Abstracts of Impact Track Presentations

Jonathan Gerber, Mark Cieliebak, Don Tuggener, Manuela Hürlimann

Zurich University of Applied Sciences (ZHAW) and
Swiss Association for Natural Language Processing (SwissNLP)
[gerj, ciel, tuge, hueu]@zhaw.ch

1 Assessing the Trustworthiness of Large Language Models on Domain-specific Questions

Sandra Mitrovic, Roberto Larcher and Jérôme Guzzi

Pre-trained Large Language Models (LLMs) can be leveraged to answer domain-specific questions using prompt engineering and retrieval-augmented generation. However, ensuring the trustworthiness of such systems remains a critical challenge. In this work, we propose a general methodology to evaluate the reliability of LLM-based modules by constructing large, representative, and unbiased datasets of questions and answers through automated variation generation. We define key metrics to assess correctness, robustness, and explainability. We apply our approach to a real-world use case in which a smart wheelchair provides answers about its functioning, exploiting RAG with ChatGPT as the underlying LLM. Our experimental results, based on a dataset of over 1,000 questions, reveal that while correctness and robustness are generally strong, the model struggles with open-ended questions, negations, and idiomatic expressions, with explainability being the most challenging aspect. Beyond the specific results (which heavily depend also on dataset at hand), we emphasize the generalizability of our methodology, which can be adapted to various domains. We are currently working on automating the evaluation pipeline to reduce reliance on human assessment and extending the methodology for real-time monitoring of LLM responses.

2 Building commercial GenAI-based solutions: Emerging use cases and best practices

Olmo Barberis and Keibel

Over the past three years, LLMs have impressed the world with their powerful capabilities to understand and generate human language. As with most technological innovations, it takes time and significant efforts to build successful productive solutions (and not just prototypes) around LLMs which generate real-world business revenues and ultimately render the upfront investment profitable. Companies and organizations around the world are still at a fairly early stage in exploring how to best leverage LLMs productively, but some trends and best practices are emerging as to which types of use case are worthwhile to pursue and how LLM-based solutions should be built. In this talk, we will pick up on some of these trends and best practices and sketch out what we believe most commercial GenAI projects might look like in a few years from today. In doing so, we take the perspective of projects which do not have unlimited budget. We will look more closely at some common challenges and ways to mitigate them. We will give some examples from real-life projects that focus on automating business processes.

3 Enhancing Qualitative Content Analysis via LLM Multi-Agent Systems

Norman Süsstrunk, Caroline Dalmus and Albert Weichselbraun

Despite the growing popularity of using large language models (LLMs) such as ChatGPT for qualitative content analysis, current approaches often rely on overly simplistic prompting strategies.

As Mayring (2025) highlights in his field report, the primary issue lies in the inadequacy of many prompt designs. General requests such as "Do a qualitative content analysis according to Mayring" result in superficial outputs that lack adherence to the step-by-step methodology central to rigorous qualitative analysis. Even with more structured prompts, ChatGPT frequently fails to follow essential procedural elements such as inductive category formation, abstraction level calibration, and coder agreement testing. The outcomes typically resemble rough summaries rather than methodologically grounded categorizations, leading to what Mayring refers to as "rough approximations and gross errors". These limitations are further exacerbated when applied to larger datasets or when theoretical grounding is required. Prompt-based approaches, even when refined, struggle to maintain the iterative and transparent logic required by Mayring's qualitative content analysis. As a result, the reliability and reproducibility of the outcomes remain questionable. To overcome these limitations, we propose the integration of multi-agent systems that mirror the structured, procedural logic of Mayring's methodology. Rather than relying on single, monolithic prompts, a system of specialized LLM agents can be deployed, with each agent responsible for a specific task aligned with Mayring's distinct techniques (e.g., inductive category formation, summarization, explication). For instance, individual agents can be designated to handle:

- Category definition
- Calibration of abstraction levels
- Identification and validation of coding units
- Verification of coder agreement

Crucially, these agents would operate under human oversight, ensuring interpretive validity and adherence to ethical and methodological standards. This agent-based architecture is inspired by recent advances in the design of LLM agents, where specialized agents collaborate under human guidance to plan, execute, and optimize complex experiments. By adapting this collaborative structure to qualitative content analysis, we can reflect Mayring's method not only in output, but also in process – step-by-step, transparent, and verifiable. This hybrid system presents a promising way to elevate current practices from surface-level approximations toward structured, scientifically grounded qualitative content analysis.

