SAND2018-3166C

**Title:**
Protecting Sensitive Textual Information Using Information Extraction and Semantic Technologies

**Abstract:**
Protecting sensitive information is of vital importance to the business, legal, and national security sectors. Previous approaches have focused on the use of text classification and information retrieval techniques to identify sensitive information in text. While fast and able to identify a body of text as sensitive, these techniques are often black box methods and therefore lack the ability to provide users with insights into why the text was deemed sensitive. This may be unacceptable in mission critical situations where organizations need to prohibit the inadvertent release of information such as intellectual property, attorney-client privilege information, or state secrets. This paper outlines an approach leveraging domain-specific information extraction to instantiate an ontological representation of input text that is then enriched by mapping it to a domain-specific knowledge base. SPARQL queries are then used to reason on the sensitivity of the original input text. This process is not only capable of suggesting whether a textual document is sensitive but it also provides the requisite information needed to enable quick human verification of the output.