Nandan Thakur

Ph.D. Student in Computer Science @ University of Waterloo
[Website] [Google Scholar] [Twitter] [GitHub] [LinkedIn] [Email]

Education

- **University of Waterloo** — Waterloo, Canada
  Ph.D. Student in Computer Science, Supervisor: Prof. Jimmy Lin — *Sep 2021 - Jul 2026*
- **Birla Institute of Technology and Science, Pilani (BITS Pilani)** — Goa, India
  B.E.(Hons.) Electronics & Instrumentation + Minor in Finance — *Aug 2014 - May 2018*

Employment

- **Google Research** — California, USA
  Student Researcher, Supervisors: Dr. Daniel Cer, Dr. Jianmo Ni — *Sep 2022 - May 2023*
- **UKP Lab, Technical University of Darmstadt** — Darmstadt, Germany
  Research Assistant, Supervisors: Prof. Iryna Gurevych, Dr. Nils Reimers — *Nov 2019 - Aug 2021*
- **KNOLSKAPE** — Bengaluru, India
  Data Scientist, Manager: Mr. Chaithanya Yambari — *Sep 2018 - Oct 2019*
- **(EMBL) European Molecular Biology Laboratory** — Heidelberg, Germany
  Research Trainee, Supervisors: Dr. Toby Gibson, Dr. Manjeet Kumar — *Jun 2018 - Aug 2018*

Publications

Peer-Reviewed Conference and Workshop Papers

[C7] SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval.
Nandan Thakur, Kexin Wang, Iryna Gurevych, Jimmy Lin.
To appear in SIGIR 2023 Resource Track.

[C6] Domain Adaptation for Memory-Efficient Dense Retrieval. [pdf]
Nandan Thakur, Nils Reimers, Jimmy Lin.
To appear in ReNeuIR 2023 Workshop.

[C5] Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages. [pdf] [code]
Xinyu Zhang*, **Nandan Thakur***, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, Jimmy Lin. (* denotes equal contribution)
To appear in TACL 2023. *Coverage*: [WSDM Cup 2023]

[C4] Evaluating Embedding APIs for Information Retrieval. [pdf]
Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, **Nandan Thakur**, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Jimmy Lin.
ACL 2023 Industry Track.

[C3] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. [pdf] [code]
Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych.
NAACL-HLT 2022. *Coverage*: [Pinecone.ai]

[C2] BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. [pdf] [code]
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych.
NeurIPS 2021 - Datasets and Benchmark Track.
Coverage: [Stanford CS224U] [Open-NLP Meetup] [Transformers-at-Work] [ML News]

[C1] Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. [pdf] [code]
Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych.
NAACL-HLT 2021. *Coverage*: [Pinecone.ai] [Blogpost]

Arxiv Preprints

[P2] Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard. [pdf] [code]
Ehsan Kamalloo, **Nandan Thakur**, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin. 2023.

[P1] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval. [pdf] [code]
Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, **Nandan Thakur**, Jheng-Hong Yang, Xinyu Zhang. 2023.

Open Source Packages

[G3] SPRINT: https://github.com/thakur-nandan/sprint (2023). Developer and Maintainer.
[G3] INCOME: https://github.com/thakur-nandan/income (2022). Developer and Maintainer.
[G2] BEIR: https://github.com/beir-cellar/beir (2021). Developer and Maintainer. Over 1K+ stars.
[G1] Augmented SBERT: https://github.com/UKPLab/sentence-transformers (2021). Developer.

Research Experience

- **Google Research** — California, USA
  Student Researcher — Mentors: Dr. Daniel Cer, Dr. Jianmo Ni — *Sep 2022 - May 2023*
  - Developed SWIM-IR, a large-scale multilingual synthetic dataset containing over 29 million synthetic training pairs for 18 languages using the "summarize-then-ask" few-shot prompting technique with the PaLM-2 model for training fully synthetic (i.e. unsupervised) multilingual retrieval systems.
  - Experience with prompt engineering and few-shot learning with large language models (LLMs) present internally within Google such as FLAN, PaLM-2, and Chinchilla models.
  - Evaluated robust and efficient unsupervised multilingual pertaining strategies for pre-training both encoder-only and encoder-decoder multilingual retrieval models involving contrastive and generative loss.

- **UKP Lab, Technical University of Darmstadt** — Darmstadt, Germany
  Research Assistant — Supervisors: Prof. Iryna Gurevych, Dr. Nils Reimers — *Nov 2019 - Aug 2021*
  - Developed a heterogeneous zero-shot IR benchmark (BEIR) containing 18 datasets spanning diverse retrieval tasks and domains. Conducted quantitative analysis on out-of-distribution generalization of large pre-trained language models and traditional IR systems for robust text retrieval. [link]
  - Quantitatively analyzed cross- and bi-encoder attention networks for pairwise sentence tasks such as semantic similarity (STS). Developed a novel data augmentation technique to distill knowledge from high-performing and inefficient cross-encoders to improve efficient bi-encoder performances. [link]
  - Created and open-sourced a real-world pairwise argument similarity dataset using best-worst scaling (BWS) annotation for eight controversial topics via crowdsourcing on Amazon MTurk. [link]
  - Developed a scalable entity-aware dashboard to search and cluster similar arguments for various data collections using Flask, PyTorch, Docker, VueJS, and SQL. Led an industrial project on automatic failure detection of diaper complaints with consumer argument detection and clustering. [link]

