My Desktop
Prepare & Submit Proposals
Prepare Proposals in FastLane
New! Prepare Proposals (Limited proposal types)
Proposal Status
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1619028 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1619028
Project Title:	III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR)
PD/PI Name:	Edward A Fox, Principal Investigator Andrea L Kavanaugh, Co-Principal Investigator Chandan K Reddy, Co-Principal Investigator Donald J Shoemaker, Co-Principal Investigator
Recipient Organization:	Virginia Polytechnic Institute and State University
Project/Grant Period:	01/01/2017 - 12/31/2019
Reporting Period:	01/01/2018 - 12/31/2018
Submitting Official (if other than PD\PI):	Edward A Fox Principal Investigator
Submission Date:	01/27/2019
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Edward A Fox

Accomplishments

* What are the major goals of the project?

We will ingest tweets and Web-based content from social media and the general Web, including news and governmental information. In addition to archiving materials found, we will build an information system that includes related metadata and knowledge bases, consistent with the 5S (Societies, Scenarios, Spaces, Structures, Streams) framework, along with results from our intelligent focused crawler, to support comprehensive access to event related content. With the support of key partners, the GETAR team will undertake research, education, and dissemination efforts, to achieve three complementary objectives:
1. Collecting: We will spot, identify, and make sense of interesting events and trends. We also will accept specific or general requests about types of events. Given resource and sampling constraints, we will integrate methods to identify appropriate URLs as seeds, and specify when to start crawling and when to stop, with regard to each event or subevent. We will integrate focused crawling and filtering approaches in order to ingest content and generate new collections, with high precision and recall.
2. Archiving & Accessing: Permanent archiving, and access to those archives, will be ensured by our partner, Internet Archive (IA). Immediate access to ingested content will be facilitated through big data software built on top of our Hadoop cluster.
3. Analyzing & Visualizing: We will provide a wide range of integrated services beyond the usual (faceted) browsing and searching, including: classification, clustering, summarization, text mining, topic identification, trend analysis, and visualization.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

The GETAR project, with more than 50 collaborators and 12 collaborating institutions, developed tweet and webpage collections, datasets, services, software, systems, and methods. The related activities include: collecting event-related content, software and system development and refinement, experimentation, evaluation, and working with diverse users (representing key stakeholder groups). In June 2018, PI Fox met with Internet Archive PI Bailey at the Internet Archive, to report on progress, synchronize activities, and plan further collaboration.

The Internet Archive (see also its separate report for IIS-1619371) has expanded its collections and technology support, as well as outreach activities. It has expanded its R&D staffing. It hosts, preserves, and provides public access with attribution to web collections created by the project team through its public Wayback Machine interface and Archive-It service. The latter may be browsed by descriptive metadata and searched through Archive-It's full-text Elasticsearch engine. New or updated Internet Archive and Archive-It API documentation and workshops provide project stakeholders with several means to query the data from and about these collections, and to derive datasets for further textual and visual analyses. Studies have proceeded of important events, including integration of survey and analysis approaches, and publishing findings.

Specific Objectives:

Doctoral dissertation research led to improved methods for focused crawling and for assigning locations to tweets. More than 22 computers are connected, mostly in a Hadoop cluster. This network was constructed to support collection, processing, and access already of over 3.5 billion tweets across over 1600 collections, along with millions of webpages, covering hundreds of important events. Regarding collections, prior collections were extended, new ones were launched as events occurred or requests were made by users, the event focused crawler was deployed, and diverse related curation efforts proceeded. Master's thesis and class (independent study, undergraduate. and graduate courses) research led to improved tweet and webpage techniques for content cleaning and processing, information extraction, classification, clustering, topic analysis, sentiment analysis, indexing, searching, browsing, and visualization.

Significant Results:

Advances have been made in big data handling, computational linguistics, digital libraries, information retrieval, information visualization, machine learning, and Web archiving. These have been integrated into a large system built around a Hadoop cluster, that works with growing numbers of expanding collections of tweets and webpages, supplemented by cleaning, information extraction, and adding value through advanced analysis.

The GETAR project has developed novel methodology and workflows, tailored to addressing the challenging problem of working with events and trends. At a high level, for collection building, is a workflow to collect tweets about each event or event class, extract URLs, use the URLs present therein as seeds to our event focused crawler, and add resulting webpages to our Web collection. The event focused crawler workflow uses the extracted URLs as seeds to construct an event model that guides the selection and focused crawling for webpages. A new method to detect events, with a more elaborate event model, was devised using both tweets and news. Key new methods were developed to analyze and accordingly add value (and metadata) to the collected content. Regarding our processing of tweets, a new framework was extended to streamline a variety of tweet analysis and transformation workflows.

Regarding building tweet classifiers for the hundreds of events studied, a learning optimizer method employing iterative processing with minimal human effort yielded high quality classification of tweets into collections for particular real world events. Regarding the problem that few tweets have associated latitude and longitude values, our methodology for associating locations with tweets based on location indicative words was leveraged, and collaboration launched with a team in Belgium to further refine the methods. This work was applied to help a sister project funded through NSF's CRISP program regarding recent hurricanes.

Key outcomes or Other achievements:

A number of new methods were developed for check-in time prediction, short text topic modeling, summarizing reviews, simulations of disease dynamics, and reciprocal link creation. A variety of methods were evaluated with 11 events to automatically construct summaries from large collections of related webpages.

