Nupoor Gandhi

phone: (408) 893-3764 | email: nmgandhi@cs.cmu.edu | web: nupoorgandhi.github.io

Research Statement

My research focuses on **low-cost structured information extraction** through the development of **efficient annotation methods** in combination with **task decomposition**. I am particularly interested in low-resource settings where large-scale expert annotation is prohibitively expensive.

Education

Carnegie Mellon University Expected Graduation 2026
Ph.D., Machine Learning and Public Policy *Pittsburgh, PA*
Advisor: Emma Strubell
Thesis: Center-based Decomposition for Efficient Annotation of Structured Representations

University of Illinois at Urbana-Champaign August 2016 – May 2020
Bachelor of Science, Computer Science; Minor in Mathematics *Urbana, IL*
Advisor: ChengXiang Zhai
Senior Thesis: Forecasting Public Health Outcomes with Social Media Data

Peer-Reviewed Publications

1. **Decomposing Unitization and Typing for Efficient and Consistent Span-Bound Concept Annotation**.
N. Gandhi, M. Bada, E. Strubell
Findings of the Association for Computational Linguistics (ACL), 2026.

2. **SynthTextEval: Synthetic Text Data Generation and Evaluation for High-Stakes Domains**.
K. Ramesh, D. Smolyak, Z. Zhao, **N. Gandhi**, R. Agarwal, M. Bjarnadóttir, A. Field
Proceedings of EMNLP – Demo Track, 2025.

3. **Beyond Text: Characterizing Domain Expert Needs in Document Research**.
S. Gururaja, **N. Gandhi**, J. Milbauer, E. Strubell
Findings of the Association for Computational Linguistics (ACL), 2025.

4. **Evaluating Differentially Private Synthetic Data Generation in High-Stakes Domains**.
K. Ramesh, **N. Gandhi**, P. Madaan, L. Bauer, C. Peris, A. Field
Findings of Empirical Methods in Natural Language Processing (EMNLP), 2024.

5. **Challenges in End-to-End Policy Extraction from Climate Action Plans**.
N. Gandhi, T. Corringham, E. Strubell
Proceedings of the ClimateNLP Workshop at ACL, 2024.

6. **Mention Annotations Alone Enable Efficient Domain Adaptation for Coreference Resolution**.
N. Gandhi, A. Field, E. Strubell (***Selected for Oral Presentation***)
Proceedings of the Association for Computational Linguistics (ACL), 2023.

7. **Examining Risks of Racial Biases in NLP Tools for Child Protective Services**.
A. Field, A. Coston, **N. Gandhi**, A. Chouldechova, E. Putnam-Hornstein, D. Steier, Y. Tsvetkov
Proceedings of ACM FAccT, 2023.

8. **Improving Span Representation for Domain-adapted Coreference Resolution**.
N. Gandhi, A. Field, Y. Tsvetkov
Proceedings of the EMNLP Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC), 2021.

9. **Predicting Opioid Abuse with Text-based Twitter Features**.
N. Gandhi, A. Morales, S. Chan, D. Albarracin, C. Zhai
Proceedings of the AAAI Conference on Artificial Intelligence – Student Abstract, 2020.

10. **Multi-Attribute Topic Feature Construction for Social Media-based Prediction**.
A. Morales, **N. Gandhi**, S. Chan
IEEE International Conference on Big Data (Big Data), 2018.

Non-archival Workshops and Preprints

1. **Task Decomposition for Efficient Annotation**.
N. Gandhi, E. Strubell
Under Review, 2026.

2. **Extracting Structured Policy Information from Climate Action Plans**
T. Corringham, **N. Gandhi**, B. Flores, and E. Strubell, S. Gururaja, T. Romanov, J. Dunafon
NeurIPS 2025 Workshop on Tackling Climate Change with Machine Learning, 2025.

3. **LLMs for Few-shot Information Extraction with Climate Action Plans**
N. Gandhi, T. Corringham, and E. Strubell
Text-as-Data Workshop 2023

Invited Talks

1. **Centering as a Critical Challenge for Efficient Annotation of Structured Representations**
Poster Presentation at Midwest Speech and Language Days 2025

2. **Challenges in End-to-End Policy Extraction from Climate Action Plans**
Poster Presentation at ClimateNLP Workshop, ACL 2024
Oral Presentation at Climate NLP Reading Group 2024

3. **LLMs for Few-shot Information Extraction with Climate Action Plans**
Poster Presentation at Text-as-Data Conference 2023

4. **Mention Annotations Alone Enable Efficient Domain Adaptation for Coreference Resolution**
Oral Presentation at ACL 2023

5. **Risk Prediction and Information Extraction for Child Welfare in Allegheny County**
Guest Lecture for Computational Ethics in NLP 2022

6. **Improving Span Representation for Domain-adapted Coreference Resolution**
Oral Presentation at EMNLP CRAC Workshop 2021

