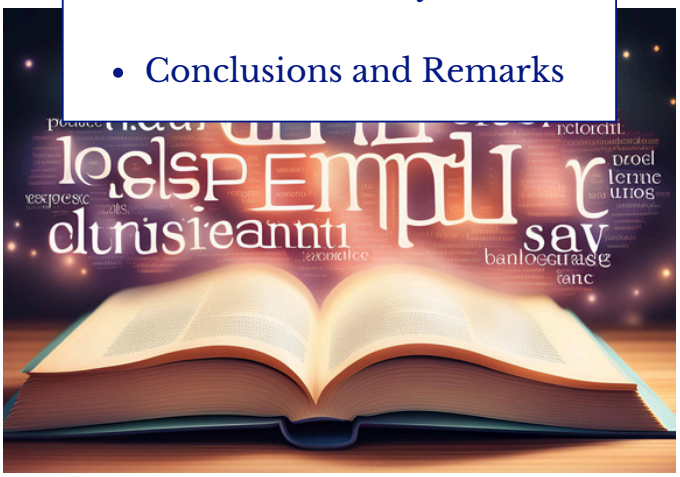
Behind the Code

software campus

WHAT'S IN THIS MONTH'S ISSUE:



Written by Dr. Natalia Teixeira Silva

Expected reading time: 20 minutes

The Software Campus is an accelerator program for future leaders. It provides young researchers who aspire to take leadership and management positions with novel experiences as project managers.

After being thoroughly selected, participants pursuing their PhDs in our Research Partner institutions (RP) across Germany are paired with an Industry Partner (IP). The companies support their mentees with an experienced leader and assist them in leading an IT-related project over two years. These young talents, financially supported by the Bundesministerium für Bildung und Forschung (BMBF), undergo a set of competence trainings while developing their projects and leading their teams. The Software Campus learning-by-doing approach enables them to gain the skills required for their future in a variety of roles, such as a researcher in an academic environment, a tech manager in industry, or an entrepreneur launching a start-up.

In autumn 2023 the Software Campus launched a new strategy to promote collaborative work among our participants: new Communities of Practice (CoPs) organized by topic in Computer Sciences.

With our CoPs we aim to create a cooperative environment where theoretical insights are fused with practical industry perspectives, reinforcing our main ambition: to expose our participants to the intersection between research and industry demands and support the tech industry with talented and highly trained young minds. Within the CoPs, our participants can engage with their peers as well as with our industry and research partners in dedicated activities.

Our eight Communities are:

- **Natural Language Processing**
- **Computer Vision**
- **Cybersecurity and Privacy**
- **Hardware and Systems Engineering**
- **Data Science and Analysis**
- **Agile and Software Development**
- **IoT and Distributed Systems**
- **Interdisciplinary Machine Learning**

In parallel, and in an attempt to raise awareness of the Software Campus outcomes among our stakeholders, we will periodically publish the Behind the Code series. These articles will address a specific CoP, their underlying Software Campus projects, and the participants' professional development. We will provide you with an insider's view of the methodologies applied, the challenges from behind the scenes, and the application power these initiatives have in shaping the future of the tech scenario. The Behind the Code series will not only capture the technical intricacies but also the human experiences, their passion, and their shared commitment to science and innovation.

NLP INSIGHTS:
UNREVEALING INNOVATION IN TEXTUAL INTELLIGENCE

Our first CoP, launched as a pilot initiative in October 2023 is the Natural Language Processing (NLP) Community. Today, we will navigate the Software Campus NLP projects, where novel ideas with real-world applications are transforming knowledge into solutions.

In the era of GPTs and chatbots all around, NLP has gained substantial attention over the past few years. This research area is a Computer Sciences subfield with intricate grounds in linguistics. This CoP is formed by members covering topics such as text summarization, argument mining, and domain adaptation. They exploited the best of Neural Networks, Machine Learning (ML) architectures, Retrieval-Augmented Generation (RAG), and of course, the new and trendy Large Language Models (LLMs). Among our currently ten NLP members, we have participants partnering their projects with multiple companies, such as Holtzbrinck Publishing Group, DATEV eG, Huawei, and Software AG.

