SimPPL: A Research Program for Indian Undergraduates from Underserved Communities

Asst. Professor Pranit Bari

Dept. of Computer Engineering, DJ Sanghvi College of Engg. (DJSCE), Faculty In-charge for Unicode (student software engineering training program) and affiliate at SimPPL (research training pilot program)

Project Purpose

Provide a brief description of your concept that could be used to explain your work to a new audience. (140 Characters or Fewer)

AI for Social Good technical research program offered to 350+ undergraduates from colleges within Mumbai University, housed at DJSCE.

Approach

Please describe how you would approach the project methodologically and how you envision its implementation. How does your project help engage new and diverse perspectives in conversations about socially responsible technologies and computing? (2500 Characters or Fewer)

This project will provide a series of workshop-based training sessions (in person and remote) to pursue coursework-adjacent responsible computing research projects. Student participants will be placed on teams to complement their in-classroom experiences via mentored, collaborative research projects. The research ecosystem at Tier II and Tier III institutes is limited by compute and hardware availability, as well as access to quality mentorship considering the low teacher:student ratio at most colleges within Mumbai University. This proposed program will host in-person workshops and online sessions to train students towards developing a research mindset, offer topical lectures from experts in responsible AI, data science, machine learning, and motivate them to apply their newfound skillset towards developing concrete research projects, publications, and open-source software tools. It will substantively include weekly check-ins, 1:1 mentor meetings, project updates, documentation, deployment, literature review, reading groups, and research writing sessions. The in-person sessions will be housed at DJSCE, but the funding will allow us to expand this to participation from 1 sister college and 3 other colleges in Mumbai University where we have had students successfully working with us already, via the pilot program we have run since 2021 called SimPPL (https://simppl.org). SimPPL is a research collective that publishes and builds open-source research projects using AI for social good, reducing online harms and improving trust and safety. In the past 6 years launching and growing Unicode and SimPPL, we have created an inclusive and supportive ecosystem for students that face a lack of adequate computing resources, travel limitations, socio-cultural discrimination, and financial hardship, which is demonstrated by the self-sustaining Unicode program that caters to hundreds of students each year. We are well-placed to pursue the expansion of this program with access to the resources, mentors, and student

channels necessary to identify those who might benefit most from such a program. This includes nonprofit partners in India and Bangladesh who have benefitted from student projects in the last year and are excited to partner with us. It also includes experts at Google, Meta, the Oversight Board, the Integrity Institute, and govt. agencies like the UK's DCMS, FCDO, who have provided us with 1:1 guidance on how to develop impactful research in the past year.

Deliverables

What will be produced as a result of your activities through this award (e.g., syllabi changes, classroom material, teaching methodology adjustments, new courses)? Your response should be a numbered or bullet point list. (1250 Characters or Fewer)

1. 5-7 open-source research tools built by student teams comprising 5-15 members each, that will focus on: responsible AI, multilingual and multimodal machine learning (using cutting-edge, open-source, computer vision and natural language processing models), web and app platforms that will cater to use cases important to the local community and nonprofit partners (https://www.djunicode.in/projects, as a running example of tools).

2. 2-4 research publications at workshops in top-tier conferences (e.g. https://openreview.net/forum?id=YBk2jG7MEaX from our past work) in similar topical areas, led by student teams based on their interests and career goals.

3. A report on the methodologies and course materials utilised to teach these students a research-augmented course alongside their primary education. It will delineate the successes and open opportunities for scaling peer-based research in countries like India where the quality of education at non-premier institutions suffers due to a lack of research resources.

4. Recommendations for syllabi updates to include research-based coursework as a graded part of their academics. Suggestions for teaching methodology updates from our learnings to align research with academic coursework.

Suitability

Why are you ideally placed to advance this work? What else have you done in this space? Include links to other projects, grants, and previous work as applicable. Please also note if you are working with any team members or partners, explain how these collaborations advance your concept. (2500 Characters or Fewer)

The project will be undertaken in partnership with Dr. Swapneel Mehta, postdoctoral researcher at Boston University and MIT, and Board of Studies Member at DJSCE. Dr. Mehta has co-founded DJ Unicode as a pilot program with 15 students in 2017. By 2020, they expanded activities to over 200 students in-person in Mumbai as well as online. In 2021, his team provided a similar project-based ML course to 100+ students across 27 Tier-II and Tier-III Indian colleges funded by Google Research India (https://djunicode.github.io/umlsc-2021/). The subsequent

interest birthed SimPPL (https://simppl.org), a ‘research collective’ that are focused on the research and development of open-source tools to advance trust and safety on the social internet. We aim to advance the SimPPL research-oriented learning model since our students have demonstrated a keen interest in project-oriented research whereby they partner with nonprofits and solve real-world problems whilst conducting meaningful research and learning about new topics relevant to their projects.

