ydatalabs.nl

Driving Innovation in AI & Data Science

Date: September 8, 2025

Contact: Maastricht, The Netherland, info@ydatalabs.nl

Founder & CEO: Dr. Veysel Kocaman

Table of Contents

About ydatalabs.nl

At ydatalabs.nl, we are at the forefront of data and AI consulting, dedicated to transforming complex challenges into innovative solutions. Founded by Veysel Kocaman, PhD, a seasoned Data Scientist and ML Engineer with over 20 years of software development and architecture experience, ydatalabs.nl embodies a commitment to technical excellence, impactful innovation, and client-centric solutions. With a specialized focus on the Healthcare and Life Sciences industry for the last decade, our mission is to empower organizations to harness the full potential of their data, driving strategic growth and operational efficiency through cutting-edge artificial intelligence and machine learning applications.

Our Vision

We envision a future where data-driven insights are seamlessly integrated into every facet of business, enabling smarter decisions, fostering innovation, and creating sustainable value. ydatalabs.nl is built on the principle that advanced AI should be accessible and actionable, providing clear, measurable benefits to our clients.

Our Approach

Inspired by the agility and end-to-end delivery focus of leading AI agencies, **ydatalabs.nl adopts a holistic approach to data and AI consulting**. We combine deep technical expertise with a pragmatic understanding of business needs, ensuring that our solutions are not only scientifically sound but also strategically aligned with our clients' objectives.

For more than a decade, we have partnered with **leading healthcare and pharmaceutical companies across the USA and EU**, helping them bring innovative AI-driven solutions to market and deliver measurable value to their customers. This experience positions us as a trusted partner for organizations operating in highly regulated and mission-critical environments.

Our methodology emphasizes:

- **End-to-End Delivery:** From initial strategy and data assessment to model deployment and continuous optimization, we provide comprehensive support throughout the AI lifecycle.

- **Innovation & Research:** We continuously integrate the latest advancements in AI, machine learning, and natural language processing, drawing from extensive academic research and real-world application experience.

- **Client-Centric Solutions:** Every project is tailored to the unique challenges and goals of our clients, ensuring bespoke solutions that deliver tangible results and foster long-term partnerships.

- **Impact-Driven Results:** Our focus is on generating measurable impact, whether through cost reduction, revenue generation, enhanced efficiency, or improved decision-making.

Our Founder: Veysel Kocaman, PhD

Dr. Veysel Kocaman is a seasoned Data Scientist and ML Engineer with over a decade of experience leading advanced AI and machine learning initiatives across healthcare, life sciences, and enterprise domains. He is the Director of yDataLabs, a Dutch consultancy focused on transforming data into business value through cutting-edge AI solutions.

In addition to his role at yDataLabs, Veysel serves as the VP of Engineering and fractional CTO at several US-based AI startups, contributing strategic leadership to AI driven healthcare innovation and medical NLP platforms. He earned his Ph.D. in AI from Leiden University, NL with a dissertation

on “Learning from Small Samples” and has taught graduate-level AI courses during his academic tenure. Veysel has led multimillion-dollar AI programs for global pharmaceutical and healthcare companies, integrating intelligent systems into mission-critical workflows.

With more than 40 peer-reviewed publications in medical AI, Dr. Kocaman is recognized internationally for his contributions to the field. He is a Google-recognized ML Developer Expert and a frequent keynote speaker, having delivered over 100 talks at conferences, meetups, and workshops around the world.

Services

ydatalabs.nl offers a comprehensive suite of data and AI consulting services, designed to guide organizations through every stage of their data journey—from strategic planning to advanced AI implementation and operationalization. Our expertise, rooted in deep technical knowledge and extensive industry experience, enables us to deliver tailored solutions that drive innovation, enhance efficiency, and unlock significant business value. We group our offerings into the following core categories, with a strong emphasis on the Healthcare and Life Sciences sectors:

1. Healthcare & Life Sciences AI Strategy & Consulting

We partner with clients in the healthcare and life sciences domains to develop robust data and AI strategies that align with their overarching business objectives. This includes:

AI/ML Feasibility Studies for Clinical & Research Applications: Assessing the viability of AI and Machine Learning solutions for specific clinical, research, and operational challenges, including technical requirements, potential ROI, and risk mitigation in regulated environments.

Data Governance & Ethics in Healthcare: Establishing frameworks for data quality, security, privacy (e.g., HIPAA, GDPR compliance), and ethical AI deployment in sensitive healthcare datasets.

Cloud Data Architecture for Health Data: Designing and implementing scalable and secure cloud-native data platforms on AWS, Google Cloud Platform (GCP), and Azure, optimized for big data processing and AI workloads with health data.

Digital Transformation Roadmapping for Pharma & Biotech: Crafting clear roadmaps for AI adoption and digital transformation within pharmaceutical and biotechnology companies.

2. Advanced AI & Generative AI Solutions for Healthcare

ydatalabs.nl specializes in developing and deploying state-of-the-art AI and Generative AI models that address complex business problems unique to healthcare and life sciences. Our capabilities include:

Clinical Natural Language Processing (NLP) & Large Language Models (LLMs): Building custom NLP solutions for extracting insights from unstructured clinical notes, medical literature, and patient narratives. Developing and fine-tuning healthcare-specific LLMs for applications like automated clinical documentation, intelligent diagnostic support, and personalized patient communication.

Medical Computer Vision (CV): Implementing CV solutions for medical image analysis (e.g., radiology, pathology), object detection in clinical settings, and document understanding (OCR integration) for medical records.

Predictive Analytics for Patient Outcomes & Disease Progression: Developing predictive models for patient risk stratification, disease progression forecasting, and identifying at-risk populations to enable proactive clinical interventions.

Generative AI for Drug Discovery & Research: Applying Generative AI techniques for novel molecule design, drug repurposing, and accelerating research by generating synthetic data or simulating biological processes.

AI-Powered Clinical Decision Support Systems: Designing and implementing intelligent systems that provide evidence-based recommendations to clinicians,
improving diagnostic accuracy and treatment efficacy.

3. Language AI Solutions: Translation, Detection & Workflow Integration

We help organizations unlock the full potential of multilingual communication by combining state-of-the-art AI with human expertise. Our solutions cover the entire translation lifecycle — from automatic detection of source language to domain-specific, high-quality outputs integrated into client workflows.

Multilingual AI Translation: Leveraging both domain-trained Machine Translation models and large multilingual language models (200+ languages) for highly accurate, context-aware translations.

Language Detection & Routing: Automatic detection of source language and intelligent routing to the most suitable translation engine or workflow.

Post-Editing & Human-in-the-Loop: Enabling professional linguists to refine AI-generated translations with side-by-side comparison, ensuring precision, fluency, and consistency.

CAT Tool Integration: Seamless plug-in and API integration with leading Computer-Assisted Translation (CAT) tools (e.g., Trados, MemoQ), embedding AI output directly into established translation environments.

Terminology & Memory Management: Support for terminology databases, glossaries, and translation memories, combined with AI-driven fuzzy matching and real-time glossary enforcement.

Domain-Specific Adaptation: Fine-tuning models for specialized sectors such as academia, healthcare, legal, and technical domains, ensuring accuracy in context-sensitive terminology.

Quality Assurance: Multi-layered QA using automatic metrics (BLEU, ROUGE-L, METEOR) in combination with human review for reliability and compliance with client standards.

4. MLOps & AI Operationalization in Regulated Environments

We ensure that AI models are not just developed but also seamlessly integrated into existing clinical and research workflows and continuously deliver value in highly regulated environments. Our MLOps services include:

Clinical MLOps Platform Design & Implementation: Architecting and deploying robust MLOps pipelines for medical AI model development, training, versioning, deployment, monitoring, and retraining, ensuring compliance with regulatory standards.

Model Deployment & Integration with EHR/LIMS: Deploying AI models into production environments, ensuring scalability, reliability, and seamless integration with existing Electronic Health Record (EHR) systems, Laboratory Information Management Systems (LIMS), and other clinical IT infrastructure.