4 Entity Extraction, Linking, and Disambiguation Pipeline for News Documents

Tsvetan Rangelov, Yannick Suter and Guillaume Comte

At RepRisk we maintain a large database of news incidents coupled to com- panies accused of ESG issues violations. This data is used by asset managers, investors or institutions to make informed decisions about entities they are inter- ested in. This data is multilingual, with a long history, enriched by new compa- nies created every day and combines both human analysis and machine learning. Our pipeline addresses the complex challenge of associating news documents with corporate entities, a critical need for clients who rely on accurate, timely data. Faced with the absence of a single source of truth, duplicate records, and disparate naming conventions—where legal names, journalistic aliases, and outdated entries coexist—we developed a robust, multi-faceted solution. We leverage our unique dataset where texts are associated with IDs to identify entities corresponding to both legal names and the commonly-used variants as well as custom transliteration routines to address our varied multilingual data. Our approach integrates advanced entity extraction with candidate generation, recall-based linking for candidate selec- tion, and precision-based verification for optimal results. To enhance multilingual performance, we incorporate all this contextual information into cutting-edge trans- former models combined with large language models through tailored prompting. This comprehensive system not only resolves data inconsistencies across heteroge- neous sources but also sets a new benchmark for technical rigor and operational efficiency in real-time news content processing.

5 ErrorCatcher: LLM-Powered Editorial Quality Assurance for Reuters News

Luca Malagutti, Guilherme Thomaz and Claudia Schulz

In the fast-paced environment of news production, ensuring editorial quality while maintaining tight publication schedules remains a significant challenge. We present ErrorCatcher, an LLM-powered editorial quality assurance system devel-

oped at Reuters News to help journalists identify and correct both syntactic errors and style guide violations before publication. ErrorCatcher leverages a suite of specialized prompts designed in collaboration with experienced journalists to analyze news articles across multiple dimensions: grammatical correctness, adherence to in-house style guidelines, consistency in terminology, and in-story factual coherence. The system offers targeted feedback, identifying errors and suggesting corrections which reference relevant style guidelines. Our system addresses a significant challenge in integrating sizable organizational style guidelines by developing a hierarchical approach that categorizes style elements by priority and relevance, enabling the system to focus on the most pertinent rules, lowering costs and improving response coherence. We evaluate several leading LLMs as the backbone of our system, revealing that LLMs optimized for complex reasoning demonstrate superior capabilities in identifying subtle style inconsistencies and nuanced grammatical issues across journalistic content. Our preliminary deployment of ErrorCatcher as an internal tool has shown promising results, with journalists reporting improved workflow efficiency and heightened awareness of recurring style issues. We outline our approach to developing ErrorCatcher, discuss the technical and practical challenges of implementing AI editorial assistance in a global news environment, and share our progress in extending ErrorCatcher with additional capabilities while evaluating its performance.

6 Exploring NLP-Driven Personalized Support for Type 1 Diabetes Management: A Preliminary Study

Sandra Mitrovic, Federico Fontana, Andrea Zignoli, Christian Berchtold, Sam Scott and Laura Azzimonti

