ForHumanity Auditing AI System

Jacqueline Fraley
Master of Data Science Candidate
University of Virginia, School of Data Science
Charlottesville, U.S.
jjf4rp@virginia.edu

Andrew Chaphiv
Master of Data Science Candidate
University of Virginia, School of Data Science
Charlottesville, U.S.
ac2gq@virginia.edu

Patrick Dunnington
Master of Data Science Candidate
University of Virginia, School of Data Science
Charlottesville, U.S.
pe9zgx@virginia.edu

Ian Yung
Master of Data Science Candidate
University of Virginia, School of Data Science
Charlottesville, U.S.
icy4r@virginia.edu

***Abstract*—This capstone project, in partnership with ForHumanity, utilizes the AI Incident Database (AIID) to analyze and address risks associated with the deployment of artificial intelligence (AI) across various sectors. By applying data analysis techniques such as topic modeling, word embedding, and clustering, the study identifies key issues like misuse patterns and the impact of rapid AI deployment on societal biases and misinformation. The findings highlight the need for stringent regulatory measures and informed policy interventions. The research not only provides actionable insights to ensure the ethical use of AI but also supports policymakers in crafting strategies for safer AI integration.**

I. Introduction

Artificial intelligence is becoming increasingly common in everyday life, but it also brings risks and challenges that need careful management. The 'ForHumnaity Auditing AI System' uses data science to check AI systems, aiming to reduce these risks and make AI more transparent and safe. Our research analyzes data related to these issues to investigate problems caused by harmful incidents in AI. We use the AI Incident Database, a public source that gathers reports of AI harms, to help understand and prevent these problems.

A. Project Goal

Our main goal is to address challenges posed by harm and incidents in artificial intelligence by thoroughly analyzing AI-related data. Leveraging data science techniques, we uncovered correlations and directional trends in harm and incidents, delivering meaningful insights to ForHumanity. We aimed to provide significant insights regarding problems related to AI harms and incidents. A comprehensive analytical framework that identifies patterns and emerging risk in AI use, to clarify the complexities in a manner that facilitates actionable insights.

B. Research Questions

This project was driven by the critical need to address the expanding risks associated with increasingly complex AI systems that are becoming more common. As these technologies advance, the potential for incidents increases. It is crucial to understand these incidents. By leveraging data-driven insights, risks can be proactively identified to ensure safer integration of AI across various sectors.

II. Literature Review

In AI incident analysis, significant contributions have been made by adopting advanced NLP techniques and Word2vec applications. Studies such as " A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports" have showcased the capability of NLP to systematically sift through vast amounts of unstructured text data to discern patterns and trends in AI-related incidents [2]. This approach facilitates a deeper understanding of the frequency and nature of AI failures and helps categorize them into actionable insights. Similarly, the research outlined in "A machine learning analysis using Word2Vec and GloVe embeddings" demonstrates how semantic analysis tools like Word2vec can map out relationships between different AI failures [1]. By analyzing the vector spaces where words associated with AI incidents cluster, these studies provide a nuanced understanding of underlying causes and effects, enabling stakeholders to address potential risks preemptively. Together, these works form a crucial part of the academic foundation that supports ongoing efforts to enhance AI safety and governance.

III. Data Description

The Artificial Intelligence Incident Database (AIID) is a resource for tracking and analyzing incidents involving AI systems. Predominantly utilized in the United States, at the time of writing the database records a total of 599 incidents, accompanied by 3016 reports. Each incident in the AIID is meticulously cataloged with details, including an Incident ID, incident date, links to verifying reports, the entity deploying the AI, the developer, a brief description, and a title. Reports associated with these incidents provide additional information such as a unique report number, the author's name, publication date, the source domain (e.g., news websites, social media), linked incident number, the full text of the report, and a URL. This comprehensive dataset is regularly

maintained and cleaned by dedicated engineers, ensuring high data integrity and usability.

