Hassan Sajjad

Current Role	Associate Professor, Director of HyperMatrix lab
Address:	Faculty of Computer Science, Dalhousie University, Halifax, Canada
Email:	hsajjad@dal.ca
Webpage:	https://hsajjad.github.io/
Google Scholar:	https://goo.gl/H5XPzb
Membership:	ACM, IEEE

Research Interests

Develop reliable, safe and trustworthy AI systems

- Safe and Trustworthy AI - interpretability, explainability, robustness, generalization, compositionality, model safety, model editing and steering, and LLM evaluations
- Natural Language Processing and Applications - machine translation, summarization, language modeling, automatic evaluation

Education

- **University of Stuttgart** — **Germany**
  PhD in Computer Science (Magna Cum Laude) — *2008–2012*
  Thesis: Statistical models for unsupervised, semi-supervised and supervised transliteration mining
  Advisor: Prof. Dr. Hinrich Schütze
  Committee: Prof. Dr. Alex Waibel, PD Dr. Helmut Schmid, Dr. Alexander Fraser, Prof. Dr. Albrecht Schmidt, Prof. Dr.-Ing Stefan Funke

Professional Experience

- **Dalhousie University** — **Canada**
  Associate Professor — *Aug. 2022*

- **Qatar Computing Research Institute, HBKU** — **Qatar**
  Senior Research Scientist — *Jan. 2021–Aug. 2022*
  Managed the projects on machine translation and the interpretation of deep neural networks

- **Qatar Computing Research Institute, HBKU** — **Qatar**
  Research Scientist — *Jan. 2014–Dec. 2020*
  Managed the research, development and commercialization of the machine translation project, and international collaborations – MIT and H2020 SUMMA project

- **Qatar Computing Research Institute, HBKU** — **Qatar**
  Post-doctoral Researcher — *Feb. 2013–Jan. 2014*

Worked on dialectal Arabic machine translation

- **University of Stuttgart** — **Germany**
  Research Assistant and PhD Scholar — *Feb. 2008–Oct. 2012*
  Worked on unsupervised and semi-supervised methods

- **Microsoft Research** — **United States**
  Research Intern — *Jul. 2011–Oct. 2011*
  Worked on efficient vertical search. US Patent 9,767,144

- **Center for Research in Urdu Language Processing, NUCES** — **Pakistan**
  Research Assistant — *Feb. 2006–Feb. 2007*
  Worked on Pan Asian Networking Localization project

Publications

A complete list of my research work including the under-review articles can be found at my Google Scholar page (https://goo.gl/H5XPzb).

Peer-reviewed Journal Papers

Muhammad Irzam Liaqat et al. "Chameleon: A Multimodal Learning Framework Robust to Missing Modalities". In: *International Journal of Multimedia Information Retrieval* (2025).

Nadir Durrani, Fahim Dalvi, and **Hassan Sajjad**. "Discovering Salient Neurons in deep NLP models". In: *Journal of Machine Learning Research (JMLR)* 23-0074 (2023).

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. "On the Effect of Dropping Layers of Pre-trained Transformer Models". In: *Computer Speech Language (CSL)* 77 (2023), p. 101429. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2022.101429.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. "Neuron-level Interpretation of Deep NLP Models: A Survey". In: *Transactions of the Association for Computational Linguistics (TACL)* 10 (2022), pp. 1285–1303. DOI: https://doi.org/10.1162/tacl_a_00519.

Reem Suwaileh, Tamer Elsayed, Muhammad Imran, and **Hassan Sajjad**. "When a Disaster Happens, We are Ready: Location Mention Recognition from Crisis Tweets". In: *International Journal of Disaster Risk Reduction (IJDRR)* 78 (2022). (link), pp. 103–107.

Prakhar Ganesh et al. "Compressing Large-Scale Transformer-Based Models: A Case Study on BERT". In: *Transactions of the Association for Computational Linguistics (TACL)* 9 (2021). (link).

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, **Hassan Sajjad**, and James Glass. "On the Linguistic Representational Power of Neural Machine Translation Models". In: *Computational Linguistics (CL)* (2020). (link).

Abdul Rafae, Asim Karim, **Hassan Sajjad**, Faisal Kamiran, and Jia Xu. "A Clustering Framework for Lexical Normalization of Roman Urdu". In: *Natural Language Engineering (NLE)* (2020). (link).

Shafiq Joty, Nadir Durrani, **Hassan Sajjad**, and Ahmed Abdelali. "Domain Adaptation using Neural Network Joint Model". In: *Computer Speech and Language (CSL)* 45.C (2017). (link). ISSN: 0885-2308.

Hassan Sajjad, Helmut Schmid, Alexander Fraser, and Hinrich Schütze. "Statistical models for unsupervised, semi-supervised and supervised transliteration mining". In: *Computational Linguistics (CL)* 43.2 (2017). (link).

Walid Magdy, **Hassan Sajjad**, Tarek El-Ganainy, and Fabrizio Sebastiani. "Bridging social media via distant supervision". In: *Social Network Analysis and Mining* 35.5 (2015). (link).

