

Center for Teaching Old Models New Tricks (TOMNET) A USDOT Tier 1 University Transportation Center

PROJECT PROPOSAL

Title: Real-time Transportation Social Media Analytics using Pulse (Pulse-T)

Principal Investigator: Srinivasa Srivatsav Kandala PhD, Senior Research Manager, Decision Theater Network, Arizona State University

Co-Principal Investigator: Vikash Bajaj, Research Engineer, Decision Theater Network, Arizona state University

1. Introduction/Problem Statement

As city planners and transportation system planners consider changes and upgrades to transportation systems and infrastructure, they require models that accurately reflect communities' needs. Planners need access to advanced activity-travel demand analysis models that are responsive and sensitive to emerging transportation technologies; models are needed that not only provide insights into communities' current travel demands and behaviors, but also help understand people's attitudes and expectations toward a change — or a proposed change — in a community's transportation infrastructure or transportation options. For example, using this system public sentiment can be tracked when accidents involving autonomous vehicles occur or when transportation milestones are achieved in this field.

However, the data on which the current models rely has limitations that prevent planners and policymakers from tapping into residents' attitudes and perceptions widely, across the population and across time. Current models utilize surveys or opinion polls and yield a regimented set of responses to fixed questions. Moreover, the surveys reach a relatively small number of self-selecting individuals. They measure attitudes or self-reports of behavior at a single point in time, and to update them with new research topics or at new points in time is laborious and expensive. And, while some research requires datasets that extend over time, other, critical research requires real-time data that allows gauging current community sentiment around a topic.

Policymakers and researchers are increasingly recognizing the need to delve into emerging data sources like social media for use in transportation planning [1]. Social media provides large amounts of rapidly refreshing rich data that differs by location and by geographic, demographic, and socio-economic factors, which, when subjected to machine learning algorithms and, powerful, innovative analytics, can help researchers recognize important patterns and model communities' perceptions and sentiments, discovering how specific people are influenced by or are influencing other groups of people [1]. Social media can also provide live data about the impact of a policy change; in particular, Twitter offers a large volume of publicly available data in which people and groups broadcast their feelings and preferences far more widely than what a survey instrument could capture [2].

In this project we build the Pulse-T, which will exponentially expand the access of TOMNET researchers and other organizations to an up-to-date, filtered dataset of public opinion and discussions around virtually any transportation research area. Researchers and organizations will have user perceptions on transport demand at their fingertips, enabling them to take appropriate measures and actions and undertake planning projects much more effectively than is possible today.

2. Project Objectives

With successful analysis of social media data guiding outreach strategies that are useful for TOMNET, transport planning and necessary interventions can be done at the right time.

Objective 1: To **increase TOMNET's visibility** through a public-facing dashboard that provides real-time access to users' reactions or opinions about specific transportation policies or topics.

Objective 2: To give TOMNET researchers the ability to **mine large amounts of data** on different topics through filtering by keywords, location, user accounts, geographical attributes, and language of search. Researchers will be able to access topic-specific data and analyze it through an interactive dashboard.

Objective 3: To **provide TOMNET researchers with crowd-sourced, location-specific enriched data** on residents' attitudes, values, perceptions, and preferences around transportation options or changes, which can be employed in current and future research.

3. Proposed Methodology and Data

Pulse-T will conduct sentiment analysis on streamed tweets and perform network analysis on the millions of Twitter users who voice opinions, discuss policies, and indicate preferences around transportation topics each day. By filtering for the right hashtag, this tool can generate the precise data that researchers are looking for. The development of the Pulse-T involves the following steps: collection, storage, processing, and analysis of data, and responses to TOMNET user queries from the front end such as filtering or requesting raw data.

Step 1: Storage and processing. We will use distributed systems to collect data through Twitter streaming API and store it, together with its metadata, in the database. Processing will involve filtering out tweets only for required languages; separating URLs, hashtags, mentions, and images.

Step 2: Extraction of sentiment. Sentiment analysis is a technique which seeks to identify the viewpoint(s) underlying a text span — a powerful tool for interpreting textual data [3, 4]. Sentiment analysis will be at the heart of this product. We will build models using the open source models DeepMoji [5] and Unsupervised Sentiment Neuron [6]. The DeepMoji model is used to predict emojis; this model is trained from 1.2B tweets that are filtered from 55B tweets. We will be extending the model to extract sentiment and emotion and to detect sarcasm. Unsupervised

Sentiment Neuron is trained to predict next letter in Amazon reviews, and its makers assert that it has 91.8% accuracy in detecting sentiment.