The AIID began accepting reports in 2020 but also includes incidents that occur before its launch to provide a comprehensive history. The first report was around 1984 and AIID is still accepting reports. Some assumptions about the data are that all the data is either metadata or text data and that the data is correctly processed and verified. One limitation is that the AIID currently lacks first-hand complaints, and most incident records are from third parties, slightly decreasing data reliability.

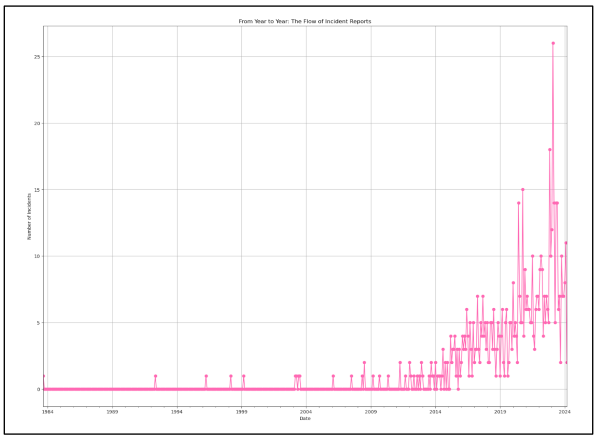

Fig. 1. Incident Number Reported on AIID Over Time

IV. Methodology

A. Data Processing

We began our data analysis by doing a comprehensive overview of the data and data cleaning before moving into our exploratory data analysis. Our dataset was sourced from the AI Incident Databases. We first checked each row for missing values and duplicate values. We had no missing values, but we did have some duplicate entries. After removing the duplicate rows, we transformed some data. Several columns containing Python lists, such as the “Alleged Deployer of AI System” column, were converted from string representation of lists into readable comma-separated strings.

Our exploratory analysis focused on understanding the distribution of entries in key columns. We described the statistical distribution of the 'Alleged harmed or nearly harmed parties' column, noting the most frequently occurring entries. We assessed the uniqueness of categories within the 'Alleged developer of AI system' and 'Alleged deployer of AI system', finding 281 and 370 unique entries, respectively. We analyzed the frequency of occurrences in the 'Alleged deployer of AI system' and 'Alleged developer of AI system', identifying the most common and singular occurrences. For instance, Tesla and Facebook were among the most frequently mentioned deployers and developers. We created a new column, 'Match', to check if the entries in 'Alleged deployer of AI system' matched those in 'Alleged developer of AI system'. This comparison revealed 372 matches and 255 mismatches, offering insights into the overlap between deployers and developers within reported incidents.

Before conducting further analysis, we ensured that data entries were correctly formatted. Lists encoded as strings in specific columns were parsed and converted to comma-separated strings, enhancing the dataset's readability and usability for subsequent analyses.

B. Techniques

To extract insights from the AI Incident Database, we utilized several data analysis techniques, each contributing to our understanding of AI–related incidents. These techniques included topic modeling, word embedding, clustering, and locational analysis.

We applied Topic Modeling, specifically Latent Dirichlet Allocation (LDA), to identify prevalent themes within the incident reports. This technique analyzes the co-occurrence patterns of words in the text to reveal latent topics that dominate the discussions in the reports. For instance, we discovered significant discussions around privacy violations, ethical abuses, and technological failures. We found a relation between harmed parties and generative technology malfunctions like unintended behaviors in AI-driven content generation and recommended systems. For example, we identified incidents where generative models produced inappropriate or harmful content, such as deepfakes or biased news articles, which directly affected the integrity and safety of the users' digital interactions. By grouping similar incidents, LDA provided insights into common underlying issues across different reports, revealing patterns such as the correlation between harmed parties and specific failures in generative technologies.

