Avisha Das

dasavisha10@gmail.com ◇ dasavisha.github.io ◇ Google Scholar Profile

RESEARCH INTERESTS

Applications of Natural Language Processing (with a focus on Deep Learning) to Security Analytics and Biomedical/Clinical Informatics.

EDUCATION

Ph.D. in Computer Science 2014 – 2020
University of Houston, Houston, TX
Dissertation Title: Proactive Defense through Automated Generation of Targeted Attacks

B.Tech. in Electronics and Communication Engineering 2010 – 2014
West Bengal University of Technology, Kolkata, India

EXPERIENCE

Research Fellow November 2023 – Present
Arizona Advanced AI & Innovation (A3I) Hub
Mayo Clinic Arizona, Phoenix, AZ

Postdoctoral Research Fellow April 2021 – November 2023
School of Biomedical Informatics
University of Texas Health Science Center (UTHealth), Houston, TX

Data Science-NLP Intern May 2019 – August 2019
Occidental (Oxy) Petroleum Corporation
The Woodlands, TX

Summer Research Intern June 2018 – August 2018
Production Solutions Team, Halliburton Energy Services
Houston, TX

Data Science Intern June 2017 – August 2017
2H Offshore Inc.
Houston, TX

Graduate Research and Teaching Assistant August 2014 – December 2020
Department of Computer Science
University of Houston (UH), Houston, TX

TEACHING EXPERIENCE

Teaching Assistant, University of Houston

1. Artificial Intelligence (COSC 6368) [Summer 2020]
2. Software Design (COSC 4353/6353) [Spring 2020]
3. Machine Learning (COSC 6342) [Fall 2019]
4. Computer Organization and Architecture (COSC 6323) [Fall 2018]
5. Security Analytics (COSC 4397/COSC 6346) [Spring 2018, Spring 2019]
6. Software Design (COSC 4353/6353) [Fall 2017]
7. Data Structures and Algorithms (COSC 3320) [Fall 2016, Spring 2017]

Guest Lectures

1. Foundations of BMI Methods II (BMI 505), Arizona State University
   [Spring 2024, Topic: Introduction to NLP and Regular Expressions]
2. Advanced Natural Language Processing (COSC 7336), University of Houston
   [Fall 2022, Topic: Introduction to Generative Language Models]

PUBLICATIONS

Journal Papers

1. **Das, A.**, Talati, I., Manuel, J., Rubin, D., and Banerjee, I. (2025). **Weakly Supervised Language Models for Automated Extraction of Critical Findings from Radiology Reports.** *npj Digital Medicine. [IF: 15.2]*
2. Li, Z., Wei, Q., Huang, L.C., Li, J., Hu, Y., Chuang, Y.S., He, J., **Das, A.**, Keloth VK, Yang Y, and Diala CS. (2024). **Ensemble pretrained language models to extract biomedical knowledge from literature.** *Journal of the American Medical Informatics Association (JAMIA) [IF: 4.7].*
3. Yang, Y., Zuo, X., **Das, A.**, Xu, H., and Zheng, W. Jim (2024). **Representation Learning of Biological Concepts: A Systematic Review**. *Current Bioinformatics [IF: 2.4].*
4. **Das, A.** and Verma, R. (2020). **Can Machines Tell Stories? A Comprehensive Comparison of Pre-Trained and Fine-Tuned Deep Neural Language Models**. *IEEE Access [IF: 3.4].*
5. El Aassal, A., Baki, S., **Das, A.**, and Verma, R. (2020). **An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs.** *IEEE Access [IF: 3.4].*
6. **Das, A.**, Baki, S., El Aassal, A., Verma, R., and Dunbar, A. (2019). **SoK: A Comprehensive Reexamination of Phishing Research from the Security Perspective.** *IEEE Communications Surveys & Tutorials [IF: 35.6].*
7. Karimi, S., Moraes, L., **Das, A.**, Shakery, A., and Verma, R. (2018). **Citance-based retrieval and summarization using IR and machine learning.** *Scientometrics [IF: 3.8].*

