Ben Hagag

AI RESEARCHER | DATA SCIENTIST

Profile

AI researcher and data scientist with over ten years of experience in developing algorithms and leveraging ML, NLP, and deep learning for security, justice, and litigation technology. Proven leader across academia and industry, with a focus on interdisciplinary collaboration and ethical AI solutions.

Employment History

Head of Research, Darrow.ai, Tel Aviv

JUNE 2021 — PRESENT

- Leading the research department in developing proprietary methodologies and algorithms, leveraging NLP, large language model (LLM) adaptation, and machine learning to transform unstructured data into actionable legal intelligence.
- Directing the development of the LegalLens system, a pioneering AI framework that enhances accuracy in detecting legal violations, driving Darrow's growth and profitability.
- Managing innovative projects using advanced NLP and machine learning techniques, presenting results at leading conferences such as EACL and EMNLP, and organizing a global shared task at NLLP 2024 with participation from over 100 researchers. Promoting high standards in open-code research and publishing practices.
- Raising Darrow's profile within AI-driven legal technology by presenting key advancements at AI and NLP conferences and establishing collaborative research initiatives that advance sophisticated legal violation detection methods.
- Strengthening partnerships with a diverse network of legal, technological, and social experts, collaborating with world-class universities. Aligning advanced AI research with practical legal applications to enhance Darrow's strategic objectives and industry reputation.

Lead Data Scientist (Major, Res.), the Israeli Intelligence Department | 8200 Signal Intelligence Unit, Tel Aviv

DECEMBER 2012 — JULY 2021

- Overseeing a high-performing team of data scientists and engineers in exploring and applying advanced AI techniques in network security and cyber threat intelligence. Developed proprietary algorithms and innovative solutions that advanced military technology.
- Managing intelligence-focused projects using Apache Spark, Hadoop, and TensorFlow, creating data-driven models critical for national security. Leveraged advanced machine and deep learning algorithms to detect threats and suspicious behaviors, enhancing detection capabilities and decision-making processes.
- Designing and implementing large-scale data science projects, using machine ML, NLP, and deep learning to extract actionable insights from vast, structured, and unstructured datasets, improving detection capabilities for signal anomalies and potential threats.
- Leading an organization-wide initiative to develop trustworthy, transparent, and explainable AI models, supporting responsible decision-making processes at the highest levels of national security.

Details

Tel Aviv, Israel
+972 52-622-2500
benhagag10@gmail.com

Links

LinkedIn

Skills

- Machine Learning
- Artificial Intelligence
- Natural Language Processing
- Statistics
- Leadership
- Public Speaking

Languages

- English
- Hebrew

Publications

NLLP 2024 (EMNLP) - "LegalLens Shared Task 2024: Legal Violation Identification in Unstructured Text"

EACL 2024 - "LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text"

EMNLP 2023 (findings) - "The Truth, The Whole Truth, and Nothing but the Truth: A New Benchmark Dataset for Hebrew Text Credibility Assessment"

NLLP 2022 (EMNLP) - "Class Action Prediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US"

ISCOL 2020- "Classifying Hebrew Political Statements for their Credibility Based on Linguistic Features"

Education

MSc Computer Science, Bar Ilan University, Tel Aviv

SEPTEMBER 2020 — JUNE 2022

- Graduating with distinction at the top 5% of the class.
- Research Fellow at the NLP Lab under the supervision of Prof. Dagan, Pro. Goldberg, and Prof. Tsarfati.
- EMNLP-published dissertation, presenting a benchmark Hebrew dataset for credibility classification, featuring annotated statements from Israeli public figures, and shows that combining contextual embeddings and evidence retrieval with transformer-based models significantly enhances the accuracy of credibility prediction.

BSc Computer Science, the Open University of Israel, Tel Aviv

SEPTEMBER 2015 — JUNE 2019

- Graduating with distinction at the top 5% of the class on the Dean's List.

Honors and Awards

Best AI-Based Solution for Legal | AI Breakthrough Awards

2024

National Security Award | The Chief of the Military Intelligence Directorate

2020

Dean's List | The Open University

2015 — 2019

Volunteering

Committee Member, Natural Legal Language Processing (NLLP) Group

JANUARY 2024

Chief NLP Scientific Adviser, Middat

JANUARY 2020 — PRESENT

Promoting public health by making reliable medical information accessible using innovative machine-based approaches to increase public trust and countering misinformation.

References

References available upon request

With over a decade of experience in technology leadership and research, my work has focused on leveraging NLP and AI to promote fairness, detect misinformation, and drive advancements in legal technology, contributing to a safer and more transparent world. As the Lead Data Scientist in Israeli intelligence, I developed scalable intelligence systems that integrated vast datasets for counterterrorism, directly enhancing national security. Additionally, my contributions to AI technologies in the international litigation industry have revolutionised case analysis and decision-making, demonstrating how data and AI can address complex societal challenges while remaining ethically sound, lawful, and scalable. Looking ahead at the research challenges and opportunities of the field, I am delighted to apply for the PhD Computer Science programme at Columbia University to take my research in NLP for social good to the next level.