Peer-reviewed Conference Papers

Enes Altinisik, Safa Messaoud, Husrev Taha Sencar, **Hassan Sajjad**, and Sanjay Chawla. "Explaining the role of Intrinsic Dimensionality in Adversarial Training". In: *International Conference on Machine Learning (ICML),* Vancouver, Canada, 2025.

Elahe Rahimi, **Hassan Sajjad**, Domenic Rosati, Abeer Badawi, Elham Dolatabadi, and Frank Rudzicz. "Not Lost After All: How Cross-Encoder Attribution Challenges Position Bias Assumptions in LLM Summarization". In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Suzhou, China, 2025.

Hammad Rizwan, Domenic Rosati, Ga Wu, and **Hassan Sajjad**. "Resolving Lexical Bias in Model Editing". In: *International Conference on Machine Learning (ICML),* Vancouver, Canada, 2025.

Mahtab Sarvmaili, **Hassan Sajjad**, and Ga Wu. "Data-Centric Prediction Explanation via Kernelized Stein Discrepancy ". In: *International Conference on Learning Representations (ICLR),* (link). Singapore, 2025.

Paolo Gajo, Domenic Rosati, **Hassan Sajjad**, and Alberto Barrón-Cedeño. "Dependency Parsing is More Parameter-Efficient with Normalization". In: *Conference on Neural Information Processing Systems (NeurIPS)*. (link). Vancouver, Canada, Dec. 2025.

David Arps, Laura Kallmeyer, Younes Samih, and **Hassan Sajjad**. "Multilingual Nonce Dependency Treebanks: Understanding how LLMs Represent and Process Syntactic Structure". In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. (link). Mexico City, Mexico, June 2024.

Domenic Rosati et al. "Long-form Evaluation of Model Editing". In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. (link). Mexico City, Mexico, June 2024.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and **Hassan Sajjad**. "SUGARCREPE++ Dataset: Vision-Language Model Sensitivity to Semantic and Lexical Alterations". In: *Conference on Neural Information Processing Systems, Dataset Track (NeurIPS),* (link). Vancouver, Canada, Dec. 2024.

Domenic Rosati et al. "Representation Noising: A Defence Mechanism Against Harmful Finetuning". In: *Conference on Neural Information Processing Systems (NeurIPS),* (link). Vancouver, Canada, Dec. 2024.

Domenic Rosati et al. "Immunization against Harmful Fine-tuning Attacks". In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Miami, Florida, Nov. 2024.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and **Hassan Sajjad**. "Latent Concept-based Explanation of NLP Models". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Miami, Florida, Nov. 2024.

Enes Altinisik, **Hassan Sajjad**, Husrev Taha Sencar, Safa Messaoud, and Sanjay Chawla. "Impact of Adversarial Training on Robustness and Generalizability of Language Models". In: *Proceedings of the Findings of Association for Computational Linguistics (ACL)*. (link). Toronto, Canada, July 2023.

Yu Yu, **Hassan Sajjad**, and Jia Xu. "Learning Uncertainty for Unknown Domains with Zero-Target-Assumption". In: *International Conference on Learning Representations (ICLR)*. (link). Kigali Rwanda, May 2023.

Yimin Fan, Fahim Dalvi, Nadir Durrani, and **Hassan Sajjad**. "Evaluating Neuron Interpretation Methods of NLP Models". In: *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*. New Orleans, US, Dec. 2023. URL: https://openreview.net/forum?id=YiwMpyMdPX.

David Arps, Younes Samih, Laura Kallmeyer, and **Hassan Sajjad**. "Probing for Constituency Structure in Neural Language Models". In: *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Abu Dhabi, 2022.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and **Hassan Sajjad**. "Discovering Latent Concepts Learned in BERT". In: *International Conference on Learning Representations (ICLR)*. (link). Online, 2022.

Nadir Durrani, **Hassan Sajjad**, Fahim Dalvi, and Firoj Alam. "On the Transformation of Latent Space in Fine-Tuned NLP Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Abu Dhabi, 2022.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. "Analyzing Encoded Concepts in Transformer Language Models". In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. (link). Seattle, US, 2022.

Hassan Sajjad, Firoj Alam, Fahim Dalvi, and Nadir Durrani. "Effect of Post-processing on Contextualized Word Representations". In: *Proceedings of the International Conference on Computational Linguistics (COLING)*. (link). Gyeongju, Republic of Korea, Oct. 2022.

Nadir Durrani, **Hassan Sajjad**, and Fahim Dalvi. "How Transfer Learning Impacts Linguistic Knowledge in Deep NLP Models?" In: *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*. (link). Online, Aug. 2021.

Firoj Alam, **Hassan Sajjad**, Muhammad Imran, and Ferda Ofli. "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing". In: *International Conference on Web and Social Media (ICWSM)*. (link). Online, June 2021.

Firoj Alam et al. "Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms". In: *International Conference on Web and Social Media (ICWSM)*. (link). Online, June 2021.

Esther Seyffarth and Younes Samih, Laura Kallmeyer, and Hassan Sajjad. "Implicit Representations of Event Properties within Contextual Language Models: Searching for "Causativity Neurons"". In: *International Conference on Computational Semantics (IWCS)*. (link). Groningen, Netherlands, June 2021.