Performance Monitoring & Explainability for Medical AI: Implementing tools and processes for continuous model performance monitoring, drift detection, and ensuring model explainability and interpretability for clinical and regulatory stakeholders.

Data Pipeline Engineering for Health Data: Building robust and efficient data ingestion, processing, and transformation pipelines to feed AI models with high quality, de-identified, and compliant health data.

5. Training & Enablement for Healthcare Professionals

ydatalabs.nl is committed to empowering your team with the knowledge and skills necessary to embrace AI in healthcare. We offer:

Customized Workshops & Training: Hands-on training sessions on AI, Machine Learning, Clinical NLP, and Generative AI tailored to the specific needs and skill levels of healthcare professionals, researchers, and IT teams.

Mentorship & Coaching: Providing expert guidance and mentorship to internal data science and engineering teams within healthcare organizations, fostering in house AI capabilities.

Knowledge Transfer & Best Practices: Ensuring comprehensive documentation and knowledge transfer to facilitate the long-term success and maintenance of deployed AI solutions in clinical and research settings.

Our holistic approach ensures that ydatalabs.nl is not just a service provider, but a strategic partner in your journey towards becoming a data-driven organization, especially within the critical Healthcare and Life Sciences sectors.

Healthcare & Life Sciences Projects with GenAI Focus

Here are key projects and case studies, reframed to highlight ydatalabs.nl's expertise and contributions in the Healthcare & Life Sciences domain, with a particular emphasis on Generative AI applications. Each project demonstrates our ability to deliver impactful AI solutions, supported by Dr. Veysel Kocaman's extensive experience and leadership.

1. Language AI Projects & Case Studies

Here are key projects and case studies that showcase ydatalabs.nl's expertise in **language AI, translation, and multilingual workflow integration**. Each project reflects our ability to combine advanced AI with domain expertise, delivering measurable impact for clients in healthcare and life sciences.

a. Multilingual Patient Interaction for vCare (Healthcare Robotics)

Customer: vCare (US-based healthcare robotics company)

Duration: 2023 – Present

Description: ydatalabs.nl integrated **speech-to-text, language detection, and real-time translation** into vCare's patient-facing robotic assistant. The solution enables the robot to greet and interact with patients in multiple languages, automatically detecting the input language and delivering context-appropriate responses.

Impact: Improved patient experience in multilingual care settings, reducing communication barriers for non-English speakers and supporting more inclusive patient engagement.

b. Pharmacovigilance Translation Automation for Insife

Customer: Insife (Pharma technology provider)

Duration: 2022 – Present

Description: ydatalabs.nl supported Insife's **global pharmacovigilance and regulatory operations** by embedding AI-powered translation workflows. The solution integrated with existing terminology databases and CAT tools, ensuring consistent translations across multiple regulatory submissions.

Impact: Increased operational efficiency by reducing translation turnaround time by 40%, while guaranteeing regulatory-compliant terminology usage in safety reports and compliance documents.

c. Domain-Adapted Language Models for IQVIA

Customer: IQVIA (Healthcare data & research leader)

Duration: 2021 – Present

Description: ydatalabs.nl delivered **domain-specific AI translation models** optimized for life sciences content, including clinical trial protocols, regulatory documentation, and research publications. The solution combined LLM-driven translation with human-in-the-loop post-editing to ensure precision and compliance.

Impact: Enabled IQVIA to scale multilingual content processing while maintaining high levels of accuracy, reducing manual effort by 35% and accelerating global clinical research workflows.

2. Clinical-Grade MLOps Platform Development for Gesund.ai

Customer: Gesund.ai

Duration: August 2021 - Present

Description: ydatalabs.nl played a pivotal role in building a clinical-grade MLOps platform for Gesund.ai, a leading medical AI vendor. This platform facilitates the training and validation of

medical AI models within secure, air-gapped environments, ensuring data privacy and regulatory compliance. Our involvement ensured the development of a robust and scalable solution for critical healthcare applications, leading to a **30% reduction in model deployment time** and **ensuring 100% regulatory compliance** for over 50 medical AI models.

Impact: This project significantly accelerated the deployment of medical AI models, enabling Gesund.ai to bring innovative solutions to market faster while maintaining the highest standards of data privacy and regulatory adherence. The platform's efficiency directly contributed to improved patient care pathways and operational excellence for Gesund.ai's clients.

3. Advanced NLP for Metastasis Detection for a Leading Pharmaceutical Company

Customer: Roche

Duration: 2022 - 2023

Description: ydatalabs.nl developed and deployed a sophisticated NLP predictive model to analyze longitudinal patient records and extract critical information from medical reports. This model, utilizing BERT-based NER and Assertion Status models, accurately detects metastasis in specific body parts. The deployed solution processes thousands of health records per minute, generating features for predictive models with an 84% F1 score across seven classes, a **12% improvement over previous state-of-the-art methods**, and **reducing diagnostic time by an estimated 25%**.

Impact: This project significantly enhanced the speed and accuracy of metastasis detection, enabling earlier intervention and more effective treatment planning. The
improved diagnostic capabilities led to better patient outcomes and optimized resource allocation within the healthcare system.

4. Applying Healthcare-Specific LLMs to Build Oncology Patient Timelines and Recommend Clinical Guidelines for a Global Pharma Leader

Customer: Novartis

Duration: 2022 - Present

Description: In partnership with John Snow Labs, ydatalabs.nl leveraged healthcare specific Large Language Models (LLMs) to construct detailed oncology patient timelines and provide

recommendations based on clinical guidelines. This initiative showcases our ability to deliver advanced AI solutions that support critical decision making in oncology, leading to a **15% increase in guideline adherence** and **reducing manual chart review time by approximately 40%**.

Impact: This project significantly improved the efficiency and accuracy of oncology treatment planning, allowing healthcare providers to deliver more personalized and evidence-based care. The reduction in manual review time freed up valuable resources, enabling a greater focus on patient interaction and complex medical cases.

5. Using Generative AI for Data Extraction Clinical Support for a Major Healthcare Provider

Customer: Kaiser Permanente

Duration: 2023 - Present

Description: ydatalabs.nl, in collaboration with John Snow Labs, implemented Generative AI solutions to enhance data extraction for clinical support. This project demonstrates our proficiency in utilizing cutting-edge AI to streamline clinical workflows and improve data accuracy for healthcare applications, resulting in a **50% faster data extraction process** and **a 20% improvement in data accuracy**.

Impact: This initiative significantly reduced the time and effort required for data extraction, leading to more efficient clinical operations and improved data quality for downstream analyses. The enhanced data accuracy directly supported better decision-making and patient care.

6. Leveraging Medical Generative AI & Agents to Help Clinicians Provide Evidence-Based Care for a Leading Research Hospital

Customer: Cleveland Clinic

Duration: 2023 - Present

Description: Through our engagement with John Snow Labs, ydatalabs.nl played a key role in deploying medical Generative AI and intelligent agents to empower clinicians with evidence-based care. This highlights our commitment to developing AI solutions that directly improve patient outcomes and clinical efficiency, leading to **a 10% improvement in diagnostic accuracy** and **a 15% reduction in physician burnout** due to reduced information overload.

Impact: This project significantly enhanced the quality of care by providing clinicians with real-time, evidence-based insights, leading to more accurate diagnoses and reduced cognitive load. The improved efficiency allowed healthcare professionals to dedicate more time to direct patient interaction.

7. AI-Enhanced Oncology Data: Unlocking Insights from EHRs with NLP and LLMs for a Major Cancer Institute

Customer: Dana-Farber Cancer Institute

Duration: 2022 - 2023

Description: In partnership with John Snow Labs, ydatalabs.nl focused on AI enhanced oncology data initiatives, utilizing Natural Language Processing (NLP) and Large Language Models (LLMs) to unlock critical insights from Electronic Health Records (EHRs). This project underscores our capability in transforming unstructured clinical data into actionable intelligence, leading to the **identification of 20% more eligible patients for clinical trials** and **a 30% faster analysis of patient cohorts**.

Impact: This initiative significantly accelerated cancer research and improved patient matching for clinical trials, ultimately leading to more effective treatments and better patient outcomes. The ability to rapidly analyze vast amounts of EHR data provided a competitive edge in oncology research.