- **European Molecular Biology Laboratory, (EMBL)** — Heidelberg, Germany
  Research Trainee — Advisors: Dr. Toby Gibson, Dr. Manjeet Kumar — *Jun 2018 - Aug 2018*
  - Single-handedly developed a prediction toolkit using a weighted Logistic Regression (scikit-learn) model to computationally predict kinase substrate phosphorylation sites with (CAMK) protein sequences.
  - Researched heavily over dataset debiasing techniques, particularly focused on oversampling techniques such as SMOTE, nested cross-validation to avoid overfitting during hyperparameter tuning and Wilcoxon rank-sums test (α=0.05) to help us identify significant protein binding regions.
  - Conducted a thorough analysis on metrics used for evaluation of heavily-biased classification models using precision, f-measure, sensitivity and specificity along with ROC curves.

Work Experience

- **KNOLSKAPE** — Bengaluru, India
  Data Scientist — Manager: Mr. Chaithanya Yambari — *Sep 2018 - Oct 2019*
  - Designed and developed Krawler, an enterprise product for effectively searching a company's large messy content libraries. Developed the back-end architecture for efficient indexing. Implemented search and processing of data using Flask, Apache-Airflow, Elasticsearch and MongoDB. [link]
  - Worked on segmenting unstructured multimodal content into multiple subtopic segments. Particularly focused on unsupervised learning algorithms. Implemented and experimented with lexical (TextTiling) and semantic neural architectures for text segmentation, and conducted an error analysis.
  - Constructed an approximate content deduplication pipeline to identify near-duplicate multimodal contents. Computed hashes at scale and used Locality-sensitive hashing (LSH) and Perceptual hashing (PH) algorithms for detecting near duplicates within similar buckets for textual documents and images.

Honor and Awards

- Received University of Waterloo (UW) Graduate Scholarship for Doctoral Study in Computer Science (2021)
- BEIR: Only preprint publication to be included in teaching material in CS224U at Stanford University [link]
- Created and designed both the ELLIS NLP 2021 [link] and SustaiNLP 2021 workshop websites [link] (2021)
- Got Selected to speak for PyCon Italia titled "Extract or Replace Keywords in sentences 28x times faster than Regex - FlashText" (Cancelled due to Covid-19) (2020)
- Finalists in Technology Premier League (TPL), India's top IT strategy contest amongst fifty select corporate teams held by CIO & Leader, IT Next. (2019)
- Only UG student to receive a fully-funded Machine Learning (ML) Fellowship in EMBL, Heidelberg (2018)

Teaching Experience

- **Teaching Assistant** — Waterloo, Canada
  University of Waterloo — *2021 - Present*

  1) CS 135 (Designing Functional Programs) - Fall 2021
  2) CS 136 (Elementary Algorithm Design and Data Abstraction) - Winter 2022, Spring 2023
  3) CS 241 (Foundations of Sequential Programs) - Spring 2022
  4) CS 479/679 (Introduction to Artificial Intelligence) - Winter 2023

Reviewer/Program Committee

- **Reviewer (*CL/NLP conferences):** ACL Rolling Review: Oct-Nov (2021), Jan-Apr (2022)
- **Reviewer (ML conferences):** NeurIPS: June and July (2023)

Coursework

- **University of Waterloo:** CS 889: Data Sources for Emerging Tech (Ongoing), CS 886: Graph Neural Networks, CS 886: Robustness of Machine Learning, CS 679: Neural Networks, CS 848: Information Retrieval, CS 649: Human-Computer Interaction, CS 854: Experimental Performance Evaluation.
- **BITS Pilani:** Machine Learning, Neural Networks & Fuzzy Logic, Data Structures & Algorithms, Probability & Statistics, Linear Algebra, Econometric Methods, Discrete Mathematics.
- **Independent Study:** NLP by Deeplearning.ai (Coursera), Deep Learning for NLP (CS224d-Stanford), Machine Learning (Andrew NG), Django Introduction (Mike Hibbert), SWIRL (John Hopkins University)

Technical Skills

- **Programming:** Python, Flask, Pytorch, FastAPI, Tensorflow, VueJS, Django, JavaScript, ReactJS, R, C, C++, HTML, CSS, VBA, Advanced Excel, MATLAB, Racket, LaTeX.
- **Skills:** SQL, MongoDB, Docker, Elasticsearch, Redis, RabbitMq, Pub/Sub, Apache-Airflow, Postman.

Services

- **Machine Learning Volunteer** — Bangalore, India
  Knolskape Solutions Pvt. Ltd — *2018 - 2019*

  Organized a ML learning workshop for my colleagues, designed weekly assignments and took classes on Machine Learning. Implemented traditional ML models from scratch using pandas and scikit-learn.

- **Chief Coordinator, Mime Club** — Goa, India
  BITS Pilani KK Birla Goa Campus — *2014 - 2018*

  Led a team of 30 performers in one of the most active and popular clubs in college. Was actively involved in acting, sound mixing and creating stories for more than 12 shows over a span of 4 years for an audience of more than 2000 college students. [link]