Madhumita Sushil, PhD

San Francisco, CA 94110 • madhumita.sushil.k@gmail.com
`https://madhumitasushil.github.io/` • `https://github.com/MadhumitaSushil/`

Experience

University of California, San Francisco — San Francisco, CA
Postdoctoral researcher — Mar '21 – present
Mentor: Atul Butte, MD, PhD (Atul.Butte@ucsf.edu)

- Trained a BERT model from scratch on 40 billion tokens from electronic health records.
- Created a benchmark for evaluating LLMs for oncology (NIH grant co-lead for this research).
- Supervised creation of guidelines for effective use of generative LLMs in medicine.
- Supervised investigation of socioeconomic conditions and disparities from discussions in clinical notes.
- Patent pending for FDA collaboration on use of LLMs for adverse event detection from clinical notes.

Google (Brain Applied team) — Zürich, Switzerland
Research intern — Dec '19 – Mar '20
Host: André Susano Pinto (andresp@google.com)

- Investigated reasoning skills of BERT language models.
- Neural network training on distributed GPU and TPU infrastructure using xManager.
- Developed retrieval-augmented natural language inference methods.
- Contributed to the open source code base for BERT language models in Tensorflow 2.
- Added question answering datasets to open source Tensorflow datasets.
- Evaluated new tensorflow-hub LLMs on natural language inference tasks.

German Research Center for Artificial Intelligence — Germany
Research assistant — Apr '16 – Dec '17
Mentor: Günter Neumann, PhD (neumann@dfki.de)

- Machine learning models for natural language inference, using word embeddings for text alignment.

Education

University of Antwerp — Belgium
PhD, Computational Linguistics — Apr '16 – Mar '21
Thesis: Exploring and Understanding Neural Models for Clinical Tasks
Topics: Neural network interpretability, Representation learning, Retrieval-augmented inference
Advisors: Prof. Walter Daelemans (Walter.Daelemans@uantwerpen.be)
Simon Šuster, PhD (simon.suster@unimelb.edu.au)

Saarland University — Germany
MSc, Language Science and Technology — Oct '13 – Feb '16
Thesis: Recognizing Textual Entailment
Relevant Coursework: Language modeling, Pattern recognition and machine learning, Statistical NLP

VIT University, Vellore — India
BTech, Computer Science and Engineering — July '09 – Mar '13
Student Achiever '12, Finalist in Intel India Embedded Challenge '12.
Internship project: Developing financial question answering system for an incubating startup

Skills

Frameworks: Pytorch, Tensorflow, Scikit-learn, Pandas, SQL, PySpark, Numpy, Scipy
Infrastructure: Unix-based SGE high performance computing cluster, AWS, XManager

Research and Travel Grants

National Cancer Institute, National Institute of Health Sept '23 – May '24
Co-lead, Large language models for treatment and pathology data extraction ($300,000)
Grant number P30CA082103, supplement award to Helen Diller Comprehensive Cancer Center

Merck Feb '24 – Mar '25
Co-Investigator, LLMs for investigating real-world evidence for MASH liver disease.
Study ID: NIS103452

Google Inc. Apr '20
Google Cloud Platform research credit ($1000 in kind).

Dutch Research Agenda
Collaborator, Responsible use of free text notes in electronic health records to improve medical prediction research.
Grant number NWA.1418.22.008 (no salary support)

Google intern travel grant for Grace Hopper Celebration Dec '19

Mentorship and Thesis Committee

UC Berkeley MEng Capstone project Sept '22 – May '23
LLMs for information extraction from breast cancer pathology reports. *Fung Institute Mission Award.*

Rotation and PhD students June '21 – present
Reinforcement learning with AI feedback for oncology note inference.
Training a Mamba model on UCSF clinical notes.
Investigating concepts encoded within clinical LLM representations.
Large language models for clinical note inference.
Social determinants of health inference from social work notes.

Google Summer of Code: CLiPS research center May '19 – Aug '19
Adversarial approaches for gender and racial bias recognition in neural networks.
    `https://github.com/clips/gsoc2019_bias`.

Master's thesis committee 2019
Extracting Drug, Reason, and Duration Mentions from Clinical Text Data — a comparison of approaches

Activities

Student board, European Association of Computational Linguistics 2019 – 2020
Co-organizer, student research workshop at EACL '21.

