8. Data Management Plan

8.1. Expected Data

In the course of the project, we expect to collect or generate three types of data. The first type is social media text data, which will be downloaded either from public social media platforms or from private sources through a paid subscription or with the owner's permission. We will keep a copy of all of the collected social media text data on our project server so that we can reuse them for further analysis. The second type is the software of our tools for generating evolution models, for analyzing dissemination patterns, and for parsing and identifying texts that have evolved from the same text but have linguistic or structural variations. This type of data includes source code, binary executable, design documents and tutorials. The third type are the evolution models for identifying social media texts belonging to the same discourse/dialogue and constructing their temporal evolutionary paths. Variations of the evolution models with annotations of relevant events, such as whether they are deemed "real news" or "fake news," which could inspire particular transformations of their grammatical structure, lexical forms, or semantic content, will also be saved for verification purpose.

8.2. Period of Data Retention

All relevant data will be retained on our data storage server permanently at the University of South Carolina. As new texts are added to the dataset over time, previous texts will still be needed for the computation and refinement of the evolution models.

8.3. Data Formats and Dissemination

All data will be stored electronically. The archived social media texts as well as state files will be converted into an XML file format in both text and binary forms and stored on our server. The source code of the crawling, parsing, and evolution model construction tools will be stored in the version control system (CVS) installed on the server. The binary format will be stored on the server as well. Moreover, the source code and binary format will be posted on open source repositories such as http://sourceforge.net. All tutorials and documents will be in PDF and HTML format and will be available on our project website. The sources of our collected social media text data will be clearly stated in our publications as well as in our software.

Users will be able to download social media text data obtained from public sources and to redistribute them to other parties without restriction. Since we use public sources to build our tools and social media discourse evolution model, we believe there should not be any restrictions on distributing our model. However, if a user wants to access restricted data (such as social media text data obtained from private sources), then the redistribution of such data and generated results will be subject to restrictions imposed by the owner of restricted source. We will put this requirement as disclaimers on our website. In the disclaimer, we will also require the users to cite our Evolutionary Analysis of Social media Texts (EAST) project if they publish results generated

from the data hosted by our project. We will also acknowledge all of the sources, including financial support that made the development of system and tools possible.

Importantly, in order to provide timely access to both the research community and the general public, we will set up an Evolutionary Analysis of Social media Texts (EAST) website to interface users and the data storage server. Although it may take a substantial amount of time to establish a meaningful evolution model of any given event's social media discourse, we will update our website on a weekly basis such that public users can access most recently archived related texts and the latest version of evolution model constructed by our tool. Moreover, we will make the website interactive by designating specific well-known events as its focus and soliciting contributions of related social media texts from general public users. Other data, including social media discourse evolution models and source code of developed tools, documents and tutorials will also be available for free on our website. In subsequent phases, we will expand the scale and construct a public cloud to house all the data.

8.4. Data Storage and Preservation of Access

We have budgeted for one workstation and two 10TB external drives, which will be dedicated for the storage and maintenance of collected and generated data. Moreover, the IT team at University of South Carolina will provide enterprise backup solutions to perform daily backup for all of the data on the servers we use, and the backup will be stored in separate locations that are fire and waterproof. We will require all of the project participants to check in their documents and source code to CVS at the end of the day so new developments can be saved and backed up. The backed-up data are typically stored for up to 30 days, which should be sufficient to recover the data in the event of a data loss.