Step 3: Network analysis. Analyzing data from all users is not necessary to find prominent conversational issues and extract meaningful sentiment [7]. We will use network analysis to understand influential Twitter users' statements, predictions, opinions, and impact. Picking the highly active users gives us information about the most-talked-about issues in the Twittersphere, helping us access voices related to our topic of interest, i.e., transportation in the Twittersphere. First, we will make use of specific metrics to capture prominent users. Then, we will analyze the sentiment of tweets and the sentiment of the reactions to the influential tweets from prominent users and depict the results in a multidimensional data representation [8].

Step 4: Indexing of results. Lastly, we will index our data and results on a full indexed distributed database such as Elasticsearch. Usage of a distributed database provides high-speed query, search, and aggregation capabilities by serving parts of queries from the nodes in the distributed system along with providing redundancy in case of downtime [9].

Step 5: Development of dashboard. We will build the user-facing real-time dashboard that includes data analytics, sentiment and network analysis, visualizations, and charts.

4. Work Plan

Task 1: Project Conception and Planning

Sub-task 1.a. Create a design based in part on requirements identified by reviewers of the project abstract. TOMNET researchers will be consulted regularly during the design and planning process of the project.

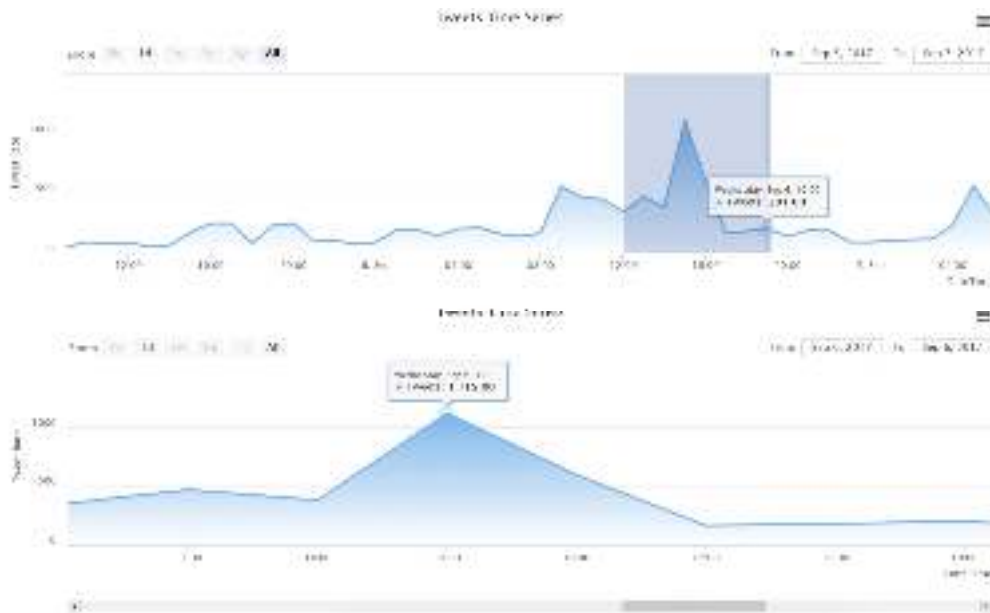
Sub-task 1.b. Identify social media platform and type of data including language, geography, and time frame. Twitter is our preferred social media platform for the Pulse-T. However, we are currently exploring ways to integrate other social media platforms, namely, Facebook and Reddit, in case of any restrictions on Twitter API. In this sub-task we will decide on technology stack, plan the pipeline, and select machine learning algorithms for sentiment analysis, metrics, and topology for network analysis and representation.

Task 2: System Design and Architecture

Sub-task 2.a. Build a distributed system framework to collect live real-time streams of tweets from Twitter and carry out filtering based on filters set by the user through the Pulse-T dashboard (Fig. 1).

Sub-task 2.b. Perform parallel processing of streaming data and list guidelines to develop modular components to perform various analytics (see Task 3).

