Suraj Maharjan

http://www2.cs.uh.edu/~suraj/
https://github.com/sjmaharjan, https://bitbucket.org/sjmaharjan/

Email : mhjsuraj@gmail.com
Mobile : +1-205-948-7085
Skype : suraj_maharjan1

Education

- **University of Houston** — *Advisor: Dr. Thamar Solorio*
  Ph.D. in Computer Science; GPA: 4.00 — *Jan 2015 – May 2018*
  Dissertation: Stylistically Aware Representations of Books
- **University of Alabama at Birmingham** — *Advisor: Dr. Thamar Solorio*
  Masters of Science in Computer Science; GPA: 3.9 — *Aug 2012 – Dec 2014*
- **Tribhuvan University, Institute of Engineering**
  Bachelors in Computer Engineering; GPA: 86.30/100; Ranked 2^{nd} ***in the University*** — *2005 – 2010*

Experience

- **Pacific Northwest National Laboratory** — *Mentor: Dr. Svitlana Volkova*
  Post Doctoral Research Associate — *Aug 2018 –Present*
- **Pacific Northwest National Laboratory** — *Mentor: Dr. Svitlana Volkova*
  Research Intern — *Feb 2018 – May 2018 and May 2017 – Aug 2017*
  - **Towards Anticipatory Analytics using Dynamic Knowledge Graphs**: Built a user-aware attention model to predict users' email sending and receiving behavior for the next day using their last n days' behavior, day of the week information, and the whole knowledge graph. Summarized each day's knowledge graph information using GCN. [**Keras**, **sklearn**]
  - **Deep learning on Dynamic Knowledge Graphs**: Used a hierarchical method to first learn a summary vector using CNN/GCN on knowledge graphs (GDELT) per day and then applied RNN over the sequence of day vectors to predict the instability of a given country in the future. [**Keras**, **sklearn**]
- **University of Houston (RiTUAL Lab)** — Houston, TX
  Research Assistant — *Jan 2015 – May 2018*
  - **Genre-aware Attention Model**: Proposed a multimodal, genre-supervised neural attention model to combine feature representations from different aspects of books (book covers, content, sentiment) to improve the likability prediction of books. [**Keras**, **sklearn**]
  - **Emotion Flow**: RNN with attention model to capture an author's dexterity in the use of emotion flow across books. Results significantly improved with the inclusion of emotion flow model for books' likability prediction and movie tags prediction. [**Keras**, **sklearn**]
  - **Author Style Embeddings**: Proposed a method to learn an author's general style embeddings by using a language model to the sequence of stylistic traits (annotated character n-grams) generated from multiple samples of books written by the author. Results significantly improved with the addition of author style embeddings for the likability prediction task. [**Gensim**, **sklearn**]
  - **Stylistic Analysis of Books**: Proposed different hand-crafted and neural representations to extract the style embedded in books. Showed that adding genre as an auxiliary task to the primary task of likability prediction (multitask setting) improves results. [**Keras**, **sklearn**]
  - **Book Recommendation Engine**: Designed and implemented a web-based prototype for *Booxby* that recommends similar books by matching the style encoded in a book's content to that of other books. http://solorio.uh.edu:5000 [**AMT**, **flask**, **sklearn**, **celery**, **flower**, **RabbitMQ**, **Mongodb**]
  - **Sentiment Analysis of Financial data**: Combined hand-crafted sentiment, lexical, word embedding, and meta-data features to neural representations learned using CNN and RNN with attention to predict sentiment scores. [**Keras**, **sklearn**]
  - **Named Entity Recognition**: Used multitask setting by defining and adding an auxiliary task of predicting if a token is a named entity (NE) or not to the main task of predicting fine-grained NE (BIO) labels in noisy social media data. [**Keras**]
  - **Preventing and Deterring Cyberbullying**: Built a new corpus of invectiveness by collecting posts from ask.fm. Used CrowdFlower to annotate the posts into two classes: *invective* and *neutral*. Used lexical, syntactic, LIWC lexicon, topic model, and word embedding features to classify invective posts. [**Gensim**, **sklearn**, **CrowdFlower**]
  - **Fine-grained Semantic Similarity of Words**: Used multitask architecture with CNN applied on GloVe word embeddings. [**Keras**, **sklearn**]