When it comes to text summarization and argument mining, we can say that these two very common practices among NLP specialists, somehow, walk together. Text summarization, on the one hand, focuses on extractive approaches combined with content shortening to automatically condense multiple source files, such as text documents, video recordings, and voice-based content, like podcasts, for instance. These techniques utilize powerful engines based on Deep Learning and Machine Learning models. On the other hand, argument mining focuses on identifying and analyzing arguments, viewpoints, and different speakers within a text. Researchers and developers use Deep Learning architectures and rule-based systems to leverage linguistic patterns and classify relevant information.

The correlation and interconnection between the two tasks lie in their shared

goal of extracting meaningful and trustworthy insights. In this first publication, we will navigate four exciting projects, currently under development, illustrating such techniques and their potential applications.

DEEP DIVE:

THE SOFTWARE CAMPUS NLP PROJECTS AND THEIR JOURNEYS IN LEADERSHIP AND MANAGEMENT

Fine-tuning with elegance: enhancing precision in domain-specific extensive texts

The first project we will herein cover is managed and executed by **Anum Afzal**. Anum is a PhD researcher at the chair for Software Engineering of Business Information Systems (SEBIS) at **TU München**, under Professor Dr. Florian Matthes' supervision. Her project at the Software Campus, entitled "Abstractive Text Summarization of Domain-specific Documents" – AteSD – is developed in partnership with the Holtzbrinck Publishing Group. At Holtzbrinck's strategic growth division, Digital Science, Anum is provided technical advice with periodic meetings with her mentor Dr. Tim Steuer. Interestingly, Dr. Steuer is a Software Campus alumnus, and according to Anum, by understanding the program's goals he can deliver meaningful insights and useful suggestions beyond the project's technicalities.

AteSD focuses on simultaneously refining LLMs for domain-specific text summarization across multiple subject areas.

In language modeling, domain adaptation presents a challenging yet vital frontier for NLP research. Anum's project targets the difficulties of domain shift – when the models struggle to comprehend data falling outside their training scope, especially when dealing with domain-specific documents. By focusing on text summarization, she aims to bridge the gap between academic research and current industry needs, where large volumes of complex texts pose a challenge to the analysis and decision-making in segments such as science, medicine, and politics. These fields are particularly challenging as they require high accuracy in both data processing and summary generation to avoid inaccurate conclusions. Her work evaluates the models' ability to shift and adapt to a new domain without the necessity of recurrent re-training.

Anum initially employed language models like BART and Pegasus-X and gained substantial input from her results. As we all might be wondering at this point, the tech landscape's rapid evolution presented her unforeseen challenges, both as a young researcher and as a new manager. When the new LLMs started being brought to the public at the end of 2022, Anum had already one of her Ph.D. papers finalized and ready to be submitted. Although her first impression was that her thesis results became obsolete overnight, to keep her project innovative and aligned with the NLP advances, she had to incorporate multiple new experiments and reshape her project's focus to include the new LLMs. She is now evaluating the LLMs' capacity to adapt to multiple domains

and shift in the context of text summarization, a subject yet urging to be further explored. She uses her insights from BART and Pegasus-X, together with GPT-4 as baselines, and then assesses multiple open-source state-of-the-art models like Llama 2, Falcon, Vicuna, and Mistral. In case this happens, which are the best approaches to adapt them to a new domain? Anum's team is comparing different extents of few-shot learning and fine-tuning to leverage the model's performance.

Anum demonstrated a great sense of resilience and flexibility during her rerouting phase. Her leadership journey had transitioned from previous solo projects to her first time leading a team. The Software Campus trainings "Change Management" and "Basics for Future Managers", offered by TRUMPF and the Holtzbrinck, respectively, came in very handy. With the newly acquired knowledge, she evaluated the possibilities and their outcomes with a positive mindset and adjusted her strategy. Meanwhile, the learnings from another training – "Better understanding yourself and others", organized by Merck – improved Anum's abilities to handle difficult situations with her team. With the training she could build a new perspective on the importance of collaborative work, effective communication, and conflict resolution, to better manage projects amid uncertainty.