Since 3 years, Dr. Mehta has also led the NYU AI School (https://nyu-mll.github.io/nyu-ai-school-2023/schedule/) educating 900+ undergraduates primarily identifying with historically marginalized groups (from over 7000 applicants), in-person and online, sponsored by Genentech and DeepMind, at New York University. He continues to spearhead educational efforts in India such as a massive online seminar at IIT Madras Data Science (online) to hundreds of attendees, and teaches at SimPPL growing their offerings to solve multilingual and multimodal challenges in responsible AI, online harms, and trust and safety in India, Bangladesh, and other so-called global south countries.

Suitability

Why are you ideally placed to advance this work? What else have you done in this space? Include links to other projects, grants, and previous work as applicable. Please also note if you are working with any team members or partners, explain how these collaborations advance your concept. (2500 Characters or Fewer)

We have 6 years of prior work developing a similar program in software engineering called DJ Unicode (https://djunicode.in) offered to over 100 students annually, for free. The project will be undertaken under the leadership of Dr. Swapneel Mehta, postdoctoral researcher at Boston University and MIT, also an alumnus and Board of Studies Member at DJSCE. Dr. Mehta co-founded DJ Unicode (https://djunicode.in) as a pilot program with 15 students in 2017. By 2020, they expanded activities to over 200 students in-person in Mumbai as well as online; activities like Google Summer of Code preparatory sessions, hackathon preparatory sessions, and workshops in git and Github, web, and app development (https://www.youtube.com/@djsceunicode5127/videos). In 2021, his team provided a similar project-based ML course to 100+ students across 27 Tier-II and Tier-III Indian colleges funded by Google Research India (https://djunicode.github.io/umlsc-2021/). The subsequent interest birthed Unicode Research (https://unicode-research.netlify.app) and SimPPL (https://simppl.org), a ‘research collective’ focused on the research and development of open-source tools to advance trust and safety on the social internet in the global sout.

He has also led the NYU AI School (https://nyu-mll.github.io/nyu-ai-school-2023/schedule/) educating 900+ undergrads identifying with historically marginalized groups (from over 7000 applicants), in-person and online, sponsored by Genentech and DeepMind, at New York University. He continues to spearhead educational efforts in India such as a massive online seminar at IIT Madras Data Science (https://www.youtube.com/watch?v=ZLcPFzxCiVg) to hundreds of attendees, and teaches at SimPPL growing their offerings to solve multilingual and

multimodal challenges in responsible AI, online harms, and trust and safety in India, Bangladesh, and other so-called global south countries.

We aim to advance the SimPPL research-oriented learning model since our students have demonstrated a keen interest in project-oriented research whereby they partner with nonprofits and solve real-world problems whilst conducting meaningful research and learning about new topics relevant to their projects. All projects at SimPPL are student-led, and have won prestigious research awards from Google Cloud, Wikimedia Foundation, the Anti-Defamation League, Amazon AWS, the NYC Media Lab, the Sunday Times, and Deutsche Welle.

Primary Contact

Swapneel Mehta

https://linkedin.com/in/swapneelm

https://twitter.com/swapneel_mehta

https://github.com/swapneelm

+1 551 328 7074

swapneel.mehta@djsce.edu.in

Full Application

Concept Description

Describe your concept in detail. What problem are you looking to address and what is your intervention? How does the project engage new and diverse perspectives in conversations about just technologies? Does your approach connect socially responsible computing to industry and wider tech ecosystems? If so, how? (5000 Characters or Fewer)

We propose a year-long structured research program that complements the coursework delivered to undergraduate students at colleges within Mumbai University (MU). This intervention creates research projects pursued by groups of like-minded students working jointly towards conducting applied research in an area of their interest. The proposal addresses the lack of structured research opportunities to apply and advance student understanding of the concepts taught as part of coursework. It offers them an opportunity to work jointly towards meaningful applications of these concepts as part of technical research projects. Dwarkadas J. Sanghvi College of Engineering (DJSCE) is an autonomous institute affiliated with MU that will serve as the host institution for the proposed intervention.