Firoj Alam et al. "Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society". In: *Empirical Methods in Natural Language Processing (EMNLP)*. Online, Nov. 2021.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. "AraBench: Benchmarking Dialectal Arabic-English Machine Translation". In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. (link). Online, Dec. 2020.

Reem Suwaileh, Muhammad Imran, Tamer Elsayed, and **Hassan Sajjad**. "Are We Ready for this Disaster? Towards Location Mention Recognition from Crisis Tweets". In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. (link). Online, Dec. 2020.

Fahim Dalvi, **Hassan Sajjad**, Nadir Durrani, and Yonatan Belinkov. "Analyzing Redundancy in Pretrained Transformer Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Online, Nov. 2020.

Nadir Durrani, **Hassan Sajjad**, Fahim Dalvi, and Yonatan Belinkov. "Analyzing Individual Neurons in Pretrained Language Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. links. Online, Nov. 2020.

John Wu, Yonatan Belinkov, **Hassan Sajjad**, Nadir Durrani, Fahim Dalvi, and James Glass. "Similarity Analysis of Contextual Word Representation Models". In: *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*. (link). US, June 2020.

Nadir Durrani, Fahim Dalvi, **Hassan Sajjad**, Yonatan Belinkov, and Preslav Nakov. "One Size Does Not Fit All: Comparing NMT Representations of Different Granularities". In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. (link). Minneapolis, US, June 2019.

Hamdy Mubarak, Ahmed Abdelali, **Hassan Sajjad**, Younes Samih, and Kareem Darwish. "Highly Effective Arabic Diacritization using Sequence to Sequence Modeling". In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. (link). Minneapolis, US, June 2019.

D. Anthony Bau*, Yonatan Belinkov*, **Hassan Sajjad**, Fahim Dalvi, Nadir Durrani, and James Glass. "Identifying and Controlling Important Neurons in Neural Machine Translation". In: *International Conference on Learning Representations (ICLR)*. (link). New Orleans, US, May 2019.

Fahim Dalvi*, Nadir Durrani*, **Hassan Sajjad***, Yonatan Belinkov, D. Anthony Bau, and James Glass. "What is one Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. (link). Honolulu, US, Mar. 2019.

Fahim Dalvi, Nadir Durrani, **Hassan Sajjad**, and Stephan Vogel. "Incremental Decoding and Training Methods for Simultaneous Translation in Neural Machine Translation". In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. (link). New Orleans, US, June 2018.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, **Hassan Sajjad**, and James Glass. "What do Neural Machine Translation Models Learn about Morphology?" In: *Proceedings of the 55th Conference of the Association for Computational Linguistics (ACL)*. (link). Vancouver, Canada, Aug. 2017.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. "Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging". In: *Proceedings of the 55th Conference of the Association for Computational Linguistics (ACL)*. (link). Vancouver, Canada, Aug. 2017.

Dat Tien Nguyen, Kamla Al-Mannai, Shafiq Joty, **Hassan Sajjad**, Muhammad Imran, and Prasenjit Mitra. "Robust Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks". In: *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. (link). Montreal, Canada, May 2017.

Yonatan Belinkov, Lluís Màrquez, **Hassan Sajjad**, Nadir Durrani, Fahim Dalvi, and James Glass. "Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks". In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*. (link). Taipei, Taiwan, Nov. 2017.

Fahim Dalvi, Nadir Durrani, **Hassan Sajjad**, Yonatan Belinkov, and Stephan Vogel. "Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder". In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*. (link). Taipei, Taiwan, Nov. 2017.

Hassan Sajjad et al. "Eyes Don't Lie: Predicting Machine Translation Quality Using Eye Movement." In: *Proceedings of the 15th Annual Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL-HLT)*. (link). San Diego, US, June 2016.

Nadir Durrani, **Hassan Sajjad**, Shafiq Joty, and Ahmed Abdelali. "A Deep Fusion Model for Domain Adaptation in Phrase-based MT". In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. (link). Osaka, Japan, Dec. 2016.

Hassan Sajjad, Francisco Guzmán, and Stephan Vogel. "An Empirical Study: Post-editing Effort for English to Arabic Hybrid Machine Translation". In: *Proceedings of the Association for Machine Translation in the Americas (AMTA)*. (link). Austin, US, Oct. 2016.

Shafiq Joty, **Hassan Sajjad**, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. "How to Avoid Unwanted Pregnancies: Domain Adaptation using Neural Network Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Lisbon, Portugal, Sept. 2015.

Abdul Rafae, Abdul Qayyum, Muhammad Moeen Uddin, Asim Karim, **Hassan Sajjad**, and Faisal Kamiran. "An Unsupervised Method for Discovering Lexical Variations in Roman Urdu Informal Text." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Lisbon, Portugal, Sept. 2015.

Walid Magdy, **Hassan Sajjad**, Tarek El-Ganainy, and Fabrizio Sebastiani. "Distant Supervision for Tweet Classification Using YouTube Labels". In: *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM)*. (link). Oxford, UK, May 2015.