The project's outcomes suggest that domain adaptation can be efficiently achieved with models of varying complexities, both if the model is smaller or larger, and her promising results are now under review for publication. By the time we published this piece, and to the best of our knowledge, Anum's project was the first one to evaluate text summarization in the context of domain adaptation for multiple domains at once by using LLMs. For Anum, one of her greatest achievements is to be able to make a significant contribution to the bigger picture of NLP text summarization research, where she can in the future provide the community a base guide or a head start to facilitate domain adaptation procedures. As a final goal, yet to be achieved in the upcoming year, she wishes to develop her project to a stage where it will also be possible to retrieve the quality of the generated summaries.

A leap towards advanced legal technologies: integrating NLP into legal tech

In the rapidly evolving landscape of technology and law, the participant **Tobias Eder** stands out by modernizing the legal tech space with advanced NLP solutions. Tobias is a PhD candidate at the **TU München** School of Computation, Information, and Technology under Prof. Dr. Georg Groh's supervision. He has been developing his Software Campus project in collaboration with DATEV in a project aiming at leveraging BERT-encoded models and LLMs for semantic analysis and context understanding of legal documents. This project emphasizes the crucial role of natural language generation in modernizing legal text analysis to support legal professionals to its higher efficiency. Historically, legal tech has predominantly relied on rule-based systems, focusing almost solely on

automating tasks such as contract generation and parsing. However, the advent of modern NLP technologies, such as BERT and later models, has opened new possibilities for speeding legal document analysis and information retrieval coupled with facilitating knowledge transfer within law firms. The project's initial goal is to develop a system capable of bringing legal professionals up to speed on complex legal procedures. Tobias' team and his collaborators aimed to develop a platform that could efficiently summarize legal documents, identify key information and speakers, and provide an interactive question-answering system.

The emergence of LLMs also posed challenges to Tobias's project and confirmed the team's adaptability and commitment to innovation. Tobias, as Anum Afzal, also incorporated the new LLMs into the project, as a baseline to evaluate their own models' performance. The larger context window and processing capacity of LLMs enhanced the quality and depth of legal document analysis. This transition to models capable of handling extensive texts and facilitating interactive chats represents a leap towards more intuitive and effective legal tech solutions. Important to notice, that the project's innovation extends to its focus on the German legal system, addressing the unique challenges of legal language and its implications. By fine-tuning models specifically for the German legal framework, the team also contributes valuable insights into the adaptability of NLP technologies across different languages and legal contexts. In this aspect, the close cooperation with DATEV was crucial, as they provided Tobias with a consistent amount and high-quality datasets, aside from project guidance, and regular feedback.

Besides the already discussed inclusion of LLMs, to improve the model's output performance, Tobias is also now using RAG models. This new architectural approach enhances the efficacy of LLMs by using customized data, and in this case, Tobias has been leveraging a plethora of relevant legal documents from his dataset as context for the model. Given the constant and rapid evolution of the legal agenda, the use of RAGs for this sector is particularly interesting, as it allows the model enrichment with up-to-date information.

Tobias's reflections on his development as a young manager reveal the impact of participating in a leadership accelerator program. The program's emphasis on leadership skills and interdisciplinary collaboration has equipped him with the tools to navigate the complexities of project management and team leadership. The participant also emphasized the program's networking power. Throughout the Software Campus training and events, he had the chance to meet with other Industry Partners and NLP researchers around Germany to exchange both technical and management subjects. Tobias emphasized his excitement with the new Communities of Practice, which provides the participants with an extended and more solid networking channel for stable professional exchange and follow-up with other participants.

Looking ahead, Tobias envisions that NLP technologies become integral in multiple sectors, yet a less visible component to the users. The NLP's potential to automate and enhance legal

work, from document analysis to client consultation, points towards a future where professionals can use technology to deliver more efficient and accurate services. Likewise, the project's implications extend beyond the legal domain, offering a blueprint for integrating NLP into other fields. The foundation research and methodologies developed through this initiative have a vast application potential on how complex information processing can be tackled.
Tobias Eder still has a few months before becoming a Software Campus alumnus. We are eager to see the outcomes of this project, and maybe even experience an interactive prototype, as disclosed by him.

A new approach to conversational search: enhancing scientific literature search through NLP and knowledge graphs

In a world where the volume of scientific literature is overwhelming, Phillip's project emerges as an inspiration for the future of literature search. **Phillip Schneider** is a young computer scientist at the chair of Software Engineering for Business Information Systems at **TU München**, and, like Anum, he has been developing his PhD under the supervision of Professor Dr. Florian Matthes. Phillip initiated the Software Campus program very early as a PhD candidate and having Springer Nature – a Holtzbrinck's company – as his industry partner was fundamental to support the search for his main dissertation topic with an innovative perspective.