Figure 1. Filtering time series



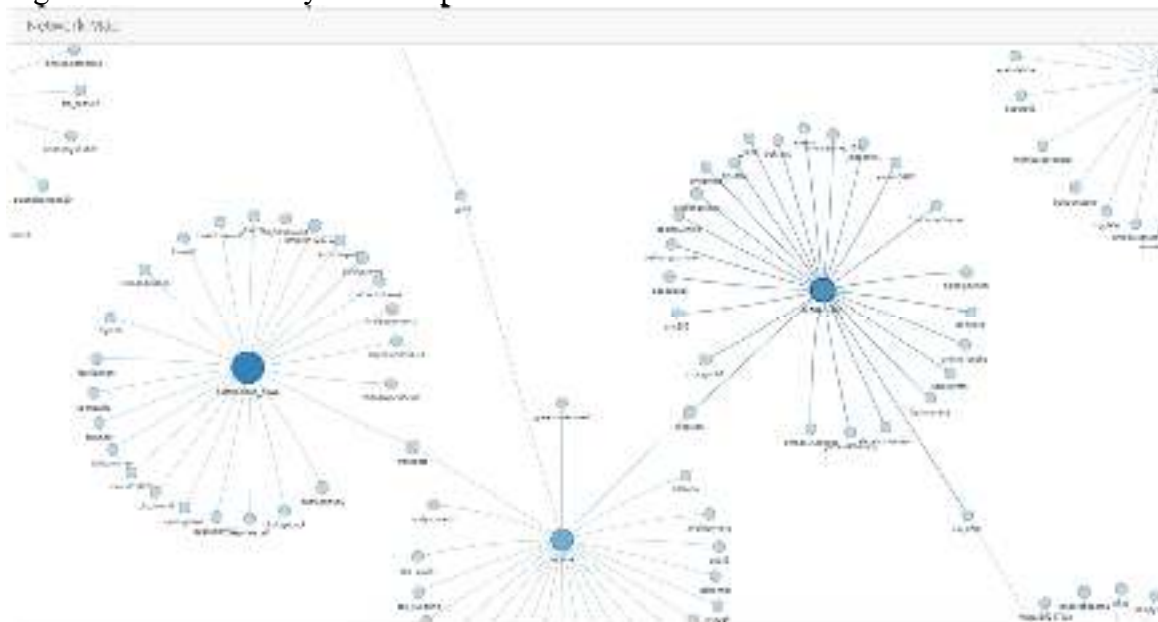
Task 3: Development and Integration of Machine Learning Algorithms and Data Analytics

Sub-task 3.a. Develop a real-time sentiment analysis model for the Pulse-T that includes training the sentiment algorithm on existing tweets followed by validation and testing.

Sub-task 3.b. Build real-time network analysis capability into the Pulse-T based on different metrics and topology (Fig. 2).

Sub-task 3.c. Build a module that will return descriptive statistics based on different queries and filters from TOMNET researchers.

Figure 2. Network analysis and representation



Task 4: Interface Development

Sub-task 4.a. Develop the user-interface for project creation, project selection, project configuration, and controls (Fig. 3).

Sub-task 4.b. Build the user-facing real-time dashboard that includes data analytics, sentiment and network analysis, visualizations, and charts.

Sub-task 4.c. Build a real-time interactive dashboard that compliments the user dashboard with interactivity for the researchers.

Figure 3. A prototype of user creating Projects

The image shows a web-based form for creating a project. At the top, there are two buttons: 'List' and 'Create'. Below this is a section titled 'Project Details'. It contains several input fields: 'Name' (with a red asterisk indicating it's required), 'End Date', 'Keywords', and 'Keywords File'. The 'Keywords File' field has a file selection interface showing 'theater file' and 'No file chosen'. Below the 'Keywords File' field, there is a link that says 'Or upload a set file with keywords in CSV format'. The 'Location' field has a link that says 'Generate Location Coordinates'. At the bottom of the form, there are four buttons: 'Save' (blue), 'Save and Add Another' (grey), 'Save and Continue Editing' (grey), and 'Cancel' (red).

Task 5: Task Integration and Testing

Sub-task 5.a. Integrate, and test the system and the accuracy of our analytics modules.

Sub-task 5.b. Solicit feedback from TOMNET including suggestions for changes in the aesthetics and usability of the user interface. Minor technical changes will be considered and incorporated into the Pulse-T in subsequent iterations of the project, if deemed feasible by the graphic design artist and software development team.