The widespread availability of wearable devices and sports monitoring applications has enabled individuals, including those with Type 1 diabetes (T1D), to easier track their physical activity. Given the importance of exercise in managing T1D, personalized feedback can play a critical role in optimizing workout routines while mitigating the risks of hypo- and hyper-glycemia. This study explores the feasibility of leveraging Natural Language Processing (NLP) models to generate tailored messages based on an individual's activity data and expert inputs. In particular, we consider two types of workouts: with negative-outcome (i.e., where the individual's glucose level went out of range, further subdivided into hypo- and hyper-glycemia) and with positive- outcome (i.e., where the individual's glucose level remained within the range). Negative-outcome workouts require a behavior change, and messages should advise the individual on how to adjust. Conversely, if the outcome is positive, the individual should be encouraged to maintain their current behavior. Driven by the potential future goal to integrate our approach into an app that prioritizes user privacy and transparency, we focus on evaluating several open-source NLP models to determine their effectiveness in producing high-quality, personalized messages. Furthermore, we consider two types of prompts. First, the simpler one, referred to as the observable prompt type, is based on the combination of a behavioral pattern (i.e., a more precise description of the out-of-range behavior selected from a pre-defined set of possibilities) and its accompanying expert-provided information. Second, the more complex one, referred to as the actionable prompt type, adds to the observable prompt type personalized actionable variables (derived by the underlying ML model1). Additionally, we implemented prompt refinement strategies to enhance message quality and safety, though further research is needed to optimize these approaches. We perform quantitative and qualitative evaluation of prompts. For example, within the qualitative evaluation we focused on prompt adherence, correctness, level of detail, emotional tone, and medical content comprehension. Contrary to expectations, our results reveal that models fine-tuned on medical data or those excelling in medical benchmarks do not necessarily generate superior messages for this application. Among the tested models, Mistral-7B-Instruct-v0.3 demonstrated the most promising performance, while others, including Starling-LM-7B-beta, gemma-2-2b-it, Llama-3.2-3B-Instruct, and JSL-MedPhi2- 2.7B, yielded suboptimal outcomes. This work serves as a proof of concept for the feasibility of using personalized NLP-driven messages in diabetes management, with the ultimate goal of driving behavior change. However, we acknowledge the limitations of our study, particularly regarding dataset size and the narrow scope of actionable variables considered. Future research should focus on expanding the dataset and refining both model selection and prompt engineering techniques to improve the relia-

bility and effectiveness of NLP-generated guidance in diabetes care.

7 GZIP-KNN for ChatGPT Text Detection: A Low-Resource Alternative to Supervised Methods

Matthias Berchtold, Sandra Mitrovic, Davide Andreoletti, Daniele Puccinelli and Omran Ayoub

With the increasing capability of Large Language Models (LLMs) to generate highly plausible and human-like text, the need for reliable AI-generated text detection has become critical. This need is additionally underpinned by recent findings of several studies showing that even adults often struggle to distinguish between human- and machine-authored content. Furthermore, misattributing authorship can lead to the spread of misinformation and the unethical appropriation of text. On the other hand, Transformer-based architectures, which power these models, are highly resource-intensive, adding another layer of complexity to their widespread use. In this study, we investigate the potential of GZIP-KNN, a recently proposed lightweight method, for detecting AI-generated text, specifically content generated by ChatGPT. We evaluate GZIP-KNN's predictive performance, training time, inference time, and memory footprint in comparison to logistic regression, eXtreme Gradient Boosting (XGB), and Gated Recurrent Unit (GRU). As our focus is on low-resource approaches, we do not consider pretrained models. Using five open datasets from different domains, we conduct two experiments. The first examines the trade-off between predictive performance and computational complexity in an in-domain setting. The second assesses performance under data and inference time constraints in an out-of-domain scenario. Experimental results indicate that GZIP-KNN achieves strong predictive accuracy, outperforming alternative methods even with limited data. However, its higher inference time limits its applicability in scenarios requiring rapid decision-making. Nonetheless, findings suggest that GZIP-KNN can match the performance of other methods when trained on only a small subset of available data in an out-of-domain context.

8 Presenting LLMs' collective intelligence approach for Multilingual Hallucination Detection

Sandra Mitrovic, Joseph Cornelius, David Kletz, Ljiljana Dolamic and Fabio Rinaldi

Hallucinations pose a crucial problem in the utilization of large language models (LLMs). The problem is even more pronounced as literature lacks the standardized definition of hallucinations. Furthermore, different LLMs may identify different parts of the same text as hallucinations and in general, different LLMs have different hallucination rates. The problem of identifying hallucinations is even more complex in the multilingual setup. In this study we present our approach to multilingual hallucination detection, as part of MuSHROOM ("Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes"), a SemEval-2025 Task-3. This task is complex as it consists in both detecting exact hallucination spans and determining the hallucination probability. Moreover, the task covers 14 different languages and provides no labeled data, apart from several validation instances for 3 different languages. Task used two evaluation metrics: intersection-over-union (IoU) and correlation (Corr). We tackle this problem simulating the original annotation process that uses multiple artificial annotators. Each artificial annotator is instantiated through a different LLM service combined with varying prompts. Subsequently, the outputs of individual annotators are aggregated into a single annotation using as final hallucination probability the ratio of annotators that denoted the span as a hallucination. We use six different LLM APIs and three different prompts, and we experimented also with different merging variants. Our approach shows great potential as it, in terms of IoU, scored 4th for French (out of 30 teams), 5 for Italian (out of 28 teams), 12 for English (out of 41 teams), and 15th for German (out of 28 teams). In terms of Corr, the results were even better as we ranked 1st , 3rd , 4th and 7th for English, German, French and Italian, respectively. Beside the quantitative results, where we established which models and prompts perform the best, we also performed extensive qualitative analysis, looking deeper in different aspects of differences between published ground truth and our system annotations.

