HIATUS

HUMAN INTERPRETABLE ATTRIBUTION OF TEXT USING UNDERLYING STRUCTURE

INTELLIGENCE VALUE

The HIATUS program aims to address various Intelligence Community (IC) and Law Enforcement (LE) needs such as combating sophisticated malicious information campaigns online, addressing counterintelligence risks, and fighting human trafficking and online fraud. It does so by developing novel Artificial Intelligence (AI) systems for attributing document authorship **(Who is the author of this text?)** in an explainable way **(What is the evidence of authorship?)**, detecting AI-generated text **(Was this document written by human or machine?)** and protecting author privacy **(How do I ensure that my writing is not attributable?)**.

Humans and machines produce vast amounts of text content every day. Text contains linguistic features that can reveal author identity. To support and protect the IC mission, the HIATUS program's objective is to develop multi-language-capable tools to attribute authorship and protect author privacy. These tools must implement novel explainable Artificial Intelligence techniques to provide trustworthy and verifiable results to human users regardless of author background or document genre, topic, and length.

The HIATUS program casts authorship attribution and privacy as different aspects of the same underlying challenge: understanding author-level linguistic variation by elucidating stable identifiers of individual authors across diverse types of text. Performers' authorship attribution and privacy systems compete with one another pushing teams to create higher fidelity author fingerprints. Attribution systems are evaluated on ability to match items by the same author in large collections of diverse text, while privacy systems are evaluated on ability to thwart attribution systems. System explainability is evaluated using a sophisticated Human-In-The-Loop protocol.

The program began in October 2022 and has a duration of 45 months. Performers have significantly exceeded the state of the art (SOTA) on author attribution in English and Russian. The effort is now broadening to include more languages, shorter texts and multi-author documents. HIATUS is transitioning its technology to several partner agencies. **Currently available deliverables** include containerized software systems for Author Attribution, Author Privacy and Machine-Generated Text Detection; annotated development/test datasets; and auxiliary tools.

PRIME PERFORMERS

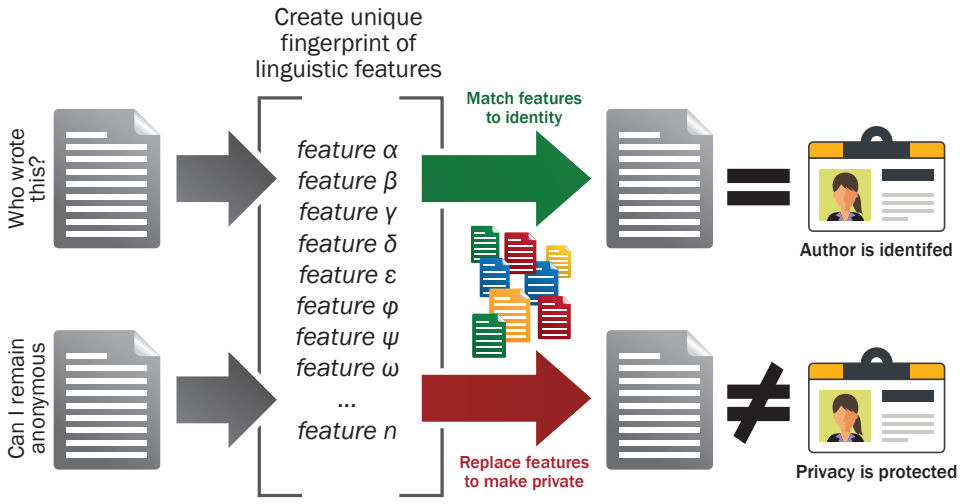
- Raytheon BBN
- University of Southern California, Information Sciences Institute

TESTING AND EVALUATION PARTNERS

- Lawrence Livermore National Labs (U.S. Department of Energy)
- Pacific Northwest National Labs (U.S. Department of Energy)
- University of Maryland's Applied Research Laboratory for Intelligence and Security

KEYWORDS

- Adversarial Machine Learning
- Authorship Attribution
- Explainable AI
- Forensic Linguistics
- Human Language Technology
- Machine Generated Text Detection
- Privacy



The HIATUS vision: A combined authorship attribution and privacy system that can be trusted and audited by human operators

PROGRAM MANAGER
Perry Sherouse Ph.D.
Phone: (301) 243-2029
Perry.Sherouse@iarpa.gov

HIATUS

www.iarpa.gov/research-programs/hiatus