Nadir Durrani, **Hassan Sajjad**, Shafiq Joty, Ahmed Abdelali, and Stephan Vogel. "Using Joint Models for Domain Adaptation in Statistical Machine Translation". In: *Proceedings of the 15th Machine Translation Summit (MT Summit XV)*. (link). Florida, USA, Nov. 2015.

Ahmed Abdelali, Francisco Guzman, **Hassan Sajjad**, and Stephan Vogel. "The AMARA Corpus: Building Parallel Language Resources for the Educational Domain". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. (link). Reykjavik, Iceland, May 2014.

Mohammad Moeen Uddin, Mohammad Imran, and **Hassan Sajjad**. "Understanding Types of Users on Twitter". In: *Proceedings of the 6th ASE International Conference in Social Computing (SocialCom)*. (link). Stanford, USA, May 2014.

Nadir Durrani, **Hassan Sajjad**, Hieu Hoang, and Philipp Koehn. "Integrating an Unsupervised Transliteration Model into Statistical Machine Translation". In: *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL)*. (link). Gothenburg, Sweden, Apr. 2014.

Kareem Darwish, **Hassan Sajjad**, and Hamdy Mubarak. "Verifiably Effective Arabic Dialect Identification." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (link). Doha, Qatar, Oct. 2014.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. "Translating Dialectal Arabic to English". In: *Proceedings of the 51st Conference of the Association for Computational Linguistics (ACL)*. (link). Sofia, Bulgaria, Aug. 2013.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. "A Statistical Model for Unsupervised and Semi-supervised Transliteration Mining". In: *Proceedings of the 50th Conference of the Association for Computational Linguistics (ACL)*. (link). Jeju, Korea, July 2012.

Hassan Sajjad, Patrick Pantel, and Michael Gamon. "Underspecified Query Refinement via Natural Language Question Generation". In: *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. (link). Mumbai, India, Dec. 2012.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. "An Algorithm for Unsupervised Transliteration Mining with an Application to Word Alignment". In: *Proceedings of the 49th Conference of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. (link). Portland, OR, USA, June 2011.

Hassan Sajjad, Nadir Durrani, Helmut Schmid, and Alexander Fraser. "Comparing Two Techniques for Learning Transliteration Models Using a Parallel Corpus". In: *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*. (link). Chiang Mai, Thailand, Nov. 2011.

Nadir Durrani, **Hassan Sajjad**, Alexander Fraser, and Helmut Schmid. "Hindi-to-Urdu Machine Translation through Transliteration". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. (link). Uppsala, Sweden, July 2010.

Hassan Sajjad and Helmut Schmid. "Tagging Urdu Text with Parts of Speech: A Tagger Comparison". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. (link). Athens, Greece, Apr. 2009.

Peer-reviewed Demo Conference Papers

Fahim Dalvi, **Hassan Sajjad**, and Nadir Durrani. "NeuroX Library for Neuron Analysis of Deep NLP Models". In: *Proceedings of the Association for Computational Linguistics (ACL)*. Toronto, Canada, July 2023.

Fahim Dalvi, Nadir Durrani, **Hassan Sajjad**, Tamim Jaban, Mus'ab Husaini, and Ummar Abbas. "NxPlain: A Web-based Tool for Discovery of Latent Concepts". In: *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*. Dubrovnik, Croatia, May 2023.

Firoj Alam, Fahim Dalvi, Nadir Durrani, **Hassan Sajjad**, Abdul Rafae Khan, and Jia Xu. "ConceptX: A Framework for Latent Concept Analysis". In: *AAAI Conference on Artificial Intelligence (AAAI)*. (link). Washington DC, USA, Feb. 2023.

Hamdy Mubarak, Ahmed Abdelali, Kareem Darwish, Mohamed Eldesouki, Younes Samih, and **Hassan Sajjad**. "A System for Diacritizing Four Varieties of Arabic". In: *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*. (link). Hong Kong, China, Nov. 2019.

Fahim Dalvi et al. "NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks". In: *AAAI Conference on Artificial Intelligence (AAAI)*. (link). Honolulu, USA, Jan. 2019.

Fahim Dalvi et al. "QCRI Live Speech Translation System". In: *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. (link). Valencia, Spain, Apr. 2017.

Renars Liepins et al. "The SUMMA Platform Prototype". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. (link). Valencia, Spain, Apr. 2017.

Peer-reviewed Workshop Papers

Sher Badshah and **Hassan Sajjad**. "Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form QA". In: *Proceedings of the Workshop on Widening NLP (WiNLP), EMNLP 2025*. Suzhou, China, Nov. 2025.

Domenic Rosati, Sebastian Dionicio, Xijie Zeng, Subhabrata Majumdar, Frank Rudzicz, and **Hassan Sajjad**. "Locking Open Weight Models with Spectral Deformation". In: *Proceedings of the Technical AI Governance Workshop, ICML*. Vancouver, Canada, July 2025.

Domenic Rosati et al. "Evaluating Defences against Unsafe Feedback in RLHF". In: *Proceedings of the Workshop on Artificial Intelligence for Cyber Security (AICS), AAAI 2025*. Philadelphia, USA, Mar. 2025.

Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, and **Hassan Sajjad**. "Post-hoc analysis of Arabic transformer models". In: *Proceedings of the BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. (link). Abu Dhabi, Nov. 2022.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. "Neural Machine Translation Training in a Multi-Domain Scenario". In: *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*. (link). Tokyo, Japan, Dec. 2017.

Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, **Hassan Sajjad**, and Prasenjit Mitra. "Applications of Online Deep Learning for Crisis Response Using Social Media Information". In: *Proceedings of the 4th International Workshop on Social Web for Disaster Management (SWDM)*. (link). Indianapolis, US, Oct. 2016.

Wajdi Zaghouani, Ahmed Abdelali, Francisco Guzmán, and **Hassan Sajjad**. "Normalizing Mathematical Expressions to Improve the Translation of Educational Content". In: *Proceedings of the AMTA 2016 Workshop Semitic Machine Translation (SeMaT)*. (link). Austin, US, Oct. 2016.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, **Hassan Sajjad**, and Stephan Vogel. "How do Humans Evaluate Machine Translation". In: *Proceedings of the Workshop on Machine Translation (WMT)*. (link). Lisbon, Portugal, Sept. 2015.

Kamla Al-Mannai, **Hassan Sajjad**, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. "Unsupervised word segmentation improves dialectal Arabic to English machine translation". In: *Proceedings of the Workshop of Arabic Natural Language Processing (ANLP)*. (link). Doha, Qatar, Oct. 2014.

Francisco Guzmán, **Hassan Sajjad**, Stephan Vogel, and Ahmed Abdelali. "The AMARA Corpus: Building Resources for Translating the Web's Educational Content". In: *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT)*. (link). Heidelberg, Germany, Dec. 2013.

Shared-task Workshop Papers

Lucia Specia et al. "Findings of the WMT 2020 Shared Task on Machine Translation Robustness". In: *Proceedings of the Conference on Machine Translation (WMT), Shared Task Papers*. Online, Nov. 2020.

Xian Li et al. "Findings of the First Shared Task on Machine Translation Robustness". In: *Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*. (link). Florence, Italy, Aug. 2019.

Houda Bouamor and **Hassan Sajjad**. "H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings". In: *Proceedings of the 11th Workshop on Building and Using Comparable Corpora (BUCC)*. (link). Miyzaki, Japan, May 2018.

Nadir Durrani, Fahim Dalvi, **Hassan Sajjad**, and Stephan Vogel. "QCRI's Machine Translation Systems for IWSLT'2016". In: *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. (link). Seattle, USA, Dec. 2016.

Mohamed Eldesouki, Fahim Dalvi, **Hassan Sajjad**, and Kareem Darwish. "QCRI @ DSL 2016: Spoken Arabic Dialect Identification Using Textual". In: *Proceedings of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects*. (link). Osaka, Japan, Dec. 2016.

Houda Bouamor, **Hassan Sajjad**, Nadir Durrani, and Kemal Oflazer. "QCMUQ@QALB-2015 Shared Task: Combining Character level MT and Error-tolerant Finite-State Recognition for Arabic Spelling Correction". In: *Proceedings of the Workshop of Arabic Natural Language Processing (ANLP)*. (link). Beijing, China, July 2015.

Hassan Sajjad et al. "QCN Egyptian Arabic to English Machine Translation System for NIST OpenMT15". In: *Workshop of NIST OpenMT15*. (link). Washington DC, US, June 2015.

Nadir Durrani, Helmut Schmid, Alexander Fraser, **Hassan Sajjad**, and Richárd Farkas. "Munich-Edinburgh-Stuttgart Submissions of OSM Systems at WMT13". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. (link). Sofia, Bulgaria, Aug. 2013.

Hassan Sajjad, Svetlana Smekalova, Nadir Durrani, Alexander Fraser, and Helmut Schmid. "QCRI-MES Submission at WMT13: Using Transliteration Mining to Improve Statistical Machine Translation". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. (link). Sofia, Bulgaria, Aug. 2013.

Marion Weller et al. "Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. (link). Sofia, Bulgaria, Aug. 2013.

Hassan Sajjad et al. "QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation". In: *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT)*. (link). Heidelberg, Germany, Dec. 2013.

Tutorials

- Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi and Nadir Durrani, "Fine-grained Interpretation and Causation Analysis in Deep NLP Models". In Proceedings of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL-HLT), June 2021.

Teaching Experience

- **NLP with Deep Learning** — **Canada**
  Dalhousie University — *Winter 2023/24/25*
- **CSCI 1108 – Experimental Robotics** — **Canada**
  Dalhousie University — *Fall 2022/24/25, Winter 2023/25*

- **NLP with Python** — **Qatar**
  Carnegie Mellon University in Qatar — *19-30th Oct. 2020*
  As part of the NLP with Python course, delivered 6 hours lectures covering the fundamentals of deep learning with practical exercises

- **Deep Learning for NLP** — **Germany**
  University of Duisburg-Essen — *9-13th Sept. 2019*
  15 hours course covering the fundamentals of deep learning from intuition to practical exercises

- **Deep Learning for NLP** — **Sri Lanka**
  International Spring School, University of Moratuwa — *16-21st Mar. 2019*
  15 hours course on the basics of deep learning with hands-on exercises involving several NLP tasks