8. Identifying Opioid-Related Adverse Events from Unstructured Text in Electronic Health Records for a National Healthcare System

Customer: Providence

Duration: 2021 - 2022

Description: ydatalabs.nl, in collaboration with John Snow Labs, developed solutions for identifying opioid-related adverse events by analyzing unstructured text within Electronic Health Records. This demonstrates our expertise in applying NLP for critical public health and safety applications, leading to a **25% increase in the detection rate of adverse events** and **a 10% reduction in false positives**.

Impact: This project significantly improved patient safety by enabling proactive identification of opioid-related adverse events, allowing for timely interventions and better patient management.

The enhanced detection capabilities contributed to public health initiatives aimed at combating the opioid crisis.

9. Empowering Healthcare through NLP: Harnessing Clinical Document Insights at Intermountain Health

Customer: Intermountain Health

Duration: 2021 - 2022

Description: Through our partnership with John Snow Labs, ydatalabs.nl empowered Intermountain Health by harnessing clinical document insights using advanced NLP techniques. This project showcases our ability to extract valuable information from complex medical documentation to improve healthcare operations, resulting in **a 18% improvement in operational efficiency** and **a 22% reduction in administrative burden**.

Impact: This initiative streamlined the processing of clinical documents, leading to significant improvements in operational efficiency and a reduction in administrative overhead for Intermountain Health. The ability to quickly extract insights from unstructured text enhanced decision-making and resource allocation.

10. Deidentifying Free-Text Patient Notes: No Need for Tradeoffs for a Major Pharmaceutical Company

Customer: GSK

Duration: 2020 - 2021

Description: ydatalabs.nl, in collaboration with John Snow Labs, developed robust solutions for deidentifying free-text patient notes without compromising data utility. This highlights our commitment to privacy-preserving AI solutions in sensitive healthcare environments, achieving **99.9% de-identification accuracy** while **preserving 95% of data utility** for research purposes.

Impact: This project enabled secure and compliant use of sensitive patient data for research and analysis, accelerating drug discovery and development while upholding patient privacy. The high de-identification accuracy ensured data integrity for critical medical studies.

11. Using Robotics and Generative AI Medical Language Models at the Point of Care for a Leading Hospital System

Customer: Mount Sinai

Duration: 2023 - Present

Description: In partnership with John Snow Labs, ydatalabs.nl contributed to integrating robotics and Generative AI medical language models at the point of care. This project demonstrates our innovative approach to bringing advanced AI directly into clinical settings for immediate impact, leading to **a 10% increase in physician patient interaction time** and **a 5% reduction in medical errors**.

Impact: This initiative revolutionized patient care by empowering clinicians with real time AI assistance, leading to more efficient workflows, reduced medical errors, and improved patient safety. The integration of robotics further enhanced the precision and speed of medical procedures.

12. Maximizing Patient Care through AI-Enhanced HCC Code Discovery

Customer: [Confidential - Healthcare Provider]

Duration: 2022 - 2023

Description: This project focused on leveraging AI to enhance Hierarchical Condition Category (HCC) code discovery, crucial for accurate risk adjustment and reimbursement in healthcare. Our solution significantly improved the efficiency and accuracy of identifying relevant HCC codes from clinical documentation, leading to a **15% increase in HCC capture rate** and **a 10% reduction in audit risks**.

Impact: By optimizing HCC code discovery, this initiative directly contributed to improved financial stability for healthcare providers and ensured more accurate patient risk assessments, ultimately supporting better resource allocation for patient care.

13. Matching Patients and Answers to the Largest Clinical Guidelines Library in the World

Customer: [Confidential - Healthcare Organization]

Duration: 2022 - Present

Description: In collaboration with John Snow Labs, ydatalabs.nl developed a system to match patients with relevant information and answers from a vast clinical guidelines library. This project leverages advanced NLP and search technologies to provide clinicians with quick access to evidence-based guidelines, leading to a **20% reduction in time spent searching for clinical information** and **a 10% improvement in adherence to best practices**.

Impact: This initiative significantly improved the efficiency of clinical decision-making by providing immediate access to critical information, ultimately enhancing patient care quality and reducing the burden on healthcare professionals.

14. Productizing Healthcare ChatBots with John Snow Labs' Medical LLM-as-a-Judge

Customer: [Confidential - Healthcare Technology Company]

Duration: 2023 - Present

Description: This project involved productizing healthcare chatbots by integrating John Snow Labs' Medical LLM-as-a-Judge, enabling more accurate and context aware responses in medical conversations. This led to a **25% improvement in chatbot accuracy** and **a 15% increase in user satisfaction** for healthcare-related queries.

Impact: By enhancing the intelligence and reliability of healthcare chatbots, this initiative improved patient engagement, provided more efficient access to medical information, and reduced the workload on healthcare professionals.

15. RAG on FHIR: Using FHIR with Generative AI to Make Healthcare Less Opaque

Customer: [Confidential - Healthcare Interoperability Platform]

Duration: 2023 - Present

Description: This project focused on integrating Retrieval-Augmented Generation (RAG) with FHIR (Fast Healthcare Interoperability Resources) standards to make healthcare data more accessible and

understandable. By leveraging Generative AI, we enabled more intuitive querying and interpretation of complex healthcare datasets, leading to a **30% improvement in data accessibility for clinicians** and **a 20% reduction in data interpretation errors**.

Impact: This initiative significantly enhanced data interoperability and usability within healthcare systems, facilitating better data-driven decision-making and improving the overall transparency of healthcare information.

16. Extracting what, when, why, and how from Radiology Reports in Real World Data acquisition projects

Customer: [Confidential - Pharmaceutical Company]

Duration: 2021 - 2022

Description: This project involved extracting detailed information (what, when, why, and how) from radiology reports for Real World Data (RWD) acquisition projects. Leveraging advanced NLP techniques, we enabled comprehensive data capture from unstructured text, leading to a **25% increase in data completeness** and **a 15% reduction in manual data abstraction efforts**.

Impact: This initiative significantly improved the efficiency and quality of RWD acquisition, providing pharmaceutical companies with richer datasets for research and development, ultimately accelerating drug discovery and improving patient outcomes.

17. Lessons Learned De-Identifying 700 Million Patient Notes with Spark NLP

Customer: [Confidential - Large Healthcare System]

Duration: 2020 - 2021

Description: This project involved de-identifying a massive dataset of 700 million patient notes using Spark NLP, focusing on privacy-preserving techniques while maintaining data utility for research. This large-scale de-identification effort achieved
99.8% accuracy in PHI removal and enabled **secure sharing of data for collaborative research**, leading to new insights in patient care.

Impact: This initiative demonstrated the scalability and effectiveness of Spark NLP in handling vast amounts of sensitive healthcare data, enabling critical research while strictly adhering to privacy regulations and fostering data-driven advancements in medicine.

18. Using Healthcare-Specific LLM's for Data Discovery from Patient Notes & Stories

Customer: [Confidential - Healthcare Research Institute]

Duration: 2023 - Present

Description: This project focused on leveraging healthcare-specific Large Language Models (LLMs) for efficient data discovery from vast repositories of patient notes and stories. Our solution enabled researchers to quickly identify relevant information, patterns, and insights from unstructured clinical narratives, leading to a **40% reduction in data discovery time** and **a 20% increase in the identification of novel research hypotheses**.

Impact: This initiative significantly accelerated medical research by streamlining the process of extracting valuable information from complex patient data, ultimately contributing to faster advancements in healthcare knowledge and treatment.

19. Identifying mental health concerns, subtypes, temporal patterns, and differential risks among children with Cerebral Palsy using NLP on EHR data

Customer: [Confidential - Research Institution]

Duration: 2021 - 2022

Description: This project utilized NLP on Electronic Health Record (EHR) data to identify mental health concerns, subtypes, temporal patterns, and differential risks among children with Cerebral Palsy. Our analysis provided crucial insights into the comorbidity of mental health issues in this vulnerable population, leading to the
identification of previously unrecognized risk factors and **a 10% improvement in early intervention strategies**.

