Madhumita Sushil

+1-4156152965
madhumita.sushil.k@gmail.com
madhumitasushil.github.io

My research interest is to develop generalizable and fair foundation models.

Education

Sep '16 – Mar '21 **PhD in Computational Linguistics**
University of Antwerp, Belgium.
Thesis — Exploring and Understanding Neural Models for Clinical Tasks
- Developed methods for patient-level representation learning from clinical notes.
- Developed methods for interpreting causal decision patterns learned by deep neural networks.
- Investigated retrieval-augmented language models for natural language inference.

Advisors: Prof. Dr. Walter Daelemans, Dr. Simon Šuster

Oct '13 – Feb '16 **Master of Science in Language Science and Technology**
Saarland University, Germany, *1.5/5 (1.0 highest, lower is better)*.
Thesis — Recognizing Textual Entailment
- Developed subword distance-based and embedded vectors-based lexical alignment algorithms to align text and hypothesis segments for textual entailment classification.

Advisors: Prof. Dr. Günter Neumann, Prof. Dr. Dietrich Klakow

July '09 – May '13 **Bachelor of Technology in Computer Science and Engineering**
VIT University, Vellore, India, *8.98/10*.

Work Experience

Mar '21 – present **Postdoctoral Scholar**
University of California San Francisco (UCSF), San Francisco, USA.
Advisor: Prof. Dr. Atul Butte
- Developed comprehensive annotation schemas and benchmarking datasets for oncology to extensively evaluate generative language models for oncology information extraction. Awarded an NIH-NCI grant as a co-lead for this work. PI of a non-expiring IRB for this research.
- In collaboration with the US-FDA, developed an end-to-end pipeline for serious adverse event detection from clinical notes, resulting into a provisional patent application with the USPTO (Serial No. 63/465,382). Developed a UCSF-BERT language model from 75 million de-identified clinical notes to facilitate this pipeline.
- Developed a pipeline to extract social determinants of health factors extraction from clinical notes, and used it to assess the reliability of NLP outputs in clinical observational studies.
- Mentored MEng students on pathology information extraction, and co-mentored a PhD student on inferring clinical disparity from social work notes at UCSF.

Dec '19 – Mar '20 **Research internship**
Google Brain Applied team, Zürich, Switzerland.
Host: André Susano Pinto
- Analyzed inductive bias in BERT representations towards linguistic reasoning skills.

Apr '16 – Dec '17 **Junior Research Developer**
Antwerp University Hospital, Antwerp, Belgium.
Natural Language Processing for clinical applications (Project *Accumulate* funded by VLAIO, Belgium).
- Developed techniques for unsupervised patient representation learning with gradient-based analysis for model interpretability.
- Developed classifiers for automated psychiatric symptom severity identification.

Dec '13 – Nov '15 **Research Assistant**
German Research Center for Artificial Intelligence, Saarbrücken, Germany.
- Designed and implemented a textual entailment engine for English (Project *Excitement* funded by the EU).
- Designed and developed the website: `http://www.qt21.eu`.

Jan '13 – May '13 **SDE Intern**
TCorpus Analytics, Technology Business Incubator, Vellore, India.
- Designed and implemented a factual question answering platform for financial textual reports.

Peer-reviewed Publications

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.

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.

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.

Brenda Y Miao, **Madhumita Sushil**, Ava Xu, Michelle Wang, Douglas Arneson, Ellen Berkley, Meera Subash, Rohit Vashisht, Vivek Rudrapatna, and Atul J. Butte. Characterization of digital therapeutic clinical trials: a systematic review using natural language processing. *Lancet Digital Health*, 2023.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Conference Abstracts

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.

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.

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.

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.

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.

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.

Preprints

Madhumita Sushil*, Travis Zack*, Divneet Mandair*, Zhiwei Zheng, Ahmed Wali, Yan-Ning Yu, Yuwei Quan, and Atul J. Butte. Developing a general-purpose clinical

language inference model from a large corpus of clinical notes. *Computing Research Repository, arXiv:2401.13887*, 2024.

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. *medRxiv (Under revision at the Journal of Clinical Pharmacology and Therapeutics)*, 2023.

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.

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

Awards and Recognition

Research grant Co-lead, National Cancer Institute (NCI) real-world data/large language model administrative supplement award to the Helen Diller Family Comprehensive Cancer Center (HDFCCC), UCSF, '23–'24.
Google Cloud Platform research credit grant (in kind, retail value $1,000), '20.
Collaborator (no salary support), Dutch Research Agenda grant on *Responsible and valid use of free text notes in electronic health records to improve medical prediction research*.

Excellent reviewer Machine Learning for Health workshop, NeurIPS '19 (among top 5% of the reviewers).

Travel grant Google intern travel scholarship for Grace Hopper Celebration '19.

Fellowship International Max Planck Research School for Computer Science (IMPRS-CS) PhD fellowship, Saarbrücken, '15. *Declined* in favour of clinical NLP research in Antwerp.

Student achiever VIT University, '12.

Finalist One of the top 22 teams across India in the *Intel India Embedded Challenge* '12 for the project *Smartphone for the visually impaired*.

Winner Hackathon, *Exebit* '12, IIT Madras.

Activities and Service

Mentoring MEng Capstone project, University of California, Berkeley, '22-'23. *Winner of the Fung Institute Mission Award.*
Google Summer of Code '19 project on bias identification in machine learning models: `https://github.com/clips/gsoc2019_bias`.

Guest lecture 5th National Big Data Health Science Conference, South Carolina, '24.
Clinical Informatics - Data Science Pathway seminar series, UCSF, '23.
UCSF-Stanford Center of Excellence in Regulatory Science and Innovation (CERSI) EHR training series, '23.
Computer Science and Engineering coursework, VIT University, '22, '23.
Blackbox@NL: Dutch workshop on interpretation of neural networks, Den Bosch, '19.

Thesis committee	Assessment of MA thesis in Computational Linguistics by Jens Lemmens, '19. Title: *Extracting Drug, Reason, and Duration Mentions from Clinical Text Data — a comparison of approaches.*
Program committee / Reviewer	Journal of American Medical Informatics Association '24. Journal of Biomedical Informatics '23. Biomedical Natural Language Processing (BioNLP) workshop at ACL '22, '23. AMIA Informatics Summit '23, AMIA Annual Symposium '22, '23. BMC Medical Informatics '22. International Conference on Information Management and Big Data (SIMBig) '21, '23. Workshop on Machine Learning for Health at NeurIPS '17, '18, '19. Widening NLP Workshop at ACL '19, '20. Student Research Workshop at ACL '19.
Invited poster	3rd Google NLP Summit, Zürich '19. Title: *Rule induction for global explanation of neural classifiers*.
Student board	European Association of Computational Linguistics (EACL) '19 – '20.
Virtual training	Google's invite-only *Get Ahead* virtual technical development program, May – Jun '19.
Summer school	Lisbon Machine Learning school (LxMLS) '16.
Organizer	Student Research Workshop, EACL '21. ATILA Workshop '19: `https://www.clips.uantwerpen.be/atila19`.
Volunteer	Chair, NLP Methods session, AMIA '22. Mentor to an under-privileged girl in rural India through the initiative *e-shishya*, Jul – Dec '18. Interspeech '15.