Pradipto Das

PhD, Computer Science, SUNY Buffalo, USA E: pradipto.das@gmail.com | T: 716 440 0610 | W: http://pradipto.com

Executive Summary

Core Proficiencies: Well-rounded technical knowledge on applying machine learning and natural language processing techniques on artificial intelligence problems involving natural language and multimodal data. A scientific thinker who works well in teams as a collaborator and leader, is highly organized, analytical, delves deep into solving problems around the following more focused areas in artificial intelligence:

- Developing information extraction and summarization systems and exploratory probabilistic browsing models (topic models)
- Developing predictive modeling systems using sparse logistic regression, support vector machines etc.

Professional and Research Skills:

- **Experience in end-to-end process of product research prototype development to its successful commercialization**
- **Deep analysis and interpretation of results**
- **Innovate over state-of-the-art analytics tools for big data**

Education

PhD, Computer Science – Fall 2006 to Fall 2013 | **State University of New York at Buffalo, USA**
MCA (Master of Computer Applications) – July 2004 | **West Bengal University of Technology, Kolkata, India**
BS (Honors) Mathematics – July 2001 | **Jadavpur University, Kolkata, India**

Professional and Research Skills Demonstrated

I. **Principal Software Engineer, NLP and Machine Learning, SmartFocus US, Inc.** [Sep 2014 – **Current**]

Accomplishments: [Coding language used: primarily C++]

- **Leading efforts** to build multi-core framework for running various optimization algorithms for Logistic Regression using both feature and data level parallelism. Improved average precision for multi-label news document classification by at least 40%
- **Built end-to-end production ready framework** for distributed memory indexing using Map-Reduce principles for machine learning tasks including complete pipeline for feature-transform, cross-validation and final model deployment
- **Built end-to-end production ready framework** for fast merging of topics from topic models over Twitter using spanning trees. **Improved recommendation click-through rates** by at least 10% in a recent proof-of-concept A/B testing phase.
- **Built production ready libraries** for Statistically Improbable Phrases (SIPs) and Ternary Search Tries for Unicode strings

II. **Principal Software Engineer, NLP and Machine Learning, Content Savvy Inc., Snyder, NY, USA** [Dec 2013 – Aug 2014]

Accomplishments: [Coding language used: primarily C++]

- **Reduced false positives** from Part-of-speech sequence tagger by 11% on English Penn Treebank test set (sect. 22-24)
- **Improved runtime of existing sequence tagging** within the source tree by 30% using profiling tools like Valgrind & gprof
- **Reduced false positives** from Named Entity sequence tagger by 7%
- **Experience in Agile development practices** for transforming Machine Learning prototypes into production ready deployable modules with the help of bug tracking and version controlling tools such as JIRA and Git
- Improvements in structured learning modules helped in the betterment of core technologies of the company leading to its eventual acquisition by UK-based SaaS Company – SmartFocus. **Helped in three US/Canadian citizens getting hired**

III. **Research Engineer, CSE Department, SUNY Buffalo, NY, USA** [Spring 2011 – Fall 2013]

A. Project: Natural Language based Multimedia Event Detection/Recounting (MED/MER)

Successfully completed a large project on translating videos to text and back without using expensive video annotation efforts

Accomplishments: [Coding language used: primarily Java and C++]

- **System ranked first** in TRECVID 2012 Multimedia Event Recounting track for matching videos on a given abstract event to specific event descriptions based purely on predicted text
- **Joint research in collaboration** with Honeywell ACS Labs (Minneapolis, MN), Kitware Inc. (Albany, NY), Stanford University, Simon Fraser University and Georgia Tech University [*Project funded by IARPA's ALADDIN program*]

B. Project: Exploratory Data Analysis and Multi-document Summarization using Topic Models

Accomplishments: [Coding language used: primarily Java and C++]

- **Successfully formulated and implemented from scratch** bi-perspective topic models that allow modeling of ubiquitous document representations – documents which incorporate both word level annotations and document level metadata

IV. **Research Intern, Janya Inc., Amherst, NY, USA** [Summer 2010]

Project: Gibbs sampling based Topic Modeling Framework for the Semantex® Text Analytics Processor Pipeline

Accomplishments: [Coding language used: primarily C++]

- **Improved product capabilities** by including corpus based solutions in addition to document/sentence centric models
- Code written while internship is now deployed in production as part of the main software pipeline

V. **Teaching Assistant, CSE Dept., SUNY Buffalo, Buffalo, NY, USA** [Fall 2006 – Fall 2010]

VI. **Visiting Research Fellow, Center for Soft Computing Research, Indian Statistical Institute, Kolkata** [Aug 2005 – Jul 2006]

VII. **Assistant Systems Engineer, Tata Consultancy Services Ltd. (TCS), Kolkata, India** [Sep 2004 – Jul 2005]
VIII. **Project Intern, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India** [Spring 2004]

Publications

[8] **P. Das**, C. Xu, R. F. Doell and J. J. Corso – "**A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching**," in Proceedings of the Twenty Sixth IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), Portland, Oregon, Jun, 2013 [**Full paper and spotlight**]