Applying to Software Campus coupled with the close collaboration with experienced mentors with different perspectives guided him in shaping his future steps into NLP.

> "When I applied for the program, I was four months into my PhD, so still quite early, and I hadn't found my research topic yet. Therefore, I can say that applying for the Software Campus program having these initial meetings with my industry partner, and writing the project proposal helped me to discover what I was most interested in."
>
> PHILLIP SCHNEIDER
> Software Campus Participant

With the COGNOSCO project (from Latin cognōscere: to discover/to know) he is proposing a new way to interact with and access scientific literature through a conversational search system. To create a more intuitive, interactive interface to explore publication data Phillip merged NLP techniques with the power of knowledge graphs. These are sophisticated knowledge representations mapping out the connection between data groups and their underlying relationship, making data integration performance significantly higher. Unlike traditional search engines that return a static list of results, Phillip's system can understand and guide users through the complexity of scientific literature repositories.

At the beginning of the project development, Phillip used a big Springer Nature database known as SciGraph, with a wide range of disciplines and based on knowledge graphs. However, a year later into the project the SciGraph was discontinued and no longer updated with the latest publications, a normal circumstance in business. With a strategic evaluation of the possibilities, Phillip's team decided to build up a new and more dedicated database for a use case focused on NLP research literature. Consequently, the development of the conversational agent would be more focused and controllable. Now, close to the end of the second project year, Phillip is already building the agent prototype and is optimistic about the prospective outcomes from the evaluation with real users.

Although the solution's target group was initially the scientific and highly specialized groups, Phillip slowly realized the potential of his tool. This system design makes it especially valuable for those who lack precise search terms or domain-specific knowledge. The idea is that by benefiting from a natural language-based retrieval chat a non-specialist can also access a greatly reduced number of relevant scientific publications to begin with. In our current scenario, where multidisciplinary projects are a growing need scientists and academics can also leverage such engines to familiarize themselves with different topics in a timely manner. This blend of technologies facilitates an exploratory search mechanism tailored for both experts and novices in the field.

To execute his Software Campus project Phillip counts on the help of his team. As a first-time manager, the trainings significantly improved his communication and negotiation skills. Towards the "Effective Leadership Communication" training organized by Zeiss and "Better understanding yourself and others", offered by Merck, he discovered the concept of empathetic leadership and gained a better understanding of the spectrum of personality types. This new perception played a crucial role in his development and his approach to collaborating with coworkers.

Phillip's thoughts on the impact of LLMs highlight the evolving challenges and opportunities in NLP research, accompanied by the necessity of critically evaluating their strengths and limitations. He argues that NLP research will be required to find a delicate balance between engineering-focused optimization and a deeper understanding of language itself. While larger models and improved algorithms can yield performance gains, there's a risk of losing sight of the core principles of computational linguistics. The computer scientist emphasizes the importance of retaining a focus on understanding language nuances and communication mechanisms. He also brings attention to the importance of an alignment with human needs and values, where the pursuit of technological advancement must be tempered with a commitment to ethical considerations.

Navigating the future of fact-checking and scientific claim verification

The young research associate **Juraj Vladika** is currently performing his PhD

SEBIS at TU München, in the same research group as his colleagues Anum Afzal and Phillip Schneider. The project entitled VeriSci – for "Scientific Claim Verification with Evidence from Text and Structured Knowledge" – exemplifies the intersection of AI and medical research and is dedicated to the automation of scientific claim verification and modeling disagreement, representing critical challenges within the medical and health sectors.

Claim verification consists of assessing the truthfulness or validity of text statements, in this case focused on scientific literature. Disagreement modeling, on the other hand, concentrates on understanding and representing differing viewpoints expressed in the texts, emphasizing a logical evaluation of its support or contradiction to the claims. This scenario is particularly interesting in medical sciences, as novel discoveries surge from time to time, changing paradigms and adapting decisions of the medical community. These engines pose a powerful advance to our understanding of textual content and promote informed decision-making in the dedicated fields. By filtering thousands of research publications, identifying and evaluating evidence supporting or refuting specific claims the project aims to transform the labor-intensive process of fact-checking – traditionally done manually by medical and health professionals – into an automated, efficient, and accurate system. Juraj's research results on comparing different knowledge sources for scientific claim verification are showcased in a recent publication at the EACL 2024 (Vladika & Matthes, EACL 2024).

