

# Tharindu Cyril Weerasooriya

☎ +1 (123) 456 7890 • ✉ thacyrweer@gmail.com • 🌐 cyrilw.com • in cyrilw  
🔑 cyrilw • 🔑 P6zvClgAAAAJ

## Research Highlights

---

### Fairness in Human Annotation and Disagreement Prediction.....

Building a representative human-annotated training datasets using machine learning for subjective classification tasks on social media (political content, offensive language, and hate speech).

**Collaborators:** Dr. Christopher M. Homan (advisor), Dr. Alexander G. Orobia, Dr. Ashique KhudaBukhsh, Dr. Sarah Luger, and Dr. Marcos Zampieri.

**Publications:** EMNLP 2023 [1], ACL 2023 [4, 2], ECAI 2020 [7], and NLPerspectivist @ LREC 2022 [6, 5]

**Data:** Subjective social media content (~ 50,000+ items) from Reddit, Twitter, and Facebook.

**Interests:** Machine learning, Label distribution learning, Deep learning, Data science, Statistics, Probabilistic learning, Human computation, Crowdsourcing, Information retrieval, Natural language processing, Large Language Models, Sociolinguistics, and Computational Social Sciences.

## Education

---

### Rochester Institute of Technology

Rochester, NY

*Ph.D. in Computer Science, GPA 3.9*

2018–2024

Machine Learning, Natural Language Processing, Fairness in AI, Data Science, and Computational Social Science.

**Thesis:** Fairness in Human Annotation. **Advisor:** Dr. Christopher M. Homan

### University of Kelaniya

Sri Lanka

*B.Sc. in Computer Science*

2013–2017

**Thesis:** KeyXtract Twitter Model - An Essential Keywords Extraction Model for Twitter. **Advisor:** Dr. Sidath Liyanage

## Experience

---

### Rochester Institute of Technology

Rochester

*Research Assistant, NY*

2018 - Present

Fairness in AI, Population level annotator learning for human disagreement prediction

Tools used - Python (Scikit-learn, Keras, and TensorFlow), BERT, MongoDB for data storage, and Google Cloud.

- Led human subjects research through Amazon MTurk and managed a budget of \$10,000.
- Research to building socially representative and privacy-centric datasets for machine learning pipelines.
- Research work accepted at EMNLP '23, ACL '23, ICLR '23 Tiny Papers, LREC '22 NLPerspectives, and ECAI '20.

### Amazon Ads

New York

*Applied Scientist Intern, NY*

2023 May - 2023 August

Project - Contextual Relevancy under weak human-in-the-loop judgments and LLM Assisted Models.

Tools used - Python (Scikit-learn and PyTorch), SBERT, MySQL, LLM (LangChain) and AWS.

- Designed methods for understanding annotator quality across different locales and languages.
- Introduced new pipelines for adjusting existing contextual relevancy pipelines to work with annotator scores.

### Meta formerly known as Facebook

Menlo Park

*Machine Learning Research Engineer Intern, CA*

May 2022 - August 2022

Facebook Creators Wellbeing Team on Public Conversations.

Project - Introduction of a Multi-Label Multi-Task model for assisting page administrations for comment management.

Tools used - Python, PHP, and Presto. Python/PyTorch for model building. Presto for data pipelines.

- Collaborations with peers at a team level as well as organization level for adopting existing single level ML models and infrastructure.
- The model bypassed the existing individual action-based models used by 40% based on ROC and PR AUC scores.
- Evaluated the model in production with A/B testing on 4% of overall global Facebook users.

## LKEMR

### Tech/Infrastructure Lead

Cloud-based Electronic Medical Record project with the Faculty of Medicine and Colombo North Teaching Hospital. Tools used - PHP, SQL, HTML, CSS, and UNIX. Python for data mining. Hosted on DigitalOcean and AWS.

- Web-based electronic patient record management system, tailored for Sri Lanka using PHP backend and MySQL.
- EMR currently holds records for over 1,000 patients across several cities in Sri Lanka.
- System is used to identify potential patients with COVID-19, as it is the only EMR widely used in Sri Lanka.
- The work also contributes to open sourced EMR Project, Open-EMR.

Sri Lanka

2018-2022

## European Chapter of the Association for Computational Linguistics

### Infrastructure Lead for EACL 2021

- Built infrastructure for hosting and deploying virtual conference platform on Amazon AWS.
- Integrated an external registration system with AWS Cognito for single-sign-on (OAuth2) Auth0.
- Led a team to manage support requests from Conference chairs, conference sponsors, and attendees.