7. **Predicting Opioid Abuse with Text-based Twitter Features**
Poster Presentation at AAAI Conference on Artificial Intelligence 2020

8. **Reducing Response Times to Citizen Legal Questions in Uganda**
Lightning Talk at KDD – Social Impact Session 2020
Oral Presentation at Data Science for Social Good: DataFest 2019

9. **Natural Selection and the Bystander Effect**
Poster Presentation at Undergraduate Research Symposium 2019

10. **Multi-dimensional Features for Prediction with Tweets**
Poster Presentation at ACM Tapia Conference 2018

11. **Predicting Sexually Transmitted Infections with Twitter Data**
Rapid-Fire Talk at Society of Women Engineers Conference 2018
Poster Presentation at Illinois Scholars in Undergraduate Research Expo 2018

Teaching Experience

Probabilistic Graphical Models (10-708) *with Andrej Risteski* Spring 2026
Machine Learning Department, CMU

Computational Ethics in NLP (11-830) *with Emma Strubell and Alan Black* Spring 2022
Language Technologies Institute, CMU

Fellowships & Awards

AAAI Travel Grant 2020
Knights of St. Patrick, *highest award for leadership by College of Engineering, UIUC* 2020
Outstanding Poster Presentation - Science & Math, *Undergraduate Research Week, UIUC* 2019
Gold Collegiate Section Award, *Society of Women Engineers Conference* 2018
Rapid Fire Research Finalist, *Society of Women Engineers Conference* 2018
Illinois Scholars Research Grant, *awarded for excellence in research, UIUC* 2017

Research Experience

Graduate Student Researcher January 2022 – Present
Carnegie Mellon University, Structure in(g) LAnguage LAB *Pittsburgh, PA*
Advisor: Emma Strubell
- Methods for data-efficient domain transfer with limited annotation schemes.
- Cross-domain alignment of structured knowledge using shallow semantic structures.

Graduate Student Researcher September 2020 – December 2021
Carnegie Mellon University, CMU TsvetShop *Pittsburgh, PA*
Advisor: Yulia Tsvetkov and Alexandra Chouldechova
- Adapting end-to-end coreference resolution model for clinical notes from child welfare (Allegheny County Dept. Human Services) and medical domains.
- Developing new methods to use external knowledge to adapt coreference models to new domains.

Undergraduate Student Researcher August 2016 – September 2020
University of Illinois at Urbana-Champaign, Text Information Management Analysis Group *Urbana, IL*
Advisor: ChengXiang Zhai
- Predicted outbreaks of sexually-transmitted infections and opioid mortality rates using LDA topic features over a Twitter corpus.
- Constructed features to represent regional colloquialisms following geolocation of tweets.

Data Science for Social Good Fellow June 2019 – August 2019
Imperial College London *London, UK*
Advisor: Rayid Ghani, Leid Zejnilovic
- Built an Information Retrieval system to assist lawyers at nonprofit Barefoot Law reach citizens 60% faster.

Industry Experience

Software Engineering Intern May 2018 – August 2018
Qualcomm *San Diego, CA*
- Achieved over 10% reduction in memory consumption for transceiver chip.
- Developed tool to eliminate redundancies in radiofrequency settings data for 5G modem.
- Modeled embedded software with Abstract Syntax Trees to automate optimization of memory layout.

Software Engineering Intern May 2017 – August 2017
Synopsys *Mountain View, CA*
- Built ML API to analyze runtime performance of EDA tools.
- Developed web-based push notification system for SQL server.

Service

Reviewing
ACL Rolling Review 2026
ClimateNLP Workshop at ACL 2024–2026
Journal of Comparative Policy Analysis 2024
EMNLP 2023

Academic
Application Reviewer, Data Science for Social Good Fellowship 2022, 2026
Mentor, CMU AI Mentoring Program 2021 – Present
Committee Member, Machine Learning Department Masters Admissions 2022

GSA Representative, *Graduate Student Assembly, CMU* 2023 – Present
Student Leadership Committee Representative, *CMU Senate* 2025–2026
Student Volunteer, ACL Conference 2023
Volunteer, *Scopeathon Data Science for Social Good, CMU* 2022, 2025
President, *Society of Women Engineers, UIUC* 2019–2020
External Vice President, *Society of Women Engineers, UIUC* 2017–2018

Community

Canvasser, *Daeja Baker Shaler County School Board Campaign* 2025
Canvasser, *Summer Lee U.S. House Congressional Campaign* (Primary) 2022, 2024
Canvasser, *Kamala Harris U.S. Presidential Campaign* 2024
Organizer, *Pittsburgh Public Parks Coalition* (campaign to plug abandoned oil wells in SW PA) 2025
Volunteer Data Scientist, *Illinois Legal Aid Online* 2020

Skills

Python, Java, C, TensorFlow, PyTorch, NumPy/SciPy/pandas, Transformers, Huggingface, NLTK, R, LaTeX