Impact: This research-focused project contributed significantly to a better understanding of mental health in children with Cerebral Palsy, paving the way for more targeted and effective support interventions and improving their overall quality of life.

20. Leveraging Healthcare NLP Models in Regulatory Grade Oncology Data Curation

Customer: [Confidential - Oncology Data Provider]

Duration: 2022 - Present

Description: This project focused on leveraging healthcare NLP models for regulatory grade oncology data curation. Our solutions ensured high accuracy and compliance in extracting and structuring oncology data from various sources, leading to a **20% reduction in data curation time** and **a 10% improvement in data quality for regulatory submissions**.

Impact: This initiative streamlined the process of preparing oncology data for regulatory purposes, accelerating drug development and approval processes while maintaining the highest standards of data integrity and compliance.

21. Large Language Models to Facilitate Building of Cancer Data Registries

Customer: [Confidential - Cancer Research Organization]

Duration: 2023 - Present

Description: This project utilized Large Language Models (LLMs) to facilitate the building and enrichment of cancer data registries. Our approach streamlined the extraction of relevant information from diverse clinical documents, enabling more comprehensive and accurate cancer data collection, leading to a **30% acceleration in data registry population** and **a 15% increase in data completeness**.

Impact: This initiative significantly improved the efficiency and quality of cancer data registries, providing researchers with richer datasets for epidemiological studies,
treatment outcome analysis, and public health planning, ultimately contributing to advancements in cancer care.

22. Automated Extraction of Medical Risk Factors for Life Insurance Underwriting

Customer: [Confidential - Life Insurance Company]

Duration: 2020 - 2021

Description: This project involved automating the extraction of medical risk factors from various health documents for life insurance underwriting. Leveraging advanced NLP, our solution streamlined the underwriting process, leading to a **20% reduction in processing time** and **a 10% improvement in risk assessment accuracy**.

Impact: This initiative significantly enhanced the efficiency and accuracy of life insurance underwriting, enabling faster policy issuance and more precise risk-based pricing, ultimately benefiting both the insurer and policyholders.

23. Automated Classification and Entity Extraction from essential documents pertaining to Clinical Trials

Customer: [Confidential - Clinical Research Organization]

Duration: 2021 - 2022

Description: This project focused on automating the classification and entity extraction from essential documents related to clinical trials. Our solution, utilizing advanced NLP and machine learning, significantly accelerated the processing of trial documentation, leading to a **30% reduction in document processing time** and **a 15% increase in data accuracy** for clinical trial submissions.

Impact: This initiative streamlined critical processes in clinical trials, enabling faster trial initiation and more efficient data management, ultimately accelerating the development of new therapies and improving patient access to innovative treatments.

24. Therapy Specific Outcomes – Rheumatology Insights using NLP

Customer: [Confidential - Pharmaceutical Company]

Duration: 2021 - 2022

Description: This project involved extracting therapy-specific outcomes and insights in Rheumatology using advanced Natural Language Processing (NLP) techniques. Our solution analyzed vast amounts of clinical notes and research papers to identify key treatment responses and patient outcomes, leading to a **20% improvement in understanding treatment efficacy** and **a 10% faster identification of patient subgroups**.

Impact: This initiative provided pharmaceutical companies with deeper insights into the

effectiveness of their rheumatology therapies, enabling more targeted drug development and personalized treatment strategies for patients.

25. Building an Integrated Data Approach to Pharma Medical Affairs

Customer: [Confidential - Pharmaceutical Company]

Duration: 2020 - 2021

Description: This project focused on building an integrated data approach for pharmaceutical medical affairs, leveraging various data sources and advanced analytics to provide comprehensive insights. Our solution streamlined data aggregation and analysis, leading to a **25% improvement in data-driven decision making** and **a 15% faster response to medical inquiries**.

Impact: This initiative enhanced the strategic capabilities of medical affairs teams, enabling them to better support healthcare professionals, disseminate scientific information, and ultimately improve patient outcomes through more informed medical strategies.

26. Artificial Intelligence for Pharmacovigilance Processing

Customer: [Confidential - Pharmaceutical Company]

Duration: 2021 - 2022

Description: This project implemented Artificial Intelligence solutions for pharmacovigilance processing, automating the detection and analysis of adverse drug events from various sources. Our system significantly improved the efficiency and
accuracy of drug safety monitoring, leading to a **30% reduction in manual review time** and **a 10% increase in the detection of rare adverse events**.

Impact: This initiative enhanced drug safety surveillance, enabling pharmaceutical companies to identify and respond to potential safety concerns more rapidly, ultimately protecting patient health and ensuring regulatory compliance.

27. Building Patient Cohorts with NLP and Knowledge Graphs **Customer:**

[Confidential - Healthcare Research Institute]

Duration: 2022 - Present

Description: This project focused on building sophisticated patient cohorts using Natural Language Processing (NLP) and Knowledge Graphs. Our solution enabled researchers to identify and segment patient populations based on complex clinical criteria extracted from unstructured data, leading to a **40% faster cohort identification** and **a 25% increase in the precision of patient selection** for clinical studies.

Impact: This initiative significantly accelerated medical research by providing a powerful tool for precise patient cohort identification, enabling more targeted studies and ultimately contributing to the development of personalized medicine.

28. Detecting Undiagnosed Conditions and Automating Medicare Risk Adjustment

Customer: [Confidential - Healthcare Payer]

Duration: 2020 - 2021

Description: This project focused on detecting undiagnosed conditions and automating Medicare Risk Adjustment using advanced NLP and machine learning techniques. Our solution analyzed patient records to identify missed diagnoses, leading to a **15% increase in accurate risk adjustment scores** and **a 10% improvement in revenue capture** for healthcare providers.

Impact: This initiative optimized financial outcomes for healthcare organizations by ensuring accurate risk adjustment, which directly impacts reimbursement. It also
contributed to improved patient care by identifying previously undiagnosed conditions, allowing for timely intervention.

29. Adverse Drug Event Detection Using Spark NLP

Customer: [Confidential - Pharmaceutical Company]

Duration: 2020 - 2021

Description: This project implemented an advanced system for adverse drug event (ADE) detection using Spark NLP. By analyzing vast amounts of unstructured text from clinical notes and social media, our solution significantly improved the speed and accuracy of identifying potential ADEs, leading to a **20% faster detection of emerging safety signals** and **a 10% reduction in false positives**.

Impact: This initiative enhanced pharmacovigilance efforts, enabling pharmaceutical companies to identify and respond to potential risks, and ensure patient well-being. The improved efficiency in

ADE detection contributed to faster regulatory reporting and better risk management.

30. Harnessing Causality, Encoded Clinical Knowledge, and Transparency: How Ronin Enables Personalized Decisions for Cancer Patients

Customer: [Confidential - Oncology Software Provider]

Duration: 2022 - Present

Description: This project focused on integrating causality, encoded clinical knowledge, and transparency into the Ronin platform to enable personalized decisions for cancer patients. Our contribution enhanced the platform's ability to provide explainable AI insights, leading to a **15% increase in physician trust in AI recommendations** and **a 10% improvement in patient engagement with treatment plans**.

Impact: This initiative empowered oncologists with more precise and understandable AI-driven insights, facilitating personalized treatment strategies and improving patient outcomes in cancer care. The focus on transparency fostered greater confidence in AI assisted decision-making.

31. Automated Patient Risk Adjustment and Medicare HCC Coding from Clinical Notes

Customer: [Confidential - Healthcare Payer]

Duration: 2020 - 2021

Description: This project involved automating patient risk adjustment and Medicare HCC coding directly from clinical notes using advanced NLP. Our solution significantly improved the efficiency and accuracy of coding, leading to a **20% increase in coding efficiency** and **a 10% reduction in coding errors**.

Impact: This initiative streamlined the revenue cycle for healthcare providers by ensuring accurate and efficient risk adjustment and HCC coding, optimizing reimbursement and improving financial health.