As a researcher in Israeli intelligence, I laid a critical foundation in data science and NLP by designing high-stakes intelligence systems that leveraged vast datasets to support national security, setting the stage for my future work in socially impactful AI. I designed and implemented data pipelines and algorithms capable of handling massive datasets, focusing on developing NLP models to meet evolving intelligence needs. My role involved adapting and refining machine learning techniques for multilingual information extraction and real-time threat detection, essential for the unit's rapid response to emerging threats and complex intelligence demands. Through advanced NLP capabilities, I transformed raw data into actionable insights, making critical information both accessible and timely. This experience sharpened my technical skills in large-scale data processing and highlighted the value of NLP-driven solutions in supporting national security and addressing dynamic intelligence challenges.

After completing my service as a data science researcher and graduating with distinction in my undergraduate studies, I pursued a master's in computer science, focusing on NLP. For my thesis, I designed the first comprehensive automatic credibility assessment system for Hebrew texts, addressing a critical gap in Hebrew-language NLP resources. This project involved building a new dataset from scratch and developing models that could identify language patterns indicating truthfulness or falsehood. Partnering with journalists and fact-checkers, I took an interdisciplinary approach to incorporate external evidence into credibility assessments, demonstrating the essential role of context for accurate predictions.. The system I developed has since been implemented in their day-to-day workflows. The dataset, models, and benchmarks from this research have become core resources for further advancements in Hebrew NLP.

Building on my experience and deep NLP expertise, I plan to collaborate with Columbia's NLP Group as a PhD student. My research agenda centres on computational justice and fairness in AI and promoting Responsible NLP. I aim to design systems that go beyond technical performance to uphold ethical principles, and cultural sensitivity, and foster social cohesion. Working alongside Columbia's world-class faculty aligns with this objective. I am particularly excited to partner with Professor McKeown on research interests such as text summarisation and credibility assessment, where I can leverage my experience building datasets and models for low-resource languages. Additionally, under the guidance of Professor Smaranda Muresan, I aim to incorporate social media interactions as contextual signals to enhance real-time misinformation detection. These collaborations would be instrumental in advancing multilingual, context-aware NLP systems that address pressing misinformation and social justice issues in digital environments.

Beyond my academic research, I collaborated across sectors as a volunteer scientific advisor with an Israeli watchdog combating public health-related misinformation. In this role, I oversee the development of models to understand pandemic crises, assess their reliability, and provide actionable insights, such as how to refute misinformation or address health-related issues. This experience highlighted how AI can support public health by identifying harmful narratives in real time and fostering trust in scientific information and public policy. Continuing this commitment to countering misinformation in the public interest, I plan to partner with Professor Elhaded to further develop models that translate complex health data into actionable knowledge, bridging my prior work in credibility assessment with impactful real-world health applications.

In my role as Head of Research at Darrow, a legal tech company, I lead the development of disruptive machine learning models that analyse unstructured legal data, powering Darrow's AI-driven litigation platform to detect legal violations, predict outcomes, and identify critical trends. By advancing NLP research, I have driven the development of proprietary technology that interprets and extracts information from complex legal documents with high accuracy, transforming how law firms approach and manage litigation. My research has led to publications in industry conferences such as EMNLP and EACL, where I partnered with the Natural Legal Language Processing (NLLP) workshop as a committee fellow. My research includes organising a global NLP shared task with over 100 participants, attracting diverse perspectives and a wide range of domain experts and researchers, raising Darrow's profile in the field and setting high standards for open-source research.

Building on my experience in developing disruptive technologies and promoting fair datasets, I aim to focus my PhD research on ethical AI and NLP at Columbia by collaborating with Professor Goldblum. Through this partnership, I would work to increase transparency in neural networks and establish safeguards that ensure AI technologies are used responsibly, aligning with my dedication to Responsible NLP for social justice. His experience in developing frameworks for ethical AI directly complements my goal of creating resilient, equitable AI systems that effectively serve diverse communities. Together, we can advance the field toward AI solutions that prioritise fairness, trust, and inclusivity in critical real-world applications.

Drawing from over a decade of experience in data science, NLP, and technology leadership, I have dedicated my career to creating ethical AI solutions that promote justice, transparency, and security. I envision research initiatives that pursue social justice with technical innovation, enhancing the accessibility and inclusivity of NLP advancements for diverse global populations. Collaborating with Columbia's faculty will allow me to integrate cutting-edge methodologies into systems that sustain social progress and human development while upholding the highest ethical standards. Ultimately, upon driving impactful advancements that address the ever-evolving legal, economic, and social challenges of our time, I aim to continue developing responsible NLP technologies during the AI revolution.