Yiğit Bekir Kaya, Ph.D.

Istanbul, Turkey • +905374349948 • yigit@tankaventures.nl • linkedin.com/in/yigitbekir • https://www.tankaventures.nl/

I recently completed my Ph.D. in Natural Language Processing (NLP) with a focus on developing deep transfer learning models. As an experienced software developer, I have a strong background in artificial intelligence, big data, and data science, as well as a passion for video game design and development.

Throughout my career, I've held various leadership roles, from setting up and managing an innovation lab at Istanbul Technical University to founding several tech startups. Currently, I'm a Technical Research Advisor at Tanka, working on innovative solutions to global challenges such as climate change and sustainability.

In addition to my work at Tanka, I led a virtual reality game startup, managing a team of 20 artists and developers. My expertise in artificial intelligence and deep machine learning helps me tackle complex problems with advanced computing and analytics.

At Tanka, we're dedicated to developing cutting-edge materials and process technologies that address some of the world's most pressing issues. Our focus areas include bioplastics, green steel, clean mining, and biotech and life sciences. By collaborating with strategic partners, we aim to scale and commercialize these technologies, creating new operating companies for a better future.

WORK EXPERIENCE

Tanka Ventures BV • Amsterdam, North Holland, Netherlands • 07/2020 - Present
Technical Research Advisor • Full-time

- Assessed the scientific foundations of startup technology firms for companies in various industries, including FinTech, Green Steel, Green Mining, Bioplastics, BioTech, Life Sciences, Sustainable Construction, and Biopolymers. Together with Can, evaluated the viability of various enterprises, their ESG impact, and their contribution to the circular economy to determine which sectors to focus on. Attended numerous meetings with the founders and investors of these companies to gain a deeper understanding of their inner workings. Additionally, helped develop state-of-the-art digital concepts for advanced manufacturing and achieving circularity objectives.

CBILABVR Studios • Istanbul, Turkey • 08/2016 - 11/2018
Founder

- In my role as the founder of a cutting-edge Virtual Reality and Augmented Reality game studio and as a former game director, my professional responsibilities revolved around creativity and leadership. I personally created the game's design and development, meticulously designing almost every single scenario to depict my vision perfectly. Enlisting the help of twenty incredibly talented professionals (software engineers and technical artists), we worked together in an environment that fostered collaboration and innovation. Consequently, we showcased our technical and artistic talents in an incredible VR game trailer (all developed in Unreal Engine).

- My team and I developed, under my management, an engaging thirty-minute demo. The YouTube streamers who tested the game raved about it. This short game demo was a testament to our creativity and our ability to take on the fast-moving world of video game industry. I led the team through the complexities of building a game, writing a narrative, and coordinating the team members. I encouraged the team to be flexible and innovative. From this experience, I learned how to approach challenges in a more measured, strategic manner. This experience strengthened my skillset and prepared me for challenges that lie ahead not just in the gaming world but also in other creative fields.

CBILAB • Istanbul, Turkey • 07/2015 - 07/2020
Principal Advisor

Co-Founder

- Spotting a gap in rapid prototyping amid emerging global trends, I co-founded Cbilab in 2015. As Co-Founder, I drove fast prototyping and highlighted standout products, especially from Kickstarter. Thus, Cbilab became a key hub for future-centric product exploration and benchmark-setting evaluations.
- Beyond my primary duties at Cbilab, I saw an innovation gap and led initiatives, delivering speeches, mentoring students, and orchestrating 33 events within a year. Collaborating with my community, we generated over 200 ideas, successfully transforming two into thriving businesses.
- To bolster community member growth and promote continuous learning, I collaborated with Prof. Dr. Metin Orhan Kaya, introducing Python and TRIZ technique certifications, reinforcing our dedication to innovation and a brighter future.

Istanbul Technical University • Istanbul, Turkey • 03/2013 - 06/2015
Data Science Researcher

- I developed a delay prediction framework using data science techniques to enhance the ATM system's resilience. This framework bolstered Resilience2050's insights into ATM's resilience metrics.
- Using historical FAA data to predict air traffic disturbances, I developed a tool merging data mining and machine learning to analyze four decades of data. This tool informed EUROCONTROL's delay predictions and earned me the Boeing Graduate grant 2014.