Conference and Workshop Papers

8. **Das, A.**, Diala, CS., Chen, G., Li, Z., Li, R., Anjum, O., and Zheng, W. (2025). **Efficient Training Corpus Retrieval for Large Language Model Fine Tuning: A Case Study in Cancer.** *20th World Congress on Medical and Health Informatics (MedINFO).*
9. Tariq, A., **Das, A.**, Nakach, F., Yu, N., Patel, B., and Banerjee, I. (2025). **Two-phase Framework for Clinical Question-Answering – Auto-correction for Guideline-concordance.** *AAAI Workshop on Health Intelligence (W3PHIAI).*
10. Joshi, V., Correa, R., **Das, A.**, and Banerjee, I. (2025).**Multi-factor debiasing for correlating confounders for 'fair' diagnostic model.** *SPIE Medical Imaging.*
11. **Das, A.**, Tariq, A., Batalini, F., Dhara, B. and Banerjee, I. (2024). **Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer.** *AMIA Annual Symposium.*
12. **Das, A.**, Li, Z., Wei, Q., Li, J., Huang, L.C., Hu, Y., Li, R., Zheng, W. and Xu, H. (2023). **Extracting Drug-Protein Relation from Literature using Ensembles of Biomedical Transformers.** *19th World Congress on Medical and Health Informatics (MedINFO).*
13. **Das, A.**, Selek, S., Warner, A., Zuo, X., Hu, Y., Keloth, V., Li, J., Zheng, W., and Xu, H. (2022). **Conversational Bots for Psychotherapy: A Study of Generative Transformer Models Using Domain-specific Dialogue.** *ACL Workshop on Biomedical Natural Language Processing Workshop (BioNLP).*
14. **Das, A.**, Li, Z., Wei, Q., Li, J., Huang, L. C., Hu, Y., Li, R., Zheng, W., and Xu, H. (2021). **UTHealth@ BioCreativeVII: domain-specific transformer models for drug-protein relation extraction.** *Workshop on BioCreative VII Challenge Evaluation.*

15. Zeng, V., El Aassal, A., Baki, S., Verma, R., Moraes, L. and **Das, A.** (2020). **Diverse Datasets and a Customizable Benchmarking Framework for Phishing**. *ACM CODASPY International Workshop on Security and Privacy Analytics (IWSPA).*
16. **Das, A.**, and Verma, R. (2019). **Automated email Generation for Targeted Attacks using Natural Language.** *Language Resources and Evaluation-LREC Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS).*
17. El Aassal, A., Moraes, L., Baki, S., **Das, A.**, and Verma, R. (2018). **Anti-Phishing Pilot at ACM IWSPA 2018: Evaluating Performance with New Metrics for Unbalanced Datasets**. *Conference on Data and Application Security and Privacy (CODASPY) Anti-Phishing Shared Task Pilot.*
18. Verma, R., and **Das, A.** (2017, March). **What's in a URL: Fast feature extraction and malicious URL detection.** *ACM CODASPY International Workshop on Security and Privacy Analytics (IWSPA).*
19. De Moraes, L. F., **Das, A.**, Karimi, S., and Verma, R. (2018). **University of Houston@ CL-SciSumm 2018.** *SIGIR Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL).*
20. Karimi, S., Moraes, L. F., **Das, A.**, and Verma, R. (2017). **University of Houston@ CL-SciSumm 2017: Positional language Models, Structural Correspondence Learning and Textual Entailment.** *SIGIR Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL).*

Posters and Abstracts

21. Tariq, A., Luo, M., Urooj, A., **Das, A.**, Jeong, J., Trivedi, S., Patel, B. and Banerjee, I. (2024). **Domain-specific LLM Development and Evaluation–A Case-study for Prostate Cancer.** *AMIA Annual Symposium.*
22. **Das, A.**, Anjum, O., Chen, G., Zheng, W., and Li, Rongbin (2024). **Efficient Training Corpus Retrieval for Large Language Model Fine Tuning** *AMIA Informatics Summit.*
23. **Das, A.**, Anjum, O., Zheng, W., and Diala, C. (2023). **A Multi-faceted Mining Tool for Knowledge and Data Discovery for Cancer Research.** *International Conference on Intelligent Biology and Medicine (ICIBM).*
24. **Das, A.**. (2019) **AskAna: Retrieval Based Virtual Assistant for Digital Operations and Field Development.** *Rice Data Science Conference.*
25. **Das, A.**, and Verma, R. (2017). **What's in a URL: Fast Feature Extraction and Detection of Malicious URLs.** *Women in CyberSecurity (WiCyS) Conference.*
26. **Das, A.**, and Verma, R. (2016). **Analyzing Phishing URLs.** *Poster at Grace Hopper Conference for Celebration of Women.*
27. **Das, A.**, and Verma, R. (2016). **Are Legit and Phishing URLs similar? Hell No! – Lexical characterization and Analysis of URLs.** *Women in CyberSecurity (WiCyS) Conference.*
28. **Das, A.**, and Verma, R. (2016). **Studying Phishing URLs the NLP way.** *Computing Research Association (CRA-W) Grad Cohort Workshop.*