- **From Theory to Practice: Deep Learning for NLP** — **Germany**
  University of Duisburg-Essen — *8-14th Apr. 2018*
  15 hours course covering theory of deep learning models from intuition to practical exercises in Keras (content)

- **Deep Learning for Machine Translation** — **Germany**
  9th Computational linguistics fall school, DGfS — *11-22nd Sept. 2017*
  15 hours course teaching fundamentals of deep learning and its application to machine translation (content)

- **Probability and Statistics** — **Pakistan**
  National University of Computer and Emerging Sciences — *Sept. 2004–Dec. 2004*
  Teaching assistant

- **Calculus** — **Pakistan**
  National University of Computer and Emerging Sciences — *Feb. 2003–Jun. 2003*
  Teaching assistant

- **Introduction to Natural Language Processing** — **Pakistan**
  ADDO AI — *Nov. 2019–Jan. 2020*
  Course content design and delivery advisor

Grants

- FCS Research Equipment Grant (FCS-REG), **Hassan Sajjad (co-PI)**, CAD 35,000 (2025)
- Personalized User Engagement based on Activity Insights, MITACS, **Hassan Sajjad (PI)**, CAD 30,000 (2024)
- Transforming Climate Action, Large Research Project Application for Cluster 1.2. (TCA -AI), Canada First Research Excellence Fund (CFREF), **Hassan Sajjad (co-applicant)**, CAD 1,065,333 (2024-2028)
- Interpreting Deep Learning Models of Natural Language Processing, Research Nova Scotia, **Hassan Sajjad (PI)**, CAD 148,553 (2023)
- Interpreting Deep Learning Models of Natural Language Processing, CFI JELF, **Hassan Sajjad (PI)**, CAD 148,553 (2022)
- Fine-grained Interpretation of Deep Neural Network Models of NLP, Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, **Hassan Sajjad (PI)**, CAD 145,000 (2022-2027)

Underreview Grants

- Human-inspired Generative AI for Psycholinguistic Research (HAIPs). Open Up – New Research Spaces for the Humanities and Cultural Studies, Volkswagenstiftung, **Hassan Sajjad (co-PI)** with two other co-PIs. Euro Euro 380,360 (2026-2027)

A few Unsuccessful Grants

- Human Inspiration for Generative AI: Insights into (Human) Language Processing (HuGAI). ERC Synergy Grant 2025. **Hassan Sajjad (co-PI)** with two other co-PIs. Euro 10,000,000 (2026-2032)
- Human-Centered AI for Industrial Preventive Maintenance (HCAI-Prev), HORIZON-CL4-2024-HUMAN-03-02 Explainable and Robust AI. **Hassan Sajjad (co-PI)** with several other institutes. Euro 15,000,000 (2025-2029)

Patent

- Latent concept analysis method, US Patent US20230325426
- Search system with query refinement, US Patent 9,767,144
- Method and system for diacritizing Arabic text, US Patent App. 17/598,633

Open Source Contributions

- NeuroX - Interpretation and manipulation of deep models (alpha version: link)
- LayerDrop for efficient transfer learning using pre-trained models (integrated into a fork of the huggingface transformer library, link) (paper)
- Transliteration Mining Tool (link)
- Linguistic Resources
  - SugarCrepe++ (link)
  - AraBench - Dialect Arabic to English evaluation suite (link)
  - The QCRI Educational Domain corpus – a collection of parallel corpora of the education domain in multiple languages (link)
  - English-Hindi and English-Arabic gold set for transliteration mining evaluation (link)
  - Most widely-used part of speech tagset of Urdu and annotated corpus of 100,000 words

Talks and Keynotes

- Latent Concept-Based Explanation of NLP Models. Workshop on Decoding Decisions: Explainability in ML & Sequential Decision Making at the Conference on Robots and Vision (May 2025)
- Navigating Latent Space for Safety, Interpretability and Explainability. New York University (May 2025)
- Are LLMs a Good Model of Human Thought? The Challenge of Compositional Learning. Atlantic Canada AI Summit (May 2025)
- Latent Space Exploration for Safe and Trustworthy AI Models. MBZUAI, Abu Dhabi (Aug. 2024)
- Latent Space Exploration for Safe and Trustworthy AI Models. Representation learning for NLP at ACL (Aug. 2024) (**Keynote**)
- Latent Space Exploration for Safe and Trustworthy AI Models. AI@Thomson Reuters (Apr. 2024)