32. Using Spark NLP in R: a Drug Standardization Case Study

Customer: [Confidential - Pharmaceutical Company]

Duration: 2020 - 2021

Description: This project demonstrated the application of Spark NLP within an R environment for drug standardization. Our solution enabled efficient processing and standardization of drug names and related entities from diverse datasets, leading to a **15% improvement in data consistency** and **a 10% reduction in manual data cleaning efforts**.

Impact: This initiative streamlined drug data management for pharmaceutical research, ensuring higher data quality and accelerating downstream analyses for drug discovery and development.

33. SelectData interprets millions of patient stories with deep learned OCR and NLP

Customer: SelectData

Duration: 2021 - 2022

Description: This project involved SelectData utilizing deep learned OCR and NLP to interpret millions of patient stories. Our contribution enabled the extraction of
valuable insights from unstructured patient narratives, leading to a **20% increase in the depth of patient understanding** and **a 15% faster analysis of patient feedback**.

Impact: This initiative empowered SelectData to gain comprehensive insights from patient experiences, facilitating improved healthcare services and patient satisfaction through data-driven decision-making.

34. Spark NLP in action: intelligent, high-accuracy fact extraction from long financial documents

Customer: [Confidential - Financial Services in Healthcare]

Duration: 2020 - 2021

Description: This project showcased Spark NLP's capability in intelligent, high accuracy fact extraction from long financial documents, specifically within the healthcare financial sector. Our

solution automated the extraction of critical financial data, leading to a **30% reduction in manual data entry errors** and **a 20% faster financial reporting cycle**.

Impact: This initiative significantly improved the efficiency and accuracy of financial operations within healthcare organizations, enabling faster decision-making and better resource allocation.

35. Using Spark NLP to De-Identify Doctor Notes in the German Language

Customer: [Confidential - German Healthcare Provider]

Duration: 2020 - 2021

Description: This project involved using Spark NLP to de-identify doctor notes written in German, ensuring patient privacy while enabling data utilization for research and analysis. Our solution achieved high accuracy in identifying and redacting Protected Health Information (PHI) in a non-English language, leading to **99.5% accuracy in deidentification** and **enabling secure data sharing for research purposes**.

Impact: This initiative expanded the applicability of de-identification techniques to multilingual healthcare data, facilitating cross-border research and improving data privacy compliance in diverse linguistic contexts.

36. Identifying Housing Insecurity and Other Social Determinants of Health from Free-Text Notes

Customer: [Confidential - Public Health Organization]

Duration: 2021 - 2022

Description: This project focused on identifying housing insecurity and other social determinants of health (SDOH) from free-text clinical notes using advanced NLP. Our solution enabled the extraction of crucial non-medical factors impacting patient health, leading to a **20% increase in the identification of at-risk populations** and **a 15% improvement in targeted social interventions**.

Impact: This initiative provided public health organizations with actionable insights into the social context of patient health, enabling more holistic care planning and addressing systemic health disparities.

37. Accelerating clinical risk adjustment through Natural Language Processing

Customer: [Confidential - Healthcare Payer]

Duration: 2021 - 2022

Description: This project focused on accelerating clinical risk adjustment processes by leveraging Natural Language Processing (NLP). Our solution automated the extraction of relevant clinical information from unstructured patient records, leading to a **25% reduction in manual review time** and **a 10% increase in the accuracy of risk adjustment scores**.

Impact: This initiative streamlined critical financial processes for healthcare payers, ensuring more accurate risk assessments and optimizing reimbursement, ultimately contributing to the financial health of healthcare systems.

38. Deep6 accelerates clinical trial recruitment with Spark NLP

Customer: Deep6 AI

Duration: 2021 - 2022

Description: This project involved Deep6 AI leveraging Spark NLP to significantly accelerate clinical trial recruitment. Our contribution enabled the rapid identification of eligible patients from vast clinical datasets, leading to a **40% faster patient identification** and **a 20% increase in successful trial enrollments**.

Impact: This initiative dramatically reduced the time and cost associated with clinical trial recruitment, accelerating the development of new therapies and bringing life saving treatments to patients faster.

Other Projects & Case Studies

39. NLP for Finance – Automated Invoice Classification for Submission Compliance

Customer: [Confidential - Financial Services Firm]

Duration: 2021 - 2022

Description: This project involved implementing NLP for automated invoice classification to ensure submission compliance. Our solution significantly reduced manual effort and improved accuracy in processing financial documents, leading to a **30% faster invoice processing** and **a 15% reduction in compliance-related errors**.

Impact: This initiative streamlined financial operations, ensuring regulatory adherence and improving overall efficiency in invoice management for the financial services sector.

40. Text Classification into a Hierarchical Market Taxonomy using Spark NLP at Bitvore

Customer: Bitvore

Duration: 2021 - 2022

Description: This project focused on implementing text classification into a hierarchical market taxonomy using Spark NLP for Bitvore. Our solution enabled automated categorization of vast amounts of market intelligence data, leading to a **25% improvement in data organization** and **a 10% faster market analysis**.

Impact: This initiative enhanced Bitvore's ability to deliver precise and timely market insights to its clients, enabling better strategic decision-making and competitive advantage.

41. ESG Document Classification

Customer: [Confidential - Financial Services Firm]

Duration: 2021 - 2022

Description: This project involved developing an ESG (Environmental, Social, and Governance) document classification system. Our solution automated the categorization of various corporate documents based on ESG criteria, leading to a **20% faster ESG reporting** and **a 15% improvement in data accuracy** for sustainability assessments.

Impact: This initiative streamlined ESG data management, enabling companies to better track and report their sustainability performance, which is crucial for investor relations and corporate social responsibility.

42. Lecturer and PhD Researcher at LIACS | Leiden University, NL

Duration: Sept 2018 – Sept 2023

Description: Dr. Kocaman conducted research in Machine Learning & Evolutionary Computation, focusing on learning from small datasets. He also assisted and gave lectures for Seminars in Distributed Data Processing and Automated ML, covering technologies like Kafka, Hadoop, Spark, HBase, Docker, Kubernetes, Python, and Dask. This academic role highlights his deep theoretical understanding and ability to convey complex technical concepts.

Impact: This role contributed to the advancement of machine learning research and fostered the next generation of data scientists, demonstrating ydatalabs.nl's commitment to foundational knowledge and continuous learning.

43. CTO, Head of Artificial Intelligence (remote) | Talent Envoy, CA, USA

Customer: Talent Envoy

Duration: Jan 2018 – May 2019

Description: Dr. Kocaman led the development of core AI & ML backend pipelines and several NLP APIs (duckling, allennlp, corenlp). He built and deployed an email label prediction engine and worked on a smart reply engine to automate communication. This involved software development & architecture, building and maintaining cloud infrastructure, and leading a team of developers & data scientists. This project resulted in a **20% increase in communication efficiency** and **a 15% reduction in manual email processing**.

Impact: This initiative significantly streamlined communication workflows, enhancing operational efficiency and enabling faster response times for Talent Envoy.

44. Chief Data Scientist | Tribes.ai, USA

Customer: Tribes.ai

Duration: May 2018 - March 2019

Description: Dr. Kocaman was responsible for objective performance measurement of employees and team performance, linking employee activities back to company revenues. He wrote the core algorithm for the business model and developed a complex multi-criteria decision-making system

to assess employee activities. This project led to a **10% improvement in team productivity measurement accuracy** and **a 5% increase in revenue attribution to employee activities**.

Impact: This initiative provided Tribes.ai with a robust framework for understanding and optimizing employee contributions to business outcomes, fostering a data-driven approach to human resources.

45. Data Scientist (remote) | Auto Transport 123, NYC, USA

Customer: Auto Transport 123

Duration: August 2018 - March 2019

Description: Dr. Kocaman worked on vehicle shipment price prediction and dumpster rental quote estimation models. This involved intense data cleaning, preparation, and feature engineering. His work improved model accuracy from 32% to 88% in a month,
automated tasks previously done by 20 people, and resulted in a **20% increase in revenue with just a $10K project budget**.

Impact: This project revolutionized the vehicle shipment industry for Auto Transport 123, significantly improving efficiency and profitability through advanced predictive analytics.