Collection building and analysis (of both tweets and webpages) has improved through advances in classification, big data workflows, focused crawling (to identify webpages focused on an event of interest), inferring the location of tweets from their text when GPS data is unavailable, topic analysis, and natural language processing (including Arabic). Insights gained have been shared regarding juvenile delinquency, school shootings, and the use of information during conflicts, crises, elections, and uprisings. Collections are available to support other research and exploration regarding important events since 2007 such as the above, as well as attacks, bombings, celebrations, climate change, collapses, community activities, crashes, disease outbreaks, earthquakes, eclipses, environmental disruptions, erosion, explosions, fires, floods, hurricanes, innovations, judicial decisions, pollution, power outages, protests, revolutions, shootings, sports, storms, summits, tornadoes, transportation failures, tsunamis, typhoons, and veteran activities. Collaboration has expanded to support a broad set of researchers.

*** What opportunities for training and professional development has the project provided?**

In the Fall 2017 class CS5604 (Information Retrieval, IR), the class-wide term project, carried out by students working in teams (each uploading deliverables into the local institutional repository), was in support of GETAR. Through project based learning they applied IR theory and methods, using our Hadoop cluster, to ingest, analyze, index, and visualize event-related tweets and webpages. In Spring 2018, six teams in CS4624 (Multimedia, Hypertext, and Information Access) worked on projects related to GETAR, also uploading deliverables (e.g., reports, presentations, data, code). In June 2018, an MS thesis was completed; Abigail Bartolome then moved to undertake doctoral research related to NLP at Dartmouth. In Fall 2018, eleven teams in CS4984/CS5984 (Big Data Text Summarization) developed collection level summarization methods, learning NLP, IR, ML, and deep learning.

*** How have the results been disseminated to communities of interest?**

Dissemination has been through the reported publications and presentations. Further dissemination was through the project website (http://eventsarchive.org) and the website connected to the tweet collections and descriptions (http://hadoop.dlib.vt.edu/). In addition, we helped lead the 2018 Web Archiving and Digital Libraries (WADL) workshop.

*** What do you plan to do during the next reporting period to accomplish the goals?**

We are proceeding with collection and data analysis activities, using software and other results from the team and related class efforts. In spring 2019 there will be some term projects in CS4624 (Multimedia, Hypertext, and Information Access), as well as several volunteer student efforts, to further extend our efforts. Several student efforts will help to consolidate and extend results from our prior research. Then in fall 2019 the CS4984/CS5984 class will again focus on team term projects related to collection level summarization with GETAR data. Three Ph.D. students who have been involved in GETAR should move to the next stage in their doctoral work, completing their prelims.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Abigail Bartolome, D. Scott McCrickard, and Edward A. Fox (2018). Exploring cultural differences in the triple crown trails. *ACM GROUP 2018 workshop on Technology on the Trail., Sanibel Island, FL USA, January 2018*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Abigail Bartolome, Edward Fox and Scott McCrickard (2018). Understanding Trail Cultures through Various Stakeholders of the Trail. *Proc. "W12: HCI Outdoors: Understanding Human-Computer Interaction in the Outdoors" Workshop at CHI 2018, 21 April 2018, Montreal, Canada*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Aman Ahuja, Ashish Baghudana, Wei Lu, Edward A. Fox, and Chandan K. Reddy (2019). A Probabilistic Spatio-Temporal Model for Event Detection. *Proc. 23rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2019), 14-17 April 2019, Macau, China*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Andrea Kavanaugh, Steven D. Sheetz, Hamida Skandrani, and Edward A. Fox (2017). Media use by young Tunisians during the 2011 revolution vs 2014 elections. *Information Polity*. 22 (2-3), 137. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/IP-170412

Andrea L. Kavanaugh, Rodrigo Sandoval-Almazan, and David Valle-Cruz (2018). The Diffusion of Social Media Among State Governments in Mexico. *Int. J. Public Adm. Digit. Age*. 5 (1), 63. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.4018/IJPADA.2018010104

Dave, V. S., Hasan, M. A., Zhang, B., & Reddy, C. K. (2018). Predicting interval time for reciprocal link creation using survival analysis. *Social Network Analysis and Mining*. 8 (1), 16. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/s13278-018-0494-1

Donald J. Shoemaker, Jason Callahan, Liuqing Li, Ziqian Song, and Edward Fox (2018). Visual Comparisons of Tweets and Urls for Ten School Shootings. *Annual meeting of the American Society of Criminology, Atlanta, Georgia, November 15, 2018*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Edward Fox (2018). Introduction to Digital Libraries. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18). ACM, New York, NY, USA, http://fox.cs.vt.edu/talks/2018/20180603FoxTutorialSlidesJCDL.pptx*. 415. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3197026.3201779

Florian J. Zach, Yufeng Ma, and Edward A. Fox (2019). A Preliminary Analysis of Images in Online Hotel Reviews. *ENTER 2019: The 26th Annual eTourism Conference, Nicosia, Cyprus, 30 January - 1 February, 2019, to appear in Review of Tourism Research (eRTR), Springer journal special issue*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Guolei Yang, Ying Cai, and Chandan K. Reddy (2018). Recurrent Spatio-Temporal Point Process for Check-in Time Prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18). ACM, New York, NY, USA*. 2203. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3269206.3272003