- Neuron Interpretation of Deep NLP Models. ITU, Lahore, Pakistan (Dec. 2023)
- Latent Concept based Explanation of Deep Learning Models. MBZUAI, Abu Dhabi (Nov. 2023)
- Latent Concepts in Transformer Models of NLP. UKP, TU Darmstadt, Germany (Jun. 2023)
- Knowledge Manifolds in Transformer Models of NLP. National Research Council (NRC), Canada (Apr. 2023)
- Knowledge Manifolds in Transformer Models of NLP. TAMALE seminar, University of Ottawa, Canada (Apr. 2023)
- Analyzing Latent Concepts in Deep Neural Network Models of NLP. STCI Microsoft, India (Jun. 2022)
- Analyzing Latent Concepts in Deep Neural Network Models of NLP. Data Science Institute, National University of Ireland Galway, (Jun. 2022)
- Exploiting Redundancy in Pre-trained Models for Efficient Transfer Learning. Machine Learning and Data Analytics Symposium, Qatar (Mar. 2021)
- Exploiting Redundancy in Pre-trained Models for Efficient Transfer Learning. Facebook, US (Feb. 2021)
- Exploiting Redundancy in Pre-trained Models for Efficient Transfer Learning. National Research Council (NRC), Canada (Nov. 2020)
- Hidden Linguistics in Deep NLP Models. Heinrich-Heine Universität Düsseldorf, Germany (Oct. 2020)
- Interpreting Deep NLP Models. University of Edinburgh, UK (Apr. 2020)
- Interpreting Deep NLP Models. University of Sheffield, UK (Mar. 2020)
- Efficient Transfer Learning of Pre-trained Model. 7th International Conference on Language and Technology (link), UET, Pakistan (Feb. 2020) (**Keynote**)
- Interpreting Deep NLP Models: A Case Study on Neural MT at Google, US (Apr. 2019)
- Interpreting Deep NLP Models: A Case Study on Neural MT at Salesforce, US (Apr. 2019)
- Interpreting Deep NLP Models: A Case Study on Neural MT at Bosch, US (Apr. 2019)
- Interpreting Deep NLP Models: A Case Study on Neural MT at Facebook, US (Apr. 2019)
- Interpreting Deep NLP Models: A Case Study on Neural MT at Amazon, US (Apr. 2019)
- Hidden Linguistics in Deep NLP Models. Symposium on Natural Language Processing. University of Moratuwa, Sri Lanka (Mar. 2019) (**Keynote**)
- Machine Translation in the Real World. King's College London, UK (Mar. 2019)
- Analyzing Individual Neurons in Deep NLP Models. University of Melbourne, Melbourne, Australia (Feb. 2019)
- Analyzing Individual Neurons in Deep NLP Models. Thomson Reuters, Toronto, Canada (Feb. 2019)
- What do Neural Machine Translation Models Learn about Morphology. Macquarie University, Sydney, Australia (Apr. 2017)
- What do Neural Machine Translation Models Learn about Morphology. University of Sydney, Sydney, Australia (Apr. 2017)
- From Phrase-based to Neural Machine Translation. Workshop on Semitic Machine Translation, AMTA, Austin, US (Nov. 2016) (**Keynote**)
- Deep Learning – Neural Machine Translation. Sixth Conference on Language and Technology

(CLT16), Lahore, Pakistan (Nov. 2016) (**Keynote**)

- Content Model Applications for Promoting Local Language Content. Workshop on Facilitating Local Language Content Access and Generation using Human Language Technologies, UET, Lahore, Pakistan (Aug. 2015) (**Keynote**)
- Statistical Machine Translation for Community Service: Translating Educational Content. Fifth Conference on Language and Technology (CLT14), Karachi, Pakistan (Nov. 2014) (**Keynote**)
- Separating Transliterations from Translations in Transliteration Mining Context. FBK, Trento, Italy (Oct. 2012)
- Unsupervised Transliteration Mining. School of Science and Engineering, Lahore University of Management and Sciences, Pakistan (Apr. 2012)
- Unsupervised Transliteration Mining, Punjab University College of Information Technology, Pakistan (Apr. 2012)

Honors and Awards

- Panelist on discussing Data Centric AI for Reliable Models at the 9th workshop on representation learning for NLP at ACL 2024 (Aug. 2024)
- Panelist at the Halifax Innovation Challenge, Canada (13-14th Oct. 2023)
- Media coverage: pioneer dialectal Arabic translation Gulf Times, MENAFN, HBKU (Aug. 2020)
- Outstanding reviewer at the EMNLP 2020 conference
- Machine translation technology deployed to BBC and Deutsche Welle as part of the SUMMA project (2019) and tech transfer to KanariAI (June 2020),
- Media coverage: 1 billion tokens translated by our machine translation system (Aug. 2020). (link1), (link2)
- Panelist on discussing NLP Research in Pakistan: Building Synergies and Collaborations at 7th International Conference on Language and Technology (Feb. 2020). (link)
- Media coverage: MIT News covers our work on analyzing and controlling deep models (Feb. 2019) (link)
- Best Innovation Award at the Annual Research Conference 2018, Qatar for our speech translation system (link)
- Media coverage: MIT News covers our work on analyzing representations in deep models (Dec. 2017) (link)
- Qatar Science and Technology Park speed pitching (Nov. 2015) – selected among the top 3 teams out of 26 teams
- Second position in the Qatar Annual Research Conference 2014 for a student research project
- Google sponsorship for participation in the Second Lisbon Machine Learning School
- Full PhD merit-based scholarship from HEC-DAAD for doctoral studies in Germany (4 years)
- Four year undergraduate scholarship

