Sanya Bathla Taneja
Website | LinkedIn | GitHub | Semantic Scholar

PhD candidate and computer scientist with research experience in natural language processing, machine learning, and knowledge representation, and biomedical informatics. Proficient in Python and SQL, with strong communication, writing, and interdisciplinary team science skills.

EDUCATION

PhD Intelligent Systems (Major: Artificial Intelligence) | University of Pittsburgh | 2020-2024
MS Intelligent Systems | University of Pittsburgh | 2020
B.Tech. Computer Science and Engineering | Indira Gandhi Delhi Technical University | 2018

SKILLS AND INTERESTS

Natural Language Processing – Named Entity Recognition, Information Retrieval, Large Language Models (LLMs); Data Mining; Literature-based Discovery; Machine Learning; Knowledge Graphs; Graph Representation Learning; OMOP Data Model; ETL of Electronic Health Records data; Ontologies
Technologies: Python, Cypher, Neo4j, R, SQL, Git, C++, OHDSI Toolkit, Langchain, GPT, Spacy

WORK EXPERIENCE

National Library of Medicine (NLM), NIH | PI: Zhiyong Lu | Research Intern | ***May-July 2023***
- Developed BERT-based entity linking methods for diseases identified in PubMed articles.
- Implemented natural language processing pipeline in Python with LLMs including ChatGPT API prompting with Langchain.

University of Pittsburgh | Malawi, Africa | Research Intern | ***June – August 2019***
- Developed Bayesian networks and machine learning models with decision tree analysis to diagnose and manage childhood malaria in Malawi. Collaborated with experts at health centers and Global Health Informatics Institute in Malawi and UPMC Children's Hospital.

Amazon India | Software Development Engineer (SDE) Intern | ***February – July 2018***
- Developed backend APIs for the Seller and Retail website using Java, Spring MVC, Coral, JavaScript.

RESEARCH EXPERIENCE

University of Pittsburgh | Graduate Student Researcher | ***February 2020 – 2024***
- Led knowledge graph (KG) research and development with 1M nodes and 8M edges, combining biomedical literature and ontologies for generation of mechanistic hypotheses for natural product-drug interactions and adverse event reporting systems. Presented related work at 7 conferences, with 1 Best Poster Award and 2 peer-reviewed publications.
- Implemented pipelines for end-to-end named entity recognition, entity linking, text mining, KG inferences and embeddings with Python.
- Led development and evaluation of prototype system for evidence synthesis with databases, KG, large language models for retrieval augmented generation (RAG) and summarization.
- Responsible for EHR data extraction, data analysis, and technical development of machine learning and case-control epidemiological analyses OMOP Common Data Model for Alzheimer's disease risk factors using OHDSI methods in R and Python and biomedical terminologies (SNOMED, ICD), with 1 peer-reviewed publication and 2 conference presentations.

University of Pittsburgh, School of Medicine | Research Assistant | ***September 2018 – February 2020***
- Developed natural language processing and machine learning pipelines for twitter surveillance using Python and resources at the Pittsburgh Supercomputing Center.

- Responsible for RITHM software framework maintenance, documentation, and upkeep of the GitHub repository for real-time Twitter data mining, with 3 peer reviewed publications.

SELECTED PEER REVIEWED PUBLICATIONS (Full list on Google Scholar)

- **Taneja SB**, Callahan TJ, Paine MF, Kane-Gill SL, Kilicoglu H, Joachimiak MP, Boyce RD. Developing a Knowledge Graph Framework for Pharmacokinetic Natural Product-Drug Interactions. *Journal of Biomedical Informatics*. 2023. DOI: doi.org/10.1016/j.jbi.2023.104341.
- Malec SA, **Taneja SB**, Albert SM, Shaaban CE, Karim HT, Levine AS, Munro PW, Callahan TJ, Boyce RD. Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: a use case studying depression as a risk factor for Alzheimer's disease. *Journal of Biomedical Informatics.* 2023. DOI: doi.org/10.1016/j.jbi.2023.104341.
- **Taneja, S.**, Boyce, R., Reynolds, W., & Newman-Griffis, D. "Introducing Information Retrieval for Biomedical Informatics Students." *Proceedings of the Fifth Workshop on Teaching NLP, Association for Computational Linguistics, 2021, pp. 96–98. ACLWeb*.
- **Taneja, S.B.**, Douglas, G.P., Cooper, G.F., Michaels, M.G., Druzdzel, M.J., Visweswaran, S. Bayesian network models with decision tree analysis for management of childhood malaria in Malawi. *BMC Med Inform Decis Mak 21, 158 (2021).* https://doi.org/10.1186/s12911-021-01514-w
- Visweswaran S, Colditz JB, O'Halloran P, Han NR, **Taneja SB**, Welling J, Chu KH, Sidani JE, Primack BA, Machine Learning Classifiers for Twitter Surveillance of Vaping: Comparative Machine Learning Study, *J Med Internet Res 2020;* https://www.jmir.org/2020/8/e17478.

SELECTED CONFERENCE PRESENTATIONS

- **Taneja SB,** Sivarajkumar S, Wang Y, Boyce RD. Information Extraction from Unstructured Text using Large Language Models for Natural Product-Drug Interactions. *Poster presentation, Pacific Symposium of Biocomputing; January 5, 2024.*
- **Taneja SB**, Chapin MR, Li X, Kane-Gill SL, Boyce RD. Generating Mechanistic Hypotheses for Pharmacovigilance Signals using a Natural Products Knowledge Graph. *Poster presentation, Advances in Pharmacovigilance for Herbal Medicines Conference; April 12-14, 2023.*
- **Taneja SB,** Paine MF, Kane-Gill SL, Boyce RD. Extending the OMOP Standard Vocabulary to Include Botanical Natural Products. *Poster presentation, Observational Health Data Sciences and Informatics (OHDSI) Symposium; October 14-16, 2022.* https://www.ohdsi.org/2022showcase-24/.
- **Taneja SB**, Joachimiak MP et al. Evaluation of Named Entity Recognition Systems to Improve Ontology Concept Annotation for Biomedical Knowledge Graphs. *Oral and poster presentation, ISMB Bio-ontologies COSI; July 10-14, 2022.* doi.org/10.5281/zenodo.6941350.
- **Taneja SB**, Ndungu PW, Paine MF, Kane-Gill SL, Boyce RD. Relation Extraction from Biomedical Literature on Pharmacokinetic Natural Product-Drug Interactions. *Poster presentation, AMIA Informatics Summit 2022; March 21-24, 2022.*

OTHER PROFESSIONAL ACTIVITIES

- **Editorial Activities:** JAMIA Student Editorial Board Member | *2022-2023*
- **Peer Review:** BMC Bioinformatics; Bioinformatics; Journal of Biomedical Informatics**;** ISMB Bio-Ontologies Group (2022); AMIA Informatics Summit (2021)
- **Volunteering:** Moderator, ISMB conference (2022); Translational Bioinformatics Year-in-Review team, AMIA Informatics Summit (2021, 2022)**;** Co-editor, AMIA Student Working Group Newsletter (2021-2022); Student Volunteer, AMIA Annual Symposium 2021
- **Awards:** Provost Fellowship, University of Pittsburgh (2023); ISMB/ECCB Best Poster Award (2021)