Guolei Yang, Ying Cai, and Chandan K. Reddy (2018). Spatio-temporal check-in time prediction with recurrent neural network based survival analysis. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18), Jérôme Lang (Ed.). AAAI Press*. 2976. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.24963/ijcai.2018/413

Kavanaugh, Andrea L and Song, Ziqian (2018). Engaging a community through social media-based topics and interactions. *First Monday*. 23 (4), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: https://doi.org/10.5210/fm.v23i4.8146

Li, Liuqing; Fox, Edward A. (2019). Understanding patterns and mood changes through tweets about disasters. *Proc. ISCRAM 2019, 16th International Conference on Information Systems for Crisis Response and Management*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Liuqing Li and Edward Fox (2018). A Study of Historical Short URLs in Event Collections of Tweets. *Proc. WADL 2018, Web Archiving and Digital Libraries Workshop at JCDL 2018, Fort Worth, TX, USA WADL 2018*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Martin Klein, Zhiwu Xie, and Edward A. Fox (2018). Web Archiving and Digital Libraries (WADL). *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18). ACM, New York, NY, USA*. 425. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3197026.3200209

Rodrigo Sandoval and Andrea Kavanaugh (2018). Social media and government. *First Monday*. 23 (4), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: https://firstmonday.org/ojs/index.php/fm/issue/view/592

Sangho Suh, Sungbok Shin, Joonseok Lee, Chandan K. Reddy, and Jaegul Choo (2018). Localized user-driven topic discovery via boosted ensemble of nonnegative matrix factorization. *Knowl. Inf. Syst.*. 56 (3), 503. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/s10115-017-1147-9

Sheetz, Steven; Kavanaugh, Andrea; Fox, Edward; Hassan, Riham; Yang, Seungwon; Magdy, Mohamed; Donald, Shoemaker (2019). Information Uses and Gratifications Related to Crisis: Student Perceptions since the Egyptian Uprising. *Proc. ISCRAM 2019, 16th International Conference on Information Systems for Crisis Response and Management*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy (2018). Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. *Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland*. 1105. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3178876.3186009

Ting Hua, Chandan K. Reddy, Lei Zhang, Lijing Wang, Liang Zhao, Chang Tien Lu, and Naren Ramakrishnan (2018). Social media based simulation models for understanding disease dynamics. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18), Jérôme Lang (Ed.). AAAI Press*. 3797. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.24963/ijcai.2018/528

Vineeth Rakesh, Weicong Ding, Aman Ahuja, Nikhil Rao, Yifan Sun, and Chandan K. Reddy (2018). A Sparse Topic Model for Extracting Aspect-Specific Summaries from Online Reviews. *Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland*. 1573. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3178876.3186069

