Shubhanshu Mishra

shubhanshu.com · shubhanshumishra · napsternxg · 013OA04AAAAJ
shubhanshu.mishra@gmail.com

Experience

Doordash USA
Staff Machine Learning Engineer, Tech Lead Search *Dec 2023 - Present*

- Developed first foundational content understanding model at Doordash using language models and heterogenous graphs.
- Leading ML efforts for improving autocomplete, search result ranking.

Instacart USA
Machine Learning Engineer (L6), Search Machine Learning *Feb 2023 - Dec 2023*

- Developed LLM based Question Answering. Deployed in 2 months. **10x** cost reduction and improved QnA content moderation approval. Drove adoption of QnA artifacts across additional product surfaces.
- Leading ML efforts for AskInstacart Conversational Search. Reduced costs by **90%**.
- Developed Prompt Engineering and Evaluation framework supporting LLM APIs. Used in 4+ projects.
- Developed multi-modal entity search. Won **best ML Innovation** & **best accessibility feature**.
- Developed recipe ingredient extraction and product retrieval using search logs (10% improvement).
- Developed query tagging and understanding models.
- Filed 5 patents.

Twitter, Inc. USA
Senior Machine Learning Researcher, Content Understanding Research *Aug 2019 - Jan 2023*

- Improved candidate generation for Home Timeline (**+8.5M** UAM) and Notifications (**+300K** mDAU).
- Developed contextual language models which utilize spatio-temporal and social graph context.
- Led entity linking project with new model and service, released public datasets & papers.
- Developed python demo and serving library. Used for 20+ demos and 1 shipped project.
- Improved ads classification, misinformation claim matching, query expansion, and multi-lingual NER.
- Worked on bias assessement in NER, and image cropping algorithm (200+ users).
- Mentored 4 interns with projects deployed and/or published.
- Published 8 research papers at Neurips, EMNLP, CSCW, and AKBC.

Twitter, Inc. USA
Software Engineering Intern, Content Understanding and Applied Deep-learning *Jun 2018 - Aug 2018*

University of Illinois at Urbana-Champaign USA
Research Assistant, Information Extraction from Networks and Texts *Aug 2013 - July 2019*

Citrix India
Software Engineer, NetScaler Infra Team *Jul 2012 - Jul 2013*

Improved authentication and authorization for NetScaler and developed a real time collaborative canvas app.

Barclays Capital Singapore
Global Technology Analyst, Commodities *May 2011 - Jul 2011*

Global Venture Lab Finland
Lead Web Developer *Dec 2009 - Jan 2010*

National University of Singapore Singapore
Research Assistant at Institute of Systems Science *May 2009 - Jul 2009*

Skills

Machine Learning: Numpy, Tensorflow, PyTorch, Transformers, spaCy, SciKit-Learn
Data: SQL, BigQuery, Google Cloud Storage, Hadoop, Apache Spark, Dataflow, Elasticsearch, Snowflake
Infra: Linux, Docker, Windows, AWS, GCP
Programming: Python, Javascript, Java, HTML, CSS, C, Scala, PHP, Rust

Education

University of Illinois at Urbana-Champaign USA
Doctor of Philosophy (Ph.D.) Library and Information Science *Aug 2013 - May 2020*
Thesis: Information extraction from digital social trace data with applications to social media and scholarly communication data

- Social Media Information Extraction - Multi-task learning for Tagging, and Classification.
- PyTAIL - Interactive and Incremental Learning of NLP Models with Human in the Loop.
- Profiling authors and articles based on novelty, expertise and self-citation
- ConText - Tool for extracting and analyzing network data from text

Indian Institute of Technology Kharagpur India
Bachelors and Masters in Science Mathematics and Computing *Jul 2007 - May 2012*
Thesis: Analysis of Social Media Data to determine Positive and Negative Influential Nodes in the Network

Selected Publications

- **S. Mishra**, A. Saini, R. Makki, S. Mehta, A. Haghighi and A. Mollahosseini, "TweetNERD – End to End Entity Linking Benchmark for Tweets", in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2 (NeurIPS Datasets and Benchmarks 2022), arXiv, 2022
- R. Eskander, **S. Mishra**, S. Mehta, S. Samaniego and A. Haghighi, "Towards improved distantly supervised multilingual named-entity recognition for tweets", in Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), Association for Computational Linguistics, 2022, pp. 115–124
- J. Li, **S. Mishra (equal)**, A. El-Kishky, S. Mehta and V. Kulkarni, "NTULM: Enriching social media text representations with non-textual units", in Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), Association for Computational Linguistics, 2022, pp. 69–82
- **S. Mishra** and A. Haghighi, "Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction", in Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, 2021, pp. 381–388
- K. Yee, U. Tantipongpipat and **S. Mishra (equal)**, "Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency", Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, pp. 1–24, 2021
- **S. Mishra** and J. Diesner, "Semi-supervised Named Entity Recognition in noisy-text", in Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, 2016, pp. 203–212

