ai PHASE II

NATIONAL SECURITY STANDARDS FOR ARTIFICIAL INTELLIGENCE

2019
PUBLIC-PRIVATE
ANALYTIC EXCHANGE PROGRAM

ABSTRACT

Recent advances in artificial intelligence (AI) have led to an explosion of multimedia applications (e.g., computer vision (CV) and natural language processing (NLP)) for different domains such as commercial, industrial, and intelligence. In particular, the use of AI applications in a national security environment is often problematic because the opaque nature of most systems leads to an inability for a human to understand how the results came about. A reliance on "black boxes" to generate predictions and inform decisions is potentially disastrous.

This paper explores how the analytic tradecraft standards outlined in Intelligence Community Directive (ICD) 203 can provide a framework for assessing how well an AI system can explain the basis for its decisions (i.e. "analysis"). The paper creates and describes a Multisource AI Scorecard Table (MAST) to gauge explainability and transparency. Providing the developer and user of an AI system with standards focused on the principles of good analysis adopted by the intelligence community (IC) can help promote the development of more understandable systems and engender greater trust in AI outputs.

Disclaimer Statement. This document is provided for educational and informational purposes only. The views and opinions expressed in this document do not necessarily state or reflect those of the U.S. Government (USG) or the Public-Private Analytic Exchange Program Partners, and they may not be used for advertising or product endorsement purposes. All judgments and assessments are solely based on unclassified sources and are the product of joint public and USG efforts.

Acknowledgements. The authors appreciate the support from the DHS-ODNI Public-Private Analytic Exchange Program and all of the individuals who volunteered their time to provide us with valuable insights.

contents

INTRODUCTION..... 4

MOTIVATION..... 6

STANDARDS FOR THE INTELLIGENCE COMMUNITY..... 7

AI DEVELOPMENT PROCESS..... 9

- AI DEVELOPMENT AND DEPLOYMENT..... 9
- NATURAL LANGUAGE PROCESSING..... 10

CASE STUDIES..... 12

CONCLUSIONS..... 21

REFERENCES..... 22

TEAM MEMBERS..... 24

APPENDIX 25

I. INTRODUCTION

Artificial intelligence's (AI) popularity from multimedia applications (e.g., natural language processing (NLP), computer vision) creates both opportunities and challenges for national security. Each of these examples uses AI to support a decision maker in their tasks, especially for situations involving huge, multimodal, and novel situations requiring the fusion of information [1, 2, 3].

Current AI methods, such as NLP, have reached the threshold necessary for developing a set of standards or best practices for the development and deployment of the products. In particular, user understanding of machine decisions and trust in AI outputs are not well addressed. Inherent in the discussion of the future deployment of AI systems are trust metrics [4, 5] that address security, safety, and transparency [6] in order for users and customers to widely accept AI decisions. A key question is "What are the standards that would apply to the stages of AI deployment shown in Figure 1 to promote greater trust in AI?"

We propose the use of the analytic tradecraft standards outlined in Intelligence Community Directive (ICD) 203 [7] as a framework for guiding the development of AI standards in the national security space.

Section II reviews the motivation for this effort. Section III discusses ICD 203 standards and Section IV examines the AI development process. Section V recommends a notional evaluation system that could provide a framework to judge the utility (e.g., trust) of an AI system and leverages representative use cases to show the evaluation system in action. Section VI draws conclusions and provides recommendations.

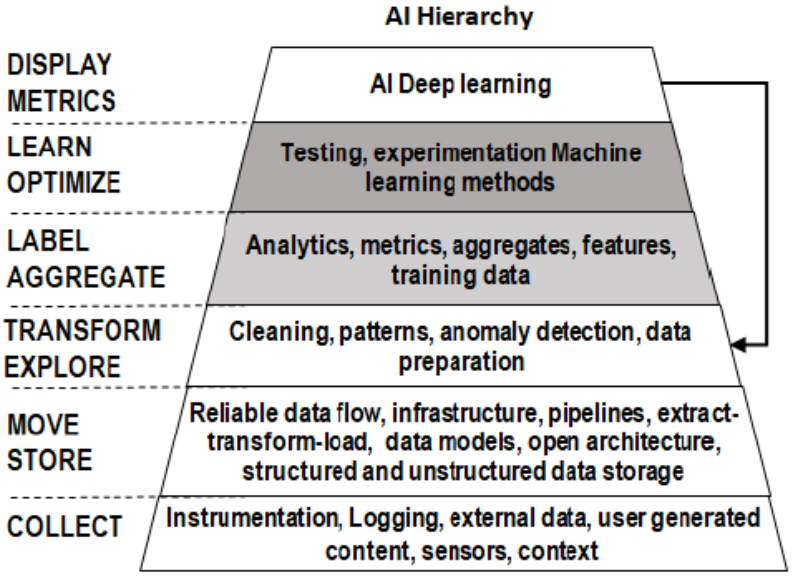

Figure 1. AI Hierarchy for Stages of AI Deployment

II. MOTIVATION

As AI applications grow in sophistication, they are becoming increasingly prevalent in the national security space. We focus on NLP, a quickly developing component of AI that uses machines to analyze, understand, and generate natural language to:

- **Manage the Flood of Information.** The amount of unstructured text data generated daily is exponentially increasing. This presents challenges for analysts of various types to classify, triage, and examine all relevant information for their specific problem area of interest, with potential applications to multiple languages. For example, to keep pace with this ever-increasing amount of information, the Intelligence Advanced Research Projects Activity's (IARPA) BETTER program seeks to develop new tools to enable personalized extraction of semantic information from text for triage and retrieval [8].

- **Create and Employ New Means of Generating Insights**. AI has enabled different sectors to tackle complex security issues. For example, work from MIT researchers in 2019 evaluated an automated fake-news detection system, revealing how machine-learning models catch subtle but consistent differences in the language of factual and false stories. Facebook, in 2017, experimented with AI to understand text that might be advocating for terrorism. They are experimenting with analyzing text that they have already removed for praising or supporting terrorist organizations such as ISIS and Al Qa'eda so they can develop text-based signals for content that may be terrorist propaganda [9].

The use of AI applications in a national security environment, however, is often problematic because the opaque nature of most systems leads to an inability for human beings to understand how the results came about. Reliance on "black boxes" to generate predictions and inform decisions is potentially disastrous. The ability to trust a machine's output is central to the continued development of beneficial systems. The February 2019 Executive Order (EO) on Maintaining American Leadership in Artificial Intelligence highlights trust as one of the five guiding principles for future AI development [10].

II. MOTIVATION (continued)

As of this writing, a number of researchers and groups have proposed strategies for making the AI development process more transparent, understandable, and, in a word, trustworthy. These efforts include:

- Datasheets [11], data statements [12], and "nutrition labels" [13] to better understand the data that underlies a machine learning system;

- Cards to clarify the intended use cases of machine learning algorithms and minimize their usage in contexts for which they are not well suited [14];

- A "nutritional label" to communicate details of the ranking methodology or of the output to the end user [15];

- Factsheets to increase trust in AI services as a whole by documenting purpose, performance, safety, security, and provenance for customer examination [16]; and

- A focus on providing thorough documentation, including but not limited to how AI systems were designed and for what purposes. Other areas of focus include where the data came from and why that data was chosen; how they were trained, tested and corrected; and what purpose they're not suitable for? [17]

To build upon previous work and extend it for the national security community, we created a Multisource AI Scorecard Table (MAST) checklist based on the principles of good analysis adopted by the IC. MAST would subject an AI system and its outputs to standards similar to the rigorous ones used to evaluate analysis performed by humans. We took two approaches to illustrating how the analytic standards outlined in ICD 203 can generate greater transparency and trust. First, we examine a representative NLP application and show how applying ICD 203's nine standards can improve system results. We then examine three use cases of AI tools that benefit from the application of the MAST checklist.

III. STANDARDS FOR THE INTELLIGENCE COMMUNITY

The intelligence community (IC) has a variety of best practices of which the 2015 ICD 203 provides a good starting point for the use of AI in the national security environment. ICD 203 was part of a congressionally mandated Act to ensure IC products "are timely, objective, independent of political considerations, based on all sources of available intelligence, and employ the standards of proper analytic tradecraft." [19] The driving principle of ICD 203 is to engender trust in the IC's analysis. The elements of building this trust includes ensuring analytic integrity, rigor, objectivity, relevance, accuracy, timeliness, and assurance for privacy while guarding against bias and politicization. To be ICD 203 compliant, IC analysis must reflect the nine standards.

- **Standard 1. Properly describes quality and credibility of underlying sources, data, and methodologies (Sourcing)**: This standard requires statements on factors affecting source quality and credibility. Such factors can include accuracy and completeness, possible denial and deception, timeliness of information, technical elements of collection, source pedigree and priority, evidence analysis, and assumptions.

- **Standard 2. Properly expresses and explains uncertainties associated with major analytic judgments (Uncertainty)**: Output assessments should indicate data and decision uncertainties. Analytic confidence may be based on the logic and evidentiary base that underpin it, including the quantity and quality of source material. Analytic outputs should note causes of uncertainty and explain how uncertainties affect analysis.

