NIKITA SONI

Stony Brook, NY, 11790
nisoni@cs.stonybrook.edu, +1-631-202-7042
https://www.linkedin.com/in/nikita-soni-640ba137/
https://www3.cs.stonybrook.edu/~nisoni/
https://scholar.google.com/citations?user=1w2rduoAAAAJ&hl=en

EDUCATION

Stony Brook University, New York, U.S.
Doctor of Philosophy *in Computer Science -Natural Language Processing* – GPA: 3.92/4.0 Aug 2020 – Feb 2025
Research Focus: Language Modeling and Understanding
Advisors: Prof. Niranjan Balasubramanian and Prof. H. Andrew Schwartz

Master of Science *in Computer Science (Thesis)* – GPA: 3.92/4.0 Aug 2019–
(NLP, Machine Learning, Data Science, Probability & Statistics, Theory of Database)

Bocconi University, Milan, Italy
Visiting Researcher in Computing Science Department at MilaNLP Lab Sep 2022 - Dec 2022
Working with Prof. Dirk Hovy

Krishna Engineering College, Uttar Pradesh Technical University, India Jul 2008 - June 2012
Bachelor of Technology *in Computer Science and Engineering* – 76.86%
Final Project: Rule Based Cancer Detection System (Java); *Courses*: Databases, Data Structures & Algorithms; *Advisor:* Prof. Avinash Dwivedi

RESEARCH PUBLICATIONS

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. *Large Human Language Models: A Need and the Challenges.* To appear in NAACL 2024. (*Under Review*)

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, **Nikita Soni**, Rajath Rao, Kevin Lanning, Isabella Valejo, Lucie Flek, H Andrew Schwartz, Charles Welch, and Ryan L Boyd. *Archetypes and Entropy: Theory-Driven extraction of Evidence for Suicide Risk*. CLPsych workshop in EACL 2024.

Nikita Soni, Niranjan Balasubramanian, H. Andrew Schwartz, and Dirk Hovy. *Comparing Pre-trained Human Language Models: Is it Better with Human Context as Groups, Individual Traits, or Both?.* arXiv 2024 (*Under Review*).

Salvatore Giorgi, David M. Markovitz, **Nikita Soni**, Vasudha Varadarajan, Siddharth Mangalik, and H Andrew Schwartz. *"I Slept Like a Baby": Using Human Traits To Characterize Deceptive ChatGPT and Human Text.* IACT workshop at ACM SIGIR conference 2023.

Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V Ganesan, Shashanka Subrahmanya, **Nikita Soni**, Sean AP Clouston, and H Andrew Schwartz. *Robust language-based mental health assessments in time and space through social media*. arXiv 2023.

Vasudha Varadarajan, **Nikita Soni**, Weixi Wang, Christian Luhmann, H Andrew Schwartz, and Naoya Inoue. *Detecting Dissonant Stance in Social Media: The Role of Topic Exposure.* NLP+CSS workshop in EMNLP 2022.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, **Nikita Soni**, Sharath Chandra Guntuku, Johannes Eichstaedt, and H Andrew Schwartz. *WWBP-SQT-lite: Difference Embeddings and Multi-level Models for Moments of Change Identification in Mental Health Forums*. CLPsych workshop in NAACL 2022.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. *Human Language Modeling.* ACL-Findings 2022.

Matthew Matero, **Nikita Soni**, Niranjan Balasubramanian, and H. Andrew Schwartz. *MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection.* EMNLP-Findings 2021.

RESEARCH INTERNSHIPS

Bloomberg AI Group – *NLP Research Intern* Jul 2023 – Oct 2023

Designed and experimented for *dynamic* conditional text generation by modifying GPT2, T5, and HaRT models.

Capital One – *PhD NLP (Data Science) Intern* Jun 2021 – Aug 2021
Designed and experimented a high performing solution to identify customer sentiment during a call by smart preprocessing of call transcripts data and using it to pre-train a large (transformers based) language model over the auto-loans calls domain and fine-tune it for sentiment analysis.

McAfee LLC – *Data Science Intern (MVISION Cloud Business Unit)* May 2020 – Aug 2020
Designed and implemented training features using machine learning techniques from historical data to identify anomalous sequences in user activities for different cloud services.