Part I - 3 months

We will host 5 in-person workshop sessions that will teach students about research pedagogy and methodology with the aim of facilitating conversations between groups of students to help them identify groups that share their interests. The aim of these sessions is the successful creation of aligned student groups, each with their mutually agreed project pitch and objectives listed therein.

The five modules we will rely on broadly for the sessions include the following:

Workshop Session 1: Introduction to Research and Problem Identification

- Module 1: Introduction to Research - Understanding the significance of research in responsible computing.
- Module 2: Identifying Research Problems - Developing skills to spot research questions and real-world problems.
- Module 3: Literature Review - Learning how to find and summarize existing research.

Workshop Session 2: Narrowing Down Research Topics

- Module 4: Ethical Considerations - Discussing the ethical dimensions of technology and research.
- Module 5: Brainstorming Research Ideas - Encouraging idea generation and selection based on existing expertise.

Workshop Session 3: Developing Research Proposals

- Module 6: Research Proposal Structure - Understanding the components and structure of research proposals.
- Module 7: Research Methodology - Choosing suitable research methods and ethical practices.

Workshop Session 4: Literature Review and Project Planning

- Module 8: In-Depth Literature Review - Expanding on the literature review process and gap identification.
- Module 9: Project Planning - Creating detailed project plans, milestones, and timelines.

Workshop Session 5: Finalizing Research Projects

- Module 10: Presentation and Communication - Developing effective research communication skills.
- Module 11: Mentorship and Resources - Exploring mentorship and available research resources.
- Module 12: Final Project Pitches - Presenting and refining research project proposals.

These modules will help guide students through the process of identifying research problems, developing proposals, conducting literature reviews, and preparing for responsible computing research projects at their colleges in a manner that is aligned with classroom teaching.

Part 2 - 6 months

For a period of 6 months thereafter, we will host dedicated remote sessions (with optional in-person, informal student hangouts) to deliver technical workshops to students that need to understand git, Github, web development, app development, data analysis, data visualization, data infrastructure, web scraping, and other topics relating to research in responsible computing, with the target areas being applied machine learning and social media analysis.

We will tap into the TAs, SimPPL Mentors, Research Lead, and Program Lead to set up discussions with individual teams to track progress, highlight blockers, walk through the research plan and changes therein, and support teams advance their research agenda in sync with classwork.

We will measure the success of our intervention through the following metrics:

1. Anonymized student feedback collected after each workshop, after office hours, and at the end of the program.
2. Peer-feedback collected from within groups to measure how aligned they are and to preemptively reduce any potential issues resulting in friction between team members.
3. Achievement of Project Objectives: The objectives will comprise of both, a technical output, and a research output. This is because we anticipate students to already possess technical knowledge through classroom teaching and labs so defining technical objectives serves as a reasonable baseline metric to track their contributions and progress. The research objectives include the subjective reviews that mentors will provide for measuring their progress and tracking their final research output.

Our goal is to seed the student project ideas with social-media related topic areas. This is because:

1. This area is diverse: conducting social networks research includes large scale data collection and analysis, text as data problems, AI, ML, and NLP problems, data processing and database management, web and app development for data donations research, and other areas that are both intuitive and have sufficient depth to garner student interest while remaining generalizable to provide them with skillsets applicable beyond social media-related research areas.

2. Through running our pilot program called SimPPL, we have gained significant expertise and have published past research in this area. We have mentors trained in this area who can guide student research projects actively, on a day-to-day basis.

3. We can support the students to find external partners to conduct their research project with: We work with the One Fact Foundation in the USA, various nonprofits and newsrooms including The Sunday Times (UK), Deutsche Welle (Germany), the NYC Media Lab (USA), and others.

Final:

We propose a year-long structured research program that complements the coursework delivered to undergraduate students at colleges within Mumbai University (MU). This intervention creates research projects pursued by groups of like-minded students working jointly towards conducting applied research in an area of their interest. The proposal addresses the lack of structured research opportunities to apply and advance student understanding of the concepts taught as part of coursework. It offers them an opportunity to work jointly towards meaningful applications of these concepts as part of technical research projects. Dwarkadas J. Sanghvi College of Engineering (DJSCE) is an autonomous institute affiliated with MU that will be the host.