Licenses

Other Conference Presentations / Papers

Edward A. Fox (2018). *Applications of Big Data Analysis for the Worldwide Collection of ETDs*. Invited keynote for ETD 2018, Sept. 26-28, 2018, http://fox.cs.vt.edu/talks/2018/20180926FoxETD2018keynote.pptx. Taiwan. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Liuqing Li, Ziqian Song, Xuan Zhang, Edward A. Fox (2018). *A Hybrid Model for Role-related User Classification on Twitter*. arXiv.org, arXiv:1811.10202 [cs.SI]. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Kingery, Ryan; Yellapantula, Sudha Ravali; Xu, Chao; Huang, Li Jun; Ye, Jiacheng (2018). *Abstractive Text Summarization of the Parkland Shooting Collection*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-12; http://hdl.handle.net/10919/86370. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Wanye, Frank; Ganguli, Samit; Tuckman, Matt; Zhang, Joy; Zhang, Fangzheng (2018). *Automatic Summarization of News Articles about Hurricane Florence*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-07; http://hdl.handle.net/10919/86399. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Gallagher, Colm; Dyer, Jamie; Liebold, Jeanine; Becker, Aaron; Yang, Limin (2018). *Big Data Text Summarization - Attack Westminster*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-14; http://hdl.handle.net/10919/86419. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Geissinger, Jack; Long, Theo; Jung, James; Parent, Jordan; Rizzo, Robert (2018). *Big Data Text Summarization - Hurricane Harvey*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-12; http://hdl.handle.net/10919/86358. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Arora, Anuj; Miller, Chreston; Fan, Jixiang; Liu, Shuai; Han, Yi (2018). *Big Data Text Summarization for the NeverAgain Movement*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-10; http://hdl.handle.net/10919/86357. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Big Data and Machine Learning in Virginia Tech’s Digital Library Research Laboratory*. STEP Faculty Seminar, 6 July 2018, Virginia Tech, http://fox.cs.vt.edu/talks/2018/20180706FoxSTEP.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Big Data and Machine Learning in Virginia Tech’s Digital Library Research Laboratory*. Yahoo! Research, Sunnyvale, CA, 21 June 2018, http://fox.cs.vt.edu/talks/2018/20180621FoxYahoo.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Eason, Andrew D.; Cianfarini, Kevin M.; Hansen, Marshall C.; Davies, Shane J. (2018). *Blog and Forum Collection for Trail Study*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech, 2018-05-07; http://hdl.handle.net/10919/83214. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Baghudana, Ashish; Li, Guangchen; Liu, Beichen; Lasky, Stephen (2018). *CS4984/CS5984: Big Data Text Summarization Team 10 ETDs*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-14; http://hdl.handle.net/10919/86418. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Khaghani, Farnaz; Marin Thomas, Ashin; Patnayak, Chinmaya; Sharma, Dhruv; Aromando, John (2018). *CS4984/CS5984: Big Data Text Summarization Team 17 ETDs*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-15; http://hdl.handle.net/10919/86420. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ward, Ryan; Lee, Jun; Beard, Stuart; Edwards, Skylar; Su, Spencer (2018). *Event Trend Detector*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech,2018-05-07; http://hdl.handle.net/10919/83205. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Hamilton, Leah; Robb, Esther; Fitzpatrick, April; Goel, Akshay; Nandigam, Ramya (2018). *Generating Text Summaries for the Facebook Data Breach with Prototyping on the 2017 Solar Eclipse*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-13; http://hdl.handle.net/10919/86395. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Goldsworthy, Michael; Tran, Thoang; Asif, Areeb; Gregos, Brendan (2018). *Hurricane Matthew Summarization*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-14; http://hdl.handle.net/10919/86408. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Chen, Xiaoyu; Wang, Haitao; Mehrotra, Maanav; Chhikara, Naman; Sun, Di (2018). *Hybrid Summarization of Dakota Access Pipeline Protests (NoDAPL)*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-14; http://hdl.handle.net/10919/86401. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Information - Friendly*. Hypatia Seminar, 18 September 2018, Virginia Tech, http://fox.cs.vt.edu/talks/2018/20180918HypatiaFox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Information Research*. ENGR 1014: Engineering Research Seminar, 5 October 2018, Virginia Tech, http://fox.cs.vt.edu/talks/2018/20181005ENGR1014Fox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Information Research*. Galileo Seminar, 12 September 2018, Virginia Tech, http://fox.cs.vt.edu/talks/2018/20180913GalileoFox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Edward A. Fox (2018). *Integrating Research and Education Regarding Global Event and Trend Archiving*. Computer Science Colloquium, Old Dominion University, Norfolk, VA, 18 January 2019, http://fox.cs.vt.edu/talks/2019/20190118ODUseminarFox.pptx. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Khawas, Prapti; Banerjee, Bipasha; Zhao, Shuqi; Fan, Yiyang; Kim, Yoonjin (2018). *Summarization of Maryland Shooting Collection*. CS4984/CS5984: Big Data Text Summarization; Virginia Tech, 2018-12-12; http://hdl.handle.net/10919/86407. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Chenault, Kirk P.; Keener, Chris L.; Chang, Brandon P.; Widrig, Joseph (2018). *Tweet Collections*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech,2018-05-07; http://hdl.handle.net/10919/83211. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Liyan; Lyu, Kehan; Sun, Guoxin (2018). *Tweet URL Analysis*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech,2018-05-02; http://hdl.handle.net/10919/83219. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bridges, Chris; Chun, David; Tat, Carter (2018). *Tweet URL Extraction Crawling*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech,2018-05-02; http://hdl.handle.net/10919/83215. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Woodson, Tianna; Simmons, Gabriel; Park, Peter; Doan, Tomy; Keys, Evan (2018). *Visual Displays of School Shooting Data*. CS4624: Multimedia, Hypertext, and Information Access; Virginia Tech,2018-05-02; http://hdl.handle.net/10919/83216. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations

Bock, Matthew. *A Framework for Hadoop Based Digital Libraries of Tweets*. (2017). MS Thesis, Virginia Tech, 2017-07-17, http://hdl.handle.net/. Acknowledgement of Federal Support = Yes

Chakravarty, Saurabh. *A Large Collection Learning Optimizer Framework*. (2017). MS Thesis, Virginia Tech, 2017-06-30, http://hdl.handle.net/. Acknowledgement of Federal Support = Yes

Abigail Bartolome. *Describing Trail Cultures through Studying Trail Stakeholders and Analyzing their Tweets, http://hdl.handle.net/10919/84528*. (2018). Virginia Tech. Acknowledgement of Federal Support = Yes

Lee, Sunshin. *Geo-Locating Tweets with Latent Location Information*. (2017). Doctoral Dissertation, Virginia Tech, 2017-02-13, http://hdl. Acknowledgement of Federal Support = Yes

Websites

Events Archiving
http://eventsarchive.org

Homepage for GETAR, as well as related prior NSF-funded projects including CTRnet and IDEAL.

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Fox, Edward	PD/PI	1
Kavanaugh, Andrea	Co PD/PI	1
Reddy, Chandan	Co PD/PI	1
Shoemaker, Donald	Co PD/PI	1
Bailey, Jefferson	Co-Investigator	1
Agozino, Onwubiko	Faculty	0
Angermeier, Paul	Faculty	0
Coleman, Shane	Faculty	0
Deligiannis, Nikolaos	Faculty	0
Elmongui, Hicham	Faculty	0
Farag, Mohamed	Faculty	0
Horning, Mike	Faculty	0
Jelesko, John	Faculty	0