Journal Reviewing
Journal of American Medical Informatics Association
Journal of Biomedical Informatics
BMC Medical Informatics

Conference and Workshop Reviewing
Biomedical Natural Language Processing (BioNLP) workshop at ACL
Workshop on Machine Learning for Health at NeurIPS (Excellent reviewer, 2019)
Widening NLP Workshop at ACL

Student Research Workshop at ACL
AMIA Annual Symposium
AMIA Informatics Summit
International Conference on Information Management and Big Data (SIMBig)

Invited Talks

5th National Big Data Health Science Conference, South Carolina	2024
Clinical Informatics - Data Science Pathway seminar series, UCSF	2023
UCSF-Stanford Center of Excellence in Regulatory Science and Innovation (CERSI) EHR training series	2023
Computer Science and Engineering coursework at VIT University, India	2022, 2023
Blackbox@NL: Dutch workshop on interpretation of neural network	2019
3rd Google NLP Summit, Switzerland	2019

<div align="center">

Peer-reviewed Publications

</div>

[1] **Madhumita Sushil**, Vanessa E. Kennedy, Divneet Mandair, Brenda Y. Miao, Travis Zack*, and Atul J. Butte*. CORAL: Expert-curated medical oncology reports to advance language model inference. *New England Journal of Medicine (NEJM)-AI*, 2024.

[2] Anna L Silverman*, **Madhumita Sushil***, Balu Bhasuran*, Dana Ludwig, James Buchanan, Rebecca Racz, Mahalakshmi Parakala, Samer El-Kamary, Ohenewaa Ahima, Artur Belov, Lauren Choi, Monisha Billings, Yan Li, Nadia Habal, Qi Liu, Jawahar Tiwari, Atul J. Butte, and Vivek A. Rudrapatna. Algorithmic identification of treatment-emergent adverse events from clinical notes using large language models: a pilot study in inflammatory bowel disease. *Journal of Clinical Pharmacology and Therapeutics*, 2024.

[3] **Madhumita Sushil**, Atul J. Butte, Ewoud Schuit, Maarten van Smeden, and Artuur M. Leeuwenberg. Cross-institution natural language processing for reliable clinical association studies: a methodological exploration. *Journal of Clinical Epidemiology*, page 111258, 2024.

[4] Shenghuan Sun, Travis Zack, Christopher Y K Williams, **Madhumita Sushil***, and Atul J Butte*. Topic modeling on clinical social work notes for exploring social determinants of health factors. *JAMIA Open*, 7(1):ooad112, 01 2024.

[5] Brenda Y Miao, **Madhumita Sushil**, Ava Xu, Michelle Wang, Douglas Arneson, Ellen Berkley, Meera Subash, Rohit Vashisht, Vivek Rudrapatna, and Atul J Butte. Characterisation of digital therapeutic clinical trials: a systematic review with natural language processing. *The Lancet Digital Health*, 6(3):e222–e229, 2024.

[6] Michelle Wang, **Madhumita Sushil**, Brenda Y Miao, and Atul J Butte. Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data. *Journal of the American Medical Informatics Association*, 30(7):1323–1332, 05 2023.

[7] **Madhumita Sushil**, Simon Šuster, and Walter Daelemans. Are we there yet? Exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 41–53, Online, June 2021. Association for Computational Linguistics.

[8] **Madhumita Sushil**, Simon Šuster, and Walter Daelemans. Contextual explanation rules for neural clinical classifiers. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 202–212, Online, June 2021. Association for Computational Linguistics.

[9] **Madhumita Sushil**, Simon Šuster, and Walter Daelemans. Rule induction for global explanation of trained models. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 82–97. Association for Computational Linguistics, 2018.

[10] Simon Šuster, **Madhumita Sushil**, and Walter Daelemans. Revisiting neural relation classification in clinical notes with external information. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 22–28. Association for Computational Linguistics, 2018.

[11] **Madhumita Sushil**, Simon Šuster, Kim Luyckx, and Walter Daelemans. Patient representation learning and interpretable evaluation using clinical notes. *Journal of Biomedical Informatics*, 84:103 – 113, 2018.

[12] **Madhumita Sushil**, Simon Šuster, Kim Luyckx, and Walter Daelemans. Unsupervised patient representations from clinical notes with interpretable classification decisions. *Workshop on Machine Learning for Health, NeurIPS, arXiv preprint arXiv:1711.05198*, 2017.