ORGANIZATION, TALKS, TEACHING

Workshop on *Human-Centered Large Language Modeling, ACL 2024.* *08/2024*
(https://hucllm-workshop.github.io/)

Tutorial on *From Text to Context: Contextualizing Language with Humans, Groups, and Communities for Socially Aware NLP, NAACL 2024.* *06/2024*

Talk on *Human Language Modeling*

World Well-Being Project (WWBP) Consortium	03/2022
All Things Language and Computation, Stony Brook University	04/2022

Stony Brook University

CSE 538: Graduate Natural Language Processing – *Guest Lecture*	Spring, 2024
CSE 357: Undergraduate Probability and Statistics for Data Science – *Guest Lecture*	Fall, 2021
CSE 538: Graduate Natural Language Processing – *Guest Lecture*	Fall, 2020
CSE 538: Graduate Natural Language Processing – *Teaching Assistant*	Fall, 2020
CSE 354: Undergraduate Natural Language Processing – *Teaching Assistant*	Spring, 2020

PROGRAM COMMITTEE & REVIEWING

ARR (ACL)	Feb, 2024
ARR (NAACL)	Dec, 2023
EMNLP	2023
EMNLP	2022

- *Language Modeling & Analysis of Language Models Track*
- *The 5th workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*

SERVICE

Volunteer (Diversity & Inclusion Committee) NAACL Conference 2022

COURSE PROJECTS and ASSISTANTSHIPS

- **Topped** the NLP class of ~**130** students working on projects like **Dialogue AgeNt Consistency Evaluation metric**, **Relation Extraction**, **Transition Parsing** with Neural Networks, **Sentence Representation** using GRU and implementing DAN, and **Skip-gram** model implementing **Cross Entropy** and **NCE** loss along with **word analogy**.
- **Ranked 1 of 96** in the class Kaggle competition for **Predicting goodness points of a wine** using **Ridge Regression**; and **Ranked 2 of 82** in another one to classify images by implementing **K means clustering** and using **LibSVM.**
- Rich data analysis and visualizations for Data Science projects to Predict Future Sales and Predict TMDB Box Office Revenue: https://github.com/soni-n/Data-Science-Projects
- Implemented **Logistic Regression** with Stochastic gradient descent in **Python** for **Crowd Image Classification**, **SVM** using quadratic programming in **Matlab** for **object detection**.
- Developed a sentiment analysis tool using Google Cloud Language APIs in **YHack 2019** Hackathon sponsored by Google Cloud, Facebook etc.

SOFTWARE INDUSTRY EXPERIENCE

[24]7.ai, India – *Lead Software Development Engineer in Test* Apr 2018 – Aug 2019

- Developed **Date Range Widget, Vertical alignment for text fields** and **Vertical alignment of multiple fields** features of the Digital Card Composer product used to build sliders for all the products of the organization.
- Introduced **Hibernate Envers** for audit logging and **Browserstack** for efficient **parallel testing**.
- Developed a comprehensive automation test suite with **90% coverage** that **reduced** regression testing **time** by **5 times.**

PegaSystems, India - *Senior Software Development Engineer in Test* Mar 2015 – Apr 2018

- Developed **Update** and **Delete** functionality for the **Asynchronous Job Scheduler** module supported in cluster mode, and its interface on **Pega** platform.
- Evangelist for **Test Driven Development** in the development of the Job Scheduler and Queue Processor modules.
- Developed the integration test suite for **NativeSQL** module with **90% code coverage** and configured parallel runs on **5 databases** using Jenkins thus **reducing** the regression testing **time** by **10 times** and reducing bug leaks.
- Performance tested **Id generator** module and **Data Upload utility** module for Oracle RDS supporting insert and upsert, with synthetic data of varied sizes, maximum being 5GB.
- Represented the Data Engine team in a **summit** held across all the geographical locations of Pega.

McAfee Software, India - *Software QA Engineer* Aug 2012 - Mar 2015

- Owned Junit testing and Hudson setup for **Cloud Connector** product, using **AWS**, **Azure**, **OpenStack**, **Rackspace** and **HP cloud**, having built unit and integration test suite from **scratch to 85%+** code coverage.
- Released **Role Based Key Management** with **zero defect leaks** as the component owner. Awarded as the best performer in the team for the same.

TECHNICAL SKILLS

Languages:	Python (~3 yrs), Java (~5.5 yrs), SQL.
Frameworks:	PyTorch, TensorFlow, Numpy, Pandas, Matplotlib, Junit (~5 yrs), EasyMock & Mockito (~3 yrs), Selenium, React.JS (~2 months), Enzyme & Jest (~2 months), Spring.
Databases:	MySQL, Oracle, Postgres, Microsoft SQL Server, SQLite.
Tools:	Jupyter, Git, SVN, Bugzilla, SonarQube, JaCoCo, Coverity, Jira, Confluence.
Platforms:	Amazon Web Services (AWS), VMware vSphere, Docker, Pega, Android.