Task 6: Training

After the development of the Pulse-T is complete, Decision Theater Network will provide training materials to TOMNET.

5. Project Schedule

Month	1	2	3	4	5	6	7	8	9	10	11	12
Task 1	■											
Task 2	■	■										
Task 3			■	■	■							
Task 4						■	■	■				
Task 5								■	■	■	■	
Task 6												■

6. Relevance to the Center Theme/Mission

Pulse-T brings the expertise of the Decision Theater Network to TOMNET by offering a new, powerful, and flexible method for gaining access to people’s attitudes, values, preferences, and perceptions regarding transportation options, events, and infrastructure.

TOMNET’s ability to integrate information about people’s attitudes, values, preferences, and perceptions in transportation demand forecasting models is hindered by the type, amount, and quality of data available for use in those models. Survey instruments capture data at a single point in time from a fairly limited number of people, and the information captured is filtered through people’s own reasoning processes — what individuals believe they do or would do under specified conditions. However, about seven out of ten Americans use social media to connect with one another, engage with news content, and share information [10], making social media a rich data source (unfiltered by the glitches inherent in self-reporting) for social research, both quality- and quantity-wise. Analysis of social media data can reveal how many people are discussing any topic at a given time and what their feelings are about that topic [11]. Transportation infrastructure planning is a prime opportunity for the use of this tool since this domain can leverage Pulse’s capabilities in large data collection from various data sources as well as running machine learning algorithms on these datasets in real time.

TOMNET researchers will find the Pulse-T to be flexible and applicable to virtually any research topic, facilitating their understanding of user perceptions on transport demand and enabling them to take appropriate measures or actions and perform related planning more effectively.

7. Anticipated Outcomes and Deliverables

Product: The central product of this research is the **Pulse-T**, which leverages social media data to measure people’s preferences and derive valuable insights into people’s attitudes and opinion on transportation topics using advanced machine learning and pattern recognition algorithms and techniques. The raw dataset and the enriched data with Pulse-T analytics will be available for download in the form of CSV or JSON to TOMNET researchers to carry out their own research. They will be able to export the dataset to Excel or any other analytics software and run their own analytics, customizing the tool to their research needs by changing keywords, location, user accounts, geographical attributes and language of search.

Benefits: Pulse-T will cater to customized and evolving research interests of the users. On setting the right customizations (by changing keywords, location, user accounts, geographical attributes, and language of search), the tool will generate specific datasets that can be used to investigate very specific transportation-related research questions. The real-time data analytics of the Pulse-T will aid TOMNET researchers in evaluating public opinion, preferences, and sentiments around a specific transportation topic in a given place and time, and enable them to make better-informed decisions and do more effective planning related to transportation systems in that location.

Deliverables

- A public-facing real-time dashboard hosted on Pulse, accessible through TOMNET's website, where the viewer can follow public attitudes and reactions on a topic.
- A researcher-accessible real-time interactive dashboard hosted on Pulse, accessed through TOMNET's website where the researcher can download data — raw and enriched with analytics.
- An administrative interface to Pulse-T that will allow the creation, selection, and configuration of projects based on transportation topics of interest to TOMNET.

8. Research Team and Management Plan

The research team will be led by Dr. Srinivasa Srivatsav Kandala and Vikash Bajaj. It will consist of a project manager and two Graduate Research Assistants. The research team will be based out of ASU DTN.

The Decision Theater Network (DTN) includes two ASU facilities in Tempe, Arizona and at the McCain Institute for International Leadership in Washington, DC and facilities at several other domestic and international partner institutions of higher education. The DTN enables rapid, cross-disciplinary collaboration, access to ASU DT's expert analytics, data visualization, and predictive modelling, and facilitated decision-making sessions with policymakers, subject matter experts, and other stakeholder groups at multiple connected locations across the globe.

Communication Plan

Meetings will be organized at DT facility periodically to provide updates on the progress.

9. Technology Transfer Plan

The Pulse-T will be hosted on DTN Pulse and accessible through TOMNET's website. After the completion of the project, DTN will provide training to the researchers at TOMNET. Training will include a step-by-step demonstration of how to use the tool. It will also include explanation of the methods used, and the inputs and outputs, including demonstrations of how these inputs become the resultant outputs. Training will be conducted at DTN facilities and through teleconference.

After project implementation, DTN will provide data, server, and hosting maintenance over the predetermined agreed period.