Name	Most Senior Project Role	Nearest Person Month Worked
Kanan, Tarek	Faculty	0
Krometis, Leigh	Faculty	0
Lee, Sunshin	Faculty	0
Murray-Tuite, Pamela	Faculty	0
Nesbitt, Sterling	Faculty	0
North, Chris	Faculty	0
Pereira, Denilson	Faculty	0
Salehi-Isfahani, Djavad	Faculty	0
Sandoval-Almazan, Rodrigo	Faculty	0
Sheetz, Steven	Faculty	1
Skandrani, Hamida	Faculty	0
Smith, Eric	Faculty	0
Tedesco, John	Faculty	0
Wimberley, Dale	Faculty	0
Xie, Zhiwu	Faculty	0
Yang, Seungwon	Faculty	0
Zach, Florian	Faculty	0
Moneim, Riham	Other Professional	0

Name	Most Senior Project Role	Nearest Person Month Worked
Holzmann, Helge	Staff Scientist (doctoral level)	0
Klein, Martin	Staff Scientist (doctoral level)	0
Mather, Paul	Staff Scientist (doctoral level)	0
Sforza, Peter	Staff Scientist (doctoral level)	0
Ahuja, Aman	Graduate Student (research assistant)	0
Alazmi, Huda	Graduate Student (research assistant)	0
Bartolome, Abigail	Graduate Student (research assistant)	1
Bock, Matthew	Graduate Student (research assistant)	0
Callahan, Jason	Graduate Student (research assistant)	1
Chakravarty, Saurabh	Graduate Student (research assistant)	0
Chandrasekar, Prashant	Graduate Student (research assistant)	1
Do, Tien	Graduate Student (research assistant)	0
Li, Liuqing	Graduate Student (research assistant)	3
Ma, Yufeng	Graduate Student (research assistant)	0
Malpani, Ashish	Graduate Student (research assistant)	0
Niu, Shuo	Graduate Student (research assistant)	1
Patil, Supritha	Graduate Student (research assistant)	1
Song, Ziqian	Graduate Student (research assistant)	3

Name	Most Senior Project Role	Nearest Person Month Worked
Wang, Xinyue	Graduate Student (research assistant)	0
Zhang, Xuan	Graduate Student (research assistant)	0
Conte, Philip	Undergraduate Student	0
Ganotra, Ayush	Undergraduate Student	0
Tewes, Stephen	Undergraduate Student	0
Hsiao, Bethany	High School Student	1
Karajeh, Ola	Other	1

Full details of individuals who have worked on the project:

Edward A Fox
Email: fox@vt.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: PI

Funding Support: This project

International Collaboration: No
International Travel: No

Andrea L Kavanaugh
Email: kavan@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI. Supervising Bethany Hsiao. Supervising aspects related to digital government, international collaborations, tweet analysis, social science applications, etc.

Funding Support: This project

International Collaboration: Yes, Egypt, Mexico, Saudi Arabia, Tunisia
International Travel: No

Chandan K Reddy
Email: reddy@cs.vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI. Guiding machine learning and deep learning applications, and developing related technologies; applying to related data.

Funding Support: This project

International Collaboration: No
International Travel: No

Donald J Shoemaker
Email: shoemake@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI. Coordinating sociology related aspects. Leading studies of school shootings. Liaising with social scientists.

Funding Support: This project.

International Collaboration: No
International Travel: No

Jefferson Bailey
Email: jefferson@archive.org
Most Senior Project Role: Co-Investigator
Nearest Person Month Worked: 1

Contribution to the Project: Internet Archive is a collaborative partner, also receiving funds on this project from NSF, through IIS-1619371. We use their equipment and services and data, and collaborate on research.

Funding Support: This project, i.e., IIS-1619371

International Collaboration: No
International Travel: No

Onwubiko Agozino

Email: agozino@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Paul Angermeier
Email: biota@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Shane Coleman
Email: shanec4@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Nikolaos Deligiannis
Email: ndeligia@etrovub.be
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding geolocating tweets.

Funding Support: His local funding. He is an assistant professor at the Electronics and Informatics department at Vrije Universiteit Brussel (VUB) and principal investigator in Data Science at the imec institute in Belgium.

International Collaboration: Yes, Belgium
International Travel: No

Hicham Galal Elmongui
Email: elmongui@alexu.edu.eg
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaboration regarding data re Egypt

Funding Support: Local support

International Collaboration: Yes, Egypt
International Travel: No

Mohamed Farag
Email: mohamedmagdy@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding the software he developed as part of his earlier doctoral work supported by a prior NSF project, IDEAL, related to this research, as well as this project.

Funding Support: Local support

International Collaboration: Yes, Egypt
International Travel: No

Mike Horning
Email: mhorning@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

John Jelesko
Email: jelesko@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Tarek Kanan
Email: tarek.kanan@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate regarding NLP and handling of Arabic texts, extending his doctoral work completed earlier at VT

Funding Support: Local support

International Collaboration: Yes, Jordan
International Travel: No

Leigh Anne Krometis
Email: lehenry@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Sunshin Lee
Email: slee116@radford.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Worked as GRA on this project, then as postdoc, now as faculty at Radford, collaborating to extend his doctoral research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Pamela Murray-Tuite
Email: pmmurra@clemson.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Formerly a VT faculty member, now on faculty at Clemson, she is PI on another NSF project in which PI Fox serves at co-PI, and is helping with curation and analysis of data related to disasters that effect both transportation and power systems.