- **Standard 3. Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments (Distinguishing)**: Assumptions are defined as suppositions used to frame or support an argument and affect analytic interpretation. Judgments are defined as conclusions based on underlying intelligence information, analysis, and assumptions. Product reports should explicitly state assumptions of an argument.

- **Standard 4. Incorporates analysis of alternatives (Analysis of Alternatives)**: Analysis of alternatives is the systematic evaluation of differing hypotheses to explain events or phenomena, explore near-term outcomes, and imagine possible futures. Analysis should present plausible alternatives, especially when outcomes are uncertain, complexity is high, or when low probability events could produce high-impact results. The standard also looks for explanation of the reasoning and evidence that underpin the alternatives and the alternative's likelihood or implications.

III. STANDARDS FOR THE INTELLIGENCE COMMUNITY

(continued)

- **Standard 5. Demonstrates customer relevance and addresses implications (Customer Relevance)**: Provide information beyond what is generally known, addressing near-term, direct, or first order implications based on customer requirements. Address prospects, context, threats, or factors affecting opportunity for action.

- **Standard 6. Uses clear and logical argumentation (Logical Argumentation)**: Present a clear analytic message with clear reasoning with no flaws in logic, combining evidence, context, and assumptions effectively.

- **Standard 7. Explains change to or consistency of analytic judgments (Consistency)**: Note how major analytic judgments compare to previous production. Explain how new information or reasoning supports changing or maintaining an analytic line.

- **Standard 8. Makes accurate judgments and assessments (Accuracy)**: Analysis should apply expertise and logic to make the most accurate judgments and assessments possible, based on the information available and known information gaps.. Outputs should express judgements clearly and precisely as possible, reducing ambiguity by addressing the likelihood, timing, and nature of the outcome or development.

- **Standard 9. Incorporates effective visual information where appropriate (Visualization)**: Present visuals that are pertinent to the analysis, using visual information to clarify, complement, or augment data or analytic points.

IV. AI DEVELOPMENT PROCESS

There are many emerging developments in AI, which require standards of processing, deployment, and use. A current focus on AI includes the discussion of trust and transparency of mechanisms. If the AI tools support prediction, there is a need understand where along the AI development process developers and users should interject to better enable trust and measures of effectiveness [20].

Building on the Department of Defense's Joint Artificial Intelligence Center, the pipeline for AI development generally includes data, models, and products in Figure 3. The first stage considers data management including collecting entities from empirical analysis to form a corpus of data. As shown in Figure 2, the pipeline requires ingesting, cleaning, and labelling the data. Since most data is unstructured and unlabeled, there is an inherent tradecraft in data acquisition, especially where there is sparse or limit data available for a national security scenario.

The second stage uses model building that assumes some labelled data, from which the power of many deep learning methods have shown promise. Since not all data has a label, some theoretical understanding of context can help (e.g., theoretical sensor model [21]). Once the AI/Machine Learning (ML) performs data analytics, the visualization and deployment of results require human-machine interfaces to afford refinement and adaptation over model drift.

The third phase provides product use. Since various users are involved (man-machine systems), cognitive models determine the value of the systems to help make decisions and predictions. Human awareness supports the appreciation of the AI/ML minimum viable product to the work domain for relevance.

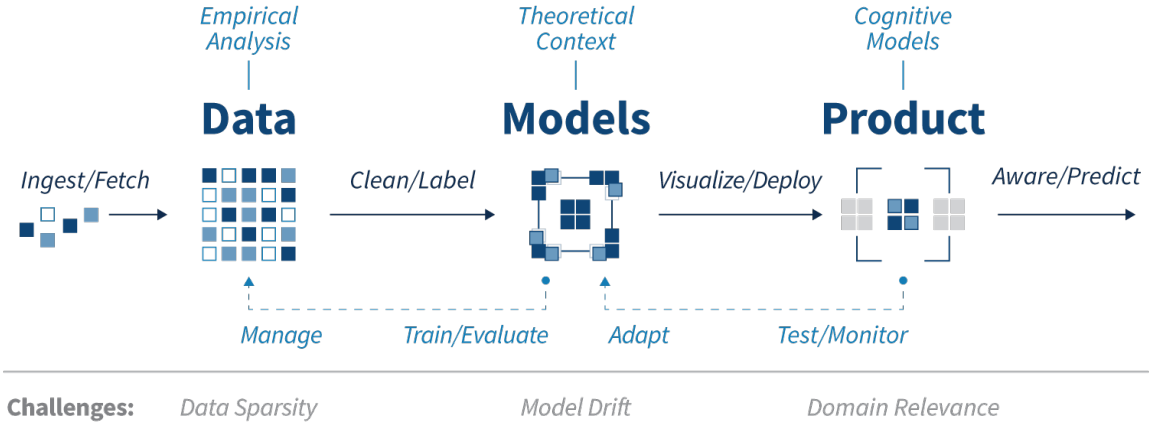

Figure 2. AI Deployment from Data to Decisions

IV. NATURAL LANGUAGE PROCESSING

There are many developments in NLP that have emerged over a variety of areas that include [22]:

Entity Analysis:

- Character Recognition
- Text Classification and Categorization
- Named Entity Recognition

Relationship determination:

- Part-of-Speech Tagging
- Machine Translation
- Speech Recognition

Event Processing:

- Semantic Parsing and Question Answering
- Paraphrase Detection
- Language Generation
- Multi-document Summarization

Each of these NLP applications has a different set of data corpus, decision-making support, as well as measures of importance. The applications follow the NLP pipeline of entity analysis, relationship determination, and event processing. At the simplest level is determining the language and keywords (e.g., entities) from documents or handwriting. These techniques support simple applications of doing a search for relevant documents that the user is pointed to for further analysis. A more complex example is question-answering [23].

The second general category would be for relationships that link together information within or among a corpus (e.g., statistical relational learning [24]). Moving from tagging to translation and speech recognition has shown promise; however mistakes are still common. Hence, what are the policies for using such techniques in decision-making? Can a translated quote be a meaningful representation of what the original speaker indicated? Hence, there are many upstream opportunities, but downstream implications of the output.

The last category builds on the previous, which includes determining the syntactic, semantic, and sentiment [25] events. For example, multi-document summarization includes many advanced techniques [26] that are not well repeatable (or consistent) such that users should be aware of blindly using the outputs without checking the source information.

In all of these NLP techniques, there is a need for standards agreed upon by academic, industrial, and government organizations. The academic community has determined some corpus data sets that support challenge problems which could be used for standard processing. Likewise, some common metrics to support trust include: credibility, precision-recall, and F score, but there are not consistent standard methods to determine the maturity of the evaluation metrics for general utility. For example, the NLP metrics only answer ICD 203 standard 2 for output uncertainty. To move the AI community towards some standard metrics, we propose a checklist scorecard.

To demonstrate the usefulness of the Multisource AI Scorecard Table (MAST) approach, we present initial examples of use cases from a series of applications (a.k.a. Apps) that utilize methods of AI towards a common set of analytics tools [27]. The future MAST content can be adapted towards the domain and use of the AI technology for different applications.

V. CASE STUDIES: ICD 203 STANDARD FOR EVALUATING THE TRUSTWORTHINESS OF AI SYSTEMS FOR NATIONAL SECURITY

Scenario: *You are an intelligence analyst and the AI system is analyzing material and deriving results for you. Given your reporting needs, the AI system will help you sort through the material and assess major topics present, provide a descriptive title for the topics, and generate a summary of the material in the topic area. This capability will help you sort through the large amounts of material available to you every day, obtain salient content, and get up-to-date on the topic of interest.*

Question: *What can the system or developers do to increase your trust in system outputs? See Appendix 1.*

To help users and developers assess how well the above fictitious AI system can explain the basis for its decisions (i.e. "analysis"), we have designed a series of scorecard ratings modeled on the analytic tradecraft standards described in ICD 203, which are used both to guide and evaluate the analysis performed by human analysts in the IC. See Appendix I for more detail on the checklist.

Taken together, the nine ICD 203 standards provide guidance to evaluate a given AI system's "analysis" in order to provide transparency and explainability, the underpinnings of trust. Standards 1 (Sourcing), 4 (Analysis of Alternatives), and 8 (Accuracy) promote transparency. Standards 2 (Uncertainty), 3 (Distinguishing), 5 (Customer Relevance), 6 (Logical Argumentation), 7 (Consistency), and 9 (Visualization) better enable explainability.

For AI systems, we propose using a modified set of standards on the back-end to evaluate the capabilities and documentation associated with a system and its outputs to provide feedback to stakeholders. For each of the categories, there is a 0-3 rating with 1-3 indicating that the application meets the ICD 203 standards to varying degrees, and 0 indicating that the application does not meet the standard. Note: many applications are designed for specific functions, so it is most likely each application would not meet all of the standards; however, in some cases in which many apps are integrated, the generalized system would meet most of them (e.g., use case 3, [24]).

V. CASE STUDIES: ICD 203 STANDARD FOR EVALUATING THE TRUSTWORTHINESS OF AI SYSTEMS FOR NATIONAL SECURITY (continued)

Figure 3 on page 14 depicts an application scored by MAST that is good at the explainability criteria but needs improvement in transparency. Below is a sample from Appendix I explaining why the system achieved the scores it did for Standard 3 ("Good") and Standard 4 ("Poor").