Virtual

January 2021 - July 2021

## Boston College

### Research Assistant Intern

Automatic Speech Recognition for low resource languages

- Studies on low resource datasets for ASR and their fairness for all speakers.
- Advised an REU student preparing for Computer Science and Machine Learning courses.

Virtual

May 2021 - August 2023

## HEALS Project (part of IBM AI Horizon)

### Research Assistant, Rensselaer Polytechnic Institute, NY

Food Data Aggregation, Project to aggregate nutritional information for crowdsourced recipes from FDA

Collaborative project with the Faculty of Medicine, and Colombo North Teaching Hospital, Sri Lanka.

Tools used - Python (document processing and SpaCy+doc2vec for NLP).

- Research based on natural language processing and information retrieval with the 1M Recipe (1 million) dataset.
- Investigated aggregating food-related data sources for the food knowledge graph, used with IBM Watson.
- The tool was able to generate the nutritional content per FDA guidelines from a crowdsourced recipe.

Troy

May 2019 - August 2019

## Skills

**Data Science:** Tableau, Jupyter Notebooks (Colab), SQL (Presto) and Excel

**Programming:** Python, R, Java, C, and PHP

**Infrastructure:** Google Cloud, AWS, and DigitalOcean

**ML/NLP:** PyTorch, Keras, TensorFlow, SpaCy, NLTK, HuggingFace (BERT) and sklearn

**Crowdsourcing:** Qualtrics and Amazon Mechanical Turk

**Design:** Photoshop, Illustrator, and Lightroom

## Conference Presentations

- [1] **Weerasooriya, Tharindu Cyril**, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M. Homan, and Ashiqur R. KhudaBukhsh. "Vicarious Offense and Noise Audit of Offensive Speech Classifiers". In: *Accepted to appear at EMNLP 2023*.
- [2] **Weerasooriya, Tharindu Cyril**, Alexander G. Ororbia II, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher M. Homan. "Disagreement Matters: Preserving Label Diversity by Jointly Modeling Item and Annotator Label Distributions with DisCo". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 4679–4695.
- [3] **Weerasooriya, Tharindu Cyril**, Sarah Luger, and Christopher M. Homan. "Offensiveness as an Opinion: Dissecting Population-level Label Distributions". In: *Proceedings of the 11th International Conference on Learning Representations*. 2023.
- [4] **Weerasooriya, Tharindu Cyril**, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher M. Homan. "Subjective Crowd Disagreements for Subjective Data: Uncovering Meaningful CrowdOpinion with Population-level Learning". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 950–966.

- [5] Christopher Homan, **Weerasooriya, Tharindu Cyril**, Lora Aroyo, and Chris Welty. "Annotator Response Distributions as a Sampling Frame". en. In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, 2022, p. 10.
- [6] **Weerasooriya, Tharindu Cyril**, Alexander G Ororbia, and Christopher M Homan. "Improving Label Quality by Joint Probabilistic Modeling of Items and Annotators". en. In: *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, 2022, p. 5.
- [7] **Weerasooriya, Tharindu Cyril**, Tong Liu, and Christopher M. Homan. "Neighborhood-based Pooling for Population-level Label Distribution Learning". In: *Proceedings of the 24th European Conference on Artificial Intelligence*. 2020.
- [8] Kumara Mendis, **Weerasooriya, Tharindu Cyril**, Supun Withana, Prabath Liyanage, Aruni Weerakoon Silva, Rajitha Wickramasinghe, and Chaminda Weerabaddana. "Cloud-Based Open Source Primary Care Electronic Patient Record System for Sri Lankan Citizens". In: *2019 National Information Technology Conference, NITC 2019*. 2019, pp. 8–10.
- [9] **Weerasooriya, Tharindu**, Nandula Perera, and S. R. Liyanage. "KeyXtract Twitter Model - An Essential Keywords Extraction Model for Twitter Designed using NLP Tools". In: *10th KDU International Research Conference*. Ratmalana, 2017. arXiv: 1708.02912.
- [12] **Weerasooriya, Tharindu**, Nandula Perera, and S. R. Liyanage. "A Method to Extract Essential Keywords from a Tweet using NLP Tools". In: *16th International Conference on Advances in ICT for Emerging Regions, ICTer 2016 - Conference Proceedings*. 2016, pp. 29–34.