Book Chapters

29. Tariq, A., Luo, M., Urooj, A., **Das, A.**, Jeong, J., Trivedi, S., Abdul-Muhsin, H., Ghaffar, U., Yu, N., Patel, B., and Banerjee, I. (2024). **Development Of LLM For Prostate Cancer - The Need for Domain-Tailored Training.** *National Cancer Institute.*

Preprints/Under Review

30. Talati, I., **Das, A.**, Manuel, J., Rubin, D., and Banerjee, I. (2025). **Detection and Classification of Critical Findings in Radiology Reports Using Large Language Models .** *Under Review at Lancet Digital Health.*

31. Tariq, A., Luo, M., Urooj, A., **Das, A.**, Jeong, J., Trivedi, S., Patel, B. and Banerjee, I. (2024). **Domain-specific LLM Development and Evaluation–A Case-study for Prostate Cancer.** *medRxiv preprint.*
32. **Das, A.**, Tariq, A., Batalini, F., Dhara, B., and Banerjee, I. (2024). **Framework for Exposing Vulnerabilities of Clinical Large Language Model: A Case Study in Breast Cancer.** *Under Review at npj Precision Oncology.*
33. **Das, A.**, Anjum, O., Chen, G., and Zheng, W. Jim (2023). **Efficient Training Corpus Retrieval for Large Language Model Fine Tuning**. *Under Review.*
34. **Das, A.**, Jin, K., Keloth, V., Selek, S., and Xu, H. (2023). **A Methodological Review of Deep Learning-based Virtual Assistants for Healthcare**. *Under Review.*
35. **Das, A.** and Verma, R. (2020). **Modeling Coherency in Generated Emails by Leveraging Deep Neural Learners**. *ArXiv preprint.*

Submitted Grants

36. National Institute of Health (NIH) Pathway to Independence Award (K99/R00). **"Title: A Privacy-preserving Framework for Large Language Models for Clinical Use."** *Submitted, under review.*
37. Cancer Prevention and Research Institute of Texas (CPRIT)-McWilliams School of Biomedical Informatics at UTHealth Houston, Genomics and Translational Cancer Research Training Program (BIG-TCR) Postdoctoral Trainee Grant. **Title: "Building an Automated Tool for Knowledge and Data Discovery for Cancer Research: A Multi-Faceted Approach by Biomedical Literature Mining."** *2022-2024.*

INVITED TALKS

1. **Framework for Exposing Vulnerabilities of Clinical LLMs: Breast Cancer.**
   Stanford MedAI Group Exchange Sessions, Stanford University, 2024.
2. **Large language models and their application in Biomedical Domain.**
   DSICCR Tuesday Seminar Series, UTHealth Houston, 2023.
3. **Domain-specific Transformer Models for Drug-Protein Relation Extraction.**
   CPH Seminar in Precision Medicine, UTHealth Houston, 2022.
4. **Leveraging NLP for Mining Biomedical Data: Named Entity Recognition and Content Recommendation.**
   CPRIT-BIG-TCR Undergraduate Summer Internship Seminar, UTHealth Houston, 2022.
5. **Natural Language Understanding and Generation**
   Advanced Natural Language Processing Course, University of Houston, 2022.

MEDIA COVERAGE

Automated Email Generation for Targeted Attacks. AD-Tech, DataSkeptic Podcast, 2022. Link.

AWARDS, HONORS AND OTHERS

Awards and Honors

1. **CPRIT BIG-TCR Postdoctoral Training Program Fellowship**,[1] 2022-2024.
   Cancer Prevention and Research Institute of Texas, UTHealth Houston.
2. **Second place, Litcoin NLP Challenge**,[2] March 2022.
   National Center for Advancing Translational Sciences (NCAT), UTHealth Houston.
3. **Cullen Graduate Success Fellowship**, Fall 2020.
   UH Alumni Association, University of Houston.