- The NLP system received a "good" for Standard 3 (Distinguishing). This means that it did well in clearly distinguishing between the derived results (i.e., topic areas, titles, and summaries) and the reports, sentences, and words it used to derive them. For example, by explicitly stating that the summary was derived from a certain number of reports and providing a link to the reports. The system also comes with a datasheet with information on the underlying assumptions that framed the choice of data, a model card that details the type of model and assumptions inherent in its development, and a factsheet with a statement of purpose.

- The system did poorly in Standard 4 (Analysis of Alternatives). The NLP system did not provide alternative topics, titles, or summaries even though system complexity, noise, and lack of data warranted their inclusion.

To expand on the notional example, three use cases are presented from applications that use AI NLP methods for tactical assessment, operational analysis, and strategical forensics. The MAST analysis results from prior experience reported in the literature. The scorecard establishes a standard rating and a benchmark of the rating value. For example, if a new user gets a scorecard with most of these ratings certified, they could reasonably trust that the tool would meet their needs. Future ratings systems could be expanded at each category to provide further refinement of the maturity of AI NLP systems.

V. CASE STUDIES: ICD 203 STANDARD FOR EVALUATING THE TRUSTWORTHINESS OF AI SYSTEMS FOR NATIONAL SECURITY (continued)

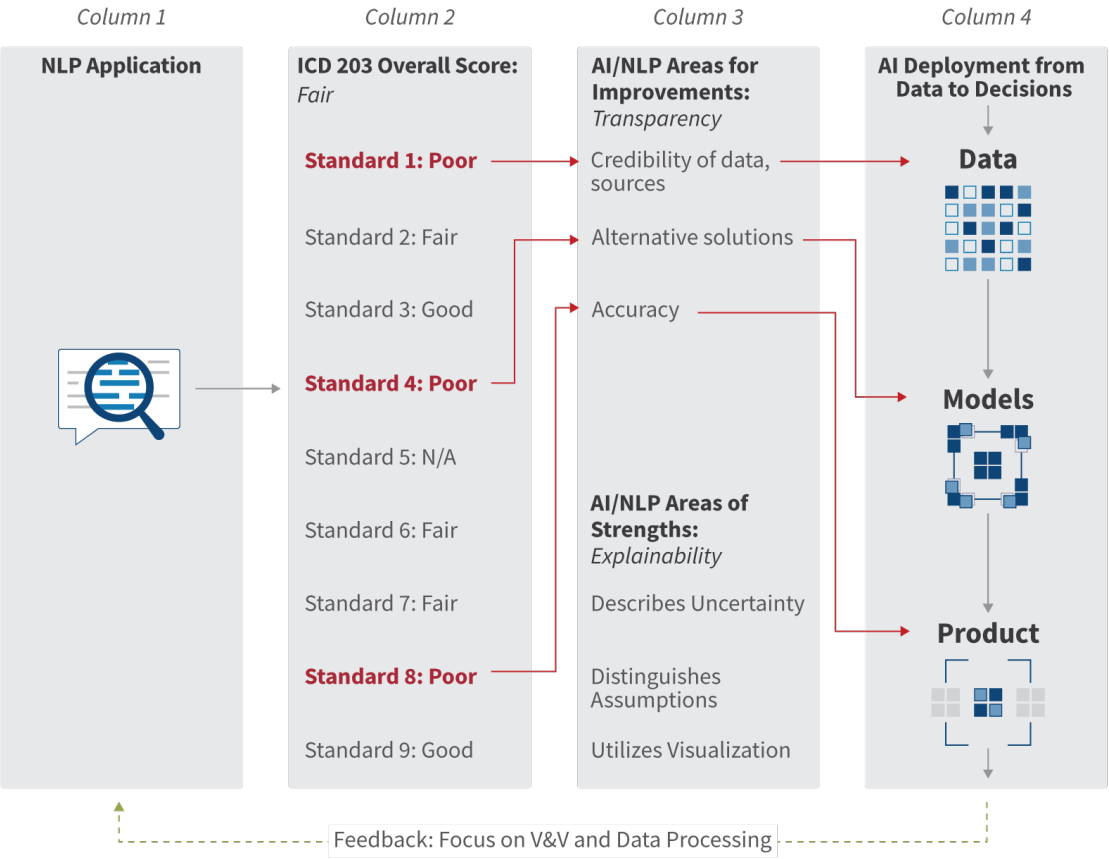

Figure 3. ICD 203 Standards Scoring Process

In Figure 3, the first column represents a representative AI/NLP system and under its "hood", in the fourth column, are the different stages of AI development. Column two presents the nine ICD 203 standards and a score derived from a modified version of the criteria found in DHS's ICD 203 evaluation form for evaluating human analysis. The ICD 203 scores range from 0 to 3 indicating how "poor," "fair," "good," or "excellent" the AI system does in meeting the standards. The third column contains modified language appropriate for evaluating the system aspects for AI improvements. The improvement suggestions point to the appropriate aspects of the AI development process shown in column four. At the end of the process, the scores provide information on the capability of a system to explain its decisions.

V. USE CASE 1: TACTICAL ASSESSMENT (SITE MONITORING)

Real time social media and intelligence sources determine the activity at a certain location for safety, security, and surveillance. As representative of a public location, information sources about the site's activities include online reporting, published documents, and real-time feeds (e.g., chat) [28]. Given a location, the fixed security devices enable video analysis, text analytics, and event reporting as shown in Figure 4 [29]. The challenge was to associate the reported chat with the video tracking.

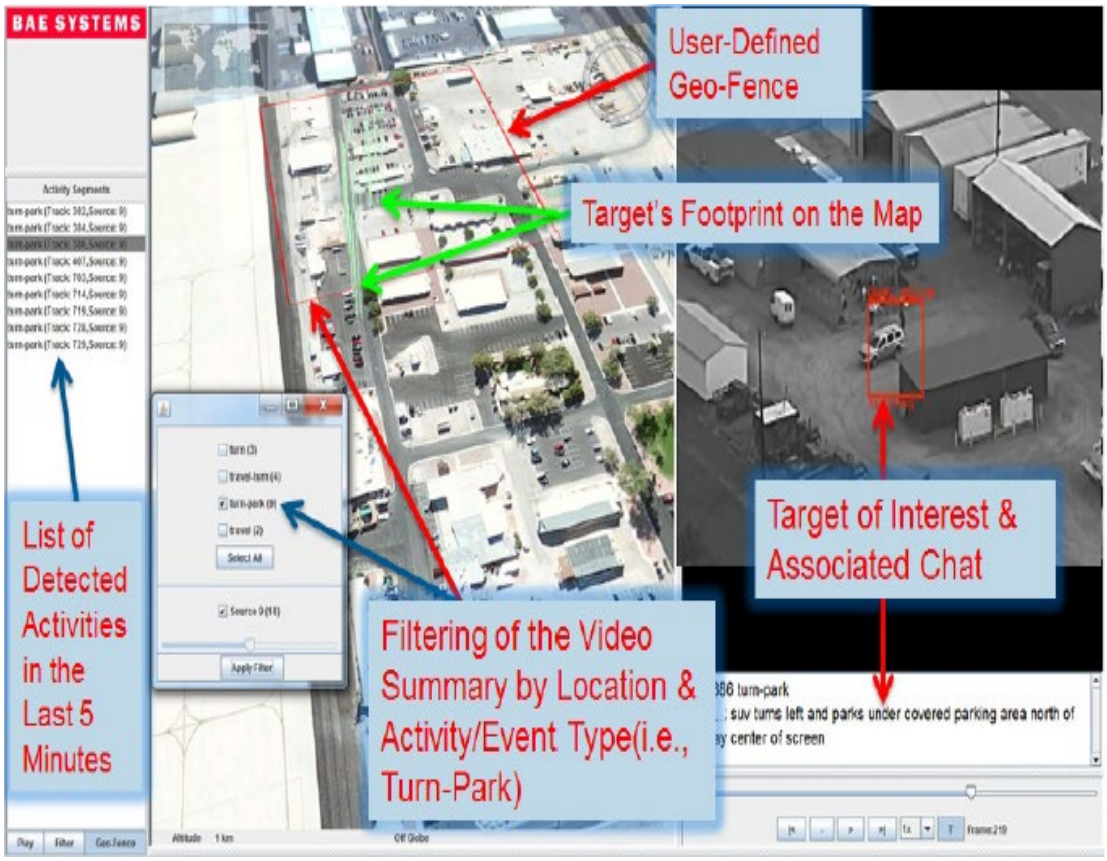

Figure 4. Association of Chats to Tracks (ACT) App [29]

V. USE CASE 1: TACTICAL ASSESSMENT (SITE MONITORING) (continued)

The Association of Chats to Track (ACT) AI system is both transparent and able to explain its results but also demonstrates that not all the standards are applicable or necessary in certain circumstances. For example, it receives a 0 (i.e., "Poor") for Standard 4 but its overall functions are sufficient for this case of real time reporting when limited time is available to provide options (Table 1). The goal was to link the video and chats in time and space, which required easy to exploit NLP methods over micro text from semi-structured content.