Research Competitions

- Workshop on Building and using Comparable Corpora (2018), French-English parallel corpus extraction
- **Best system** – International Workshop on Spoken Language Technology (2016), Arabic-English

translation task

- **Second best system and best system description paper** – NIST Open Machine Translation Evaluation Workshop (2015), Egyptian Arabic to English translation task
- **Second best system** – Shared Task on Automatic Arabic Error Correction (2015)
- **Best system** – International Workshop on Spoken Language Technology (2013), Arabic-English translation task
- **Second best constrained system** – Eighth Workshop on Statistical Machine Translation (2013), English-Russian translation task

Professional Services

- Tutorial chair: EMNLP 2023
- Area chair: NeurIPS 2025/24/23, ARR (frequent), NAACL 2024/22/21, EMNLP 2024/23/21, AACL 2022, ACL 2020
- Senior PC: AAAI 2025/22/21, ACL 2022
- Workshop/shared task organization: MT robustness 2019/20 (WMT), BlackBoxNLP 2020/21 (EMNLP)
- Conference reviewer: regular reviewer at ACL/ICML/ICLR/EMNLP/NAACL/EACL/COLING, NeurIPS 2020/21, CVPR 25, AAAI 2019/20, CONLL 2021, IJCAI 2018/19, EAMT 2017
- Workshop reviewer: IWSLT, WANLP, Blackbox NLP 2022, Rep4NLP 2024
- Journal reviewer: TACL, JAIR, ML, NLE, LRE, CSL, PAMI, PLOS-ONE, IEEE TNNLS
- Journal review editor: Language and computation
- Challenge22: proposal reviewer for phase 1 and phase 2 (2017)
- Judge at the annual Hackathon at Carnegie Mellon University - Qatar (27-28th Jan. 2017)
- Judge at the first Alice Competition at Carnegie Mellon University - Qatar (23rd May 2016)

Advising

- Ph.D. advisor (2023 –): Hammad Rizwan
- Ph.D. advisor (2022 –): Xuemin Yu
- Ph.D. advisor (2022 –): Sher Badshah
- Ph.D. co-advisor with Janarthanan Rajendran (2025 –): Biruk Abere Ambaw
- Ph.D. co-advisor with Ga Wu (2025 –): Naihe Feng
- Ph.D. co-advisor with Frank Rudzicz (2023 –): Domenic Rosati
- Ph.D. co-advisor with Alberto Barrón-Cedeño (2023 –): Paolo Gajo
- Ph.D. co-advisor with Laura Kallmeyer (2023 –): David Arps
- BCS Thesis advisor (2025): Sebastian Dionicio

Completed

- MCS advisor (2023–24): Manpreet Singh (graduated)
- BCS Thesis advisor (2024): Enyu Ye (graduated)

- BCS Thesis advisor (2023): Megan Cao (graduated, joined masters at the University of Toronto)

External PhD/MS Committee

- Ph.D.: Out-of-distribution Robustness. Yu Yu. Stevens Institute of Technology, NJ, US. Nov 2022
- MCS: Human-in-the-loop Classification for Multi-page Administrative Documents, Rakshit Makan. Dalhousie University, Canada. Dec 2022
- MCS: Multi-Modal Consensus Clustering to Identify Kidney Transplant Donor and Recipient Phenotypes, Kranthi Kiran Jalakam. Dalhousie University, Canada. Nov 2022
- Ph.D.: Jamal Abdul Nasir. Lahore University of Management Sciences, Pakistan. 2014

Masters and Bachelor Advising

- Laura Lehmann, LMU Munich, Germany (External examiner of Bachelor thesis titled MulTraLit: a multilingual transliteration system). 2020
- Kamela Ali Al Mannai, HBKU, Qatar (External advisor, Master thesis on Dialect Identification), 2018
- Abdul Rafae, LUMS, Pakistan (External advisor, later did PhD from the Hunter College City University, New York)

Research Internship

- Sasidhar Kunapuli, Mar. – Aug. 2025
- Jarrod Conrad, April – Aug. 2025
- Afif Asad, April – Aug. 2025
- Marzia Nouri, April 2024 –
- Moamen Moustafa, July 2024 –

Completed

- Manuj Malik, May – Sept. 2024
- BCS internship (2022–23): Neelam Uppal
- Sherin Rasheed, Aug. 2021 - Feb 2022
- Kamran Janjua, Jan. - Aug. 2021 (now a master student at the University of Alberta, Canada)
- Fahim Dalvi, 2013 (graduated from Stanford University, now working at QCRI)
- Dat Tien Nguyen (now a PhD student at the University of Amsterdam)
- Kamela Ali Al Mannai (now a PhD student at Hamad Bin Khalifa University, Qatar)
- Alaa Khader (now a research coordinator at NASA BHP, US)

Participations

- Deep Learning for Machine Translation, Winter School, Dublin, Ireland (18-24th Oct. 2015)
- Dagstuhl Seminar on Statistical Techniques for Translating to Morphologically Rich Languages (2-7th Feb. 2014)
- Taming the social web, 2nd Lisbon Machine Learning School (19-25th Jul. 2012)
- 5th MT Marathon, Le Mans France (13-17 Sept. 2010). Worked with Adam Lopez on Minimum

Risk Decoding in Cdec

- Summer School in Asian Language Processing (Jun.-Aug. 2006)