To explore semantic relationships and contextual similarities across incident reports, we utilized Word Embedding techniques, specifically Word2Vec. This method transforms text data into numerical vectors, allowing computers to understand and process words in context. By analyzing the vector representations of words from incident summaries, we identified strong semantic links between terms associated with minors and various types of harmful content. This highlighted the data's recurrent themes and aided in visualizing complex relationships between different incident types. We revealed strong connections between terms related to minors and various harmful content types. This analysis enabled us to uncover how certain words and phrases frequently co-occur in incident reports involving minors. For instance, terms like 'exploitation,' 'inappropriate content,' and 'privacy breaches' were commonly linked, suggesting that AI systems might be inadvertently exposing or failing to protect minors from such risks.

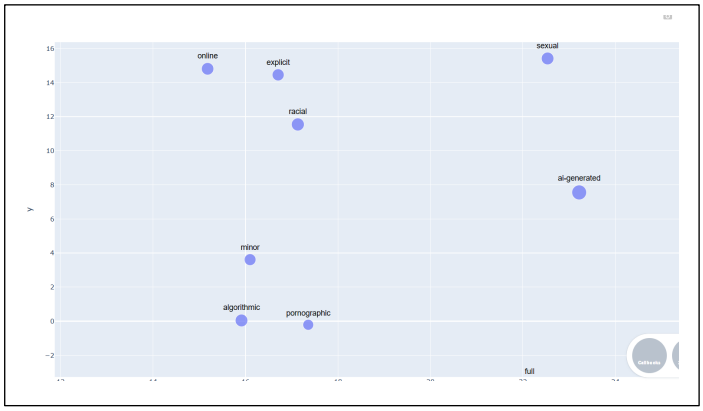

Fig. 2. AIID Report Word Embedding Graph

We applied Clustering techniques to organize the incidents into distinct groups based on their characteristics and severity. Utilizing K-Means clustering (with k=10), we categorized the incidents into distinct groups based on their inherent characteristics. This method was crucial in identifying and categorizing the types of AI-related issues, enabling us to develop tailored responses for each category. Clustering provided a clear overview of the problems, making it easier to devise effective strategies to address the diverse challenges presented by AI incidents. Our clustering analysis using K-Means highlighted critical areas in AI applications, identifying distinct categories such as transportation incidents, journalistic bias, facial recognition issues, and algorithmic bias in social media.

We conducted a detailed location analysis to better understand the geographic origins of the AI incident reports since the initial data lacked specific location classifications. Using Named Entity Recognition (NER) technology from SpaCy, we extracted geographic locations from the descriptions and titles of the incidents and converted these into precise geographical addresses. We categorized each incident into three groups: those occurring within the United States, those outside the United States, and those linked to U.S. companies without a specific location.

This allowed us to see which incidents got classified with U.S. companies but might have been reported outside the U.S. and involved a major U.S. company. Our findings revealed that a significant majority of the reports—nearly 500—originated from the United States, likely reflecting the dense concentration of AI-focused tech companies and the nation's leading role in AI development. The concentration could have also indicated less knowledge of the AIID outside the U.S., so many international incidents may have never been recorded on the website.

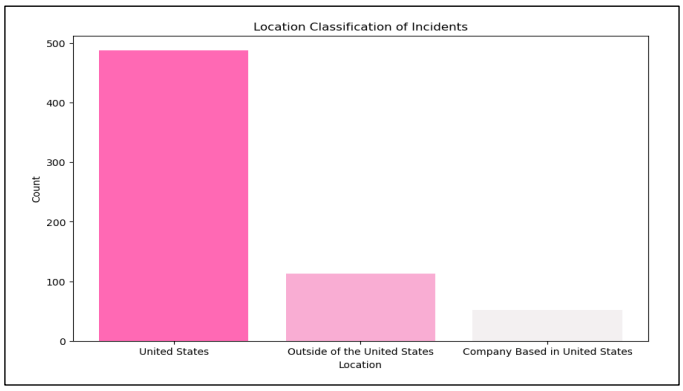

Fig. 3. Location Classification Histogram

V. Results

We observed a noticeable correlation between specific companies and increased incident reports, suggesting a need for more rigorous scrutiny of their AI deployment practices. Additionally, terms commonly found across reports frequently related to issues of misidentification and bias, particularly within facial recognition and surveillance applications. Notably, the swift incorporation of AI technologies, such as ChatGPT, has been closely associated with a surge in incident reports, indicating that rapid technological adoption is outpacing the necessary oversight mechanisms.