Table 1. ACT Scorecard

Std		Score	How Determined
1	Sourcing	3	Has availability of credibility info, relevance, and data
2	Uncertainty	2	Provides site details and chat (term) clarifications
3	Distinguishing	1	System trained to associate key words (e.g., dictionary) with objects, behaviors, and event triggers
4	Analysis of Alternatives	0	No summary of AoA as the system seeks to use NLP for change detection and object analysis (e.g., tracking).
5	Customer Relevance	3	System designed to meet user needs
6	Logical Argumentation	2	Graph matching is based on association and word-to-word association combined with word-to-vect for analysis
7	Consistency	2	Similarity verified from multiple use case studies
8	Accuracy	3	Timing, likelihood, and accuracy visual verifiable
9	Visualization	3	Visualization tested with special forces

V. USE CASE 2: OPERATIONAL ANALYSIS (SCENARIO DEVELOPMENT)

Various structured documents from vetted sources provide a baseline analysis. Situation reports (SITREPs), intelligence production reports (IPRs) and other documents provide historically relevant information. The real-time updates from human intelligence (HUMINT) and other reports provide content for various analytical methods in support of emerging scenario developments. Example cases include Egypt Uprising, Haiti Disaster Relief, Ukrainian firefights, and Puerto Rico real-time response. Figure 5 shows the Multi-INT Data Association Tool (MIDAT) app [30], which utilizes human-derived data within Multi-INT Activity Pattern Learning and Exploitation (MAPLE) for pattern of life (POL) analytics.

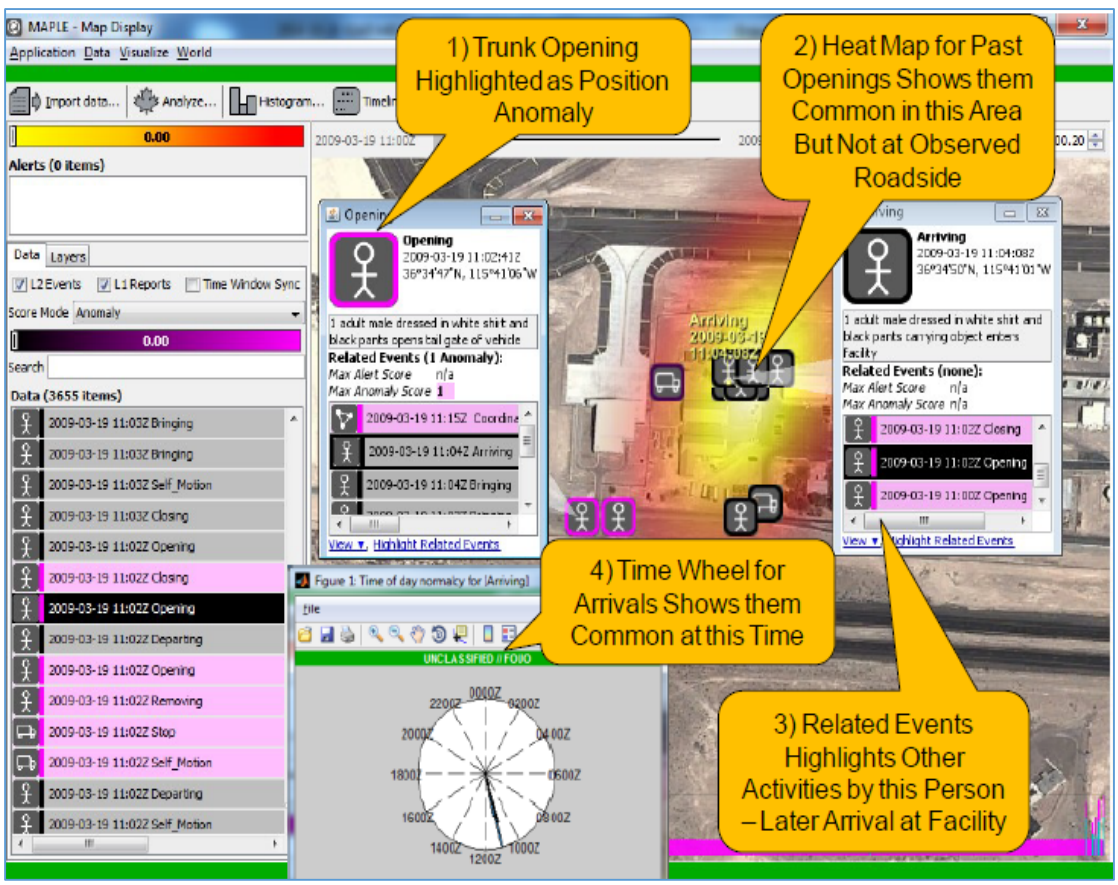

Figure 5. MIDAT App [30]

V. USE CASE 2: OPERATIONAL ANALYSIS (SCENARIO DEVELOPMENT) (continued)

Overall, the MIDAT AI system did well, scoring mostly "fair" across the board. It did poorly presenting alternative possibilities (Standard 4) for derived results when uncertainties, system complexity, noise, lack of data, mismatched data, etc. warrant their inclusion. Unlike use case 1, depending on the context and user, a poor score for Standard 4 may prevent it from being deployed. The checklist has surfaced an opportunity for developers and users to discuss system needs and make improvements to future versions.

Table 2. MIDAT Scorecard

Std		Score	How Determined
1	Sourcing	1	Has availability of credibility info, relevance, and data; but includes unstructured data for analysis
2	Uncertainty	1	Provides site details and chat (term) clarifications (extended from video event segmentation by text – VEST)
3	Distinguishing	1	System trained to associate key words (e.g., dictionary) with objects, behaviors, and event triggers for both structured data (vetted) and unstructured data (chat)
4	Analysis of Alternatives	0	Does not provide alternative explanations
5	Customer Relevance	1	System designed to meet user needs, but design varies based apps being aggregated
6	Logical Argumentation	1	Uses Dirichlet method
7	Consistency	1	Similarity verified from multiple use case studies
8	Accuracy	1	Timing, likelihood, and accuracy visually verifiable, but not well connected
9	Visualization	3	Plotted in JVIEW UDOP [31]

V. USE CASE 3: STRATEGIC FORENSICS (Social Pathway)

A corpus of documents covering a region of engagement over a multi-year time period. The collected documents contain vetted information and databases exist with additional news sources, historical records, tweets, open source intelligence (OSINT)/ public domain information (PDI) collections. Along with the OSINT information, there exists others sources collected and transcribed using a variety of NLP methods. The linked material from graphical methods include 30+ intelligence sources. The integration of various applications enhances the WATCHMAN™ Analytics tool shown in Figure 6 [32].

From a series of tests, along with user-developed needs, the WATCHMAN™ AI system provides federated search, content recommendations, analytics, and a user-defined operational picture (UDOP) for each user to determine which attributes support their mission needs. The tool supports distributed collaborations and dynamic ontological representations from which products generated from man-machine productions reside in a database for analysis of alternatives forensic analysis.

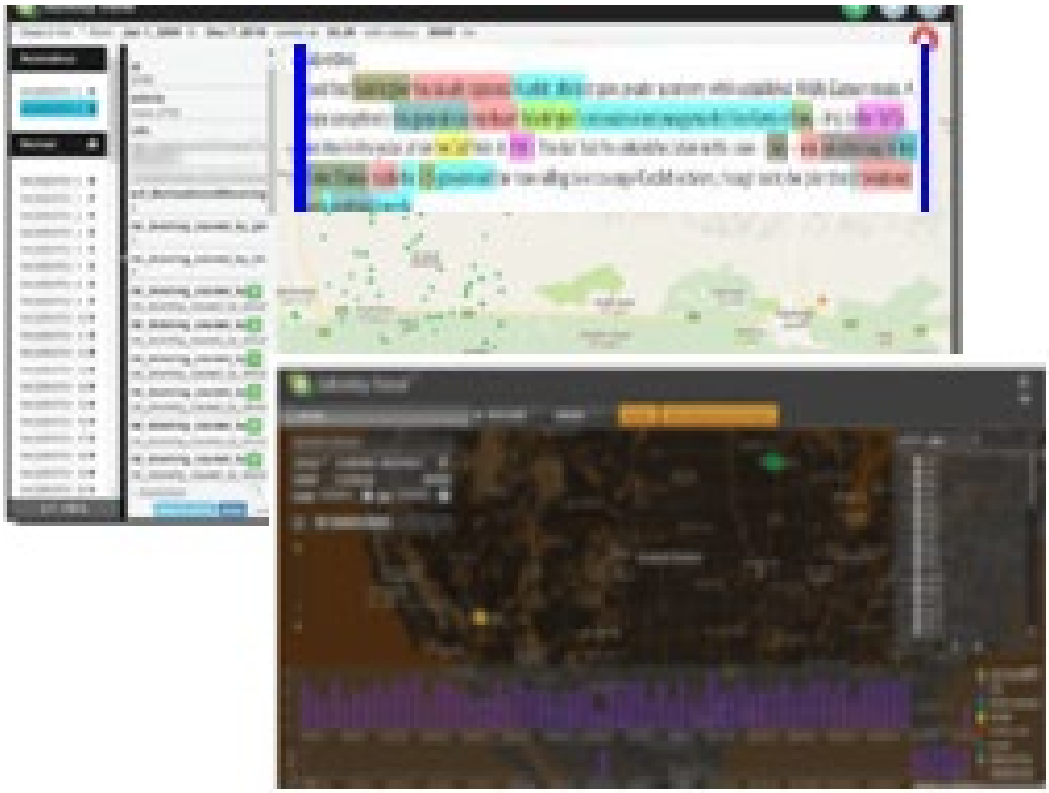

Figure 6. WATCHMAN™ Tool [27]

V. USE CASE 3: STRATEGIC FORENSICS (Social Pathway) (continued)

The AI scored well across all standards, representing close collaboration with IC members and developers. Among the highlights include how well it scored on standard 9. For example, the NLP system visually represents the quantity and quality of links between categories that have been grouped together and does not require the customer to sift through a multitude of definitions and other text sections to decipher any one axis.

