Saurav Ghosh

December 25, 2016

Ph.D. Candidate, Dept. of Computer Science
Discovery Analytics center, Virginia Tech, USA
sauravcsvt@vt.edu
https://sites.google.com/a/vt.edu/saurav-ghosh/

Objective

My area of specialization is Data Science with specific focus on Text Mining, Knowledge Discovery from Text and Deep Learning for NLP (word2vec). I am expecting to graduate mid June 2017 and therefore, seeking a Data Scientist position starting from June-July 2017.

Education

Fall 2012 - **Current**	**PhD Student** in Computer Science, Virginia Tech, USA. **Overall GPA:** 3.61/4.00 **GPA (Data Science Courses):** 3.93/4.00 **Advisor:** Dr. Naren Ramakrishnan
2008 - 2012	**B.E.** in Electronics and Telecomm. Engg., Jadavpur University, India.

Research Statement

Research Focus: Text Mining, Knowledge Discovery from Text, Deep Learning for NLP (word2vec)

- *Application area:* **Text Analytics Methods for Public Health Surveillance**

Broad Focus: Data Science, Machine Learning and Pattern Recognition.

Research Problems: Mining online news media for modeling spread of infectious diseases

- **Forecasting rare disease outbreaks from multiple news sources**
  - Real-time forecasts of hantavirus outbreaks in multiple countries of Latin America using unsupervised spatio-temporal topic modeling.
  - *Designed and implemented the entire big-data pipeline for sending hantavirus forecasts to IARPA in real-time*
  - *Projects:* EMBERS for IARPA OSI Program (winning team)
- **Assessing temporal associations between news trends and infectious disease outbreaks**
  - EpiNews: Designed and implemented a supervised temporal topic model to quantify media interest during infectious disease outbreaks.
  - *Projects:* EMBERS for IARPA OSI Program.
- **Neural word embeddings for automated disease taxonomy generation**
  - Dis2Vec: Designed and implemented a vocabulary driven word2vec method for building disease taxonomies from online news corpora.
  - *Projects:* EMBERS for IARPA OSI Program.
- **Neural word embeddings for automated generation of epidemiological line lists**
  - Designed and implemented a method (word2vec + dependency parsing) for automatic extraction of line lists from semi-structured WHO DONs with specific focus on emerging diseases.
  - *Projects:* EMBERS for IARPA OSI Program.

Publications

Current Publications

2017 S. Ghosh, P. Chakraborty, E. O. Nsoesie, E. Cohn, S. R. Mekaru, J. S. Brownstein, and N. Ramakrishnan. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. *To Appear in Nature Scientific Reports*, 2017

T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan. Forecasting rare disease outbreaks from open source indicators. *To Appear in Statistical Analysis and Data Mining, Best of SDM Special Issue*, 2017

S. Ghosh, P. Chakraborty, B. Lewis, E. Cohn, M. Majumder, J. S. Brownstein, M. V. Marathe, and N. Ramakrishnan. Guided deep list: Automating the generation of epidemiological line lists from open sources. In *Prepraration for submission to PLOS Medicine*, 2017

2016 S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, and N. Ramakrishnan. Characterizing diseases from unstructured text: A vocabulary driven word2vec approach. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM), Indianapolis, USA, October 24-28*, 2016, **Recipient of ACM SIGIR Student Travel Award**

2015 H. Wu, P. Chakraborty, S. Ghosh, and N. Ramakrishnan. Forecasting influenza in senegal with call detail records. In *NETMOB, MIT Media Lab, April*, 2015

T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan. Sourceseer: Forecasting rare disease outbreaks using multiple data sources. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM), Vancouver, BC, Canada, April 30 - May 2*, pages 379–387, 2015, **Recipient of Best Research Paper Award**

S. Ikbal, A. Tamhane, B. Sengupta, M. Chetlur, S. Ghosh, and J. Appleton. On early prediction of risks in academic performance for students. *IBM Journal of Research and Development*, 59(6), 2015

2013 S. Ghosh, T. Rekatsinas, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor, and N. Ramakrishnan. Forecasting rare disease outbreaks with spatio-temporal topic models. In *NIPS workshop on Topic Models*, 2013

S. Roy, S. M. Islam, S. Das, S. Ghosh, and A. V. Vasilakos. A simulated weed colony system with subregional differential evolution for multimodal optimization. *Engineering Optimization*, 45(4):459–481, 2013

2012 S. M. Islam, S. Das, S. Ghosh, S. Roy, and P. N. Suganthan. An adaptive differential evolution algorithm with novel mutation and crossover strategies for global numerical optimization. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 42(2):482–500, 2012

S. Ghosh, S. Das, S. Roy, S. M. Islam, and P. N. Suganthan. A differential covariance matrix adaptation evolutionary algorithm for real parameter optimization. *Information Sciences*, 182(1):199–219, 2012

2011 S. Ghosh, S. Roy, S. M. Islam, S. Zhao, P. N. Suganthan, and S. Das. Non-uniform circular-shaped antenna array design and synthesis - a multi-objective approach. In *Swarm, Evolutionary, and Memetic Computing: Second International Conference, SEMCCO 2011, Visakhapatnam, Andhra Pradesh, India, December 19-21, 2011, Proceedings, Part II*, pages 223–230. Springer Berlin Heidelberg, 2011

S. Roy, M. Islam, S. Ghosh, S. Das, A. Abraham, and P. Kromer. A modified differential evolution for autonomous deployment and localization of sensor nodes. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation (GECCO)*, pages 235–236. ACM, 2011

Technical Skills

Programming	**Python** (Preferred), Java, C, Matlab
OS	**Ubuntu** (Preferred), Arch Linux, Windows
Frameworks	NOSQL: MongoDB

Professional Experience

- **Virginia Tech** Arlington, VA

  GRA in Discovery Analytics Center. Advised by Dr. Naren Ramakrishnan *Fall 2012 - Current*

  – Worked on Disease Forecasting. Implemented big-data rare disease forecasting pipeline for EMBERS in *python* over EMBERS *AWS* cluster framework. Forecasts sent in real-time without human supervision and continuously evaluated

  – Interfaced with several agencies such as IARPA and CDC for disease forecasting problems. Collaborated with several institutes such as *Biocomplexity Institute of Virginia Tech*, *Harvard Medical School* and *University of Maryland, College Park*

- **IBM Research** Bangalore, IN

  Research Scientist Intern, advised by Dr. Shajith Ikbal *June-August 2014*

  – Worked on missing data and class imbalance problems for early prediction of risks in academic performance of students

  – *Application Area*: **Educational Data Mining**

- **Nanyang Technological University** Singapore

  Research Intern, advised by Prof. P. N. Suganthan *Summer 2011*

  – Worked on application of heuristic optimization for non-uniform circular-shaped antenna array and design

Participation in Conferences

- **NIPS 2013**, NIPS 2013 Topic Modeling Workshop, Lake Tahoe, Nevada, December 2013.
- **DDD 2013**, 2nd International Conference on Digital Disease Detection, San Francisco, USA, September 2013.
- **CDC Influenza Forecasting Workshop**, Atlanta, USA, August-September 2016.
- **Health Care Informatics and Analytics Conference**, NVTC Big Data and Analytics Committee, Fairfax, VA, May 2016.
- **CIKM 2016**, 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, Oct 2016

Awards and Honors

Award	Recipient of Best Research Paper Award, SDM	2015
	SIGIR Student Travel Award, CIKM	2016
	Placed within the top 0.7 percent in All India Engineering Entrance Examination (AIEEE), 2008	
	Recipient of National Merit Scholarship from Ministry of HRD, Government of India, 2008	