[1] https://www.uth.edu/big-tcr/people/trainees.htm

[2] Part of the UTHealth-SBMI Team (Result)

4. **Govt. of India Merit-based Scholarship for Undergraduate Education**, 2010 -2014. Ministry of Human Resources-India (MHRD), India.

Travel Grants

1. Annual Meeting of the Association for Computational Linguistics (ACL), 2020, 2022
2. Grace Hopper Conference for Women in Computing (GHC), 2015, 2016, 2018
3. International Workshop on Security and Privacy Analytics (IWSPA), 2017, 2018
4. Empirical Methods in Natural Language Processing Conference (EMNLP), 2016
5. Women in CyberSecurity Conference (WiCyS), 2016, 2017
6. Computing Research Association for Women (CRA-W), 2015

Other

1. First Place (Winner), CodeRED Discovery (2018), University of Houston
2. Third Place, CodeRED Exploration (2017). University of Houston.
3. Winner, Social Track at HackRice 7 (2017), Rice University.

PROFESSIONAL/ACADEMIC SERVICE

Professional Memberships

- Member, American Medical Informatics Association (AMIA), 2021 -
- Member, Association for Computational Linguistics (ACL), 2016 -

Journal Club

- Organizer, MedAI Group Exchange Sessions, Stanford University-Mayo Clinic Arizona.

Editorial Services

- Review Editor, Text-mining and Literature-based Discovery, Frontiers in Research Metrics and Analytics Journal.

Reviewing Services

- **Journals**
  1. Computational and Structural Biotechnology (IF: 6.2)
  2. NPJ Digital Medicine (IF: 15.2)
  3. European Journal of Radiology (IF: 3.5)
  4. Artificial Intelligence in Medicine Journal (IF: 7.011)
  5. Journal of Biomedical Informatics (JBI) (IF: 8.0)
  6. Computers & Security Journal (IF: 5.105)
  7. Journal of Information Security and Applications (IF: 4.96)
  8. IEEE Open Access Journal (IF: 3.475)
  9. Neural Computing and Applications (NCAA) (IF: 5.102)
  10. PLOS Digital Health (IF:4.01)
- **Conferences and Workshops**
  1. North American Association for Computational Linguistics (NAACL), 2024
  2. Association for the Advancement of Artificial Intelligence (AAAI), 2024
  3. Conference on Neural Information Processing Systems (NeurIPS), 2024
  4. Empirical Methods in Natural Language Processing (EMNLP), 2021, 2022, 2023
  5. Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL), 2021, 2022, 2023
  6. International Joint Conference on Natural Language Processing (IJCNLP), 2022, 2023
  7. International Conference on Bioinformatics and Biomedicine (BIBM), 2022

8. Annual Meeting of the Association for Computational Linguistics (ACL), 2019, 2018, 2024
9. ACM International Workshop on Security and Privacy Analytics (Co-located with CODASPY), 2018, 2019

Program and Organizing Committee

- Program committee member, AI for HEALTHCARE and LIFE SCIENCES 2025
- Program committee member, Workshop on Multimodal4Health 2024 (co-located with ICHI)
- Program committee member, Workshop on Natural Language Processing for Bangla 2023 (co-located with EMNLP)
- Program committee member, EMNLP 2022 (Tracks include Language Modeling and Analysis of Language Models, Natural Language Generation, and Summarization tracks)
- Program committee member, AACL-IJCNLP 2022-2023
- Chair, Organizing committee, Security and Privacy Analytics Anti-Phishing Shared Task 2018 (co-located with CODASPY)

Mentoring

- **Mentor**, Machine Learning for Health (ML4H) Workshop (Co-located with NeurIPS 2022).
- **Graduate Students**
  1. Vedant Joshi (Ph.D. candidate), Arizona State University, Phoenix (at Mayo Clinic).
  2. Rongbin Li (Ph.D. candidate), UTHealth, Houston (at UTHealth-Houston).
  3. Ayman El Aassal (Ph.D. candidate), University of Houston, Houston (at UH).
- **Undergraduate Students**
  1. Boddhisattwa Dhara, BITS-Pilani (Hyderabad Campus), India (at Mayo Clinic).
  2. Gal Egozi, University of Houston, Houston (at UH).