Table 3. WATCHMAN™ Scorecard

Std		Score	How Determined
1	Sourcing	3	Has availability of credibility info, relevance, and data. Over 30 sources including books/manuals, documents, social media, emails, chat (twitter)
2	Uncertainty	2	Includes machine uncertainty while human uncertainty is manual inputted when report is filed
3	Distinguishing	3	Affords analyst-marked updates to AI results. AI supports text/document discovery; but does not create new sources
4	Analysis of Alternatives	3	Provides top 20 assessments
5	Customer Relevance	3	Tailored to user needs
6	Logical Argumentation	3	Uses high-order Bayesian methods
7	Consistency	3	Provides details for near real-time product production such as predefined multi-INT captioning
8	Accuracy	2	Provides some narratives based on multi-INT sourcing
9	Visualization	3	User Defined Operating Picture (UDOP)

VI. CONCLUSIONS

This paper presents the need to increase trust in AI systems deployed to support national security and provides a recommendation based on the ICD203 analytic standards accepted and adopted by the IC. The proposed standards demonstrate a method to successfully rate AI machine "analytics" tools that support users conducting multi-source discovery and reporting. MAST demonstrates a benchmark for comparative assessment of AI systems to provide analysts with trustworthy outputs, guide developers toward improvements, and assist purchasers of available products.

Some of the proposed standards measure system capabilities while others potentially refer to supplemental information that should come with the system (e.g., datasheets or model cards) or actions that should be performed by the user or data owners. We acknowledge that all use cases are unique and leave it up to the end user to decide what level of performance is acceptable. Efforts to build transparency into a system will likely result in tradeoffs, such as greater development time and cost, a more complicated user interface, the display of more information, and/or slower system performance. Having a benchmark scorecard standard will support certification, testing, and sustainment requirements.

Future efforts require discussions on meaningful MAST standards from which refinements and considerations will improve the eventual process. Moving forward, the scorecard is a first step in the dialog between users, developers, and purchasers of AI systems meant for national security applications. We encourage stakeholders to take the MAST checklist and test it on systems they are developing or have already deployed and make suggestions and improvements. The MAST methods can only improve from academic, industry, and government testing and evaluation (T&E) for operational sustainment and maintenance of AI systems.

VII. REFERENCES

[1] E. P. Blasch, E. Bosse, and D. A. Lambert, High-Level Information Fusion Management and Systems Design, Artech House, Norwood, MA, 2012.

[2] L. Snidaro, J. Garcia Herrero, J. Llinas, et al. (eds.), Context-Enhanced Information Fusion: Boosting Real-World Performance with Domain Knowledge, Springer, 2016.

[3] E. Blasch, R. Cruise, A. Aved, U. Majumder, T. Rovito, "Methods of AI for Multi-modal Sensing and Action for Complex Situations," AI Magazine, Winter 2019.

[4] E. Blasch, "Trust Metrics in Information Fusion," Proc. SPIE, Vol. 9091, 2014.

[5] J-H. Cho, K. Chan, S. Adali, "A Survey on Trust Modeling," ACM Computing Surveys, 48, (2), Article 28, Oct 2015.

[6] W. Yu, H. Xu, J. Nguyen, et al., "Public Safety Communications: Survey of User-side and Network-Side Solutions and future Directions," IEEE Access, 6(1): 70397-70425, December 2018.

[7] "Intelligence Community Directive 203," 02 Jan 2015; https://www.dni.gov/files/documents/ICD

[8] IARPA, https://www.iarpa.gov/index.php/research-programs/better

[9] Facebook, https://fbnewsroomus.files.wordpress.com/2018/05/understanding_the_community_standards_enforcement_report.pdf

[10] The White House, https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/, 11 February 2019.

[11] T. Gebru, J. Morgenstern, B. Vecchione, et al., "Datasheets for Datasets," http://arxiv.org/abs/1803.09010, 14 April 2019.

[12] E. M. Bender, B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," https://openreview.net/pdf?id=By4oPeX9f, 2019.

[13] S. Holland, A. Hosny, S. Newman, J. Joseph, et al., "The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards," May 2018. https://arxiv.org/ftp/arxiv/papers/1805/1805.03677.pdf

[14] M. Mitchell, S. Wu. et al., "Model Cards for Model Reporting," Proc. Conf. on Fairness, Accountability, and Transparency, 220-229, 2019.

[15] K. Yang, J. Stoyanovich, et al., "A Nutritional Label for Rankings," Proc. Int'l. Conf. on Management of data, (SIGMOD), pp. 1773-1776, 2018.

[16] M. Arnold, R. K. E. Bellamy, M. Hind, et al, "FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity," https://arXiv.org/pdf/1808.07261v2 , 7 Feb 2019.

[17] Partnership on AI, Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML), https://www.partnershiponai.org/about-ml/, accessed 9/5/2019.

VII. REFERENCES (continued)

[18] https://www.nist.gov/programs-projects/knowledge-extraction-and-application-manufacturing-operations
[19] ODNI, https://www.intelligence.gov/mission/our-values/342-objectivity.
[20] E. Blasch, R. Breton, P. Valin, "Information Fusion Measures of Effectiveness (MOE) for Decision Support," Proc. SPIE 8050, April 2011.
[21] E. Blasch, S. Ravela, A. Aved, Handbook of Dynamic Data Driven Applications Systems, Springer, 2018.
[22] O. Davydova, "10 Applications of artificial Neural Networks in Natural Language Processing," Data Monsters, https://medium.com, Aug 17, 2017.
[23] M. Malec, T. Khot, J. Nagy, et al., "Inductive Logic Programming meets Relational Databases: An Application to Statistical Relational Learning," Inductive Logic Programming (ILP), 2016.
[24] P. Rajpurkar, J. Zhang, K. Lopyrev, et al., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," https://arxiv.org/pdf/1606.05250, 11 Oct. 2016.
[25] B. Liu, et al., "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier," IEEE Intl Conf. on Big Data, Oct. 2013.
[26] O. Vasilyev, T. Grek, et al., "Headline Generation: Learning from Decompostable Document Titles," https://arxiv.org/pdf/1904.08455, 10 May 2019.
[27] E. Blasch, J. Nagy, S. Scott, W. D. Pottenger, H. Holloway, P. DiBona, J. Ritchie, A. Basharat, J. Okoth, R. Mann, "Resolving events through Multi-Intelligence Fusion," J. of DoD Res. and Eng., 2(1):2-15, 2019.
[28] E. Swears, A. Basharat, A. Hoogs, E. Blasch, "Probabilistic Sub-Graph Matching for Video and Text Fusion," National Symposium on Sensor and Data Fusion, 2014.
[29] R. I. Hammoud, C. S. Sahin, E. P. Blasch, B. J. Rhodes, and T. Wang, "Automatic Association of Chats and Video Tracks for Activity Learning and Recognition in Aerial Video Surveillance," Sensors, 14, 19843-19860, 2014.
[30] M. K. Schneider, O. Babko-Malaya, E. Blasch, L. Chen, V. Crespi, J. C. Handuber, P. J. Haney, J. Nagy, M. S. Richman, G. Von Pless, H. Zhu, B. J. Rhodes, "Learning patterns of life from intelligence analyst chat," Proc. SPIE, Vol. 9842, 2016.
[31] E. Blasch, "Enhanced Air Operations Using JView for an Air-Ground Fused Situation Awareness UDOP," AIAA/IEEE Digital Avionics Systems Conference, Syracuse, NY, Oct. 2013.
[32] http://www.intuidex.com/papers/SBIR_Success_Story.pdf

VIII. AI TEAM MEMBERS

Name	Organization
James S (Co-Champion)	DHS
Tao N (Co-Champion)	DIA
Erik Blasch	Air Force Research Lab
Chandra Pauline (Dinakar) Daniel	UNAIDS
Karl G	DHS
Alisa Paige Mason	Guidepost Solutions LLC

IX. APPENDIX – Multisource AI Scorecard Table (MAST)

ICD 203-Based Standard for Evaluating the Trustworthiness of AI Systems for National Security

NLP Example: *You are an intelligence analyst and the AI system is analyzing material and deriving results for you. Given your reporting needs, the AI system will help you sort through the material and assess major topics present, provide a descriptive title for the topics, and generate a summary of the material in the topic area. This capability will help you sort through the large amounts of material available to you every day, obtain salient content, and get up-to-date on the topic of interest.*

Question: *What can the system or developers do to increase your trust in system outputs?*

To help stakeholders (e.g. users, developers, and purchasers of AI systems) assess how well the above representative AI system can explain the basis for its decisions (i.e. "analysis"), we have designed a series of checklists—the Multisource AI Scorecard Table (MAST)—modeled on the analytic tradecraft standards described in Intelligence Community Directive (ICD) 203. ICD 203 standards are used both to guide and evaluate the analysis performed by human analysts in the Intelligence Community.