EDUCATION

Doctor of Philosophy (Ph.D.) Candidate in Large Language Models
Istanbul Technical University • Istanbul, Turkey • 06/2024

Master of Science (M.Sc.) in Aerospace, Aeronautical and Astronautical Engineering
Istanbul Technical University • Istanbul, Turkey • 01/2015

Bachelor of Science (B.Sc., Double Major) in Aeronautical Engineering
Istanbul Technical University • Istanbul, Turkey • 01/2014

Bachelor of Science (B.Sc., Double Major) in Computer Engineering
Istanbul Technical University • Istanbul, Turkey • 01/2013

Gymnasium (Abitur) in Physics, Mathematics
Istanbul Gymnasium (German) • Istanbul, Turkey • 01/2009

CERTIFICATIONS

Sequences, Time Series and Prediction (Coursera) • 11/2021
DeepLearning.AI

DeepLearning.AI TensorFlow Developer Specialization (Coursera) • 11/2021
DeepLearning.AI

Natural Language Processing in TensorFlow (Coursera) • 02/2021
DeepLearning.AI

Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization (Coursera) • 10/2020
DeepLearning.AI

Convolutional Neural Networks in TensorFlow (Coursera) • 10/2020
DeepLearning.AI

Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning (Coursera) • 09/2020
DeepLearning.AI

Neural Networks and Deep Learning (Coursera) • 09/2020
DeepLearning.AI

Business Strategy (Coursera) • 10/2017
University of Illinois Urbana-Champaign

Managing the Organization (Coursera) • 09/2016
University of Illinois Urbana-Champaign

Designing the Organization (Coursera) • 06/2016
University of Illinois Urbana-Champaign

Applications of Everyday Leadership (Coursera) • 05/2016
University of Illinois Urbana-Champaign

Foundations of Everyday Leadership (Coursera) • 04/2016
University of Illinois Urbana-Champaign

Reproducible Research (Coursera) • 08/2015
The Johns Hopkins University

Exploratory Data Analysis (Coursera) • 08/2015
The Johns Hopkins University

Statistical Inference (Coursera) • 08/2015
The Johns Hopkins University

The Data Scientist's Toolbox (Coursera) • 07/2015
The Johns Hopkins University

R Programming (Coursera) • 07/2015
The Johns Hopkins University

Getting and Cleaning Data (Coursera) • 06/2015
The Johns Hopkins University

Pattern Discovery in Data Mining (Coursera) • 04/2015
University of Illinois Urbana-Champaign

Cloud Computing Concepts (Coursera) • 03/2015
University of Illinois Urbana-Champaign

AWARDS & SCHOLARSHIPS

Course Mastery Hall of Fame (Pattern Mining) • 04/2015
University of Illinois at Urbana-Champaign

Course Mastery Hall of Fame (Cloud Computing) • 04/2015
University of Illinois at Urbana-Champaign

Boeing Graduate Grant • 06/2014
Boeing

PROJECTS

BERT Tokenizer (Rust) • 02/2023 - 02/2023
Istanbul Technical University
The crate provides the port of the original BERT tokenizer from the Google BERT repository.

ManMade: A cinematic adventure VR game • 08/2016 - 11/2018
CBILABVR Studios
In a dynamic world of VR and AR, where captivating experiences are continually sought after, I stepped in as the lead game designer, creative director, and game director for Cbilab VR Studios. I crafted the debut adventure VR game "ManMade," guiding the studio's vision and direction. My endeavors have been central to carving out the company's niche and cementing its position in the immersive tech space.

RESILIENCE2050.EU • 07/2013 - 07/2015
Istanbul Technical University
To enhance the resilience of the ATM system against disruptions, I designed and instituted a delay prediction framework using advanced data science techniques. This endeavor enabled Resilience2050 to incorporate this knowledge into the ATM, offering a refreshed understanding of its resilience indicators.

Monitoring Safety of Air Traffic Using Big Data Mining • 03/2013 - 06/2013
Istanbul Technical University
Tasked with predicting air traffic disruptions using extensive FAA historical records, I developed a robust analytical tool, integrating data mining with machine learning to analyze data covering almost forty years. This creation successfully projected upcoming disturbances, earning recognition and subsequent implementation into EUROCONTROL's mechanism for predicting delays. This accomplishment led to my being awarded the Boeing Graduate grant in 2014.