When asked how NLP can bridge the gap in multiple, sometimes completely unrelated scientific areas, Juraj emphasized:

> "It is important that computer scientists with linguistic backgrounds work on automated solutions that can ease and optimize time-consuming and costly tasks. It will always be of great significance to have a specialist to closely curate and evaluate the outputs, but if there are tailored, well-designed, and trusted solutions to facilitate and speed these tasks, multiple fields can jointly benefit from consistent decision-making and technology development."
>
> JURAJ VLADIKA
> **Software Campus Participant**

Juraj's project applies NLP technologies coupled with Machine Learning architectures, such as BERT transformer models. Facing the challenge of a domain-specific language, he incorporated specialized, pre-trained models like BioBERT, PubMedBERT, and DeBERTa to enhance the system's comprehension of the domain terminology. To support him in developing his project, Juraj has the support of the Holtzbrinck Publishing Group as industry partner, more specifically the Digital Science team. Considering the level of data protection for medical-related documents, Juraj opted to use only open-source databases, and the interaction with the company is undertaken in the advising and knowledge exchange sphere. Their team

NLP INNOVATION:
OUR INDUSTRY PARTNERS' INSIGHTS IN THE ERA OF GPTS

The **Holtzbrinck Publishing Group** is one of the world's leading publishing and media groups and has been a Software Campus Industry Partner for the past 12 years. It operates in multiple business segments including academic publications (Springer Nature), research solutions (Digital Science), educational offers (Macmillan Learning), popular publishing (Macmillan Publishers), and Journalism (DIE ZEIT).

HOLTZBRINCK PUBLISHING GROUP

"The intellectual contributions stemming from these collaborations are invaluable, aiding the creation of groundbreaking technologies, including those involving NLP. The insights, perspectives, and specialized knowledge the young scientists and researchers bring amplify our understanding of these tools and, in turn, accelerate the development of next-generation technologies."

JULIA FURTWÄNGLER
VP of Employer Branding & Talent Acquisition

With its business focus and commitment to progress and discovery, the group is increasingly harnessing the transformative potential of Artificial Intelligence (AI), specifically NLP. At Holtzbrinck the aim is to use AI as a catalyst to fuel scientific progress and discovery, craft a more customized and adaptive learning environment, and unlock the inherent creative capacity of authors and creators globally. Holtzbrinck's approach shows an unwavering emphasis on ethical, reliable, and secure AI interventions. A guiding principle within Holtzbrinck's operations is humanizing technology by placing individuals and creativity at the forefront of its initiatives.

For them, the Software Campus Program plays a significant role in sustaining this vision by fostering exchanges with researchers and scientists.

Similarly, and as one of our longest-standing industry partners, **DATEV eG** has supported the Software Campus program since its establishment in 2011. As one of the largest IT service providers in Germany, DATEV plays an active role in driving digitalization forward. Germany's tax and legal sectors are prime candidates for the transformative power of NLP. These fields struggle with vast amounts of complex documents written in a language known for its precision and complexity.

NLP offers significant potential to streamline processes and empower professionals. Imagine lawyers who can leverage NLP to work their way through mountains of contracts, automatically extracting key details or highlighting potential risks. This frees them to focus on higher-level analysis and strategic

thinking. Tax consultants, too, can benefit by using NLP to automate accounting processes, analyze regulations, and optimize tax planning from information hidden within dense legalese. NLP's ability to process and understand these particularities can significantly reduce workload and expedite tasks, allowing professionals to dedicate more time to complex issues and client needs.

DATEV is a great supporter of the Software Campus CoPs, and for them, by incorporating these Communities into the program, participants gain access to valuable networking and learning resources. They also connect students with alumni and professionals across different experience levels. This fosters knowledge sharing, career guidance, and exposure to the realities of the IT industry, including strategic development and technological trends.

DATEV EG

"Participants gain access to peer-to-peer mentoring, allowing knowledge exchange and fostering a supportive network. Industry partners benefit by having a direct line to talented students and fresh perspectives, potentially leading to stronger industry connections and innovation."