Awards & Recognition

Impact Recognition Award - ACM CSCW	Oct 2021
Best Poster Award - UIUC Student Poster Session	Mar 2020
Best student paper award - ASIST SIGMET Workshop	Nov 2018
Graduate Teacher Certificate	May 2018
University of Illinois GIS Day Runner-up (Research Quality)	Nov 2017
Kishore Vaigyanik Protsahan Yojana Scholar	2007-2012
3rd rank in Regional Mathematics Olympiad, Uttar Pradesh, India	Dec 2006

Teaching

Tutorial presenter, Multiple venues	Sep 2019 - Current
Tutorial on hands on advanced machine learning for information extraction from tweets tasks, data, and open source tools. Details at: https://socialmediaie.github.io/tutorials/	
Co-instructor - Network Analysis	Spring 2018
Teaching Assistant - Network Analysis	Summer 2017
Teaching Assistant - Foundations of Information Processing	Spring 2017
Co-instructor - Data Mining Applications	Fall 2016

Listed in Teachers Ranked as Excellent!

All Publications scholar.google.com/citations?user=013OA04AAAAJ

[1] R. Eskander, **S. Mishra**, S. Mehta, S. Samaniego and A. Haghighi, "Towards improved distantly supervised multilingual named-entity recognition for tweets", in Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), Association for Computational Linguistics, 2022, pp. 115–124.

[2] R. Eskander, **S. Mishra**, S. Mehta, S. Samaniego and A. Haghighi, "Towards improved distantly supervised multilingual named-entity recognition for tweets", in Weak, Indirect and Self Supervision for Knowledge Extraction, (Non-Archival), 2022.

[3] J. A. Fries, L. Weber, N. Seelam *et al.*, "Bigbio: A framework for data-centric biomedical natural language processing", in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2 (NeurIPS Datasets and Benchmarks 2022), arXiv, 2022.

[4] L. Hebert, R. Makki, S. Mishra, H. Saghir, A. Kamath and Y. Merhav, "Robust candidate generation for entity linking on short social media texts", in Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), Association for Computational Linguistics, 2022, pp. 83–89.

[5] J. Li, **S. Mishra (equal)**, A. El-Kishky, S. Mehta and V. Kulkarni, "NTULM: Enriching social media text representations with non-textual units", in Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), Association for Computational Linguistics, 2022, pp. 69–82.

[6] **S. Mishra** and J. Diesner, "PyTAIL: Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data", in Human in the Loop Learning (HiLL) Workshop at NeurIPS 2022, arXiv:2211.13786 [cs], arXiv, 2022.

[7] S. Mishra, A. Saini, R. Makki, S. Mehta, A. Haghighi and A. Mollahosseini, "Tweetnerd - end to end entity linking benchmark for tweets", in Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 1419–1433.

[8] ——, "TweetNERD – End to End Entity Linking Benchmark for Tweets", in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2 (NeurIPS Datasets and Benchmarks 2022), arXiv, 2022.

[9] B. Workshop, : T. L. Scao *et al.*, "Bloom: A 176b-parameter open-access multilingual language model", 2022. arXiv: `2211.05100 [cs.CL]`.

[10] V. Kulkarni, **S. Mishra** and A. Haghighi, "LMSOC : An Approach for Socially Sensitive Pretraining", in Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, 2021, pp. 2967–2975.

[11] **S. Mishra**, "Information extraction from digital social trace data with applications to social media and scholarly communication data", SIGWEB Newsl., vol. 2021, no. Spring, 2021.

[12] **S. Mishra** and A. Haghighi, "Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction", in Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, 2021, pp. 381–388.

[13] S. Mishra, S. Prasad and **S. Mishra**, "Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media", SN Computer Science, vol. 2, no. 2, p. 72, 2021.

[14] K. Yee, U. Tantipongpipat and **S. Mishra (equal)**, "Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency", Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, pp. 1–24, 2021.

[15] K. Han, P. Yang, **S. Mishra** and J. Diesner, "WikiCSSH: Extracting Computer Science Subject Headings from Wikipedia", in Workshop on Scientific Knowledge Graphs (SKG 2020), 2020.

[16] S. Mishra, "Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data", ACM SIGIR Forum, vol. 54, no. 1, 2020.

[17] ——, 'Information extraction from digital social trace data with applications to social media and scholarly communication data,' Ph.D. Dissertation, University of Illinois at Urbana-Champaign, 2020.