[13] Elyne Scheurwegs, **Madhumita Sushil**, Stéphan Tulkens, Walter Daelemans, and Kim Luyckx. Counting trees in random forests: Predicting symptom severity in psychiatric intake reports. *Journal of Biomedical Informatics*, 75:S112 – S119, 2017. A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[14] Neha Tekriwal, **Madhumita Sushil**, and P. Venkata Krishna. Integration of safety and smartness using cloud services: An insight to future. In Khaled Elleithy and Tarek Sobh, editors, *Innovations and Advances in Computer, Information, Systems Sciences, and Engineering*, pages 293–303, New York, NY, 2013. Springer New York.

Preprints

[1] **Madhumita Sushil***, Travis Zack*, Divneet Mandair*, Zhiwei Zheng, Ahmed Wali, Yan-Ning Yu, Yuwei Quan, and Atul J. Butte. A comparative study of zero-shot inference with large language models and supervised modeling in breast cancer pathology classification. *Computing Research Repository, arXiv:2401.13887 (Under review at JAMIA)*, 2024.

[2] Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, **Madhumita Sushil**, Atul J. Butte, and Ahmed Alaa. Large language models as agents in the clinic. *Computing Research Repository, arXiv:2309.10895*, 2023.

[3] Christopher Y.K. Williams, Travis Zack, Brenda Y. Miao, **Madhumita Sushil**, Michelle Wang, and Atul J. Butte. Assessing clinical acuity in the emergency department using the gpt-3.5 artificial intelligence model. *medRxiv*, 2023.

[4] Shenghuan Sun, Travis Zack, Christopher Y. K. Williams, Atul J. Butte, and **Madhumita Sushil**. Revealing the impact of social circumstances on the selection of cancer therapy through natural language processing of social work notes. *Computing Research Repository, arXiv:2306.09877*, 2023.

[5] **Madhumita Sushil**, Dana Ludwig, Atul J. Butte, and Vivek A. Rudrapatna. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. *Computing Research Repository, arXiv:2210.06566*, 2022.

[6] Simon Šuster, **Madhumita Sushil**, and Walter Daelemans. Why can't memory networks read effectively? *Computing Research Repository, arXiv:1910.07350*, 2019.

Conference Abstracts

[1] **Madhumita Sushil**, Brenda Miao, Divneet Mandair, Travis Zack*, and Atul J. Butte*. Large language models are zero-shot oncology information extractors. *American Medical Informatics Association (AMIA) Annual Symposium proceedings*, 2023.

[2] Travis Zack, **Madhumita Sushil**, Brenda Miao, Arda Demirci, Corryn Ksapp, Atul J. Butte*, and Eric Collisson*. Clinical inference of cancer trajectory from radiology reports using ChatGPT. *American Medical Informatics Association (AMIA) Annual Symposium proceedings*, 2023.

[3] Anna L Silverman*, **Madhumita Sushil***, Balu Bhasuran*, Dana Ludwig, James Buchanan, Rebecca Racz, Mahalakshmi Parakala, Samer El-Kamary, Ohenewaa Ahima, Artur Belov, Lauren Choi, Monisha Billings, Yan Li, Nadia Habal, Qi Liu, Jawahar Tiwari, Atul Butte, and Vivek Rudrapatna. Algorithmic identification of treatment-emergent adverse events from clinical notes using large language models: A pilot study in inflammatory bowel disease. *Official journal of the American College of Gastroenterology| ACG*, 2023.

[4] **Madhumita Sushil**, Dana Ludwig, Atul J. Butte, and Vivek A. Rudrapatna. Training a transferrable clinical language model from 75 million notes. *American Medical Informatics Association (AMIA) Annual Symposium proceedings*, 2022.

[5] Shenghuan Sun, Atul J. Butte, and **Madhumita Sushil**. Predicting the cancer therapy regimen from social work notes using natural language processing. *AMIA NLP Working Group Pre-symposium*, 2022.

[6] Shenghuan Sun, Atul J. Butte, and **Madhumita Sushil**. Topic modeling on social work notes for exploring social determinants of health factors. *International Society for Pharmacoeconomics and Outcomes Research (ISPOR) proceedings*, 2022.