Sanya Bathla Taneja
Email: sbt12@pitt.edu
Website | LinkedIn | GitHub | Semantic Scholar

SUMMARY

PhD candidate in Intelligent Systems and computer scientist with research experience in natural language processing, machine learning, and knowledge representation and their applications in healthcare. Proficient in Python and SQL, with strong communication and writing skills.

EDUCATION

PhD Intelligent Systems
University of Pittsburgh | 2020-Present — Pittsburgh, PA
Major: Artificial Intelligence

MS Intelligent Systems
University of Pittsburgh | 2018-20 — Pittsburgh, PA

B.Tech. Computer Science and Engineering
Indira Gandhi Delhi Technical University for Women | 2014-18 — Delhi, India

SKILLS AND INTERESTS

Skills and Interests: Natural Language Processing – Named Entity Recognition, Information Retrieval, Relation Extraction; Machine Learning; Knowledge Graphs; Graph Representation Learning; OMOP Common Data Model; ETL of Electronic Health Records data; Knowledge Representation; Biomedical Ontologies; Large Language Models; Data Mining; Literature-based Discovery; Semantic Web
Technologies: Python, Cypher, Neo4j, R, SQL, Git, C, C++
Libraries: NLTK, Spacy, Pandas, Scikit-learn, Jupyter Lab, Keras, Networkx, Tensorflow, OHDSI Toolkit

RESEARCH EXPERIENCE

University of Pittsburgh, Intelligent Systems Program | PI: Richard D. Boyce, Mary F. Paine
Graduate Student Researcher | NaPDI Center | ***February 2021 – Present***

- Led research and development of a large-scale knowledge graph combining literature-based discovery, relation extraction of full texts of scientific articles, and biomedical ontologies for natural products including semantic representation of provenance and natural products and information extraction from texts using large language models.
- Developed methods for discovery of mechanistic hypotheses for natural product-drug interactions (NPDIs) and adverse events using embeddings, graph algorithms, pharmacovigilance signals, and large language models.
- Conceptualized and implemented standardization of open-source OMOP vocabulary for natural products with embedding-based models and generation of safety signals to identify NPDIs from adverse event reporting systems.
- Evaluated named entity recognition systems for knowledge graph construction and graph representation learning methods and extended OBO ontologies to include natural products and constituents to facilitate computational research using OWL and RDF.

University of Pittsburgh, Intelligent Systems Program | PI: Richard D. Boyce
Graduate Student Researcher | ***February 2020 – 2021***

- Responsible for longitudinal EHR data extraction, cleaning, and primary technical development of analyses with OMOP Common Data Model.
- Implemented and evaluated machine learning and case-control epidemiological analyses for Alzheimer's disease onset and risk factors using open-source OHDSI methods in R and Python.
- Supported development of knowledge graph using biomedical ontologies and machine reading to discover novel associations to prevent the onset of Alzheimer's disease.

University of Pittsburgh, School of Medicine
Research Assistant | ***September 2018 – February 2020***

- Developed and designed natural language processing and machine learning pipeline for twitter surveillance of vaping at the Center for Research on Media, Technology and Health.
- Responsible for RITHM software framework maintenance, documentation, and upkeep of the GitHub repository. (https://github.com/CRMTH/RITHM).
- Performed data extraction and processing for real-time Twitter data mining for public health research and analysis using Python and resources at the Pittsburgh Supercomputing Center.

WORK EXPERIENCE

National Center for Biotechnology Information (NCBI), National Institutes of Health | PI: Zhiyong Lu
Research Intern | ***May-July 2023***

- Developed BERT-based entity linking methods for diseases identified in PubMed articles using biomedical ontologies.
- Implemented natural language processing pipeline in Python with large language models including ChatGPT API prompting with Langchain.

University of Pittsburgh, Department of Biomedical Informatics | 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.
- Consulted and collaborated with experts at health centers and Global Health Informatics Institute in Malawi and UPMC Children's Hospital to design and implement the study.

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, and Handlebars. Involved in adding order cancellation details to the Seller dashboard to supplement the seller website.

PEER REVIEWED PUBLICATIONS

- Dilán-Pantojas, I.O., Boonchalermvichien, T., **Taneja, S.B**. et al. Broadening the capture of natural products mentioned in FAERS using fuzzy string-matching and a Siamese neural network. Sci Rep 14, 1272 (2024). https://doi.org/10.1038/s41598-023-51004-4.
- **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.