- **Author Profiling**: Designed novel features to extract the style of authors for age and gender classification. Used Hadoop Map-Reduce framework to extract features and implement Naive-Bayes algorithm, which reduced the processing time from 15 days to a couple of hours. [**Hadoop**, **Mahout**, **Maven**]

- **University of Alabama at Birmingham (CoRAL Lab)** Birmingham, AL
  Research and Teaching Assistant *Aug 2012 – Dec 2014*
  - **Codeswitching**: Data collection and annotation for English-Spanish and Nepali-English codeswitched dataset using CrowdFlower for the First Workshop on Computational Approaches to Code Switching. Built baseline systems (langID, Lexical) to evaluate participants' systems. [**CrowdFlower**, **Python**]
  - **Malware Family Identification**: Used prominent strings method (TF–IDF, Jaccard coefficient) to classify malware into their respective families. [**Python**]
  - **Teaching Assistant**: Courses: *Introduction to Object Oriented Programming with Java* and *Object-Oriented Design*

- **Verisk Information Technologies** Kathmandu, Nepal
  Software Engineer *2010 – 2012*
  - **Rule Engine**: Proposed and built a prototype rule engine using And-OR Expression Tree with heuristics to improve the speed of execution. Implemented medical rules in KnowledgeWorks and Drools frameworks. [**Lisp**, **KnowledgeWorks**, **Drools**, **Spring**]
  - **Distributed Databases**: Benchmarked performance of distributed databases: *Greenplum, Teradata, and Vertica.* Transformed SQL queries into Map-Reduce programs to run in the Greenplum database. [**Java**, **Python**]

NLP Shared Tasks

- **First** position at the Emerging and Rare Entity Recognition shared task *2017*
- **Second** position at SemEval2017-Task5: Fine-Grained Sentiment Analysis on Financial Microblogs and News subtask 2 (*First position with alternate scoring approach*) *2017*
- **First** position in detecting semantic similarity and **second** position in detecting find-grained semantic similarity at the CogALex-Vshared task *2016*
- **First** position in Arabic Dialects and **second** position in Spanish-English at the Second Workshop on Computational Approaches to Code Switching shared task *2016*
- **Third** position at PAN Author Profiling 2014 shared task *2014*

Awards

- Book analysis and recommendation, US Patent Application ID: US20180107645A1 2018
- Recipient of the European Chapter of the Association for Computational Linguistics (EACL) 2017 studentship *2017*
- Travel award recipient for the 2013 Open Science Grid User School and XSEDE13 Conference *2013*
- Scholarship for Undergraduate Studies from Institute of Engineering, Nepal *2005 – 2010*

Training and Certifications

- Deep Learning Specializations, *www.coursera.org/specializations/deep-learning* *Sep 2018*
- Lisbon Machine Learning School *(LxMLS 2016)* *Jul 2016*
- Deep Learning workshop at *MindLab*, *Universidad Nacional de Colombia* *Jan 2016*
- Machine Learning, *www.coursera.org* *May 2014*
- Functional Programming Principles in Scala, *www.coursera.org* *Dec 2013*
- 2013 OSG User School, High-Throughput Computing Systems, *University of Wisconsin—Madison* *Jun 2013*

Pulications

1. **Suraj Maharjan**, Manuel Montes-y-Gómez, Fabio A. Gonzalez, and Thamar Solorio. A Genre-Aware Attention Model to Improve the Likability Prediction of Books. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium. Association for Computational Linguistics.

2. **Suraj Maharjan**, Sudipta Kar, Manuel Montes-y-Gómez, Fabio A. Gonzalez, and Thamar Solorio. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

3. **Suraj Maharjan**, John Arevalo, Manuel Montes-y-Gómez, Fabio A. Gonzalez, and Thamar Solorio. A Multi-task Approach to Predict Likability of Books. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Volume 1, Long Papers, pages 1217–1227, Valencia, Spain, April 2017. Association for Computational Linguistics.