Funding Support: Local support

International Collaboration: No
International Travel: No

Sterling Nesbitt
Email: sjn2104@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Chris North
Email: north@cs.vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on software, collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Denilson Pereira
Email: denilsonpereira@dcc.ufla.br
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaboration on publications and related research connected with analysis of tweets, classification, disambiguation, and text analysis

Funding Support: Local support

International Collaboration: Yes, Brazil
International Travel: No

Djavad Salehi-Isfahani
Email: salehi@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Rodrigo Sandoval-Almazan
Email: rsandovuaem@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding publication, analysis, and collection/curation of data related to events in Mexico.

Funding Support: Local support

International Collaboration: Yes, Mexico
International Travel: No

Steven D. Sheetz
Email: sheetz@vt.edu

Most Senior Project Role: Faculty
Nearest Person Month Worked: 1

Contribution to the Project: Collaborator on IDEAL, prior related project, as co-PI. Continuing collaboration.

Funding Support: Local support

International Collaboration: No
International Travel: No

Hamida Skandrani
Email: hamida.skandrani@gmail.com
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborated on publications and studies related to events in Tunisia and the region.

Funding Support: Local support

International Collaboration: Yes, Tunisia
International Travel: No

Eric Smith
Email: epsmith@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

John Tedesco
Email: tedesco@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Dale Wimberley
Email: wimberly@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Faculty collaborating on collecting, curating, and analyzing data.

Funding Support: Local support

International Collaboration: No
International Travel: No

Zhiwu Xie
Email: zhiwuxie@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Library researcher collaborating on web archiving; co-organizer of Web Archiving and Digital Libraries (WADL) workshop at JCDL.

Funding Support: Local support

International Collaboration: No
International Travel: No

Seungwon Yang
Email: seungwonyang@lsu.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on collecting and analyzing tweets related to events, especially related to the Gulf region. Led effort for MOU between GETAR and his group at LSU.

Funding Support: Local support

International Collaboration: No
International Travel: No

Florian Zach

Email: florian@vt.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating on research related to tourism

Funding Support: Local support

International Collaboration: No
International Travel: No

Riham Abdel Moneim
Email: riham@aucegypt.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Continuing collaboration related to collection, curation, and analysis of data re Egypt, and on technologies like crawling.

Funding Support: Works at Microsoft in Egypt

International Collaboration: Yes, Egypt
International Travel: No

Helge Holzmann
Email: holzmann@L3S.de
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating regarding ArchiveSpark and related web archiving technologies

Funding Support: L3S and then Internet Archive

International Collaboration: Yes, Germany
International Travel: No

Martin Klein
Email: martinklein0815@gmail.com
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on collecting data (e.g., tweets) and undertaking related analysis; co-organizer of Web Archiving and Digital Libraries (WADL) workshop at JCDL.

Funding Support: Local support

International Collaboration: No
International Travel: No

Paul Mather
Email: pmather@vt.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborate on managing our equipment and software, connecting with Library efforts.

Funding Support: Local support

International Collaboration: No
International Travel: No

Peter Sforza
Email: psforza@vt.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: Director of local center on GIS collaborating on spatial location.

Funding Support: Local support

International Collaboration: No
International Travel: No

Aman Ahuja
Email: aahuja@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Huda Alazmi
Email: ahuda1@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Curating and analyzing datasets and collaborating toward publication

Funding Support: Local support

International Collaboration: Yes, Kuwait
International Travel: No

Abigail Bartolome
Email: abijbart@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student completed thesis in collaboration with project.

Funding Support: Local support

International Collaboration: No
International Travel: No

Matthew Bock
Email: mattb93@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student completed thesis in support of this project; provided follow-on assistance.

Funding Support: Local support

International Collaboration: No
International Travel: No

Jason Callahan
Email: jcallaha@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Saurabh Chakravarty
Email: saurabc@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student completed thesis in support of this project; provided follow-on assistance.

Funding Support: Local support

International Collaboration: No
International Travel: No

Prashant Chandrasekar
Email: peecee@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: GRA on project part of the year; GRA on a related project collaborating with this project

Funding Support: This project, a related project

International Collaboration: No
International Travel: No

Tien Do
Email: thdo@etrovub.be
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating about geo-location of tweets

Funding Support: At his university

International Collaboration: Yes, Belgium
International Travel: No

Liuqing Li
Email: liuqing@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 3

Contribution to the Project: GRA supported by this project, working on all aspects.

Funding Support: This project

International Collaboration: No
International Travel: No

Yufeng Ma
Email: yufengma@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating on deep learning research, and on data related to tourism

Funding Support: Local support

International Collaboration: No
International Travel: No

Ashish Malpani
Email: ashish76@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Collaborating on software development and application

Funding Support: Local support

International Collaboration: No
International Travel: No

Shuo Niu
Email: shuoniu@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Supritha Patil
Email: patil93@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Ziqian Song
Email: ziqian@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 3

Contribution to the Project: Graduate student collaborating on project research, funded in part by this project.

Funding Support: Local support, this project

International Collaboration: No
International Travel: No

Xinyue Wang
Email: xw0078@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Xuan Zhang
Email: xuancs@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: Graduate student collaborating on project research.

Funding Support: Local support

International Collaboration: No
International Travel: No

Philip Conte
Email: pconte@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Starting as volunteer, leading up to doing undergraduate research in 2019 to help with project.

