenerhebung hat mehrere Anwendungen im Bereich der Verfolgung von Hassreden:

- Hass-Symbole (wie das Sonnenrad oder Pepe der Frosch) können entweder manuell oder automatisiert identifiziert werden.
- Accounts zur Umgehung von Sperren können durch Bildabgleich identifiziert werden.
- Plattformübergreifende Links können ähnlich identifiziert werden.

Während wir in einem Pilotversuch mehrere Millionen relevanter Profilbilder gesammelt haben, haben wir diese Arbeit nicht mit anderen Projektelementen oder Schnittstellen integriert. Der Code ist öffentlich verfügbar und wir hoffen auf zukünftige Arbeiten in dieser Richtung.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?

- Code-Repository der obersten Ebene: https://github.com/travisbrown/hassreden-tracker
- Hassreden-Tracker-Instanz verfolgt rechtsextreme Konten auf Twitter: https://memory.lol/app/
- Tägliche Twitter-Suspendierungsberichte: https://github.com/travisbrown/twitter-watch/
- Bericht der Washington Post, der Informationen zitiert, die durch das Projekt enthüllt wurden: https://www.washingtonpost.com/technology/2022/04/19/libs-of-tiktok-right-wing-media/
- Bericht von The News Daily über den Einsatz des Projekts bei den australischen Parlamentswahlen 2022: https://thenewdaily.com.au/news/politics/australian-politics/federal-election-2022/2022/04/29/deves-anti-trans-tweets-liberal/
- Beispiel eines CNN-Artikels, der auf einen vom Projekt erstellten Bericht verweist: https://edition.cnn.com/2022/06/23/politics/mayra-flores-capitol-riot-qanon-kfile/index.html

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Die Arbeit sollte ursprünglich von einem Entwickler über vier Monate ausgeführt werden, aber angesichts unerwarteter Verfügbarkeit wurde die Entwicklung weitere fünf Wochen über das geplante Ende des Arbeitszeitraums hinaus fortgesetzt (ohne Änderungen am Budget).

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, orin bestand dieser und wie bist Du damit umgegangen?

Nach der Steigerung der Sichtbarkeit durch die Berichterstattung in der Washington Post und die Integration des Webdienstes in ein CounterSocial-Produkt (SockWatch) äußerten einige Personen Bedenken, dass Elemente des Projekts dazu verwendet werden könnten, die Privatsphäre zu untergraben oder sogar Belästigung zu fördern. Wir hatten keine Hinweise auf einen solchen Missbrauch (und haben es bis heute nicht), aber wir haben den Zugriff auf den vollständigen Index auf vertrauenswürdige Nutzerbeschränkt, die sich über GitHub, Google oder Twitter authentifiziert haben. Nicht vertrauenswürdige oder nicht authentifizierte Benutzer haben nur Zugriff auf die letzten 60 Tage des Profilverlaufs, was ausreicht, um einige wichtige Anwendungsfälle zu unterstützen (z. B. die Identifizierung von Desinformationen), aber das Missbrauchspotenzial verringert.