10. Workforce Development

DT will hire two graduate students who will primarily be responsible for developing the technology stack required for the project under the guidance of the Principal Investigators. Undergraduate students who have expertise in graphic design will be involved in designing the front-end of the user-facing dashboard. Project manager from the DT staff will manage the project.

11. References

- [1] Yeung, D. (2018). Social Media as a Catalyst for Policy Action and Social Change for Health and Well-Being: Viewpoint. *Journal of Medical Internet Research*, 20(3), e94. <http://doi.org/10.2196/jmir.8508>
- [2] Link, M., Muphy, J., Schober, M. F., Buskirk, T. D., Childs, J. H., Tesfaye, C. L., ... & Pasek, J. (2014). Social media in public opinion research.
- [3] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1-135.
- [4] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 1320-1326).
- [5] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.
- [6] Lengerich, B. J., Konam, S., Xing, E. P., Rosenthal, S., & Veloso, M. (2017). Visual explanations for convolutional neural networks via input resampling. *arXiv preprint arXiv:1707.09641*.
- [7] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million-follower fallacy. *Icwsn*, 10(10-17), 30.
- [8] Himelboim, I. (2017). Social Network Analysis (Social Media). *The International Encyclopedia of Communication Research Methods*, 1-15.
- [9] Gormley, C., & Tong, Z. (2015). *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. " O'Reilly Media, Inc."
- [10] Pew Research Center. (2017). Social media fact sheet. *Pew Research Center: Internet, Science & Tech*.
- [11] Bright, J., Margetts, H., Hale, S., & Yasseri, T. (2014). The use of social media for research and analysis: A feasibility study. *Report to the Department of Work and Pensions, September*.

12. Qualifications of Investigators

SRINIVASA SRIVATSAV KANDALA

Professional Preparation

- Jawaharlal Nehru Technological University, India, Civil Engineering, Bachelor of Science, 2007
- Arizona State University (ASU), Tempe, AZ, Transportation Engineering, Master of Science, 2009
- Arizona State University (ASU), Tempe, AZ, Statistics, Graduate Certificate, 2011

- Arizona State University (ASU), Tempe, AZ, Transportation Engineering, Doctor of Philosophy, 2014

Appointments

- 2007 – 2009, Graduate Research Assistant, Transportation Systems, Group, ASU
- 2011-2012, Project Aide Lead, Decision Theater, ASU
- Sept 2013 - Nov 2013, Transportation Engineering Intern, Maricopa Association of Governments (MAG)
- 2009 - 2014, Graduate Research Associate, Transportation Systems Group, ASU
- 2014 - 2016, Senior Research Analyst, Decision Theater Network, ASU
- 2016- Present, Senior Research Manager & Research Professional, Decision Theater Network, ASU.

Synergistic Activities

Srinivasa Srivatsav Kandala (Sri) is currently working as a Senior Research Manager at the Decision Theater Network (DTN), Arizona State University (ASU). He worked as Senior Research Analyst at DTN from 2014-16. He leads the research team at the Decision Theater and his responsibilities include managing and leading the research development, reporting processes, analysis and information systems for decisions based on key findings of research initiatives. He has his PhD in transportation systems and modeling at ASU. He also has a graduate degree in statistics from ASU and a Master's degree in transportation engineering (ASU) with emphasis on traffic modeling. He worked research aide lead for MapStory project, an online social cartography platform, and played a key role in the ASU development team for the MapStory foundation. He has proficiency in social network analysis, social media mining, regression analysis, categorical data analysis, multivariate statistics, data mining, network flows, design of engineering experiments, wavelet transforms, geographical information systems, spatial modeling, and agent-based modeling. He has expertise on visualization tools like Tableau and ArcGIS. He won the Simon best paper award at the 9th annual design science research in information systems and technology (DESRIST) conference, SUN award and Commitment award from ASU for excellent performance and fostering cooperation and is a fellow of International Road Federation (IRF).

