

AI Safety, Inclusion, and Cybersecurity Through Advanced Text Analytics

Derrick L. Cogburn
American University
dcogburn@american.edu

Haiman Wong
Purdue University
wong424@purdue.edu

Theodore A. Ochieng
American University
to9648a@american.edu

Abstract

This minitrack for HICSS-58 recognizes the increasing importance of natural language processing (NLP), text mining, and text analytics to better understand and improve decision making for safety, inclusion, and cybersecurity in artificial intelligence (AI). Papers welcomed in this minitrack include, but are not limited to, analysis of AI risk management frameworks designed to promote the development of responsible and trustworthy AI, such as those developed in the United States by the National Institute of Standards and Technology (NIST), and the EU, Australia, Japan, Singapore, G20, China, Africa, and the OECD. These frameworks could be evaluated on their level of human-centeredness and stakeholder involvement, how they mitigate risks and harness the benefits of AI, how they approach “fairness”, their explainability, and support for privacy, security, safety, reliability, and accountability of AI. Also, how these interdisciplinary techniques support cyberthreat detection, and other applications of ML/DL/AI related to cybersecurity.

Keywords: AI Safety, Inclusion, Cybersecurity, NLP, Text Analytics

1. Introduction

Since HICSS 49, we have been leading successful minitracks on big data analytics and text mining. Initially these minitracks were in the Collaboration Systems and Technologies Track, focusing on analyzing the voluminous text data generated by distributed collaboration systems supporting global virtual teams specifically and information systems broadly. Also, since HICSS 48, this minitrack has been linked with an annual tutorial on text analytics, helping to build the interdisciplinary text mining, and natural language processing community within HICSS. Beginning with HICSS 56, we forged a “Fast Track” to journal publishing opportunity with *Data & Policy*, the

prestigious peer-reviewed and open access journal published by Cambridge University Press. This relationship provides an exciting publishing opportunity for our minitrack researchers who are examining the impact of text analytics on policy and governance. Another publishing opportunity for this minitrack is through Palgrave MacMillan to publish “Pivot Books” which are slightly longer than a typical journal paper, but shorter than a normal monograph. These Palgrave Pivot vehicles provide the author with rapid turnaround times and a shortened peer review.

At HICSS 57, we repositioned the minitrack to the Decision Analytics and Service Science Track, reflecting our broader focus on large scale text analytics to support decision making processes with Natural Language Processing (NLP) and supervised and unsupervised machine learning tools.

This year at HICSS 58, we have pivoted slightly to focus the minitrack on artificial intelligence and cybersecurity. With this shift, we recognize the increasing importance of NLP, text mining, and text analytics to better understand and improve decision making for safety, inclusion, and cybersecurity in artificial intelligence. Very large-scale text mining and predictive analytics are at the heart of the Large Language Models (LLMs) driving Generative AI (GenAI) tools such as ChatGPT, Claude, CoPilot, and others. We recognize the increasing importance of supporting researchers interested in understanding how to harness the power of AI, particularly GenAI and LLMs, to conduct text analytics. So, we have revised our annual text analytics tutorial to include a focus on using GenAI along with NLP and supervised and unsupervised machine learning in R and Python.

2. Minitrack Topics and Themes

Papers welcomed in this year’s minitrack included, but were not limited to, analysis of AI risk management frameworks designed to promote the development of responsible and trustworthy AI, such as those developed in the United States by the National

Institute of Standards and Technology (NIST), and the EU, Australia, Japan, Singapore, G20, China, Africa, and the OECD. These frameworks could be evaluated on their level of human-centeredness, how they mitigate risks and harness the benefits of AI, how they approach “fairness”, their explainability, privacy and security, safety, reliability, and accountability of AI. Moreover, analysis of these frameworks could assess how they consider context-specific risk management techniques; how stakeholders are involved, and how they promote secure innovation. Papers can also examine emerging AI, ML, and DL applications in various industries, such as automated fraud detection in the finance sector or enhanced cyber threat detection and incident response in the information technology sector.

Potential papers for this minitrack may deploy any number of text analytics techniques, ranging from statistical bag-of-words and rule-based approaches to syntactic parsing and natural language processing approaches, including Named Entity Recognition (NER), text embeddings, and Bidirectional Encoder Representation Transformers (BERT). Papers may also use unsupervised machine learning (ML) approaches, including topic modeling (using Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), or other topic modelling techniques; k-means clustering, as well as supervised machine and deep learning approaches, such as predictive regression and classification models, Deep Neural Networks (DNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs). Other papers may address methodological challenges such as text summarization, classification, and clustering, using generative large language models (e.g. ChatGPT, Gemini/Bard) to create synthetic data, overcoming API limitations, and working on distributed, high-performance computers. We also seek papers on enhanced explainability in text analytics (particularly AI/ML) relative to results and the detection and mitigation of bias in analytics.