9 Public Unveiling ESG Insights in Real-Time: A Live Demo of RepRisk's ML Pipeline

Guillaume Comte, Tsvetan Rangelov and Yannick Suter

At RepRisk, a leading ESG data provider, we harness the power of Natural Language Processing (NLP) and Machine Learning (ML) to extract critical ESG insights from news articles worldwide. Our advanced ML pipeline processes vast amounts of unstructured text to identify key ESG-related events, assess company involvement, and generate structured, actionable insights.

In this live demo, attendees will have the opportunity to select news articles of their choice, which will then be processed in real time through RepRisk's multi-stage ML pipeline. The system will extract ESG-relevant information, classify incidents, map companies to their identifiers, and generate predictive insights, all displayed dynamically in our interactive UI. Each prediction and extracted entity will be clickable, allowing users to explore related incidents and navigate company profiles directly on RepRisk's platform.

This session will not only showcase the sophistication of RepRisk's NLP-driven ESG analytics but also allow participants to experience firsthand the accuracy and depth of our AI models in transforming raw news into meaningful ESG intelligence.

10 RAG vs Long-Context LLMs: Choosing the Right Approach for NLP Applications

Elena Nazarenko

The landscape of Natural Language Processing (NLP) has been dramatically reshaped by the rise of Large Language Models (LLMs). Two key architectural approaches have emerged to address the challenges of integrating external knowledge and processing large volumes of text: Retrieval-Augmented Generation (RAG) and Long-Context LLMs. RAG systems excel at incorporating external knowledge sources into the generation process. By retrieving relevant documents or passages based on user queries, RAG enables LLMs to provide contextually accurate and up-to-date responses, mitigating the limitations of pre-trained models. This approach significantly expands an LLM's access to vast amounts of information at minimal cost. It is particularly valuable in applications requiring access to dynamic or proprietary information, such as question answering over internal knowledge bases, document summarization, and personalized recommendations. Conversely, recent models like Gemini 2.0, Claude 3, and GPT-4.5, with extended context windows (120K–2M tokens), have demonstrated exceptional capabilities in processing extensive text directly. This eliminates the need for external retrieval, potentially simplifying architecture and reducing latency. These models excel in scenarios where the entire relevant context is available, such as analyzing legal documents, processing scientific papers, or handling complex narrative tasks. However, the choice between RAG and Long-Context LLMs is not always straightforward. RAG systems introduce complexities related to retrieval quality, indexing, and latency, while Long-Context LLMs face challenges with computational cost, potential information dilution, and the "needle in a haystack" problem. This presentation aims to provide practical insights and guidance for NLP practitioners, enabling them to make informed decisions when selecting the most appropriate approach for their specific applications.

11 SYMBOL - Neurosymbolic AI for explainable and reliable AI in high-stake environments

Albert Weichselbraun and Norman Süsstrunk

Lack of explainability and reliability (e.g., due to hallucinations, misleading information, or biased outputs) are serious obstacles towards the adaptation of LLMs in high-stake environments. SYMBOL tackles this shortcoming by developing neurosymbolic AI models that combine embedding-based language models (sub-symbolic processing) with machine-readable domain knowledge (symbolic reasoning), organized in knowledge graphs. The project aims at bridging the semantic gap between user queries, the company's information systems (e.g., databases, customer relationship management systems, and software APIs) and its knowledge management infrastructure (e.g., domain ontologies, structured knowledge in databases and knowledge graphs, and corporate knowledge repositories). LLMs interpret user queries and translate them to the corresponding concepts in the knowledge graph. This enables processing of queries using symbolic AI which ensures very high reliability, since reasoning within symbolic AI components is deterministic. Symbolic reasoning upon domain-specific knowledge graphs also explains

query results and decisions based on (human understandable) concepts within these graphs, ensuring that system decisions are traceable and explainable to non-computer scientists. Once completed, SYMBOL will support clients in the wealth management industry by

- navigating and aiding users through complicated regulatory requirements;
- generating regulatory reports and analyses on demand, helping wealth management firms respond to audits, risk assessments, and evolving compliance mandates; and
- extracting deep business insights from their data, enabling proactive decision-making based on structured, regulatory-compliant intelligence.

