

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://www.kaggle.com/niksoni>

EDUCATION

Stony Brook University, New York, U.S.

Doctor of Philosophy in Computer Science -Natural Language Processing – GPA: 3.92/4.0 Aug 2020–May 2024

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–May 2021

(NLP, Machine Learning, Data Science, Probability & Statistics, Theory of Database)

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

Human Language Modeling (HuLM)

To appear in ACL Findings (2022)

- A hierarchical language modeling problem whereby a human-level exists to connect sequences of documents (e.g. social media messages) and capture the notion that human language is moderated by changing human state.
- Introducing, HaRT, a large-scale transformer model for the HuLM task.

MeLT: Message Level Transformer

EMNLP Findings (2021)

(<https://aclanthology.org/2021.findings-emnlp.253/>)

- A hierarchical message encoder pre-trained over Twitter and applied to the task of stance prediction.
- Trained over a variant of masked language modeling; where instead of predicting tokens, it seeks to generate an entire masked (aggregated) message vector via reconstruction loss within a sequence of messages connected by authorship.

RESEARCH INTERNSHIPS

Capital One – PhD NLP (Data Science) Intern

Jun 2021 – Aug 2021

- Developed 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

- Developed training features using machine learning techniques from historical data to identify anomalous sequences in user activities for different cloud services.

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.
- Graduate Teaching Assistant (TA) for NLP graduate and undergraduate courses at Stony Brook University: Gave a guest lecture on TensorFlow (TF) 2.0; implemented coding assignment in TF 2.0; designed exams; managed other TAs.

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 (~6 yrs).

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.