Related Publications/Products

- Alashri, S., Kandala, S., Pradhan, A., Bajaj, V., Smith, K., Desouza, K., 2017. The 2016 US Presidential Election on Facebook: An Exploratory Analysis of Sentiments. Hawaii International Conference on System Sciences (HICSS), Hawaii, USA.
- Alashri, S., Kandala, S., Bajaj, V., Ravi, R., Smith, K., Desouza, K., 2016. An Analysis of Sentiments on Facebook during the 2016 U.S. Presidential Election Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA.
- Alashri, S., Pradhan, A., Kandala, S., Desouza, K., 2016. What is the tone of the 2016 presidential campaign on Facebook? Retrieved from: www.brookings.edu/2016/08/10/what-is-the-tone-of-the-2016-presidential-campaign-on-facebook/. Brookings Institution Press, Washington, DC.
- Desouza, K., Alashri, S., Bajaj, V., Kandala, S., 2016. Which topics do presidential candidates discuss on Facebook? Retrieved from: www.brookings.edu/2016/05/19/which-topics-do-presidential-candidates-discuss-on-facebook/. Brookings Institution Press, Washington, DC.

- Desouza, K., Kandala, S., Alashri, S., Bajaj, V., 2016. Tracking presidential campaigns on Facebook, Retrieved from: www.brookings.edu/2016/04/13/tracking-presidential-campaigns-on-facebook/. Brookings Institution Press, Washington, DC.
- Fedorschak K., Kandala S., DeSouza K.C, and Krishnamurthy R. 2014. “Data Analytics and Human Trafficking”. Design Science Research in Information Systems and Technology.
- Kandala S., 2011. “Analysis of Freeway Bottleneck Capacities in Greater Phoenix”, presented at the ITS Arizona 18th annual conference.
- Ahn S., Kandala S., Uzan J. El-Basyouny M., 2009. “Impact of Traffic Data on the Pavement Distress Predictions using the Mechanistic Empirical Pavement Design Guide (MEPDG)”. Road Materials and Pavement Design.
- Ahn S., Kandala S., Uzan J., El-Basyouny M. “Comparative Analysis of Input Traffic Data and the MEPDG Output for Flexible Pavements in the State of Arizona”, presented at the Transportation Research Board annual meeting, 2009.

VIKASH BAJAJ

Professional Preparation

- National Institute of Technology, Tiruchirappalli, India, Computer Science and Engineering, Bachelor of Technology, 2015
- Arizona State University (ASU), Tempe, AZ, Computer Science, Master of Science, 2017

Appointments

- 2015 – 2017, Graduate Research Assistant, Decision Theater Network, ASU.
- 2017- Present, Asst. Research Engineer, Decision Theater Network, ASU.

Synergistic Activities

Vikash Bajaj is currently working as a Research Engineer at the Decision Theater Network (DTN), Arizona State University (ASU). Previously, He was working as a Research Assistant at DTN from 2015-17. He has expertise in full stack and distributed systems’ development and has led the technical efforts of many project at DTN.

Related Publications/Products

- Alashri, S., Kandala, S., Pradhan, A., Bajaj, V., Smith, K., Desouza, K., 2017. The 2016 US Presidential Election on Facebook: An Exploratory Analysis of Sentiments. Hawaii International Conference on System Sciences (HICSS), Hawaii, USA.
- Alashri, S., Kandala, S., Bajaj, V., Ravi, R., Smith, K., Desouza, K., 2016. An Analysis of Sentiments on Facebook during the 2016 U.S. Presidential Election Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA.
- Alashri, S., Pradhan, A., Kandala, S., Desouza, K., 2016. What is the tone of the 2016 presidential campaign on Facebook? Retrieved from: www.brookings.edu/2016/08/10/what-is-the-tone-of-the-2016-presidential-campaign-on-facebook/. Brookings Institution Press, Washington, DC.

- Desouza, K., Alashri, S., Bajaj, V., Kandala, S., 2016. Which topics do presidential candidates discuss on Facebook? Retrieved from: www.brookings.edu/2016/05/19/which-topics-do-presidential-candidates-discuss-on-facebook/. Brookings Institution Press, Washington, DC.
- Desouza, K., Kandala, S., Alashri, S., Bajaj, V., 2016. Tracking presidential campaigns on Facebook, Retrieved from: www.brookings.edu/2016/04/13/tracking-presidential-campaigns-on-facebook/. Brookings Institution Press, Washington, DC.

13. Budget Including Non-Federal Matching Funds

Total UTC (Federal) Funds Requested (US Dollars): \$60,000

Total Cost Share Provided (US Dollars): \$30,000

Source of the Cost Share: ASU DTN Cost Center/Program CC0362 PG03767.