- *For AI systems, we propose using modified standards to evaluate the capabilities and documentation associated with a system as well as system outputs to provide feedback to developers and users. The first column presents the original language found in DHS's ICD 203 evaluation form for evaluating human analysis.*
- *The second column contains modified language appropriate for evaluating an AI system based on the ICD 203 standards. The second column also provides an example showing what an AI/NLP system rated at a particular level might look like. The ICD 203 scores range from 0 to 3 indicating how "poor," "fair," "good," or "excellent" the AI system does in meeting the standards. At the end of the process, you will have a series of scores that provides information on the capability of a system to explain its decisions.*

IX. APPENDIX – Multisource AI Scorecard Table (MAST)

(continued)

Summary of Key Points

- Taken together, we believe the nine standards for a given analysis provide transparency and explainability, the underpinnings of trust. Standards 1 (Sourcing), 4 (Analysis of Alternatives), and 8 (Accuracy) promote transparency. Standards 2 (Uncertainty), 3 (Distinguishing), 5 (Customer Relevance), 6 (Logical Argumentation), 7 (Consistency), and 9 (Visualization) enable explainability.
- Some standards translate to system capabilities while others potentially refer to supplemental information that should come with the system (e.g., datasheets or model cards) or actions that should be performed by the user or data owners.
- Use of these analytic standards would help define the metrics and analytic functionality that would support user needs for work domain applications.
- We recognize that all use cases are unique and leave it up to the end user to decide what level of performance is acceptable. Efforts to build transparency into a system will likely result in tradeoffs, such as greater development time and cost, a more complicated user interface, the display of more information, and/or slower system performance.
- The scorecard is a first step in the dialog between users, developers, and purchasers of AI systems meant for national security applications. The MAST methods can only improve from academic, industry, and government testing and evaluation (T&E) for operational sustainment and maintenance of AI systems.

STANDARD 1 – Sourcing

Properly describes quality and credibility of underlying sources, data, and methodologies

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Largely lacks sourcing or describes reporting base, data, or methodologies only vaguely; OR	System does not provide a description of the data used to derive the result, or the developer describes the data used to train and validate the system and the methodology used by the algorithm(s) used to derive the results only vaguely; OR The NLP system derives a list of major topics and creates titles, but does not show a list of the reports used to generate the topic grouping or title. There is no datasheet or model card describing the training or validation data or methodology used by the algorithm. OR
Misidentifies or misrepresents cited reporting, data, or methodologies.	System or developer misidentifies or misrepresents the data used to derive the results, data used for training and validation, or methodology used by the algorithm(s). The NLP system shows reports that were supposedly used in the generation of the topics and titles, but the list is incorrect or incomplete. The description of the data used to train and validate the model and the description of the methodology used by the model are inaccurate.
FAIR (1)	
Contains basic, generic descriptions of cited reporting, data, or methodologies; AND	System shows basic, generic descriptions of the data used to derive a result; and includes a datasheet describing the data used for training and validation and a model card describing the methodology used by the algorithm(s); AND *The NLP system shows a list of the data used to derive the topics and titles, including the producing agency and serial number. The developer includes a datasheet and model card describing the data used for training and validation and algorithm methodology, respectively;* AND
The product contains source reference citation endnotes.	System allows you to view the data used to derive a result. *The NLP system provides a list of reports used to derive the topics. You are able to click on the individual reports to view the text.*
GOOD (2)	
Meets Fair standards; AND	
Detailed source descriptors; AND	System provides detailed description of sources used to derive the result; AND NLP system annotates the reports (e.g., bolded and/or colorized) to show the text that it used to derive the result; AND
Strong Source Summary Statements that assess strengths and weaknesses of reporting; AND	Developer provides a detailed datasheet describing potential biases or errors that might result from applying the model to a dataset other than the one it was trained and validated on; AND The developer provides a datasheet describing how the NLP model was trained on press reporting, and the potential errors that might result from using the system on government reporting; AND
Explains factors affecting source quality and credibility; AND	System is able to describe the body of data analyzed and provide diagnostic information about the quality of the data used to derive the result; AND *The NLP system provides the user with information about the dataset, such as the number of reports from an agency, intelligence type (HUMINT, SIGINT, etc.), and source quality. The NLP system is also able to provide you with a visualization that helps you determine where there are problems with the data; AND*
	Developer or data owner provides a datasheet describing the quality of the data used by the AI system to derive a result; AND *Data owners provide a datasheet describing the data included in the data repository searched by the NLP system. E.g. NSA provides a description of their SIGINT reporting;* AND
Cites sources as close to first-hand when possible.	User or data owner cleans the data so that only data as close to first-hand is considered for analysis by the AI system. *The user or data owner filters the reporting so the NLP system does not consider duplicates, derivative reports, and summaries for analysis.*
EXCELLENT (3)	
Addresses all of Good criteria; AND	
Discusses linkage of sources to analysis and judgments; AND	System provides a detailed description of the linkage between data and the derived result; AND *The NLP system allows you to click on text in the title or summary and it will bring up relevant passages in the underlying reporting that it used to extract the passages;* AND
Provides detailed description of factors that could affect source quality and credibility.	System provides a description of factors that could affect the source quality and credibility. *The NLP system has detailed knowledge of reporting agencies, organizations, sources, source descriptors, collection methods and techniques, and source evaluations and can provide a summary of the factors affecting the quality and credibility of sources used to derive a result.*

STANDARD 2 – Uncertainty

Properly expresses and explains uncertainties associated with major analytic judgments

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Does not indicate levels of uncertainty associated with major analytic judgments; OR	System does not indicate levels of uncertainty associated with derived results; OR *The NLP system provides you with lists of topics, a title and summary, but does not indicate in any way its confidence in the accuracy of the results;* OR
Indicates levels of uncertainty associated with major analytic judgments that are inconsistent with the basis ascribed to them; OR	System indicates levels of uncertainty associated with derived results, but are inconsistent with the underlying data; OR *The NLP system provides indicators of confidence, for example an "accuracy score" associated with the summary text, but it appears to be inconsistent when combined with visualizations or other analysis of the underlying data;* OR
System provides a set of metrics such as "accuracy" with no derivation of how determined.	System provides metrics that seem to be *ad-hoc* without any verification of how derived. *The NLP system lists metrics, such as the accuracy score, that are an incomplete representation of how it derived the summaries and also vary throughout the analytics.*
FAIR (1)	
Indicates levels of uncertainty associated with major judgments; BUT	System indicates levels of uncertainty associated with derived results; BUT *The NLP system provides you with an accuracy score associated with each topic, title, and/or summary;* BUT
Does not explain their basis (e.g. by reference to strengths and weaknesses of the information base, contrary reporting, assumptions, or the nature of the judgment).	Does not explain their basis of determination. *The NLP system provides no explanation for how it derived the accuracy score, but it appears to be generally consistent.*
GOOD (2)	
Indicates levels of uncertainty associated with major judgments; AND	System indicates levels of uncertainty associated with derived results; AND *The NLP system provides you with an "accuracy score" associated with each topic, title, and/or summary*; AND
Explains their basis.	Explains their basis. *The NLP system explains in simple terms how the accuracy score was derived for the summaries.*
EXCELLENT (3)	
Satisfies Good criteria; AND	
Provides especially thorough discussion of nature and source of uncertainties affecting major analytic judgments; OR	System can provide detailed diagnostic information about uncertainties in the derived results and why they came about; OR *The NLP system provides a detailed description of the accuracy of its topical groupings, titles, and/or summaries, including showing problems with the structure of the raw reporting, inconsistent or contrary information in the reporting, transfer bias, etc.;* OR
Identifies indicators, that if detected, would alter levels of uncertainty associated with major analytic judgments.	System can identify and describe the type of data it would need in order to increase the confidence in derived results; OR *The NLP system will tell you that more or better quality reporting in a certain area would increase the accuracy of the title or summary;* OR