- Li X, Ndungu P, **Taneja SB**, Chapin MR, Egbert SB, Akenapalli K, Paine MF, Kane-Gill SL, Boyce RD. An evaluation of adverse drug reactions and outcomes attributed to kratom in the US Food and Drug Administration Adverse Event Reporting System (FAERS) from January 2004 through September 2021. *Clin Transl Sci*. 2023; 00: 1- 10. DOI: 10.1111/cts.13505.
- Sidani, J.E., Hoffman, B.L., Colditz, J.B., Melcher, E., **Taneja, S.B.**, Shensa, A., Primack, B., Davis, E. and Chu, K.H., 2022. E-Cigarette-Related Nicotine Misinformation on Social Media. Substance Use & Misuse, pp.1-7. DOI: 10.1080/10826084.2022.2026963.
- **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*, https://www.aclweb.org/anthology/2021.teachingnlp-1.16.
- **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
- Hoffman BL, Colditz JB, Shensa A, Wolynn R, **Taneja SB**, Felter EM, Wolynn T, Sidani JE. #DoctorsSpeakUp: Lessons learned from a pro-vaccine Twitter event. *Vaccine. 2021 May 6;39(19):2684-2691. doi: 10.1016/j.vaccine.2021.03.061*.
- 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;22(8):e17478*, URL: https://www.jmir.org/2020/8/e17478, DOI: 10.2196/17478
- Abhishek, A., **Taneja, S. B.,** Malik, G., Anand, A., & Awekar, A., Fine-grained Entity Recognition with Reduced False Negatives and Large Type Coverage. *Presented at the Automated Knowledge Base Construction (AKBC) Conference, 2019*
- Gupta A, **Taneja SB**, Malik G, Vij S, Tayal DK, Jain A. SLANGZY: a fuzzy logic-based algorithm for English slang meaning selection. *Progress in Artificial Intelligence. 2019 Apr 1;8(1):111-21.*
- Thesis: **Taneja, Sanya Bathla.** Bayesian Networks for Diagnosing Childhood Malaria in Malawi. Master's Thesis, University of Pittsburgh, 2020. Available from: http://d-scholarship.pitt.edu/38993/.

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.** Application of Semantic Knowledge Representation and Natural Language Processing to Identify Pharmacologic Mechanisms. *Doctoral Consortium, International Semantic Web Conference 2023; November 6-11, 2023. Proceedings to be published at https://ceur-ws.org/.*
- **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.* https://doi.org/10.1007/s40264-023-01289-1.
- **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, Hegde H, Baumgartner Jr. WA, JH Caufield, Callahan TJ, Mungall CJ, Boyce RD. 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.*
- **Taneja SB**, Callahan TJ, Brochhausen M, Paine MF, Kane-Gill SL, Boyce RD. Designing potential extensions from G-SRS to ChEBI to identify natural product-drug interactions. *Oral and poster presentation, ISMB/ECCB Bio-ontologies COSI; July 25-30, 2021.* https://doi.org/10.5281/zenodo.5736386.
- **Taneja SB**, Boyce RD, Reynolds WT, Newman-Griffis D. Introducing Information Retrieval to Biomedical Informatics Students. *Poster Presentation, 5th TeachingNLP workshop at NAACL-HLT 2021; June 10-11, 2021.*

OTHER PROFESSIONAL ACTIVITIES

Editorial Activities

Journal of the American Medical Informatics Association (JAMIA) Student Editorial Board Member | *2022-2023*

Peer Review

BMC Bioinformatics; Bioinformatics; Journal of Biomedical Informatics
Intelligent Systems for Molecular Biology (ISMB) Bio-Ontologies Group | *2022*
American Medical Informatics Association (AMIA) Informatics Summit | *2021*

Workshops

Co-organized discussion group on Symbolic AI and Knowledge Graphs with 20 participants and 6 guest speakers | *University of Pittsburgh, June-July 2022*

Volunteering

Moderator, Intelligent Systems for Molecular Biology (ISMB) conference (Bio-Ontologies) | *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

Talks

- Knowledge Graph Framework to Generate Hypotheses for Natural Product-Drug Interactions. *Presented at: Pittsburgh-CMU Medical Informatics Colloquium; December 3, 2021. https://pcmic.github.io/*.
- Guest lecture on 'Graph Machine Learning' in Foundations of Biomedical Informatics II, University of Pittsburgh; *January 2021-2023.*

AWARDS

- Provost Fellowship, Intelligent Systems Program, University of Pittsburgh | 2023
- ISMB/ECCB Bio-ontologies Best Poster Award | *2021*