Part I - 3 months

We will host 5 in-person workshop sessions that will teach students about research pedagogy and facilitate conversations. The aim of these sessions is the successful creation of aligned student groups, each with their mutually agreed project pitch and objectives listed therein.

The five modules we will rely on broadly for the sessions include the following:

Workshop Session 1: Introduction to Research and Problem Identification

- Module 1: Introduction to Research - Understanding the significance of research in responsible computing.
- Module 2: Identifying Research Problems - Developing skills to spot research questions and real-world problems.
- Module 3: Literature Review - Learning how to find and summarize existing research.

Workshop Session 2: Narrowing Down Research Topics

- Module 4: Ethical Considerations - Discussing the ethical dimensions of technology and research.
- Module 5: Brainstorming Research Ideas - Encouraging idea generation and selection based on existing expertise.

Workshop Session 3: Developing Research Proposals

- Module 6: Research Proposal Structure - Understanding the components and structure of research proposals.
- Module 7: Research Methodology - Choosing suitable research methods and ethical practices.

Workshop Session 4: Literature Review and Project Planning

- Module 8: In-Depth Literature Review - Expanding on the literature review process and gap identification.
- Module 9: Project Planning - Creating detailed project plans, milestones, and timelines.

Workshop Session 5: Finalizing Research Projects

- Module 10: Presentation and Communication - Developing effective research communication skills.
- Module 11: Mentorship and Resources - Exploring mentorship and available research resources.
- Module 12: Final Project Pitches - Presenting and refining research project proposals.

These modules will help guide students through the process of identifying research problems, developing proposals, conducting literature reviews, and preparing for responsible computing research projects at their colleges in a manner that is aligned with classroom teaching.

Part 2 - 6 months

For a period of 6 months thereafter, we will host dedicated remote sessions (with optional in-person, informal student hangouts) to deliver technical workshops to students that need to

understand git, Github, web development, app development, data analysis, data visualization, data infrastructure, web scraping, and other topics relating to research in responsible computing, with the target areas being applied machine learning and social media analysis.

We will tap into the TAs, SimPPL Mentors, Research Lead, and Program Lead to set up discussions with individual teams to track progress, highlight blockers, walk through the research plan and changes therein, and support teams advance their research agenda in sync with classwork.

We will measure the success of our intervention through the following metrics:

1. Anonymized student feedback collected after each workshop, after office hours, and at the end of the program.
2. Peer-feedback collected from within groups to measure how aligned they are and to preemptively reduce any potential issues resulting in friction between team members.
3. Achievement of Project Objectives: These will comprise of technical and research objectives. We expect students to possess basic technical knowledge through classroom teaching and labs so technical objectives serve as a baseline metric to track their project-building contributions. The research objectives include subjective weekly reviews that mentors provide measuring their progress and their final research output.

Our goal is to seed the student project ideas with social-media related topic areas. This is because:

1. This area is both intuitive and has sufficient diversity to garner student interest while remaining generalizable to provide them with skillsets applicable beyond social media-related research areas i.e. in AI, ML, NLP, databases, systems, cryptography, privacy, and security, etc.

2. Through running our pilot program called SimPPL, we have gained significant expertise and have published past research in this area.

3. We can support the students to find external partners to conduct their research project with:

Feasibility

What are some challenges you anticipate when implementing your approach, and how do you plan to mitigate these challenges? (1250 Characters or Fewer)

We anticipate a few challenges:

1. Identifying alignment over research interests for 150 students is likely to be significantly challenging to conduct. And this is a crucial part of the pitching process since we would

like to have student groups decided before the end of the in-person workshop sessions to initiate some social activities and icebreakers for teams that will work together.
    a. The mitigation measure for this will be to utilise pre-defined roles and areas of research similar to the approach we adopted to having students self-select into predefined roles when deciding to be part of projects at SimPPL (https://miniature-stetson-1d7.notion.site/newbie-Team-Planning-0c3afe7e49ee46c08c8c9a90a40872e9?pvs=4). Doing the same for defining some sample research areas to look into helps calibrate expectations when participating in group projects, pre-defining team leads and independent responsibilities being allocated within teams. The tradeoff is the perceived reduction in flexibility but that is fixed by encouraging a conversation where a role may not be completely aligned and require some changes for fitting a student's interests.
2. Monitoring and measurement: It will be important to encourage regular filling of feedback forms and to measure attrition early on so that we can build robust teams that do not get blocked by inactive individual team members.
    a. We have already implemented such a program in place at SimPPL wherein students successfully take time off for internships, unit tests, GRE, and other exams without disrupting the team's progress on projects. This mainly depends on 1:1s with students to address any ongoing issues and lack of support to enable them to work with the rest of the team.

Final:

We anticipate a few challenges:

1. Identifying alignment over research interests for 150 students is likely to be significantly challenging to put in place and is important to achieve during the in-person workshops.
    1. The mitigation measure for this will be to utilise pre-defined roles and areas of research similar to the approach we adopted to having students self-select into predefined roles when deciding to be part of projects at SimPPL (https://miniature-stetson-1d7.notion.site/newbie-Team-Planning-0c3afe7e49ee46c08c8c9a90a40872e9?pvs=4). Doing the same for defining some sample research areas to look into helps calibrate expectations when participating in group projects, pre-defining team leads and independent responsibilities being allocated within teams.
2. Monitoring and measurement: It will be important to encourage regular filling of feedback forms and to measure attrition early on so that we can build robust teams that do not get blocked by inactive individual team members.
    1. We have already implemented such a program in place at SimPPL wherein students successfully take time off for internships, unit tests, GRE, and other exams without disrupting the team's progress on projects. This mainly relies on 1:1s with students.

Working Open

How will you document and share your concept with broader audiences? How might you engage students, scholars from other disciplines, university administrators, non-profits, start-ups, government entities, or other organizations to help shape the design of your concept? If applicable, include links to a project website, Github page, or other publicly-accessible information about your concept. (1250 Characters or Fewer)

We have continued to document all our ongoing projects initially starting with our website for Unicode (https://djunicode.in), projects listed on the website, github codebases made open-source for each project (https://github.com/djunicode) and recently a Youtube channel sharing recordings of workshops we have conducted. We will follow the same design to share our concept with broader audiences following a staggered release of all materials publicly, but a near real-time internal sharing of all these resources for student accommodation purposes.

We expect to actively work with researchers if we identify relevant collaborative projects that can benefit from the expertise offered by the One Fact Foundation, Boston University, MIT, and other nonprofitsand newsrooms previously engaged with the SimPPL team (https://simppl.org). Potential partners may include nonprofits like DigiSwasthya (Maharashtra), Aadhar Bahuddeshiya Sanstha (Maharashtra), Heal Station Foundation (Bihar), The Times and the Sunday Times (UK), Deutsche Welle (Germany), WaterAid (Bangladesh), and others.

Impact

How will your approach prepare students to understand the impact of computing work and equip them with the necessary skills to build better, more trustworthy, and less harmful technologies, policies, and tech cultures? How do you plan on evaluating the impact of the intervention? (2500 Characters or Fewer)

There is incredible talent that is limited by the lack of opportunities among undergraduate students at the universities we are hoping to target with this grant. Thie proposal unblocks them from resource and mentorship limitations and empowers them to conduct cutting edge research into topics that are very relevant for the global south – social media analysis, in particular online trust and safety! Social media platforms are consistently deploying cutting edge technologies at a breakneck pace without sufficient safeguards. But this has become a growing area of concern only recently because the harms of platform failures result in widespread ethnic violence and genocide with horrific incidents in Myanmar, Ethiopia, and India. The effects, despite often being felt by many, are not well tracked by platforms due to a lack of appropriate multilingual

infrastructure and dedicated teams to monitor these harms. By adopting this as the primary area of focus for our responsible computing research proposals we have the potential to drive massive change both in terms of technology but also in terms of awareness of the opportunities to build external tools in the trust and safety ecosystem that can cater to specific vulnerabilities and protect communities by means of student-led projects. We have set a precedent with SimPPL by creating student projects to identify coordinated information networks actively promoting misleading Russian media articles online resulting in conversations with Twitter's global trust and safety teams directly. We hope to drive a similar impact by building more responsible computing tools in the trust and safety space with this proposal.