The application of AI in areas such as surveillance and public decision-making has been shown to exacerbate existing societal biases. Moreover, AI systems have been implicated in the spread of misinformation, especially in politically sensitive contexts, posing significant threats to the integrity of democratic processes. Another major concern is the exposure of minors to inappropriate content via AI-driven platforms, emphasizing the urgent need for enhanced regulatory measures.

VI. Discussion

This study not only forecasts potential future spikes in AI incidents but also sheds light on several broader implications. The data suggests that future increases in AI incidents may be influenced by widespread technology adoption, societal shifts such as those caused by the COVID-19 pandemic, and new technological advancements. These elements act as catalysts for increased reporting of AI-related incidents. The analysis indicates a disproportionate number of reports from the United States, suggesting a higher sensitivity to reporting or possible regional variations in AI deployment. Although the tech industry shows the highest representation, significant impacts are also evident in sectors like education, law enforcement, and media services. Internationally, there is potential to enhance awareness and engagement to capture the global landscape of AI incidents more accurately. Furthermore,

Named-entity recognition (NER) analysis highlights a strong correlation between media attention and the frequency of incident reports related to specific entities, such as OpenAI. This underscores the influence of media coverage on public and regulatory perceptions of AI, pointing towards the need for comprehensive regulatory frameworks that prioritize ethical practices, transparency, and accountability in AI development and deployment.

VII. Conclusion

In our ongoing efforts to deepen our understanding of AI-related incidents, we want to further our understanding of our locational analysis, specifically within the United States. We aim to identify regional patterns and derive insights regarding the distribution of these incidents across various states. Additionally, we intend to enhance our analytical framework by integrating clustering and topic modeling techniques. This comparative analysis will help us explore whether there are underlying thematic or structural similarities between the clusters identified and the topics derived, potentially revealing new data dimensions that could inform more targeted interventions.

Our comprehensive analysis of the AI Incident Database, integrating data science techniques such as topic modeling, word embedding, and clustering, has highlighted the intricate challenges and emerging risks associated with deploying artificial intelligence technologies across various sectors. The multifaceted examination of AI incidents has revealed diverse concerns, including regional disparities, specific misuse patterns, and risks to vulnerable groups. These findings underscore the critical need for robust regulatory frameworks prioritizing ethical considerations, transparency, and accountability in AI development and deployment. Moreover, this comprehensive approach has enabled the identification of critical areas that urgently require targeted policy interventions. By pinpointing where AI deployment may exacerbate inequalities or harm, our study provides a foundation for ForHumanity to help ensure AI technologies' responsible development and deployment.

VIII. Acknowledgments

We would like to extend our sincere gratitude to Ryan Carrier, whose support was instrumental to the success of this project. His insights and expertise have significantly enriched our research, and we are grateful for the intellectual contributions provided throughout this study. Special thanks also go to ForHumanity for facilitating access to essential resources and for the collaborative opportunities that have been pivotal in our work. Furthermore, we appreciate the continuous support from our colleagues and mentors at UVA, who have provided a stimulating academic environment that has been crucial for the completion of this project.

References

The interactive visualizations containing the data analysis are available on the dashboard at ***link***.

[1] Curto, G., Jojoa Acosta, M.F., Comim, F. *et al.* Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings. *AI & Soc* **39**, 617–632 (2024). https://doi.org/10.1007/s00146-022-01494-z

[2] Ricketts J, Barry D, Guo W, Pelham J. A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports. *Safety*. 2023; 9(2):22. https://doi.org/10.3390/safety9020022