Sanya Bathla Taneja

Email: sbt12@pitt.edu

Website | LinkedIn | GitHub | Semantic Scholar

RESEARCH INTERESTS

Biomedical informatics, natural language processing, machine learning, machine reading, knowledge representation, knowledge graphs, biomedical ontologies

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

EXPERIENCE

Graduate Student Researcher | ***February 2021 – Present***
University of Pittsburgh, Intelligent Systems Program

- Developed a large-scale knowledge graph combining literature-based discovery, machine reading of full texts of scientific articles, and biomedical ontologies for discovery and generation of mechanistic hypotheses for natural product-drug interactions (NPDIs) and related adverse events.
- Extended knowledge representation in biomedical ontologies to include natural products and constituents to facilitate computational research.
- Standardization of and generation of signals to identify NPDIs from adverse event reporting systems and poison center reports to assess clinical relevance of NPDIs.
- Part of the informatics core of the NaPDI Center created by the National Institutes of Health National Center for Complementary and Integrative Health (NCCIH).

Graduate Student Researcher | ***February 2020 – 2021***
University of Pittsburgh, Intelligent Systems Program

- Responsible for longitudinal EHR data extraction, cleaning, and primary technical development of analyses with OMOP Common Data Model.
- Implemented and tested machine learning and case-control epidemiological analyses for Alzheimer's disease onset and risk factors using 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. Funded by the Pitt Momentum Teaming Grant (2020).

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

- 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.

Research Intern | ***June – August 2019***
University of Pittsburgh, Department of Biomedical Informatics | Malawi, Africa

- Developed Bayesian networks and machine learning models with decision tree analysis to diagnose and manage childhood malaria in Malawi.
- Consulted experts at health centers and Global Health Informatics Institute in Malawi and UPMC Children's Hospital to design the study and implement it as master's thesis.

Software Development Engineer (SDE) Intern | ***February – July 2018***
Amazon India

- 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.

Research Intern | ***June-July 2017***
Indian Institute of Technology, Guwahati

- Involved in mining, processing and annotating training and test sentence corpus from Wikipedia for named entity recognition and classification into 118 categories.
- Developed an end-to-end fine-grained named entity recognition system from Wikipedia and Freebase using Python.

PEER REVIEWED PUBLICATIONS

- 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 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. *Intelligent Systems for Molecular Biology/European Conference on Computational Biology (ISMB/ECCB), 2021*. https://doi.org/10.5281/zenodo.5736386.
- **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**, 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 2021; July 25-30, 2021*.
- Shaaban CE, **Taneja SB**, Witonsky KF, Malec SA, Karim HT, Pratt S, Levine AS, Munro PW, Boyce RD, Albert SM. Does clinical data capture modifiable midlife risk factors for Alzheimer's disease? *In 2021 Alzheimer's Association International Conference; 2021 Jul 26.*
- **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.*
- S. Malec, **S. Taneja**, K. Witonsky, C. Shaaban, H. Karim, A. Levine, S. Albert, P. Monro, R. Boyce. Modeling Alzheimer's Disease by Combining Knowledge Extracted from Biomedical Literature with Biomedical Ontologies. *Poster Presentation, AMIA Informatics Summit 2021; March 23, 2021.*
- Hoffman BL, Colditz JB, Sidani JE, Davis EM, **Taneja SB**, James AE, Primck BA, Morris A, Brink L, Lynch M, Rose JJ, Chu KH. Correlation of Twitter Data to Reported Cases Of E-cigarette Or Vaping Product Use-associated Lung Injury (EVALI). *Poster Presentation, 2020 American Thoracic Society International Conference; May 17, 2020 (canceled due to COVID-19).*

INVITED 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/.*
- Explaining natural product-drug interactions with biomedical knowledge graphs. *Presented at: Intelligent Systems Program AI Forum, University of Pittsburgh; April 16, 2021. http://www.isp.pitt.edu/node/2117.*
- What do you think of vaping? Machine learning methods for Twitter Stance Detection. *Presented at: Intelligent Systems Program AI Forum, University of Pittsburgh; November 8, 2019. http://isp.pitt.edu/node/1997.*
- Using Bayesian networks to diagnose childhood illness in low- and middle-income countries: case of malaria in Malawi. *Poster presented at the DBMI Annual Training Program Retreat, University of Pittsburgh; August 22, 2019.*

SKILLS AND INTERESTS

Skills and Interests: Machine Learning, Natural Language Processing, OMOP Common Data Model, ETL of Electronic Health Records data, Clinical Decision Support, Bayesian Networks, Knowledge Graph, Knowledge Representation, Biomedical Ontologies
Technologies: Python, R, SQL, Git, C, C++
Libraries: NLTK, Spacy, Pandas, Scikit-learn, Jupyter Lab, Keras, Networkx

OTHER PROFESSIONAL ENGAGEMENTS

Editorial Activities

Journal of the American Medical Informatics Association (JAMIA) Student Editorial Board Member | *2022-2024*

Peer Review

American Medical Informatics Association (AMIA) Informatics Summit | *2021*

Teaching

Guest lecture on 'Graph Machine Learning' in Foundations of Biomedical Informatics II, University of Pittsburgh | *January 2021 & 2022*

Volunteering

Co-editor, AMIA Student Working Group Newsletter | *2021-2022*
Student Volunteer, AMIA Annual Symposium 2021
Translational Bioinformatics Year-in-Review team, AMIA Informatics Summit | *2021 & 2022*

Memberships

American Medical Informatics Association (AMIA) | *2020-Present*
International Society of Computational Biology (ISCB) | *2021-Present*

OHDSI Symposium Study-a-thon | *October 20-21, 2020*

Participated in phenotype development, data quality assessment, and manuscript preparation for cardiovascular clinical prediction models using OHDSI tools.

AWARDS

- ISMB/ECCB Bio-ontologies Best Poster Award | *2021*

COURSE PROJECTS

Foundations of Biomedical Informatics | Fall 2019
Utilizing clinical notes in electronic medical records (MIMIC III) to predict mortality risk in the ICU.

Natural Language Processing | Spring 2019 | https://github.com/sanyabt/NLP-CS2731
Feature analysis and multilabel, multiclass classification of emotions in short texts using Random Forest.

Machine Learning | Spring 2019 | https://github.com/sanyabt/ML-CS2750
Comparison of supervised machine learning models to predict patient no-shows in primary care hospitals.

Undergraduate Research Project | November 2017 – May 2018
Development of algorithm for English slang meaning selection from social media using fuzzy membership functions and natural language processing with Python.