STANDARD 3 – Distinguishing

Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Does not distinguish among statements that convey underlying information, assumptions, and judgements;	The system does not provide any indicators that help the user distinguish between derived results and underlying data. The developer does not provide a datasheet, model card, or factsheet with information on the underlying assumptions that framed the choice of data, curation rationale, annotations, the collection process, or model. The developer has not tuned the data and the model to the domain. *The NLP system does not provide any indicators that help the user distinguish between titles and summaries derived from underlying intelligence reports and the raw reporting itself. There is no or limited factsheet describing the purpose of the NLP system or how the algorithm was trained on press reports so that users can tune for use on government sources or how the NLP system derives it's the topics, titles, and summaries.*
FAIR (1)	
Sometimes distinguishes among statements that convey underlying information, assumptions, and judgement; BUT	The AI system has a datasheet with information on the underlying assumptions that framed the choice of data, curation rationale, annotations, the collection process. The system provides some indicators that help the user distinguish between derived results and underlying data.; BUT *The model developer provides a datasheet describing how the NLP algorithm was trained on press reports, the origin of the reports, and the collection process, etc.;* BUT
Does not explicitly state assumptions that serve as linchpins of an argument or bridge key information gaps.	The model card and factsheet are incomplete and has limited domain expert input. *The model card and factsheet do not explain in detail how the NLP system derives the topics, titles, and summaries.*
GOOD (2)	
Consistently distinguishes among statements that convey underlying information, assumptions, and judgments; AND	The system is able to consistently provide indicators that help the user distinguish between derived results and underlying data; AND *The NLP system clearly distinguishes between derived results (i.e. topic areas, titles, and summaries) and the reports, sentences, and words it used to derive them, for example by explicitly stating that the summary was derived from a certain number of reports and providing a link to the reports;* AND
Explicitly states assumptions that serve as linchpins of an argument or bridge key information gaps.	The system has a datasheet with information on the underlying assumptions that framed the choice of data, curation rationale, annotations, and the collection process. There is also a model card that details the type of model to understand the type of assumptions, domain-relevant expert input and judgement, and current model version. There is a factsheet focused on explainability, such as statement of purpose, if there is direct or indirect use of data and algorithm, and the recommend use, users, and restriction. *The NLP system is able to explicitly state assumptions that it used to fill in the gaps when deriving summaries from limited or noisy reporting. There exists a datasheet provided by the data owners describing the type of data found in the repository used by the NLP system.*
EXCELLENT (3)	
Satisfies "good" criteria; AND	
Identifies indicators that, if detected, could validate or refute assumptions; OR	The system is able to identify indicators in the data that would validate or refute assumptions used in model development; OR *The NLP system identifies structural or semantic indicators in the reporting that, if detected, could validate or refute assumptions used to develop the model.*
Explains the implications for judgements if assumptions are incorrect	The developer includes a datasheet or factsheet that details aspects of the training data that were critical to the development of the model, how the system could react to use in unfamiliar environments, and suggestions for how to reduce risks of bias and error. *The developer of the NLP system includes detailed information that explains how press and government reporting are different, how the model might react, and the remedies that would likely need to be taken to account for the differences in the two different types of reporting.*

STANDARD 4- Analysis of Alternatives

Incorporates analysis of alternatives

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Does not present alternatives when uncertainties, complexity or low probability/high impact situations warrant their inclusion.	System does not present alternative possibilities for derived results when uncertainties, system complexity, noise, lack of data, mismatched data, etc. warrant their inclusion. *The NLP system does not provide alternative topics, titles, or summaries.*
FAIR (1)	
Presents alternatives when uncertainties, complexity, or low probability/ high impact situations warrant their inclusion.	System presents alternative possibilities for derived results when uncertainties, system complexity, noise, lack of data, mismatched data, etc. warrant their inclusion. *The NLP system demonstrates pragmatic uncertainty and provides alternative formulations of the titles and summaries for the user.*
GOOD (2)	
Presents alternatives when uncertainties, complexity, or low probability/high impact situations warrant their inclusion; AND	System presents alternative possibilities for derived results when uncertainties, system complexity, noise, lack of data, mismatched data, etc. warrant their inclusion; AND *The NLP system demonstrates pragmatic uncertainty and provides alternative formulations of the titles and summaries for the user;* AND
Explains the evidence and reasoning that underpin them; AND	System explains the evidence and reasoning that underpin the alternatives; AND *The NLP system explains how each of the alternative titles and summaries were derived, by providing a list of source documents and highlighting the most relevant text;* AND
Discusses their likelihood or implications related to U.S. interests	The system indicates the likelihood that the alternative possibilities of the derived results are correct. *The NLP system indicates which title or summary it believes to be the best and indicates the likelihood that the alternatives are actually correct.*
EXCELLENT (3)	
Satisfies "Good" criteria; AND	
Identifies indicators that would, if detected, affect the likelihood of any identified alternatives	System identifies data that would, if provided, affect the likelihood of any identified alternatives. *The NLP system indicates why it presented the alternative titles and summaries and highlights the data or keywords that, if detected, would cause it to adjust its choice for the most likely result.*

STANDARD 5: Customer Relevance

Demonstrates customer relevance and addresses implications

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Provides little or no information or analysis beyond what is generally known; OR	System provides little or no information or processing of the data beyond what is generally known; OR *The NLP system provides little or no information or processing of the text/content beyond what could easily be determined by the user from a quick visual scan*; OR
Does not respond adequately to a specific tasking	System does not provide an adequate result to a specific query. *The NLP system does not group topics adequately, or provide titles and summaries that represent the grouped topics.*
FAIR (1)	
Provides useful information and analysis but does not address implications; OR	N/A
Does not address an important issue or question raised by the analysis; OR	N/A
Satisfies a specific tasking only partially	System provides useful information and derived results, but only provides a partial answer to user queries. The NLP system satisfies a specific set of tasks on content analysis, partially. (e.g., does not extract ontologies from a requirement specification; does not process sentence segmentation or tokenization in raw text; provides groupings of topics, but does not provide a title or summary).
GOOD (2)	
Provides useful information and analysis and addresses near-term, direct, or first-order implications; AND	N/A
Satisfies a specific tasking fully but is limited towards generalization; AND	System provides useful information from derived results, but the results are limited and general; AND *The NLP system satisfies a specific tasking (e.g., sentence segmentation / entity recognition / relation recognition / automatic term recognition and automatic indexing)* ***fully***; AND
Adds value by addressing at least one of the following: trends or prospects, appropriate content, insight gained from synthesizing a large volume of information, warning of threats to U.S. interests, or factors affecting opportunities for U.S. actions (without prescribing U.S. policy)	System adds value by synthesizing a large volume of information to: identify trends or prospects for additional follow-up work by the user; or classify data to ease analysis by the user. *The NLP system is able to analyze a large number of reports, group them in topics, and provide accurate titles and summaries to assist the user in triaging relevant reporting for their account.*
EXCELLENT (3)	
Satisfies "good" criteria; **AND**	
Satisfies a specific tasking fully; OR	System provides useful information and derived results that are tailored to the needs of the system user; OR *The NLP system provides topics, titles, and summaries in the format required by the user and can provide supplementary information (e.g. sourcing, visuals, etc.) if necessary*; OR
Assesses longer-term, indirect, or second-order implications; OR	N/A
Provides exceptionally expert analysis (e.g., by drawing on multiple disciplines or presenting illuminating comparisons); OR	System adds value by synthesizing a large volume of information across multiple disciplines to draw conclusions about: trends or prospects for additional analysis, connections between entities in the data, or classification of data; OR *The NLP system is able to analyze a large number of reports, group them in topics, and provide accurate titles and summaries to assist the user in triaging relevant reporting for their account. The system can also show trends in the reporting (e.g. using timelines or other visuals), extract entities and show connections between them, or generate written reports based on its analysis of the data*; OR
Warns of threats in detail (e.g., by discussing specific indicators likelihood, or imminence) OR	N/A
Analyzes in detail factors affecting opportunities for U.S. action (e.g., by discussing the risks, benefits, or possible reactions to potential U.S. actions).	System analyzes in detail how the current derived results relate to other information of interest to the user. *The NLP system analyzes in detail how current topics relate to other topics searched by the user in the past.*
Assesses longer-term, indirect, or second-order implications; OR	N/A