[18] ——, "Non-neural Structured Prediction for Event Detection from News in Indian Languages", in Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, 2020.

[19] **S. Mishra** and D. Collier, "A Framework for Generating Annotated Social Media Corpora with Demographics, Stance, Civility, and Topicality", SSRN Electronic Journal, 2020.

[20] **S. Mishra**, S. He and L. Belli, "Assessing Demographic Bias in Named Entity Recognition", in Bias in Automatic Knowledge Graph Construction - A Workshop at AKBC 2020, 2020. arXiv: `2008.03415`.

[21] **S. Mishra** and S. Mishra, "Scubed at 3C task A - A simple baseline for citation context purpose classification", in Proceedings of the 8th International Workshop on Mining Scientific Publications, Association for Computational Linguistics, 2020, pp. 59–64.

[22] ——, "Scubed at 3C task B - A simple baseline for citation context influence classification", in Proceedings of the 8th International Workshop on Mining Scientific Publications, Association for Computational Linguistics, 2020, pp. 65–70.

[23] S. Mishra, S. Prasad and **S. Mishra**, "Multilingual Joint Fine-tuning of Transformer models for identifying Trolling, Aggression and Cyberbullying at TRAC 2020", in Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), 2020, pp. 120–125.

[24] N. N. Parulian, T. Lu, S. Mishra, M. Avram and J. Diesner, "Effectiveness of the Execution and Prevention of Metric-Based Adversarial Attacks on Social Network Data †", Information, vol. 11, no. 6, p. 306, 2020.

[25] M. V. Avram, S. Mishra, N. N. Parulian and J. Diesner, "Adversarial perturbations to manipulate the perception of power and influence in networks", in 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2019, pp. 986–993.

[26] D. Collier, S. Mishra, D. Houston, B. Hensley, S. Mitchell and N. Hartlep, "Who is Most Likely to Oppose Federal Tuition-Free College Policies? Investigating Variable Interactions of Sentiments to America's College Promise", SSRN Electronic Journal, 2019.

[27] D. A. Collier, S. Mishra, D. A. Houston, B. O. Hensley and N. D. Hartlep, "Americans 'support' the idea of tuition-free college: an exploration of sentiment and political identity signals otherwise", Journal of Further and Higher Education, vol. 43, no. 3, pp. 347–362, 2019.

[28] S. Mishra, "Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets", in Proceedings of the 30th ACM Conference on Hypertext and Social Media - HT '19, ACM Press, 2019, pp. 283–284.

[29] S. Mishra and J. Diesner, "Capturing Signals of Enthusiasm and Support Towards Social Issues from Twitter", in Proceedings of the 5th International Workshop on Social Media World Sensors - SIdEWayS'19, ACM Press, 2019, pp. 19–24.

[30] S. Mishra and S. Mishra, "3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages", in Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, 2019, pp. 208–213.

[31] S. Mishra and J. Diesner, "Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora", in Proceedings of the 29th on Hypertext and Social Media - HT '18, ACM Press, 2018, pp. 2–10.

[32] S. Mishra, B. D. Fegley, J. Diesner and V. I. Torvik, "Expertise as an aspect of author contributions", in Metrics 2018: Workshop on Informetric and Scientometric Research (SIG/MET), 2018.

[33] ——, "Self-citation is the hallmark of productive authors, of any gender", PLoS ONE, vol. 13, no. 9, e0195773, 2018.

[34] A. Addawood, R. Rezapour, S. Mishra, J. Schneider and J. Diesner, "Developing an Information Source Lexicon", in Prioritising Online Content workshop co-located at NIPS, 2017.

[35] S. Mishra, "SCTG: Social Communications Temporal Graph – A novel approach to visualize temporal communication graphs from social data", in UIUC Data Science Day, 2017.

[36] **S. Mishra** and J. Diesner, "Semi-supervised Named Entity Recognition in noisy-text", in Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, 2016, pp. 203–212.

[37] S. Mishra and V. I. Torvik, "Quantifying Conceptual Novelty in the Biomedical Literature." D-Lib magazine : the magazine of the Digital Library Forum, vol. 22, no. 9-10, 2016.

[38] S. Mishra, J. Diesner, J. Byrne and E. Surbeck, "Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization", in Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15, ACM Press, 2015, pp. 323–325.

[39] S. Mishra, S. Agarwal, J. Guo, K. Phelps, J. Picco and J. Diesner, "Enthusiasm and support: alternative sentiment classification for social movements on social media", in Proceedings of the 2014 ACM conference on Web science - WebSci '14, ACM Press, 2014, pp. 261–262.