4. Sudipta Kar, **Suraj Maharjan**, Thamar Solorio. Folksonomication: Predicting Tags for Movies from Plot Synopses using Emotion Flow encoded Neural Network. In the Proceedings of COLING 2018, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics.

5. Sudipta Kar, **Suraj Maharjan**, A. Pastor López-Monroy, and Thamar Solorio. MPST: A corpus of movie plot synopses with tags. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), May 2018.

6. **Suraj Maharjan**, Prasha Shrestha, Katherine Porterfield, Dustin Arendt and Svitlana Volkova. Towards Anticipatory Analytics: Forecasting Instability Across Countries from Dynamic Knowledge Graphs. In Proceedings of the 5th Pacific Northwest Regional NLP Workshop (NW-NLP 2018), Redmond, Washington, April 2018.

7. Deepthi Mave, **Suraj Maharjan**, Thamar Solorio. Language Identification and Analysis of Code-Switched Social Media Text. In Proceedings of The 3rd Workshop on Computational Approaches to Linguistic Code-switching, July 2018, Melbourne, Australia. Association for Computational Linguistics.

8. Gustavo Aguilar, **Suraj Maharjan**, A. Pastor López-Monroy, and Thamar Solorio. A Multi-task Approach for Named Entity Recognition in Social Media Data. In Proceedings of The 3rd Workshop on Noisy User-generated Text (W-NUT), Copenhagen, Denmark.

9. Niloofar Safi Samghabadi, **Suraj Maharjan**, Alan Sprague, Raquel Diaz-Sprague and Thamar Solorio. Detecting Nastiness in Social Media. In Proceedings of the 1st Workshop on Abusive Language Online (ALW1).

10. **Suraj Maharjan**, Sudipta Kar, and Thamar Solorio. RiTUAL-UH at SemEval-2017 Task 5: Sentiment Analysis on Financial Data Using Neural Networks. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017), Vancouver, Canada.

11. Mohammed Attia, **Suraj Maharjan**, Younes Samih, Laura Kallmeyer, and Thamar Solorio. CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings. In Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V).

12. Younes Samih, **Suraj Maharjan**, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. Multilingual Code-switching Identification via LSTM Recurrent Neural Networks. In Proceedings of the Second Workshop on Computational Approaches to Code Switching.

13. **Suraj Maharjan**, and Thamar Solorio. Using Wide Range of Features for Author profiling. In *Notebook for PAN at CLEF 2015*, Toulouse, France, 2015.

14. **Suraj Maharjan**, Elizabeth Blair, Steven Bethard, and Thamar Solorio. Developing Language-tagged Corpora for Code-switching Tweets. In The 9th Linguistic Annotation Workshop held in conjunction with NAACL 2015.

15. Thamar Solorio, Elizabeth Blair, **Suraj Maharjan**, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari et al. Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching.

16. **Suraj Maharjan**, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A Straightforward Author Profiling Approach in MapReduce. In Proceedings of the 14th Ibero-American Conference on AI (IBERAMIA 2014).

17. Prasha Shrestha, **Suraj Maharjan**, Gabriela Ramírez de la Rosa, Alan Sprague, Thamar Solorio, and Gary Warner. Using String Information for Malware Family Identification. In Proceedings of the 14th Ibero-American Conference on AI (IBERAMIA 2014).

18. **Suraj Maharjan**, Prasha Shrestha, and Thamar Solorio. A Simple Approach to Author Profiling in MapReduce. In *Notebook for PAN at CLEF 2014*, Sheffield, UK, 2014.

19. Prasha Shrestha, **Suraj Maharjan**, and Thamar Solorio. Machine Translation Evaluation Metric for Text Alignment. In *Notebook for PAN at CLEF 2014*, Sheffield, UK, 2014.

20. Rajendra Banjade and **Suraj Maharjan**. Product Recommendations using Linear Predictive Modeling. In Proceedings of the Second Asian Himalayas International Conference on Internet AH-ICI 2011, November 4-6, 2011, Nepal.