Funding Support: Self

International Collaboration: No
International Travel: No

Ayush Ganotra
Email: ayushg@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteer helping with data curation and analysis

Funding Support: Self

International Collaboration: No
International Travel: No

Stephen Tewes
Email: stewes36@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 0

Contribution to the Project: Volunteer to help with project

Funding Support: Self

International Collaboration: No
International Travel: No

Bethany Hsiao
Email: bhsiaoburg@gmail.com
Most Senior Project Role: High School Student
Nearest Person Month Worked: 1

Contribution to the Project: Volunteer helping with project

Funding Support: Self

International Collaboration: No
International Travel: No

Ola Karajeh
Email: Ola Karajeh
Most Senior Project Role: Other
Nearest Person Month Worked: 1

Contribution to the Project: Analyzing data sets, curating and classifying, such as related to heart attacks.

Funding Support: Self (preparing to start as graduate student in CS@VT)

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Al Zaytonah University of Jordan	Academic Institution	Jordan
Arab Academy for Science and Technology	Academic Institution	Alexandria, Egypt
University of Tunis - Manouba Campus	Academic Institution	Tunisia
Vrije Universiteit Brussel (VUB)	Academic Institution	Brussels, Belgium

Name	Type of Partner Organization	Location
Clemson University	Academic Institution	Clemson, SC
George Washington University	Academic Institution	Washington, D.C.
Internet Archive	Other Nonprofits	San Francisco, CA
Los Alamos National Laboratory	State or Local Government	Los Alamos, New Mexico
Louisiana State University	Academic Institution	Baton Rouge, LA
Radford University	Academic Institution	Radford, VA
Universidad Autónoma del Estado de México (UAEM)	Academic Institution	Mexico
Universidade Federal de Lavras (UFLA)	Academic Institution	Lavras, MG, Brasil

Full details of organizations that have been involved as partners:

Al Zaytonah University of Jordan

Organization Type: Academic Institution
Organization Location: Jordan

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Tarek Kanan continues to collaborate regarding NLP and handling of Arabic texts, extending his doctoral work completed earlier at VT.

Arab Academy for Science and Technology

Organization Type: Academic Institution
Organization Location: Alexandria, Egypt

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Mohamed Farag, on the faculty, is collaborating regarding the software he developed as part of his earlier doctoral work supported by a prior NSF project, IDEAL, related to this research, as well as this project.

Clemson University

Organization Type: Academic Institution
Organization Location: Clemson, SC

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Pamela Murray-Tuite, formerly a VT faculty member, now on faculty at Clemson, is PI on another NSF project in which PI Fox serves at co-PI, and is helping with curation and analysis of data related to disasters that effect both transportation and power systems.

George Washington University

Organization Type: Academic Institution
Organization Location: Washington, D.C.

Partner's Contribution to the Project:
In-Kind Support

More Detail on Partner and Contribution: We use the Social Feed Manager software from GWU Libraries, which they continue to support and enhance through collaboration.

Internet Archive

Organization Type: Other Nonprofits
Organization Location: San Francisco, CA

Partner's Contribution to the Project:
Facilities
Collaborative Research

More Detail on Partner and Contribution: Internet Archive is a collaborative partner, also receiving funds on this project from NSF, through IIS-1619371. We use their equipment and services and data, and collaborate on research. Jefferson Bailey is co-PI on GETAR.

Los Alamos National Laboratory

Organization Type: State or Local Government
Organization Location: Los Alamos, New Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We collaborate on collecting data (e.g., tweets) and undertaking related analysis. This work is led by Dr. Martin Klein.

Louisiana State University

Organization Type: Academic Institution
Organization Location: Baton Rouge, LA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We collaborate on collecting and analyzing tweets related to events, especially related to the Gulf region. This is led by Dr. Seungwon Yang, whose doctoral work was supported in part by prior NSF-funded related projects at VT.

Radford University

Organization Type: Academic Institution
Organization Location: Radford, VA

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Sunshin Lee, whose Ph.D. work was supported by this project before he started as a faculty member at Radford, continues to collaborate with our research. He served also as volunteer postdoc aiding GETAR before going to Radford.

Universidad Autónoma del Estado de México (UAEM)

Organization Type: Academic Institution
Organization Location: Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Sandoval Almazan is collaborating regarding publication, analysis, and collection/curation of data related to events in Mexico.

Universidade Federal de Lavras (UFLA)

Organization Type: Academic Institution
Organization Location: Lavras, MG, Brasil

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Dr. Pereira is continuing collaboration on publications and related research connected with analysis of tweets, classification, disambiguation, and text analysis.

University of Tunis - Manouba Campus

Organization Type: Academic Institution
Organization Location: Tunisia

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Hamida Skandrani has collaborated on publications and studies related to events in Tunisia and the region.

Vrije Universiteit Brussel (VUB)

Organization Type: Academic Institution
Organization Location: Brussels, Belgium

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Nikos Deligiannis is an assistant professor at the Electronics and Informatics department at Vrije Universiteit Brussel (VUB) and principal investigator in Data Science at the imec institute in Belgium, as well as the director of the Master in Applied Computer Science at Vrije Universiteit Brussel. He and his students are collaborating regarding tweet data analysis and geolocation.

What other collaborators or contacts have been involved?

Students who worked on related team term projects in CS4624 (Multimedia, Hypertext, and Information Access) and CS4984/CS5984 (Big Data Text Summarization); see the author list for their entries in the publications listed.

