Data Management Plan

We expect this project to generate or collect multiple types of data: source code, text data and associated metadata, derived classifier parameters and classification results, and logged usage data from evaluations. Our plan is to make the source code publicly available to other researchers and practitioners. Text and derived data will be shared only when legally and ethically permissible, in accordance with copyright law, the wishes of third-party data providers and IRB policies. For the purposes of public demonstration, we will use text data available in the public domain.

Data Preservation: All project data and code will be stored securely using facilities provided by the University of Washington's Computer Science & Engineering department. The primary backup system is disk based and utilizes the ZFS file system. The system performs snapshots of data partitions and efficient replication of data to an offsite location on a dedicated 10 Gigabit link. In addition, the department uses the UWIT/eScience hosted tape archive service for catastrophic disaster recovery.

Data Security: All sensitive data – including proprietary text and experimental data – will be kept in centralized storage that is securely maintained by professional staff. Collected experimental data will be anonymized, and stored stripped of personally identifying information.

Data Coordination: Project code and test data will be accessed through Git repositories to facilitate coordination between project members.

Data Dissemination: All code developed as part of this project will be publicly distributed under the terms of a BSD Open Source license. Any scholarly publications for this project will likewise be made publicly available free of charge. Primary distribution will be through the Interactive Data Lab website (`http://idl.cs.washington.edu`) and via GitHub.com, a popular service for open source collaboration. As applicable, we will also place project data into repositories that are maintained by various publishers. For example, papers published in ACM venues will include project data and code as supplemental materials to be archived in the ACM Digital Library.