BIPASHA BANERJEE, Ph.D.

240 971 9803 | bipashabanerjee@vt.edu | Blacksburg, VA | linkedin.com/in/bipasha-banerjee | Google Scholar

I am Assistant Professor at Virginia Tech, where I currently work as an AI Research Scientist, Digital Libraries for the University Libraries. My research mainly focuses on natural language processing, data mining, machine learning and digital libraries.

EDUCATION

Ph.D. COMPUTER SCIENCE, VIRGINIA TECH, BLACKSBURG, VA — 2024 (GPA 3.85/4)
M.S. COMPUTER SCIENCE, VIRGINIA TECH, BLACKSBURG, VA — 2022 (GPA 3.85/4)
B.Tech. COMPUTER SCIENCE & ENGINEERING
WEST BENGAL UNIVERSITY OF TECHNOLOGY, KOLKATA, INDIA — 2015 (GPA 8.63/10)

WORK EXPERIENCE

AI RESEARCH SCIENTIST — VIRGINIA TECH, UNIVERSITY LIBRARIES, BLACKSBURG, VA, AUGUST 2024- Present

Building solutions to help

- Gain better insights from data from our Virginia Tech's digital library.
- Make collections accessible and help with discovery and navigation.
- Automatic metadata curation and extraction.

GRADUATE RESEARCH ASSISTANT — VIRGINIA TECH, UNIVERSITY LIBRARIES, BLACKSBURG, VA, 2019 – August 2024

Increasing access to electronic thesis and dissertations

- Led efforts to bring computational access to over 500,000 electronic theses and dissertations (ETDs) by making chapter information more accessible using machine learning and deep learning techniques.
- Developed summarization and classification models using deep learning frameworks and large language models.
- Deployed LlaMa-2 on ARC (supercomputing cluster at VT) and fine-tuned the model (13B) on classification and summarization tasks.
- Investigated information extraction, segmentation, and bibliography parsing methods.
- Used digital library frameworks and practices to manage and organize our large (4.2 TB) digital content.
- Collaborated with ODU on data collection and organization.

Text extraction from handwritten collection

- Developed a workflow to extract text from handwritten images from archival collections using AWS Textract and lambda functions.
- Developing methods to make the extracted text searchable.
- Investigating methods to develop a asynchronous workflow to aid in ruducing concurrent execution and throttling.
- Developing methodology to integrate a human in the loop pipeline to validate and correct text extracted with low confidence.

GLOBAL PRODUCT AND TECHNOLOGY INTERN — ADP , ALPHARETTA, GA , MAY – AUG 2019

- Worked on the time and attendance, particularly the employee timesheet module.
- Built application automation and test cases by adding over 70 scenarios using Selenium and Cucumber.

GRADUATE TEACHING ASSISTANT — VIRGINIA TECH, BLACKSBURG, VA, JAN – MAY 2019

- Tutored undergraduate students (over 70) with the material taught in Cloud Software Development (CS3754) by conducting weekly office hours and grading.

ASSISTANT SYSTEMS ENGINEER - FULL TIME — TATA CONSULTANCY SERVICES, KOLKATA, INDIA APR 2016 – NOV 2017

- Developed an e-commerce Android application (reduced screen latency by 50 percent vs. the web application) using Android Studio IDE as the front and Salesforce as its backend.

- Developed Microsoft .Net web and Windows applications using Visual Basic and MVC architecture.

ACADEMIC PROJECTS

SUMMARIZATION AND CLASSIFICATION OF ETDs 2019-Present

- Dissertation topic: https://bipasha-banerjee.github.io/
- Worked on computationally extracting digital objects from book-length scholarly documents such as Electronic Theses and Dissertations (ETDs).
- Summarizing and classifying chapters ETDs using machine learning and deep learning techniques, including attention mechanisms.

INCREASING ACCESSIBILITY OF ELECTRONIC THESES AND DISSERTATIONS USING A HUMAN-IN-THE-LOOP ML APPROACH 2020

- Course project for Human-AI interaction CS 6724
- Studied the effects of using expert labels in the training of a machine-learning model.

SUMMARIZATION OF MARYLAND SHOOTING COLLECTION 2018

- Course project for Big Data Text Summarization, CS 5984, https://hdl.handle.net/10919/86407
- Studied summarization techniques to generate the summary of two shooting events from a large collection of web pages.

UNDERGRADUATE RESEARCH 2013 –2015

- Designed a Recommendation Based On-Demand Protocol for Mobile Ad-hoc Networks (MANET).
- Proposed Self-organized key management based on fidelity relationship list and dynamic path for security in MANET.
- Published a review of different routing protocols and their vulnerabilities and countermeasures.
- Published a review on attacks and secure routing protocols in MANET.

CRYPTOLOGY SUMMER INTERN – INDIAN STATISTICAL INSTITUTE MAY - JUNE 2014

- Analyzed and studied various “intractable problems” related to cryptology.
- Had hands-on training on various cryptology approaches, such as keys and ciphers.

WEB DEVELOPMENT – SAHA INSTITUTE OF NUCLEAR PHYSICS JAN- MARCH 2014

- Implemented a client-side web browser that communicates with a remote MySQL server using Common Gateway Interface (CGI) over an Apache web server.
- Reduced server-side load by using JQuery and asynchronous JavaScript (AJAX).