PUBLICATIONS

BERT2D: Two Dimensional Positional Embeddings for Efficient Turkish NLP • 05/2024

IEEE

BERT2D: Two Dimensional Positional Embeddings for Efficient Turkish NLP represents a significant advancement in NLP for morphologically rich languages like Turkish. Traditional transformer models like BERT have historically struggled with these languages due to their reliance on 1D positional embeddings, which are less effective for languages with flexible word order and complex morphological structures.

BERT2D introduces a solution by combining absolute whole word positional embeddings with relative subword positional embeddings. This two-dimensional approach improves the model's understanding of word relationships and meaning representation. Additionally, BERT2D incorporates whole word masking, whereby entire words are masked during training to preserve contextual integrity and enhance learning.
The model was extensively pretrained, fine-tuned, and evaluated using the BERTurk corpus. This robust setup ensured that BERT2D performed well across diverse linguistic contexts. The results were significant: The results demonstrated that BERT2D consistently outperformed traditional BERT models across a range of key NLP tasks, including sentiment analysis, NER, POS tagging, and question answering.
In sentiment analysis, BERT2D achieved higher accuracy and F1 scores. In NER and POS tagging, it effectively disambiguated entities and syntactic roles. For question answering, BERT2D showed marked improvements in exact match and F1 scores, maintaining the integrity of word relationships within sentences.

BERT2D's contributions include a novel two-dimensional positional embedding and the demonstration of WWM's effectiveness. This model sets a new benchmark for NLP performance in Turkish and highlights the potential for innovative architectures tailored to specific linguistic characteristics. Future work will explore the application of BERT2D to other languages and its integration into newer transformer models, enhancing NLP capabilities across diverse languages and tasks

Effect of tokenization granularity for Turkish large language models • 02/2024

Elsevier

Transformer-based language models such as BERT (and its optimized versions) have outperformed previous models achieving state-of-the-art results on many English benchmark tasks. These multi-layered self-attention-based architectures are capable of producing contextual word vector representations. However, the tokens created in the tokenization preprocessing step are not necessarily words, particularly for languages with complex morphology, such as Turkish. The granularity of the generated tokens is a feature determined by various factors related to tokenization, especially by the vocabulary size. The impact of the vocabulary size parameter on model performance has not been extensively studied. In practice, the vocabulary size is often chosen arbitrarily or through trial and error. The tokenization granularity feature is particularly important for languages with complex morphology, like Turkish. It requires careful tuning, unlike English, where the granularity feature does not significantly impact model performance. This study presents a new collection of BERT models (ITUTurkBERT) trained using various tokenization methods on Turkish language data from the BERTurk and 1BW corpora. We fine-tune these models for three downstream tasks in Turkish and achieve state-of-the-art performance on all tasks. Our empirical experiments show that increasing the vocabulary size consistently improves performance on these tasks, except for sentiment analysis.

Finding the Optimal Vocabulary Size for Turkish Named Entity Recognition •
06/2022

CEUR Workshop Proceedings

A Dynamic Bayesian Belief Network Approach for Modelling the ATM Network Delays •
05/2014

International Conference on Research in Air Transportation

SKILLS

Leadership: Innovative Problem Solving, Decisive Judgment, Mentorship, Strategic Visioning, Adaptability & Agility, Conflict Resolution, Clear Communication, Collaborative Teamwork, Stakeholder Engagement, Negotiation & Partnership Building, Project Management, Scrum

Deep Learning: Statistics, Artificial Intelligence, Statistical Data Analysis, Machine Learning, Natural Language Processing (NLP), Pre-trained Language Models, Large Language Models, Transformers, Transfer Learning, BERT, GPT, Pytorch, TensorFlow, Generative AI, TPU, Big Data, Google Cloud Platform (GCP), R (Programming Language)

Programming: Python (Programming Language), C# (Programming Language), Rust (Programming Language), Java (Programming Language), C++ (Programming Language), MySQL, C (Programming Language), MongoDB

Language: Turkish (Native), German (Professional), English (Professional)