STANDARD 6 – Logical Argumentation

Uses clear and logical argumentation

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Lacks a main analytic message OR	System does not provide derived results in a way that is understandable to the user; OR *The NLP system provides a user with a synopsis of derived results, but does not contextualize them in a way that the user can understand*; OR
Does not support analytic judgments with relevant evidence or undermines them with flawed reasoning, unclear language, or with unclear structure.	System misinterprets input data to derive unfounded analytical conclusions or supports derived results with irrelevant information. *The NLP system's input data are either ignored, not processed correctly, or are misrepresented in deriving the titles and summaries. The reports that the system displays as source material are inaccurate.*
FAIR (1)	
Presents a main analytic message; BUT	System provides a derived result that is understandable to the user; BUT *The NLP system uses input data to clearly group the analyzed reporting into topic clusters and derive titles and summaries understandable to the user*; BUT
Does not combine evidence, context, and assumptions effectively to support analytic judgments; uses weak logic; OR uses unclear language or structure that is not easily understood	System is not able to demonstrate how it derived its results. *The NLP system is able to group reports in topics and provide titles and summaries, but it is not able to show how it derived the results .*
GOOD (2)	
Presents a prominent and clear main analytic message; AND	System provides a derived result that is understandable to the user; AND *The NLP system uses input data to clearly group the analyzed reporting into topic clusters and derive titles and summaries understandable to the user*; AND
Presents clear reasoning with no flaws in logical and effectively combines evidence, context, and assumptions to support analytic judgments; AND	System output is clearly based on cited evidence, the system provides information on the logic used to derive the results; AND *The NLP system accurately shows the source documents used to derive the titles and summaries and the logic used to combine information from each source*; AND
Uses clear language and a structure that displays a logical flow appropriate for the presented argument.	Logical approach is consistent and meaningful. System uses appropriate language for intended customers and displays the results in a logical progression that supports its derived results. *The NLP system displays the topics derived from its analysis of the reports and allows the user to progressively drill down to discover the underlying titles and summaries, the reporting used to derive those titles and summaries, and other supplementary information.*
EXCELLENT (3)	
Addresses all of Good criteria; AND	
Addresses inconsistent or contrary information in a way that reconciles it with analytic judgements; OR	System provides a detailed description of contrarian or outlying data and describes why the output favored its derived result over other possibilities; OR *The NLP system highlights contrarian or outlying data in the reporting for the user to consider, but nevertheless makes its final analysis while considering the alternatives. The NLP system outputs at least one alternative and plausible result (e.g. title or summary) that is based on contrarian or outlying data in the reporting*; OR
Demonstrates notable skill or sophistication in combining evidence, context, and assumptions convincingly to support analytic judgments.	System demonstrates sophistication in providing a derived result that clearly combines data, context, and assumptions. Stated evidence and assumptions are relevant to the derived results and system describes methodological framework from which it derives judgments. *The NLP system applies notable skill or sophistication in deriving the titles and summaries, combining multiple reports, multiple reporting types (e.g. press and government), variations in the content, reporting from different viewpoints, etc.*

STANDARD 7 – Consistency

Explains change to or consistency of analytic judgments

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Does not note or explain that a major analytic judgment differs from previous production.	System does not provide a way to note change with previous analysis of the same or similar topics. *The NLP system output does not provide a way to compare titles and summaries on similar topics for searches performed at different times.*
FAIR (1)	
Does not note that a major analytic judgment is consistent with previous production; OR	System does not provide a way to note similarities between current and previously derived results; OR *The NLP system does not provide a way to compare current topics, titles, or summaries with previous results and note similarities*; OR
Notes how a major analytic judgment compares with previous production but does not explain how new information or reasoning supports changing or maintaining an existing analytic line.	System notes change or consistency with previous results, but does not explain why the data leads the system to make this determination. *The NLP system notes similarities or differences with titles and summaries derived from a previous search, but does not explain why the system determined the titles and summaries to be similar or different.*
GOOD (2)	
Notes how a major analytic judgment compares with previous production; AND	System provides a way to note how new results are similar or different from previously derived results; AND *The NLP system provides a "similarity score" between current and previous topics, titles and summaries on similar searches and notes any semantic shifts that take place*; AND
Explains how new information or reasoning supports changing or maintaining an existing analytic line.	Explains how new data that pertains to previous results leads the system to make different or consistent conclusions. *The NLP system explains in simple terms how the relevancy "similarity score" was derived and how new data has caused the system to maintain or change past derived results.*
EXCELLENT (3)	
Satisfies Good criteria; AND	
Highlights and explains how a major analytic judgment compares with judgments on the topic held within the US intelligence community, not just within the same analytic element.	System compares current results with similar results from outside systems or agencies on similar topics. *The NLP system compares results to outputs from similar systems at other agencies and describes similarities and differences to end users.*

STANDARD 8 – Accuracy

Makes accurate judgments and assessments.

First Phase: Determine whether judgments and assessments are stated clearly.

Clarity of Judgement	AI/NLP Evaluation Score Rationale
UNCLEAR JUDGMENT	
Judgment or assessment is not clearly expressed	The system's derived results are not clearly expressed and understood by the user. *The NLP system derives a list of major topics and creates titles but presents them in a way that is confusing or not actionable for a user.*
CONDITIONED JUDGMENT	
Judgment or assessment is expressed clearly and is conditioned (for example, by an "if/then" statement.)	The system's derived results are clearly expressed and understood by the user. The system explains condition-based results (e.g. based on domain relevancy). *The NLP system provides a derived result that factors in timing, likelihood, and other factors supporting a conclusion around the groupings of topics and creation of titles. For each grouping and title, the system also offers an "if/then" statement that clearly links specific conditions to components of the judgment contained within the NLP system's output.*
UNCONDITIONED JUDGMENT	
Judgment or assessment is expressed clearly and is not conditioned.	The system's derived results are clearly expressed and understood by the user and there are no other conditions. *The NLP system provides a derived result that factors in timing, likelihood, and other conditional factors supporting an actionable conclusion around the grouping of topics and the creation of titles. There are no ill-defined conditions offered by the NLP system and the human analyst is comfortable that such conditions, if they existed, would be expressed by the NLP system.*

Second Phase: Determine accuracy of judgments and assessments.

INACCURATE	
Judgment or assessment is found to be not accurate.	The user or developer determines that the system's results are not accurate. *The NLP system's grouping of topics and the creation of titles and summaries are not accurate.*
INDETERMINATE	
Accuracy could not be determined for any of several reasons. (1) Nature of judgment or assessment prevented assessment of accuracy. (2) Data needed to assess accuracy were not found. (3) Irreconcilable views among evaluators prevented agreement on a final assessment of accuracy.	The system fails to provide any assurances (e.g., % of confidence that the judgment is correct) or provide an assurance of accuracy for low confidence levels (e.g., 50% confidence level). In providing this scoring, the system cannot clearly identify the specific reason underlying a middle confidence level. *The human analyst cannot determine whether the NLP system's group of topics, titles, or summaries are accurate.*
ACCURATE	
Judgment or assessment is found to be accurate, complete, and consistent.	The system provides some assurance of accuracy (e.g., % of confidence that the judgment is correct) that is deemed sufficiently high. The user or developer is also able to independently determine that the system's results are accurate, complete, and consistent. *The human analyst is able to choose a select group of categories, titles, or summaries identified by the NLP system and test the derived results for accuracy.*

STANDARD 9 - Visualization

Incorporates effective visual information where appropriate

Original Criteria	Modified Criteria for Evaluating AI / NLP Application
POOR (0)	
Does not use visual information when its use would have substantially complemented or enhanced an analytic product; OR	The system has no capability of producing visual information. Alternatively, the system does have this capability, but the system subsequently fails to include visual information (e.g., a map to represent information linked to geography) in its output, when visual information would have been appropriate; OR *The NLP system cannot produce any visual, such as a chart, to support its grouping of topics and creation of titles. Alternatively, the NLP system can produce visuals but fails to do so for this output and provides no explanation of such failure*; OR
Presents visual information that is inconsistent with or not pertinent to the text of an analytic product, or that confuses or detracts from an analytic product's presentation.	The system has a capability of producing visual information, but the analyst is not able to direct or otherwise understand the types of visuals that can be produced. When the system does produce visual information, it is either a confusing visual given the message being conveyed to the customer or it does not otherwise pertain to the derived results. *The NLP system produces a visual, such as a regional map, that is no way pertinent to the grouping of select topics and creation of titles in the final output.*
FAIR (1)	
Presents visual information that is pertinent to an analytic product; BUT	The system can produce visual information and produces a visual that helps the customer better understand key concepts, such as a graph or scatterplot; BUT *The NLP system produces a visual that shows links between the items grouped under a particular topic*; BUT
Does little to clarify (illustrate, summarize, or provide greater detail), complement, or augment data or analytic points in an effective manner that enhances customer understanding of a product's analytics.	The system does not sufficiently define or explain the data points within the visual, such as the horizontal axis on a bar chart or the weight given to certain percentages within a pie chart. Further, the customer does not learn anything new from the visual produced by the system. *The NLP system does not visually demonstrate the quantity or quality of links between items grouped under a particular topic.*
GOOD (2)	
Presents visual information that is pertinent to an analytic product; AND	The system can produce visual information and produces a visual that helps the customer better understand key concepts, such as a graph or scatterplot; AND *The NLP system produces a visualization that shows links between the items grouped under a particular topic*; AND
Clarifies (illustrates, summarizes, or provides greater detail), complements, or augments data or analytic points in an effective manner that enhances customer understanding of a product's analytics.	The system highlights data in an effective manner that clarifies, complements, or augments derived results and enhances user understanding. *The NLP system produces a visualization that also highlights in an effective way the number and strength of links between categories grouped under a title.*
EXCELLENT (3)	
Satisfies Good criteria; AND	
Takes particularly effective advantage of visual presentational methods to convey data or analytic points in a way that enhances product's value by making complex issues more understandable, adding insight or perspective, increasing knowledge retention, or highlighting trends, drivers, or indicators.	The system is capable of choosing the most effective visual by factoring in the human analyst's goal(s) for the product and the data available to the system. The system can identify macro and micro trends and display those in a way that a customer can easily digest, along with a visual representation of why such trends are significant. The visual generated by the system does not require extensive references to addendums or other text sources. *The NLP system visually represents the quantity and quality of links between categories that have been grouped together and does not require the customer to sift through a multitude of definitions and other text sections to decipher any one axis.*

2019
PUBLIC-PRIVATE
ANALYTIC EXCHANGE PROGRAM