RECENT PUBLICATIONS & TALKS

Title: “Building Datasets to Support Information Extraction and Structure Parsing from Electronic Theses and Dissertations” January 2024
Accepted to be published in the International Journal on Digital Libraries (https://link.springer.com/journal/799).

Title: “Case Study of Analyzing the Variety of ETD Layouts” October 2023
Citation: Banerjee, Bipasha, et al. "Case Study of Analyzing the Variety of ETD Layouts." (2023).
https://ir.inflibnet.ac.in/bitstream/1944/2414/1/8.pdf

Title: “Integrated Digital Library System for Long Documents and their Elements” June 2023
Paper presented at ACM/IEEE JCDL 2023
Citation: Chekuri, S., Chandrasekar, P., Banerjee, B., Park, S. H., Masrourisaadat, N., Ahuja, A., ... & Fox, E. A. (2023, June). Integrated Digital Library System for Long Documents and their Elements. In 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 13-24). IEEE
https://doi.org/10.1109/JCDL57899.2023.00012.

Title: “Application of text analysis on scholarly long documents December 2022
Paper presented at IEEE Bigdata 2022

Citation: Banerjee, B., Ingram, W. A., Wu, J., & Fox, E. A. (2022, December). Applications of data analysis on scholarly long documents. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 2473-2481). IEEE.
https://doi.org/10.1109/BigData55660.2022.10020935

Title: "Help Me Help You - A Mixed-Initiative Approach To Explore Book-length Documents — October 2022
Talk presented at CIKM 2022 Workshop on Human-in-the-loop Data Curation

Title: "Opening scholarly documents through text analytics" — June 2022
Presented at ACM/IEEE JCDL 2022 Doctoral Consortium
Citation: Banerjee, B. (2022, June). Opening scholarly documents through text analytics. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (pp. 1-2).
https://doi.org/10.1145/3529372.3530948

Title: "Applications of mining ETDs" — November 2021
Presentation at ETD 2021 conference
https://doi.org/10.26226/morressier.614c9b8c87a68d83cb5d59b2

Title: Building A Large Collection of Multi-domain Electronic Theses and Dissertations — December 2021
Citation: S. Uddin, B. Banerjee, J. Wu, W. A. Ingram and E. A. Fox, "Building A Large Collection of Multi-domain Electronic Theses and Dissertations," 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 6043-6045
https://doi.org/10.1109/BigData52589.2021.9672058

Title: "Extracting Information from Electronic Thesis and Dissertations" — March 2021
Talk presented at ACM Capital Region Celebration of Women (CAPWIC 2021), Virtual

Title: "Summarizing ETDs with deep learning" — November 2019
Citation: Ingram, William A., Bipasha Banerjee, and Edward A. Fox. "Summarizing ETDs with deep learning." Cadernos BAD 1 (2020): 46-52
https://brapci.inf.br/index.php/res/download/134688

PROFESSIONAL ACTIVITIES

- Led breakout sessions at the workshop titled "Ensuring Scholarly Access to Government Archives and Records," organized by Virginia Tech University Libraries in partnership with Virginia Tech Center for Humanities and the U.S. National Archives and Records Administration (funded by The Andrew W. Mellon Foundation)
- Association of Women in Computing Member, Virginia Tech Chapter.
- Collaborating on a project geared toward using LLMs for University-wide analysis in partnership with University Libraries and the Office of Analytics & Institutional Effectiveness, Virginia Tech.
- DLRL student representative for CS grad council (2020-2023).
- Volunteered at VT CS grad recruiting weekend and career fair.
- Association for Computing Machinery (ACM) Member- #3768609
- Institute of Electrical and Electronics Engineers (IEEE) Member- #96669088

TEACHING EXPERIENCE

GRADUATE SUBJECT MATTER EXPERT — VIRGINIA TECH, BLACKSBURG, VA, 2022-PRESENT

- Mentored student teams in CS 4624 (Multimedia, Hypertext, and Information Access -- a capstone undergraduate course) and CS 5604 (Information Storage and Retrieval -- a graduate course).
- Responsible for setting up the project scope, providing necessary technical background, and evaluation.
- Help weekly meetings with teams to clarify doubts and check in on progress.

GRADUATE MENTOR VIRGINIA TECH, BLACKSBURG, VA, 2019-PRESENT

- Mentored graduate students pursuing an M.S. thesis and senior undergraduate students pursuing research with their research projects aligned with ETDs.
- Responsible for helping students with technical challenges, brainstorming new approaches, selecting datasets and models, and setting up experiments.

GRADUATE TEACHING ASSISTANT VIRGINIA TECH, BLACKSBURG, VA, JAN – MAY 2019

- Responsible for holding office hours (6-10 hours a week) and grading
- Tutored students to understand the material taught in class better.
- Helped with weekly assignments related to the material taught in class: Cloud Software Development

AWARDS

- Nominated for best student paper for "Integrated Digital Library System for Long Documents and their Elements" at IEEE/ACM JCDL 2023.
- Awarded a scholarship to attend the Grace Hopper Conference by the CS department at Virginia Tech in 2021 and the Association of Women in Computing (AWC, Virginia Tech Chapter) in 2022.
- Awarded free attendance at the International Conference in Pattern Recognition (ICPR), virtual 2020.
- "Student of the Year" awarded by the Institute of Engineering and Management for academics as well as contribution to the college in 2014.
- Selected by the Indian Statistical Institute, Kolkata, as one of the 60 candidates from all over India as a Summer Cryptology Intern in May 2014.