Impacts

What is the impact on the development of the principal discipline(s) of the project?

The success of CS4984/CS5984: Big Data Text Summarization, being run using the pedagogical method of problem/project based learning, with the problem of how to summarize large webpage collections in support of GETAR, should help others teaching NLP or big data to improve the learning in their classes by similar connection to research.

The event focused crawler extends the scope of Web crawling to situations where webpages are sought about an event, rather than about a topic or organization or website; this and new methods using URLs from tweet collections provided support to construct the collections for CS4984/CS5984: Big Data Text Summarization.

The methodology of using location indicative words to infer location for tweets that lack latitude and longitude values should expand the utility of tweet collections and related social network studies, by enabling analyses and visualizations that involve locations or geospatial reasoning. This is helping in a sister NSF CRISP funded project about resilience regarding hurricanes.

The integration of processing of tweets and webpages, all related to important events, in one system with linked workflows, should broaden the scope of studies that largely just use only one of these two sources for digital library and Web archiving research. Confirmation of the benefits was shown by way of event detection using both tweets and news.

What is the impact on other disciplines?

Tweet and webpage collections are of interest to many disciplines studying recent history and current events, including history, sociology, political science, economics, environmental science, linguistics, communications, government, etc. As a result of this project, scores of Virginia Tech scholars, from a variety of departments as well as University Libraries, have expressed interest in our methods and activities, and a number have worked with us on focused studies. We have collected information and shared that with them, as well as helped with related analysis. In addition, a different part of CS4984/CS5984: Big Data Text Summarization worked on chapter level summarization for theses and dissertations, across all disciplines.

This shows how broadly the impact is likely to spread to a number of other disciplines.

What is the impact on the development of human resources?

GETAR this year has led to 1 thesis, 6 student team reports in one class, and 11 student team reports in another course. The application of problem/project based learning has been very popular with students, who are highly motivated, and apply their skills in other courses as well as internships and work after graduation. Students who worked on the project are now in faculty positions at Louisiana State University and Radford University as well as universities in Egypt and Jordan. A number of those come from underrepresented groups.

The project has involved more than 50 collaborators and 12 collaborating institutions. People in diverse fields have been exposed to advanced data analytics and visualization, enhancing their appreciation of science and understanding about working with data.

Jason Callahan, working with co-PI Shoemaker, has passed his prelim, allowing him to focus further on understanding school shootings and related sociology issues.

What is the impact on physical resources that form infrastructure?

Virginia Tech Advanced Research Computing has provided essentially unrestricted access to its Cascades and Huckleberry clusters with GPUs, expanding the use of deep learning methods. A new professor in CS@VT, Jiepu Jiang, has used part of his startup package to purchase a new cluster which students involved in GETAR have started to utilize; a joint proposal to the State Council of Higher Education for Virginia (SCHEV) should lead to further expansion in 2019.

What is the impact on institutional resources that form infrastructure?

In addition to stimulating support from the Deparment of Computer Science for our infrastructure, such as providing a number of VMs, now also including some with GPUs, University Libraries has built a very similar infrastructure, and the campus IT groups have expanded the size and access to clusters to support other similar types of investigations.

What is the impact on information resources that form infrastructure?

Aided in part by the Web Archiving and Digital Libraries workshops, and other dissemination of project activities and accomplishments, other teams involved in Web archiving have engaged in related studies and efforts to devise software and methods, as well as build collections. There is a growing movement for collecting and archiving tweets and/or webpages, and to broaden the support for working with those archives. The enormous collection of over 400 billion webpages at the Internet Archive, as well as other archives, has stimulated broad interest in these information resources. Our methods to add value through analysis, and to support event-oriented studies and access, shows promise to expand the utility of the expanding information resources. A new IMLS-funded effort has been launched that will help disseminate GETAR results to aid librarians and archivists across the nation.

What is the impact on technology transfer?

The Internet Archive is a partner, working with the GETAR team, and has access to our technology, software, and data. Its actions broadly influence the rest of the worldwide Web archiving community. The June 2018 technology transfer visit by PI Fox with Internet Archive PI Bailey should enhance the effects of GETAR on Internet Archive activities.

Internet Archive staff have updated or written new documentation for stakeholders and the general public to query and use data from and about these collections and their contents through its general and collection-specific Wayback index (CDX) APIs, OpenSearch API, and “WASAPI” web archive data transfer API. Improvements to derivative dataset generation and analysis processes that the project team may use to mine and/or visualize these archives were likewise documented and formed the basis of workshops in the United States and abroad to train further librarians, archivists, and researchers to use similar resources and tools.

What is the impact on society beyond science and technology?

The collections developed can be used by any interested groups. As our software and systems mature, open access to suitable portions of our collections will be provided to the public.

Changes/Problems

Changes in approach and reason for change

Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them

Nothing to report.

Changes that have a significant impact on expenditures

We have saved some funding assigned to students, that will be spent in the remainder of the project: GRA Liuqing Li, usually working full time on GETAR, served as GTA in Fall 2018, helping with CS4984/CS5984: Big Data Text Summarization; this was sensible since 11 teams in that class worked with GETAR data, developing and evaluating new methods for collection level summarization using webpages.

Significant changes in use or care of human subjects

Nothing to report.

Significant changes in use or care of vertebrate animals

Nothing to report.

Significant changes in use or care of biohazards

Nothing to report.