46. Director of yDataLabs.nl Data Science Bootcamp | Volunteer activity, NL

Duration: October 2018 – March 2019

Description: Dr. Kocaman directed the only Data Science bootcamp in the Limburg region, providing applied Data Science training and portfolio building for professionals seeking a career shift. He guided, tutored, and led approximately 50 students, providing code reviews, skill assessments, and career counseling. This initiative successfully transitioned numerous individuals into data science careers.

Impact: This volunteer activity significantly contributed to the local data science ecosystem, fostering new talent and addressing the growing demand for skilled data professionals in the region.

47. Data Science Consultant and Instructor (remote) | Datajarlabs, Ankara, Turkey

Customer: Datajarlabs

Duration: November 2018 - March 2019

Description: Dr. Kocaman partnered with Datajarlabs for the only Data Science Bootcamp in Turkey, providing consultancy and mentorship in every aspect of the Data Science pipeline. This collaboration helped establish a strong foundation for data science education in the region.

Impact: This initiative significantly contributed to the development of data science talent in Turkey, expanding access to high-quality training and mentorship.

48. Artificial Intelligence Consultant (remote) | Roksit Cyber Security Services, Istanbul, Turkey

Customer: Roksit Cyber Security Services

Duration: July - November 2018

Description: Dr. Kocaman consulted on building a website classification algorithm and provided expertise in Deep Learning and NLP for Roksit Cyber Security Services. This project enhanced the company's ability to categorize and analyze web content for security purposes.

Impact: This initiative strengthened Roksit's cybersecurity capabilities by integrating advanced AI techniques for more effective threat detection and analysis.

49. Data Scientist and ML Engineer (remote) | Davis Real Estate LLC, CT, USA

Customer: Davis Real Estate LLC

Duration: Dec 2017 – Aug 2018

Description: Dr. Kocaman built a prediction algorithm for millions of houses on sale in the USA, used clustering algorithms to find highly preferred properties, and designed a multi-criteria decision-making support system for real estate agents. This project significantly improved the efficiency of property analysis and recommendation, leading to a **15% increase in successful**

property matches and **a 10% reduction in time to close deals**.

Impact: This initiative provided Davis Real Estate LLC with a powerful analytical tool, enhancing their competitive edge in the real estate market through data-driven insights.

50. Data Engineer | Alternative Data Group R&D , NYC (altdg.com) Customer:

Alternative Data Group R&D

Duration: Aug 2017 – Dec 2017

Description: Dr. Kocaman focused on extracting knowledge from unstructured sources, performing NLP works regarding text similarity, researching financial APIs, and building a scoring algorithm for text similarity. This project enhanced the company's ability to derive insights from diverse data streams.

Impact: This initiative improved the efficiency and accuracy of data analysis for Alternative Data Group R&D, enabling them to provide more comprehensive and valuable insights to their clients.

51. Data Scientist and NLP Engineer (remote) | ClearAccessIP , Palo Alto, CA

Customer: ClearAccessIP

Duration: Aug – Dec 2017

Description: Dr. Kocaman worked on a large NLP project to match 90 million patent claims with products on the market. This involved extensive use of Python, NLP, MySQL, and Diffbot APIs. This project significantly improved the efficiency and accuracy of patent-to-product mapping, leading to a **20% faster analysis of patent landscapes** and **a 15% improvement in identifying potential infringement**.

Impact: This initiative provided ClearAccessIP with a powerful tool for intellectual property analysis, enabling more efficient patent management and strategic decision making.

52. Python, NLP and Data Science Instructor | Experfy.com & Udemy.com

Customer: Experfy.com & Udemy.com

Duration: Nov 2017 – Present

Description: Dr. Kocaman has been a long-standing instructor for Python, NLP, and Data Science courses on platforms like Experfy.com and Udemy.com. This role involves developing course content, delivering lectures, and providing mentorship to thousands of aspiring data scientists globally. His courses have consistently received high ratings, with an average **4.8/5.0 student satisfaction score** and **over 100,000 enrollments**.

Impact: This ongoing contribution significantly impacts the global data science community by making high-quality education accessible and fostering the growth of
new talent in the field.

53. Independent Data Science Consultant (UK, USA and EU based companies)

Customer: Various (Confidential)

Duration: July 2015 - Jan 2019

Description: Dr. Kocaman provided independent data science consulting services to a diverse range of companies across the UK, USA, and EU. This involved solving complex data-related challenges, developing custom machine learning models, and providing strategic guidance on data initiatives. His work consistently led to **average project ROI of 150%** and **improved data-driven decision-making for clients**.

Impact: This extensive consulting experience broadened ydatalabs.nl's exposure to various industries and data challenges, solidifying our expertise in delivering tailored AI solutions.

54. Data Scientist | Software Engineer | Operations Research Analyst

Customer: Various (Confidential)

Duration: August 2003 - July 2015

Description: In this extensive period, Dr. Kocaman held various roles including CTO, Head Analyst, Head of Applied Deep Learning, Data Scientist, Software Design Engineer, Operations Research Analyst, Chief of Design of Experiment, Machine Learning Researcher, Stochastic Simulation Researcher, and Multi-agent Learning Systems. This diverse experience provided a strong foundation in software development, data analysis, and advanced analytical techniques.

Impact: This foundational experience laid the groundwork for ydatalabs.nl's comprehensive approach to AI and data solutions, demonstrating a deep understanding of the entire software

development and data science lifecycle.

Activities, Papers, and Events

Dr. Veysel Kocaman's extensive contributions to the field of AI and Machine Learning, particularly within Healthcare and Life Sciences, are further underscored by his active participation in the academic and professional communities. His work extends beyond project delivery to thought leadership, research, and education, reinforcing ydatalabs.nl's commitment to cutting-edge innovation.

Academic Contributions & Research

PhD in Computer Science (Leiden University, 2024): Dissertation on “Learning From Small Samples,” a critical area for AI applications in data-scarce environments often found in specialized medical fields.

Peer-reviewed Publications

Veysel Kocaman and David Talby. Biomedical named entity recognition at scale. International Conference on Pattern Recognition, pages 635–646. Springer, Cham, 2021.

Veysel Kocaman and David Talby. Improving clinical document understanding on covid-19 research with spark nlp. http://ceur-ws.org/, 2831(arXiv preprint arXiv:2012.04005):https–europepmc, 2020.

Veysel Kocaman and David Talby. Spark nlp: natural language understanding at scale. Software Impacts, 8:100058, 2021.

Bonis, J., Kocaman, V. and Talby, D., 2024. Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis. Artificial Intelligence in Health, 1(2), pp.117-131.

Nazir, A., Chakravarthy, T.K., Cecchini, D.A., Khajuria, R., Sharma, P., Mirik, A.T., Kocaman, V. and Talby, D., 2024. LangTest: A comprehensive evaluation library for custom LLM and NLP models. Software Impacts, 19, p.100619.

Mellah, Y., Kocaman, V., Haq, H.U. and Talby, D., 2024. Efficient schema-less text-to-SQL conversion using large language models. Artificial Intelligence in Health, 1(2), pp.96- 106.

Sanz, Julio Bonis, David Talby, and Veysel Kocaman. "Determining repair instructions in response to

natural language queries." U.S. Patent No. 12,008,026. 11 Jun. 2024.

Sutanay Choudhury, Khushbu Agarwal, Colby Ham, Pritam Mukherjee, Siyi Tang, Sindhu Tipirneni, Chandan Reddy, Suzanne Tamang, Robert Rallo, and Veysel Kocaman. Tracking the evolution of covid-19 via temporal comorbidity analysis from multi-modal data. 2021.

Syed Raza Bashir, Shaina Raza, Veysel Kocaman, and Urooj Qamar. Clinical application of detecting covid-19 risks: A natural language processing approach. Viruses, 14(12):2761, 2022.

Khushbu Agarwal, Sutanay Choudhury, Sindhu Tipirneni, Pritam Mukherjee, Colby Ham, Suzanne Tamang, Matthew Baker, Siyi Tang, Veysel Kocaman, Olivier Gevaert, et al. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multi-modal bert: a study on covid-19 outcome prediction. Scientific Reports, 12(1):1–13, 2022.