STEFAN MELLES
Workstream Technological Innovation

THE SOFTWARE CAMPUS NLP COMMUNITY

Aside from the clear advances in scientific knowledge and new models' proof of concept, we saw the intricacies of some of the Software Campus NLP projects, which are developing tools and solutions to specialized demands in industry and society. The rapidly evolving tech landscape directly impacts decision-making, and adjustments are constantly necessary. To manage teams and projects in this scenario, young researchers need to be equipped with skills beyond technical knowledge. It is evident that early exposure to those situations significantly impacts young professionals' growth and development. The Software Campus program is glad to play a role in this process.

At the NLP Community, three events were already conducted. At the first event, during the pilot phase, the members had the chance to introduce themselves and their project subjects to their peers and to get familiarized with this initiative's dynamics and goals. During the session, important feedback on their expectations of having a dedicated NLP CoP in the Software Campus was brought into discussion. Their perceptions and needs culminated in essential adjustments to establish the next Communities and plan future activities. In a second event, the NLP members attended their first Specialist Session: a keynote speech offered by Philipp Grandeit from DATEV. In his talk, the DATEV's Data Specialist discussed real use cases and the challenges of generative AI systems use.

Philipp's talk raised curiosity and relevant discussions could be addressed in a relaxed conversation.

To kick-start 2024, the NLP Community grew with five new members from the 2023 round. They are now starting to implement their projects and were introduced to their colleagues in our first 2024 NLP event in early April. This time we organized a Specialist Talk with our Research Partners TU Berlin and the DFKI. As part of Professor Sebastian Möller's team, Vera Schmitt and Salar Mohtaj have discussed the implications of the recently approved European AI Act for research and development and NLP innovation. The Act's nuances and potential challenges posed by the regulatory framework were also addressed, with particular emphasis on the importance of constant crosstalk and alignment due to the ever-evolving nature of AI-based tools.

The next NLP CoP activity, coming in June 2024, will for the first time be focused on peer mentoring and exchange. The participants will present their projects and get feedback from the Community members. We at the Software Campus are excited about this event and curious to witness the participants' engagement and exchange.

CONCLUSIONS AND REMARKS

Currently, we have five active Communities of Practice, where the members are all connected and ready to learn and exchange. They are the NLP, the Computer Vision, the Cybersecurity and Privacy, the Hardware and Systems Engineering, and the Entrepreneurship and Career Development CoPs. The fifth Community is a "meant-for-all" community focused on career development and the exchange of experiences for the future job market. The sessions promoted at this CoP are planned to be comprised of panel discussions with Software Campus alumni, each of whom is to share their insights on the challenges and potential of the Computer Science labor market, based on their career journeys. During the next few months, another four CoPs will be implemented: Data Science and Analysis, IoT and Distributed Systems, Interdisciplinary Machine Learning, and Agile and Software Development.

In the next issue of Behind the Code we will follow the development of the Computer Vision CoP, launched March 2024. This Community focuses on projects spanning smart cities, autonomous driving, medical robotics, biometric systems, and more. The Software Campus projects in Computer Vision are partnered with companies such as Volkswagen, Zeiss, Software AG, IAV, and Stahl-Holding-Saar. The participants develop their projects at our partner research institutions TU Darmstadt, TU Dresden, Fraunhofer IUK-Technologie, and the DFKI.

ACKNOWLEDGMENTS

The Software Campus team kindly acknowledges all the professionals who found time to contribute to this article. A special thank you to the participants who accepted being interviewed and shared their personal stories and perspectives beyond the programming. Your commit-

ment to self-development within the Software Campus program is inspiring to witness.
To Holtzbrinck Publishing Group and DATEV eG, we appreciate your feedback and openness in supporting the Software Campus Communities of Practice and the production of the Behind the Code series. We also greatly acknowledge the BMBF funding (Förderkennzeichen 01IS22096). To the DLR Projektträger, thank you for supporting the Software Campus program and additionally the Communities of Practice initiative.

LEGAL NOTICE

Published by

EIT ICT Labs Germany GmbH, Berlin

Genthiner Straße 8
D-10785 Berlin

Stefan Jazdzejewski
Managing Director