[7] **P. Das**, R. K. Srihari and J. J. Corso, "**Translating Related Words to Videos and Back through Latent Topics**," in Proceedings of the Sixth International Conference on Web Search and Data Mining (**WSDM**), Rome, Italy, Feb 2013 [**oral**]

[6] A. Perera, S. Oh, M. Pandey, T. Ma, A. Hoogs, A. Vahdat, K. Cannons, G. Mori, S. McCloskey, B. Miller, S. Venkatesh, P. Davalos, **P. Das**, C. Xu, J. J. Corso, R. K. Srihari, I. Kim, Y. C. Cheng, Z. Huang, C. H. Lee, K. Tang, L. Fei-Fei, D. Koller. "**TRECVID 2012 GENIE: multimedia event detection and recounting**," in Proceedings of **TRECVID** Workshop, Nov 2012

[5] **P. Das** and R. K. Srihari, "**Using Tag-Topic Models and Rhetorical Structure Trees to Generate Bullet List Summaries**" Extended version appears in PhD thesis: http://www.cse.buffalo.edu/tech-reports/2014-02.pdf ; Shorter version selected for presentation and appears in Proceedings of Text Analysis Conference (**TAC**), Gaithersburg, MD, Nov 2011 [**oral**]

[4] **P. Das**, R. K. Srihari and Y. Fu, "**Simultaneous Joint and Conditional Modeling of Documents Tagged from Two Perspectives**," in Proceedings of the Twentieth ACM Conference on Information and Knowledge Management (**CIKM**), Glasgow, Scotland, Nov 2011 [**oral**]

[3] **P. Das** and R. K. Srihari, "**Learning To Summarize using Coherence**," in Proceedings of Neural Information and Processing Systems (**NIPS**) Workshop on Applications for Topic Models: Text and Beyond, Whistler, BC, Dec 2009 [**poster**]

[2] **P. Das** and R. K. Srihari, "**Utterance Topic Models for Generating Coherent Summaries**," in Proceedings of NIST Text Analysis Conference (**TAC**), Gaithersburg, MD, Nov 2009 [**oral**]

[1] **P. Das**, R. K. Srihari and S. Mukund, "**Discovering Voter Preferences in Blogs using Mixtures of Topic Models**," in Proceedings of the Third Workshop on Analytics for Noisy Unstructured Text Data (**AND09**) endorsed by International Association of Pattern Recognition, Barcelona, Spain, Jul 2009 [**oral**]

Professional Service and Peer Reviewing

- **Program Committee Member:** NAACL 2015, ACL 2014
- **Primary Reviewer (Conference):** EMNLP 2015, NAACL 2015, ACL 2014, NAACL 2013
- **Primary Reviewer (Journal):** IEEE-PAMI 2015, Elsevier-Image and Vision Computing 2014
- **Secondary Reviewer:** IEEE-International Conference on Semantic Computing 2013

Computer Skills

- **Programming Languages:** C++, Java
- **Distributed and Large Data Processing Software:** Hadoop, Hive, Solr, MPI
- **Open source software:** Code for research prototypes hosted at http://pradipto.com/software/software.html

Mentions in News, Awards and Honors

- **YouCook Dataset** collected by us and our initial experiments on it for our 2013 CVPR paper has been found to be very effective in conducting state-of-the-art robotics perception research by computer vision scientists and has been **cited in Science and Technology news** – http://techxplore.com/news/2015-01-robots-kitchen-duty-cooking-video.html
- **Best poster award** for our 2013 CVPR paper titled "*A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching*" awarded at UB's Information and Computing Technology (ICT) Day Workshop held in honor of Prof. Jitendra Malik from UC Berkeley as distinguished speaker. **Mar 2013**
- **Research/Teaching Assistantship** for PhD studies at SUNY Buffalo. **Sep 2006 to Jul 2013**
- **Fellowship for the post of Visiting Research Fellow** at Indian Statistical Institute, Kolkata, India. **Aug 2005 to Jul 2006**
- **Certificate of merit and memento** for standing First Class 2nd in MCA, Kolkata, India. **Jan 2005**
- **"Top Performer Award"** in batch (T-47) for the Initial Learning Program at TCS, Trivandrum, India. **Nov 2004**
- **Govt. of India National Scholarship** based on BS results at Jadavpur University, Kolkata, India. **Aug 2003**

References:

Dr. Siddhartha Dastidar
Quant/Risk Analyst, Brigade Capital
Adjunct Asst. Prof., Columbia University
Email: sidgdastidar@gmail.com

Dr. John Chen
Principal Inventive Scientist
Interactions Corporation
Email: limacon@yahoo.com

Dr. Enrique Alfonseca
Research Tech Lead/Manager
Google Research, Zurich, Switzerland
Email: enrique.alfonseca@gmail.com

Dr. Jason J. Corso
Associate Professor
EECS Dept., UMich Ann Arbor
Email: jjcorso@eecs.umich.edu