We will measure these through the student reviews and exit surveys for the program. Ideal indicators will be to track whether students would like to continue to contribute in this direction, whether it influenced their understanding of responsible computing research and harms of online social networks, and whether they are able to apply it to relevant real-world problems they see in their communities. The sustainability of the collaborative research community we build will be the clearest indicator of long-term success of such interventions.

Field Survey

Are other individuals or organizations working on similar concepts? If yes, explain how your work complements that of others or fills a key gap. (1250 Characters or Fewer; bullet form is preferable)

Draft: There are different types of peer learning models present at the intersection of academia and industry, especially in the fields of AI and ML. The closest examples are other research collectives like ML Collective, EleutherAI, and IndabaX Conferences. These communities focus more on voluntary mentorship, with their primary goal being access to a shared knowledge of AI and ML and a platform/community for advancing conversations and unstructured collaborations on ML projects. While there is incredible value to be unlockled through these models, and our students and mentors have also been active members of these communities, a much lower fraction of their communities actively participate in projects than in our case. By targeting specific brackets like second and third-year undergraduate students, we ensure that all members are at similar stages of their careers and have a few common career goals re: higher education, research roles, internships, and industry positions. This enables us to have dedicated sessions advancing these goals and benefitting individual members much more. On the other hand, there are existing, well-structured and fully funded programs offering pre-doctoral research opportunities like those by Google Research India, Microsoft Research India, and Fellowships by Adobe Research India; AI Residencies by Apple, Google, Meta, Microsoft, Uber, and others would also fall into the structured program category. However, these are

typically full-time commitments, pursued after a formal education (undergraduate degree required), and extremely competitive, ultimately not being able to scale to a large number of interested students. We attempt to bridge the gap between structured predoctoral programs and unstructured community-run projects, by creating a robust and personalized peer-learning model for students to participate in that directly complements their academic education and benefits their career.

Final: There are different types of peer learning models present at the intersection of academia and industry, especially in the fields of AI and ML. Examples of research collectives are ML Collective, EleutherAI, and IndabaX Conferences. These communities focus more on voluntary mentorship, with their primary goal being access to a shared knowledge of AI and ML and a platform/community for advancing conversations and unstructured collaborations on ML projects. Our students have also been active members of these communities, and seen that a much lower fraction of their communities actively participate in projects than in our case. By targeting specific brackets like second and third-year undergraduate students, we ensure that our sessions can be more detailed and benefit members individually. Now there are also structured programs offering pre-doctoral research opportunities and AI Residencies (FAAMG companies). However, these are full-time post-graduation roles, and remain extremely competitive, ultimately not being able to scale to a large number of interested students. We bridge the gap between structured and unstructured peer-learning research programs that directly complements undergraduates' existing academic education.

Budget Overview

How will you use the award funds? (2500 Characters or Fewer, please provide in bullet form)

Draft: The budget will be spent primarily on 2 items:

1. Organization of in-person workshops providing training to students on the “Methodology to Conduct Scientific Research”, including topical knowledge of “Responsible Computing and Ethics”;
    a. In-person workshops are supplemented by remote sessions thereafter to advance their topical knowledge of subject matter relating to their group research projects. We already conduct workshops on software development-related topics, AI/ML, and social media analysis remotely with successful outcomes so the remote workshops would extend the same format into related topics.
    b. Sub-items include instructor and TA costs for workshops, and program lead costs for organization and management of student events in-person.

    c. We will also support all attendees with travel costs to make it easy to attend the workshops.
2. Supporting student groups to pursue independent research projects related to responsible computing that are complementary to their coursework.
    a. The budget will provide the computing resources, laptops, other equipment, cloud computing resources for online data collection, social media data APIs (Youtube, Bluesky, Whatsapp, Telegram), and other computing resources.
    b. The remainder of these resources will be used to provide a minimum stipend to the students working as part of the group projects.

Indirect Costs - DJSCE will provide staff to serve grant administratration requirements as the beneficiary institution (under the SVKM public charitable trust; the applicant) and in keeping with the Mozilla requirements, they have set a rate at or lower than 10% of direct costs.

We anticipate that the workshops will reduce the activation costs for students to form groups, identify common topic areas, and define appropriate research projects, taking on roles as part of teams, to develop a skillset complementing their coursework.