Paper for this minitrack use both theoretical and applied text mining approaches to analyze various genres of text data, including, but not limited to:

- Security alerts
- Threat intelligence feeds
- Computer logs
- Email archives (including phishing emails)
- Incident and maintenance reports
- Legal documents (patents, contracts, etc.)
- Public policies and public comments
- Online communities and social media
- Blog posts
- Published articles
- Websites

- Meeting and call center transcripts
- Speeches
- News transcripts
- Customer feedback
- Resumes and CVs
- Job Postings and Descriptions
- Employee evaluations
- Insurance claims (cyber insurance, etc.)
- Annual reports
- Case studies

In this minitrack we seek to create a highly engaging interactive forum and community for researchers and practitioners to discuss the critical issues related to text mining and analytics and contribute to the ongoing big data analytics focus and emerging AI and ML concentrations at HICSS.

We are pleased to introduce the three papers selected for the HICSS-58 minitrack on AI Safety, Cybersecurity, and Inclusion through Text Analytics. The first paper is entitled, “De-Identification of Privacy Sensitive Information in Resumes with GPT-4: A Utility Analysis for Automated Job Role Classification. The second paper is entitled “Blockchain Based Information Security and Privacy Protection: Challenges and Future Directions using Computational Literature Review. Our third and final paper is entitled “Adversarial Natural Language Processing: Overview, Challenges and Future Directions” and is our Best Paper Nominee.

As co-chairs of the minitrack, we are excited about this new direction. We received six submissions to the minitrack and after our peer review, accepted three excellent papers, including one Best Paper nomination. These papers highlight various important aspects of this emerging research community. We are excited about the potential for our minitrack and our tutorial, and we look forward to discussions of these papers.

3. Paper 1: De-Identification of Privacy Sensitive Information in Resumes with GPT-4: A Utility Analysis for Automated Job Role Classification

Our first paper is grounded in the privacy concerns arising from the sensitive information contained in resumes. As organizations face the challenge of managing large amounts of data, privacy concerns have become increasingly prevalent when sharing sensitive privacy information with machine learning experts. This paper addresses the fundamental issue of privacy-sensitive information de-identification by introducing in-prompt de-identification, an approach

that exploits the capabilities of large language models. Existing de-identification techniques often struggle to ensure complete privacy, and methods with higher privacy often result in a loss of data utility. In contrast, in-prompt de-identification can generate synthetic, human-readable data samples from given inputs and bridges the gap between privacy and utility. With this article, we contribute to the de-identification of real-world resume data using in-prompt de-identification based on OpenAI's GPT-4. Notably, our classification model, trained on GPT-4 generated data, shows no significant loss in performance compared to our baseline model trained on the original data. users.

4. Paper 2: Blockchain Based Information Security and Privacy Protection: Challenges and Future Directions using Computational Literature Review

Our second paper also focuses on privacy but sees a potential solution not in GPT-based de-identification, but in blockchain-based protection. Blockchain technology is an emerging digital innovation that has gained immense popularity in enhancing individual security and privacy within Information Systems (IS). This surge in interest is reflected in the exponential increase in research articles published on blockchain technology, highlighting its growing significance in the digital landscape. However, the rapid proliferation of published research presents significant challenges for manual analysis and synthesis due to the vast volume of information. The complexity and breadth of topics, combined with the inherent limitations of human data processing capabilities, make it difficult to comprehensively analyze and draw meaningful insights from the literature. To this end, we adopted the Computational Literature Review (CLR) to analyze pertinent literature's impact and topic modelling using the Latent Dirichlet Allocation (LDA) technique. We identified 10 topics related to security and privacy and

provided a detailed description of each topic. From the critical analysis, we have observed several limitations, and several future directions are provided as an outcome of this review.

5. Paper 3: Adversarial Natural Language Processing: Overview, Challenges and Future Directions (Best Paper Nominee).

Our third paper focuses on cybersecurity and addresses the opportunities and challenges presented by the emergence and growth of GenAI and LLM based text analysis. Natural language processing (NLP) has gained wider utilization with the emergence of large language models. However, adversarial attacks threaten their reliability. We present an overview of adversarial NLP with an emphasis on challenges, emerging areas and future directions. First, we review attack methods and evaluate the vulnerabilities of popular NLP models. Then, we review defense strategies including adversarial training. We identify key trends and suggest future directions such as the use of Bayesian methods to improve the security and robustness of NLP systems.

6. Towards a HICSS AI and NLP Text Mining Community

We believe the minitrack on AI Safety, Cybersecurity, and Inclusion through Text Analytics makes an important contribution to HICSS. It has great potential to stimulate the creation of a robust, interdisciplinary AI, NLP, text mining and cybersecurity research community within HICSS. Given the amount of unstructured textual data available to researchers, such a research community would be invaluable. The text mining papers at this 58th HICSS represent what we see as an important emergent trend, which we believe will remain for many years to come.