By allowing non-technical users to interact with SYMBOL the project will eliminate barriers to data-driven decision-making, ensuring that compliance officers, portfolio managers, and auditors can extract the necessary information when it matters most – such as to support high-stake decision-making processes and on-site regulatory reviews. Although the wealth management use case is central to the SYMBOL project, we aim at adapting the developed neurosymbolic AI components to other high-stake environments in domains such as finance, medicine, and law.

12 Scaling RAG from Pilot to Production: Evaluation, Software practices and Safety

Louis Douge and Robert Simmen

We developed and deployed Life Guide Scout, a GenAI-powered underwriting assistant to more than 3,000 Life & Health underwriters worldwide. The system uses a Retrieval-Augmented Generation (RAG) setup to integrate Swiss Re's proprietary underwriting guidance and medical knowledge, thereby speeding up information retrieval. Fully integrated in the underwriter's workflow, it enables intuitive, efficient and trustworthy interactions with highly specific knowledge. The real challenge of productively deploying an LLM-based system lies in assessing its performance over time and across versions. We present a comprehensive evaluation methodology based on synthetically generated data. For instance, on the specific task of mentioning the right underwriting rating in Life Guide Scout's answer, we achieve an end-to-end 80% hierarchical recall, a metric particularly suited to our problem. We also examine the various failure modes and suggest mitigations. In addition to this programmatic approach, human feedback played a crucial role in refining Life Guide Scout through three key approaches: expert evaluations for structured assessments, user feedback within the application for real- time insights, and surveys and interviews to gauge adoption trends. This multi-layered approach ensured continuous iteration, improving accuracy, usability, and overall user satisfaction. Developing GenAI applications also requires a blend of new and traditional engineering practices. We share insights on prompt management techniques, structured outputs, and strategies for handling frequent LLM updates, including new models and versions. While LLMs introduce novel challenges, traditional software engineering practices remain critical. We detail unit, integration, and regression testing methods, which are essential for iterating on an LLM- centric application in a production environment. Given the risks of incorrect AI-generated outputs in an insurance context, we implemented pre- and post-processing techniques to reduce inaccuracies by leveraging the specificities of our problem. Enhancing transparency, we introduced source anchoring using IDs, which not only links references but also highlights the exact section or phrase within the source that the LLM used to generate its response. This improves user trust and allows for quick verification of information. GenAI introduces new risks related to safety and security. We conducted extensive adversarial attacks, or Red Teaming, on Life Guide Scout to uncover vulnerabilities and proactively mitigate risks, ensuring alignment with responsible AI principles. By stress-testing the system against adversarial scenarios, we strengthened safeguards, improving both security and reliability. Finally, we share our approach to developing conversational memory within a RAG setup while managing token usage effectively. Maintaining context across interactions enhances the user experience but presents engineering trade-offs that we addressed through targeted optimizations.

13 SetFit for Automated Essay Scoring: Extending Longformer to a Sentence Transformer

Leon Krug, Jannik Bundeli, Jannine Meier and Elena Nazarenko

Automated Essay Scoring (AES) demands models that can evaluate student essays with human-like consistency while maintaining computational efficiency. Although standard transformer models like DeBERTa can achieve strong performance, they are often resource-intensive and constrained by a 512-token input limit, which can lead to truncated context in longer essays. This limitation hinders the model's ability to capture argument flow, coherence, and global structure, which are crucial for accurate scoring. Additionally, many existing approaches also rely on prompt engineering, further restricting practical application. To address these challenges, we present a novel prompt-free approach using SetFit for AES that achieves competitive accuracy while significantly reducing computational overhead. Unlike traditional transformer-based models, SetFit enables sentence transformer fine-tuning with contrastive learning, making it suitable for essay scoring even in low-data regimes. We extend Longformer into a sentence transformer, allowing SetFit to process full-length essays within a 4096-token window. This overcomes the 512-token restriction of traditional transformers, ensuring that the model can evaluate entire essays rather than isolated sections. Our approach integrates SetFit's lightweight contrastive learning to optimize sentence embeddings, enabling efficient, prompt-free fine-tuning with significantly lower GPU requirements compared to full transformer fine-tuning. By using contrastive learning, our model learns rich representations of essay quality without needing large-scale labeled datasets. We train our model on AES-specific datasets, so it captures the complexity of essay evaluation metrics such as coherence, grammar, and argumentation strength. Our fine-tuned model has been publicly released on Hugging Face, where it has already gained over 6,000 downloads, reflecting strong community interest in efficient, long-text NLP solutions. Our results show that SetFit with an extended Longformer sentence transformer achieves competitive accuracy and offers a cost-effective, scalable alternative to resource-heavy methods. Beyond essay scoring, our approach is applicable to other long-form NLP tasks, including legal document analysis, research paper assessment, and educational content evaluation, providing a cost-effective alternative to computationally expensive transformer-based models.