## Journal Papers

---

- [10] Nandula Perera and **Weerasooriya, Tharindu**. "The 'Racecourse' of Then and Now: Evolution of the Sri Lankan English Vocabulary Over Two Generations of SLE Speakers". In: *VISTAS Journal* 10 (2016), pp. 1–23.
- [11] **Weerasooriya, Tharindu**. "Automated Railway Ticketing System: Replacing the Paper based Ticket with the Electronic National Identity". In: *ICT for Development Working Paper Series* 6.1-2 (2016), pp. 67–77.

## Teaching

---

<b>2023:</b> Teaching Assistant for CSCI 736 - Neural Networks and ML (22 students)	<i>RIT</i>
<b>2020:</b> Adjunct Lecturer for CSCI 635 - Introduction to Machine Learning (~ 30 students)	<i>RIT</i>
<b>2020-2021:</b> Teaching Assistant for CSCI 235 - Mechanics of Programming (3 Semesters, ~ 120 students)	<i>RIT</i>
<b>2018:</b> Teaching Assistant for Programming in C (~ 130 students)	<i>University of Kelaniya</i>
<b>2017:</b> Teaching Assistant for Object Oriented Programming (~ 70 students)	<i>University of Kelaniya</i>
<b>2017:</b> Teaching Assistant for Data Structures and Algorithms (~ 70 students)	<i>University of Kelaniya</i>

## Mentorship

<b>2022:</b> Saloni Poddar, MS Student	<i>Rochester Institute of Technology</i>
<b>2022:</b> Yu Liang, MS Student	<i>Rochester Institute of Technology</i>
<b>2022:</b> Raj Bakulbhai Bhensadadia, MS Student	<i>Rochester Institute of Technology</i>
<b>2021:</b> Abhay Rajendra Dixit, MS Student	<i>Rochester Institute of Technology</i>
<b>2020:</b> Anuj Kulkarni, MS Student	<i>Rochester Institute of Technology</i>

## Service

---

### Paper Reviewer

<b>2023:</b> ACL '23	<b>2021:</b> AISTATS '21
<b>2021,2023:</b> EMNLP '21,'23	<b>2023:</b> FAccT '23

**2022:** ICWSM '22

**2019,2022:** SocInfo '19,'22

**2021-2023:** IEEE Transactions on Neural Networks and Learning Systems

**2022:** WWW '23

## Student Organizations.....

### Doctoral Student Association

**Rochester Institute of Technology**

*President, NY*

*2020 - 2023*

- Represented the Ph.D. student population at sessions for the Board of Trustees of Rochester Institute of Technology.
- Co-led projects for fundraising and distribution of awards for student conference participation.
- Aided in advocating for student concerns, managing the budget and events for the organization, and the infrastructure for it.

### HacKLN 2017

**University of Kelaniya**

*Faculty Representative, SL*

*2017*

- Advised Students in organizing and fundraising for the hackathon.
- Guided the students as the teaching assistant in charge for the Department for Vidya 50, the 50th Celebratory Exhibition of Faculty of Science.

### Gavel Club University of Kelaniya

**University of Kelaniya**

*Web Lead, SL*

*2014-2017*

- Organizing Committee Member of Inter University Best Speaker Contest 2014 - 2017.
- Designed, developed and maintained the official Club Website, Contest Website and the online account management system for the executive committee.

## Hiring Committees.....

**2022:** Search Committee for Dean for Graduate School

*Rochester Institute of Technology*

**2022:** Search Committee for Dean for College of Computing

*Rochester Institute of Technology*

**2021:** Search Committee for Assistant Dean of Graduate School

*Rochester Institute of Technology*

## Invited Talks

**2023:** CrowdOpinion: Dissecting the Opinion of the Crowd

*Rochester Institute of Technology*

**2023:** CrowdOpinion: Dissecting the Opinion of the Crowd

*Internal Amazon HCI Workshop*

**2023:** Fairness in Human Annotation

*Rochester Institute of Technology*

**2022:** Auditing Machine Learning Classifiers on Vicarious Offense

*Internal Google Disagreement Workshop*

## Awards

**2021:** AWS Startup credit for LKEMR (\$1000)

*AWS*

**2020:** Google Cloud Research Credit (\$1000)

*Google*

**2020:** Crowdfunding Campaign for DSA (raised \$1000)

*Crowdfunded*

## Certificates

**2018,2023:** Social and Behavioural Research Certificate

*Citi Program*

**2017:** Competent Gavelier Leader

*Toastmasters International USA*

**2017:** Competent Gavelier Communicator

*Toastmasters International USA*