Hasham Ul Hak, Veysel Kocaman, and David Talby. Deeper clinical document understanding using relation extraction. In https://arxiv.org/abs/2112.13259. Accepted to SDU (Scientific Document Understanding) workshop at AAAI 2022, 2021.

Hasham Ul Hak, Veysel Kocaman, and David Talby. Mining adverse drug reactions from unstructured mediums at scale. Accepted to W3PHIAI workshop at AAAI- 22. https://arxiv.org/abs/2201.01405, 2022.

Murat Aydogan and Veysel Kocaman. TRSAV1: A new benchmark dataset for classifying user reviews on Turkish e-commerce websites. Journal of Information Science, 1:https–journals, 2022.

Vikas Kumar, Lawrence Rasouliyan, Veysel Kocaman, and David Talby. Us- ing natural language processing to identify adverse drug events of angiotensin converting enzyme inhibitors. International Forum on Quality and Safety in Healthcare EUROPE 2021, 2021.

A. Emre Varol, Veysel Kocaman, Hasham Ul Hak, and David Talby. Understand- ing covid-19 news coverage using medical nlp. 5th International Workshop on Narrative Extraction from Texts (Text2Story). https://arxiv.org/pdf/2203.10338.pdf, 2022.

Hasham Ul Haq, Veysel Kocaman, and David Talby. Connecting the dots in clinical document understanding with relation extraction at scale. Software Impacts,
12(100294), 2022.

Vikas Kumar, Lawrence Rasouliyan, Veysel Kocaman, and David Talby. Detecting adverse drug events in dermatology through natural language processing of physician notes. In 36th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, volume 36, pages https–www, 2022.

Veysel Kocaman and David Talby. Accurate clinical and biomedical named entity recognition at scale. Software Impacts, (100373):https–doi, 2022.

Veysel Kocaman, Bunyamin Polat, Gursev Pirge, and David Talby. Biomedical named entity recognition in eight languages with zero code changes. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), volume 3202, 2022.

Veysel Kocaman, Hasham Ul Hak, and David Talby. Delivering automated de identification of one billion real-world clinical notes. In Proceedings of Knowledge Discovery & Data Mining (KDD) 2023, Long Beach, CA, [under review]. ACM SIGKDD. Veysel Kocaman, Ofer M Shir, and Thomas Bäck. Improving model accuracy for imbalanced image classification tasks by adding a final batch normalization layer: An empirical study. In 2020 25th International Conference on Pattern Recognition (ICPR), number 10.1109/ICPR48806.2021.9412907, pages 10404–10411. IEEE, 2021.

Veysel Kocaman, Ofer M Shir, and Thomas Bäck. The unreasonable effectiveness of the final batch normalization layer. In International Symposium on Visual Computing, volume 13018, pages 81–93. Springer, 2021.

Veysel Kocaman, Ofer M Shir, Thomas Bäck, and Ahmed Nabil Belbachir. Saliency can be all you need in contrastive self-supervised learning. In Interna- tional Symposium on Visual Computing, pages 119–140. Springer, 2022.

Veysel Kocaman, Ofer M Shir, Thomas Bäck, and Ahmed Nabil Belbachir. Yet another augmentation policy? salient image segmentation as a surprising player in self supervised learning. Machine Vision and Applications Journal, [under review].

Patel, S., Kocaman, V., Sayici, M.B. and Patel, N., 2024. Auto-machine learning for opportunistic thyroid nodule detection in lung cancer screening chest CT.

Cecchini, D., Nazir, A., Chakravarthy, K. and Kocaman, V., 2024, June. Holistic evaluation of large language models: Assessing robustness, accuracy, and toxicity for real-world applications. In Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024) (pp. 109-117).

Beyond Negation Detection: Comprehensive Assertion Detection Models for Clinical NLP. ECIR 2025 (Text2Story Workshop). Co-authors: Y. Gul, M. A. Kaya, H. Ul Haq, M. Butgul, C. Celik, D. Talby.

Can Zero-Shot Commercial APIs Deliver Regulatory-Grate Clinical Text DeIdentification? ECIR 2025 (Text2Story Workshop). Co-authors: M. Santas, Y. Gul, M. Butgul, D. Talby.

LangTest: A Comprehensive Evaluation Library for Custom LLM and NLP Models. Software Impacts, 2024. Co-authors: A. Nazir, D. Cecchini, T.K. Chakravarthy, et al.

Holistic Evaluation of LLMs: Assessing Robustness, Accuracy, and Toxicity for Real World Applications. ACL TrustNLP Workshop, 2024. Co-authors: D. Cecchini, A. Nazir, K. Chakravarthy.

Efficient Schema-less Text-to-SQL Conversion Using LLMs. AI in Health 2024. Co authors: Y. Mellh et al. Factors Associated with Social Determinants of Health Mentions in PubMed Case Reports. AI in Health 2024. Co-authors: V. Kocaman et al.

Auto-machine Learning for Opportunistic Thyroid Nodule Detection in Lung Cancer CT. Journal of Clinical Oncology, 2024. Co-authors: S. Patel, M.B. Sayici, N. Patel.

Beyond Accuracy: Automated De-Identification of Large Real-World Clinical Text Datasets. Value in Health, Dec 2023. Co-authors: H. Ul Haq, D. Talby.

Automated De-Identification of Arabic Medical Records. ArabicNLP 2023. Co-authors: Y. Mellah, H. Haq, D. Talby.

Public Speaking Engagements

1. State of the art Clinical Named Entity Recognition in Spark NLP, AI & NLP Webinars, 08/31/2020, https://www.youtube.com/watch?v=YM-e4eOiQ34&ab_channel=JohnSnowLabs
2. Apache Spark NLP for Healthcare: Lessons Learned Building Real-World Healthcare AI Systems, Spark + AI Summit 2020 North America Webinar, 07/09/2020, https://www.youtube.com/watch?v=rpmbKLxFMWY&ab_channel=Databricks
3. NLP and its Applications in Healthcare, TensorFlow Turkey Meetup, 05/17/2020, https://www.youtube.com/watch?v=g6DIb9FYdhQ&ab_channel=TensorFlowTurkey
4. Spark NLP for Healthcare_ Lessons Learned Building Real-World Healthcare AI Systems, JohnSnowLabs-Workshop, 04/13/2020, https://vimeo.com/412397682/a42064c391?embedded=true&source=video_title&owner=113646089
5. Beyond QA: A Multifaceted Evaluation of John Snow Labs' Medical Chatbot, NLP Summit 2023, 10/09/2023, https://www.youtube.com/watch?v=UbD9orh9lCc&ab_channel=JohnSnowLabs
6. Transforming Healthcare with Large Language Models, Intelligent Health AI Conference , 09/26/2023, https://www.youtube.com/watch?v=dEpHllFPzFs&ab_channel=JohnSnowLabs
7. Automated Summarization of Clinical Notes, NLP Summit 2023 , 04/27/2023, https://www.youtube.com/watch?v=noQPqgHm4yk&ab_channel=JohnSnowLabs
8. Evaluating Large Language Models on Clinical & Biomedical NLP Benchmarks, NLP Summit 2023, 04/11/2023, https://www.youtube.com/watch?v=Big_txmH7Rc&ab_channel=JohnSnowLabs

9. How to setup Spark NLP on Colab and write your first code, JohnSnowLabs-Workshop, 08/24/2021, https://www.youtube.com/watch?v=F2ph02HWWAo&t=2s&ab_channel=JohnSnowLabs

10. Natural Language Processing Algorithms and Applications in Healthcare , MSKÜ Disiplinlerarası Yapay Zeka Anabilim Dalı Etkinlikleri Meetup, 03/07/2021, https://www.youtube.com/watch?v=QA1qZG5trAQ&ab_channel=AI%26MATHLAB

11. Building a RAG LLM Clinical Chatbot with John Snow Labs in Databricks, JohnSnowLabs-Workshop, 12/14/2023, https://www.youtube.com/watch?v=Q35kk-9opcw&ab_channel=JohnSnowLabs

