Company: AI Incident Database / Digital Safety Research Institute

Project Title: Analysis of Incident Reports

Background:

The AI Incident Database (AIID) is an open-source collection of real-world harms caused by the deployment or development of artificial intelligence systems. The AIID tries to record and demonstrate these harms for the awareness and education of AI developers, researchers, policymakers, and the general public.

Check us out here: https://incidentdatabase.ai/

Read about the database on the PAI Blog, Vice News, Venture Beat, Wired, arXiv, and Newsweek among other outlets.

The AIID is a project of the Responsible AI Collaborative (RAIC), an organization chartered for the purposes of advancing the AIID and collaborating with other organizations, such as the Digital Safety Research Institute (DSRI), interested in researching and mitigating the harms of the digital system on people and society. (a lot of acronyms, I know!). Notably, the sponsor for this project is DSRI staff that develop and maintain the AIID. But for all intents and purposes, this project is for and about the AIID.

Project Overview and Problem Definition:

The problem: Computationally analyzing the changing rhetoric and trends in AI journalism.

The AIID has indexed over 600 unique reports of incidents and alleged harms, based on the evidence of over 3,500+ third-party journalistic reports. The AIID editorial staff attempt to objectively evaluate these reports in conjunction with one another when deciding whether to declare a new “incident,” name harmed, and responsible parties (if able to do so), and provide a minimally editorialized description and title. Not much decision is made to characterize the reports beyond their ability to substantiate the alleged incident

However, since beginning indexing these reports starting in 2019, the journalism around AI (not to mention AI systems themselves) has changed – the language used to describe AI systems has in some cases become more technical and less hand-wavy; corporations and developers are held in different esteem depending on the popular impact and reputation of their work; and the overall sentiment and communication techniques surrounding AI systems adapt over time.

The AIID absolutely depends on journalism. Better understanding how the _journalism surrounding AI_ (and AI incidents) has changed over time would benefit the AIID and others interested in the fair, informative discourse surrounding the technology by answering certain questions.

Have certain AI trends fallen away over time, e.g. self-driving/autonomous vehicles? Has ChatGPT starved out media attention for other types of AI systems and AI harms, like facial recognition systems? Has the sentiment of AI harms changed in particular publications or sources?

Journalistic trends and characteristics may not only differ broadly over time but also across reports of the same AI-caused harm within the database. Where do media reports agree in sentiment and nuanced language when describing AI harms? Where do they differ, contradict one another, or are less clear? Are some publishers consistently more detailed or nuanced in how they describe AI systems specifically?

Answering these questions may inform editorial staff and the broader incident research community on how to index and treat media reporting. Stakeholders in AI safety, governance, and compliance similarly rely on media reports to inform their organizations, product landscapes, and partnerships.

Lastly, the AIID may not be using media reports to their full potential. While we strive to analyze and extract structured information from the articles we index, the number of entities and relationships found in the literature is staggering. Students may also consider how to augment the AIID's hand-curated entity relationships with computational techniques for managing named entities and topics at scale.

Project Goals:

A successful project will:

- Develop a programmatic pipeline for downloading indexed media reports from the AI Incident Database (all data are free and public by the terms of the AIID);
- Formulate clear hypotheses about features of the media reports to test, including but not limited to the questions above;
- Conduct literature review (not major) on prior media analyses that have been performed over similar corpuses;
- Develop a suite of analyses (e.g. python notebooks) that use NLP techniques (such as topic modeling, sentiment analysis, and named entity recognition) or other ML techniques (text classification, feature extraction, and visualization) to provide insights into the nature of incident reports found in AIID data, answering these questions;
- Through analyses, present the methodology of these techniques, positive and negative results, and creative ways of communicating the nuances of media reporting surrounding AI in a data-driven format (e.g. visualization and story-telling) (room to expand on this story-telling);
- Prepare a blog post summarizing the results of their project with the AIID via the AIID blog, to be published conditional on completion and satisfactory nature of work.
- (Optional) Make recommendations about how the AI Incident Database, and the greater AI safety and ethics communities, can track discourse about AI and related harm :-)

Potential Roadblocks and Barriers to Success:

One realistic problem is that there might not be rich enough relationships, linguistics characteristics, or trends to extract from the dataset of reports in the incident database to answer the project questions posed – a negative result. In this case, the scope of the project can be expanded to include other types of AI-intersecting journalism beyond incidents – such as product reporting – or more emphasis can be put on the use of NLP techniques to further AIID's mission in better indexing incidents and safety data (e.g. by applying named entity recognition to reports to supplement AIID's hand-curated entity relationships).

Preferred Methodology:

This project presents a more open-ended problem that doesn't prescribe all the interesting questions or means of arriving at the answers. Rather, the methodology might require attention to exploring and iterating on techniques as students become more familiar with them. The AIID developers and staff are flexible in supporting students; they do not need to "plug-in" to a particular work management style, but be independent and consistently communicating, and agree to a cadence of meeting bi-weekly and befitting student and AIID staff schedules.

Data Requirements and Availability:

All data (indexed media reports) are publicly available and compiled in the AIID.

Should the scope of the analyses expand to data outside AIID (optional), the team will need to acquire (public) datasets from other media-type datasets and filter them for AI-related topics.

Analytics requirements:

Natural Language Processing (e.g. topic modeling, sentiment analysis, named entity recognition); other ML techniques (text classification, feature extraction, and visualization); potentially the use of LLMs and other generative AI to experimentally interact with and assess the data under test.

Preferred Tooling:

- Python is preferred to R, but ultimately whatever tools result in a reproducible and documented program and tools are most important.
- The best method of communication and presentations are left to the students. This may take the form of organized Python notebooks collated into a writeup or report, a static website, or other media, in addition to a blog post on AIID.

Project Schedule:

The sponsor has no strict requirements for the student schedule as this is not a business-critical or timely project, preferring that milestones can be adjusted to benefit students' educational needs.

- 1-3 weeks: Familiarize with AI Incident Database, data format and contents (e.g. interacting with MongoDB, formatting data for analysis in tools of preferred choice), and other exploration.
- 2-4 weeks: Producing a set of analytical questions to answer; proof of concept for particular NLP techniques resulting from explorations (e.g. topic modeling applied to a small dataset);
- 5-6 weeks: Scaling analyses to the entire dataset; iterating on techniques used, potentially revisiting questions to answer and explore. Iterating, iterating, iterating.
- 6-7 weeks: Depending on the results, begin compiling an end-to-end report of the findings, including introductory materials on techniques and a literature review of other media analyses. (Alternatively, this might be a pivot timeframe for students to deep dive into one or more particular techniques/questions, e.g. named entity recognition for entity relationships across AIID data).
- 7-9 weeks: Moving towards final presentation – polishing, story-telling, and likely still analyzing or compiling artifacts, including blog posts and analyses.

(Assuming some time for class introductions period and academic breaks)

Confidentiality Concerns:

None

Budget:

No budget was provided. The AIID is an open-source project built from freely available tools, and the scale of the data should not require advanced computing.

Contact:

Kevin Paeth <kevin.paeth@ul.org>

Sean McGregor <sean.mcgregor@ul.org>