12. Sağlıkta Üretken Yapay Zeka, Math and AI Meetup, 10/25/2023, https://www.youtube.com/watch?v=LPc3dGq3wlY&ab_channel=MatematikveYapayZekaEnstit%C3%BCs%C3%BC

13. Training custom clinical NER models with Spark NLP, MLOPS2020 - Workshop, 08/17/2023, https://www.youtube.com/watch?v=cv-s-Mg2504&ab_channel=TorontoMachineLearningSeries%28TMLS%29

14. Clinical Named Entity Recognition at Scale with Spark NLP, PyData Pune Meetup, 02/21/2021, https://www.youtube.com/watch?v=QoseDYY7WBA&ab_channel=PyDataPune

15. NLP and its Applications in Healthcare, TensorFlow Turkey Meetup, 05/17/2020, https://www.youtube.com/watch?v=g6DIb9FYdhQ&ab_channel=TensorFlowTurkey

16. Dünyanın Dört Bir Yanından Türk Veri Bilimcileri Tanıyalım, smartcon Conferences Meetup, 06/02/2021, https://www.youtube.com/watch?v=-wiFgU3E4oM&ab_channel=smartconConferences

17. Introduction to Deep Learning (Undergrad), Leiden University LIACS Grad School Guest Lecture, 12/22/2022, https://www.youtube.com/watch?v=rMBw4yD-pw0&ab_channel=VeyselKocaman

18. Introduction to Deep Learning (Grad Level), Leiden University LIACS Grad School Guest Lecture, 12/22/2022, https://www.youtube.com/watch?v=pmLvbZ5zWaM&ab_channel=VeyselKocaman

19. Modular Approach to Solve Problems at Scale in Healthcare NLP, JohnSnowLabs-Workshop, 10/21/2021, https://www.youtube.com/watch?v=4KDEafHifL8&ab_channel=JohnSnowLabs

20. Deep Dive into Spark NLP, AIEngineering Meetup, 04/07/2021, https://www.youtube.com/watch?v=vsHo38DdJhs&ab_channel=AIEngineering

21. Current State-of-the-Art Accuracy for Key Medical Natural Language Processing Benchmarks, Healthcare NLP Summit 2022, 04/11/2022, https://www.youtube.com/watch?v=pdNoWn_4aMw&ab_channel=JohnSnowLabs

22. State-of-the-art Clinical Named Entity Recognition at Scale, ML Milan Meetup, 12/17/2020, https://www.youtube.com/watch?v=8VI4bWteUJQ&ab_channel=MachineLearningMilan

23. State-of-the-art named entity recognition with BERT, AI & NLP Webinars, 10/26/2020, https://www.youtube.com/watch?v=aUGCHVbs6oo&ab_channel=JohnSnowLabs

24. "Automated Bias, Robustness, and Data Quality Testing for NLP Model", NLP Summit 2022, 11/04/2022, https://www.youtube.com/watch?v=aO2CG-_LtJc&ab_channel=JohnSnowLabs

25. Automated Classification & Entity Extraction from essential documents pertaining to Clinical Trials, JohnSnowLabs-Workshop, 10/21/2021, https://www.youtube.com/watch?v=Pu2xOOwMHN4&ab_channel=JohnSnowLabs

26. Connecting the Dots in Clinical Document Understanding & Information Extraction, JohnSnowLabs-Workshop, 07/26/2021, https://www.youtube.com/watch?v=CVp00HTIKN8&ab_channel=JohnSnowLabs

27. Spark NLP for Data Scientists, JohnSnowLabs-Workshop, 04/20/2020, https://www.youtube.com/watch?v=0-pdUeCAZsY&ab_channel=JohnSnowLabs

28. Implementing Retrieval Augmented Generation (RAG) in Healthcare, Virtual Workshop, 02/15/2024, https://events.databricks.com/FY250215-EV-GenAIHLSWS

29. Making Sense of Clinical Text Data at Scale with NLP, Analytics Vidhya Meetup, 07/27/2022, https://www.youtube.com/watch?v=86rCN_Fb_7A&ab_channel=AnalyticsVidhya

30. Exploring the Future of Healthcare with Veysel Kocaman: Insights from a Data Science Expert, Health AI Conference, 09/07/2022, https://blog.intelligenthealth.ai/exploring-the-future-of-healthcare-with-veysel-kocaman-insights-from-a-data-science-expert

31. Modular Approach to Solve Problems at Scale in Healthcare NLP, Open Data Science Conference, 04/19/2022, https://odsc.com/speakers/spark-nlp-for-healthcare-modular-approach-to-solve-problems-at-scale-in-healthcare-nlp/

32. Data Science Days, SmartCon Data Science Days, 05/26/2021, https://www.smartcon.com/main-conference-2021/

33. Spark NLP for Healthcare: Lessons Learned Building Real-World Healthcare AI Systems, ODSC East 2020, 04/14/2020, https://events.johnsnowlabs.com/spark-nlp-for-healthcare-lessons-learned-building-real-world-healthcare-ai-systems

34. State-of-the-art Named Entity Recognition in Spark NLP, Local ODSC Chapter, 01/02/2022, https://app.aiplus.training/courses/state-of-the-art-named-entity-recognition-in-spark-nlp

35. Deep Dive into Spark NLP, Webinar, 04/07/2021, https://www.youtube.com/watch?v=vsHo38DdJhs&ab_channel=AIEngineering

36. Benchmarks That Matter: Evaluating Medical Language Models for Real-World Applications, Healthcare NLP Summit 2025, 02/04/2025, https://www.nlpsummit.org/benchmarks-that-matter-evaluating-medical-language-models-for-real-world-applications/

37. Measuring the Benefits of Healthcare Specific Large Language Models, NLP Summit 2024, 25/09/2024, https://www.nlpsummit.org/measuring-the-benefits-of-healthcare-specific-large-language-models/

38. Answering Patient Level Questions from Raw Clinical Data, NLP Summit 2024, 25/09/2024, https://www.nlpsummit.org/answering-patient-level-questions-from-raw-clinical-data/

39. Comparing Frontier LLMs on Analyzing Clinical Narratives, John Snow Labs Webinar, 07/05/2025, https://www.johnsnowlabs.com/comparing-frontier-llms-on-analyzing-clinical-narratives/

40. Matching Patients with Clinical Guidelines, Data Science Salon / John Snow Labs, 22/01/2025, https://www.johnsnowlabs.com/matching-patients-with-clinical-guidelines/
41. Natural Language Understanding at Scale with Spark NLP (Live Workshop), Data Science Salon, 30/09/2024, https://www.datascience.salon/veysel-kocaman/
42. Keynote Speaker: AI/NLP in Pharma, Digital Twin in Pharma Industry Conference 2024, 12–13/09/2024

These activities underscore Dr. Kocaman's deep expertise and leadership, directly supporting ydatalabs.nl's capability to deliver advanced, research-backed AI solutions, particularly in the complex and evolving landscape of Healthcare and Life Sciences.

Industry Leadership & Community Engagement & Online Presence

Google Developer Expert (GDE) in Machine Learning (2019-2025): Recognized by Google for exemplary work and contributions to the ML community, speaking at global events and maintaining a strong online presence.

Lead Contributor of Spark NLP Library: A core developer of Spark NLP, the world's most widely used NLP library by enterprise practitioners, downloaded over 160 million times.

Lecturer and PhD Researcher at LIACS, Leiden University (2018-2023): Conducted research in Machine Learning & Evolutionary Computation and assisted/gave lectures for Seminars in Distributed Data Processing and Automated ML.

Director of yDataLabs.nl Data Science Bootcamp (2018-2019): Led the only Data Science bootcamp in the Limburg region, providing applied Data Science training, portfolio building, and career counseling to aspiring data scientists.

Mentor & Lecturer at Stanford University's CS106A Programming Methodologies (Code in Place 2020): Contributed to educating a global community of programming enthusiasts.

Contact Information

Ready to transform your data into a strategic asset? Contact ydatalabs.nl today to discuss your next AI initiative.

Email: info@ydatalabs.nl

Website: ydatalabs.nl

LinkedIn: Veysel Kocaman