14 Transforming Healthcare Documentation: Efficient AI-Powered Automation of Clinical Discharge Summaries for Inpatients

Chantal Zwick, Joseph Weibel, Daniel Olivier Perruchoud, Tristan Struja and Felice Burn

Large language models (LLMs) are widely used to speed up administrative processes across industries. In the medical sector, physicians spend up to 2/3 of their work time with administrative tasks. LLMs could substantially alleviate this burden, allowing for more time with patients. Given the complexity of summarizing information from multiple sources and the sensitivity of content contained in medical documents, LLMs need to be deployed with the utmost scrutiny on local hardware. We therefore assessed the quality and thoroughness of discharge notes generated by locally hosted state-of-the-art LLMs compared to human-written notes. Methods: History of present illness (HPI) as well as diagnoses and procedures (DXL) were extracted from patient records for three clinical scenarios: planned or elective chemotherapy (PEC), acute coronary syndrome (ACS), i.e. myocardial infarction, and acute lower back pain (ALBP). Three medium-sized LLMs, i.e. Mixtral 8x7B, Mixtral 8x22B and Llama 3.1 70B were prompted to generate discharge sum- maries based on HPI and DXL inputs. Three approaches of generating discharge notes were compared: prompting without examples (zero-shot approach), In-Context Learning (ICL) which utilized four examples of triplets consisting of HPI, DXL, and humanwritten discharge summaries (4-shot approach), and supervised fine-tuning (SFT) on Mixtral 8x7B with specific training sets (NP EC-train = 1028, NACS-train = 1920, NALBP-train = 1494). For evaluation, five simple and five complex samples were extracted for each of the three scenarios, resulting in 30 triplets of HPI, DXL and human-written dis- charge summaries. Using the different LLMs and different prompting approaches, this results in a total of 150 generated discharge summaries, which were assessed via BLEU, ROUGE-L, and BERTScore metrics. In addition, a blind panel of 6 specialists in internal medicine assessed the 150 summaries with a modified Physician Documentation Quality Instru-

ment (mPDQI-9) consisting of nine items rated on a 5-point Likert scale, with higher scores indicating better performance. Results: Our findings indicate that both ICL and SFT enhance the quality of the generated discharge summaries compared to the zero-shot approach. The improvements were most notable for SFT in the PEC scenario (median 32 vs 28 out of 45). In general, generated reports for simpler cases received higher human ratings compared to more complex cases, particularly for the PEC scenario, but hallucination was a problem. When benchmarked against their respective ground truth discharge summaries, we achieved a BERTScore of 0.75, a BLEU score of 0.18, and a ROUGE-L score of 0.35 for the simple cases with SFT, which was the best approach. Overall, zero-shot Mixtral 8x7B, 8x22B, and Llama 3.1 70B demonstrated similar performance based on the expert panel's assessment. Conclusion: Our findings demonstrate that LLMs create medical discharge summaries for simple clinical scenarios with acceptable quality, but struggle with more complex cases. This highlights the need for accurate prompting, technical solutions to hallucination, and high quality input data in training models. Addressing these challenges would alleviate much of the administrative burden for physicians, especially those in training, which currently spend only 30 % of their workdays directly with patients. This approach has the potential of enhancing workflow efficiency, reducing clinician burnout, and improving

15 Unlocking Model Potential: A Comprehensive Framework for Feature and Data Enhancement

Xavier Ferrer, Alessandro Caruso and Claudia Schulz

In the dynamic landscape of machine learning, optimizing model performance relies on a thorough analysis of feature spaces. This study introduces an innovative framework designed to refine and improve machine learning models through meticulous feature analysis. We explore the correlations between the features and the model predictions to identify areas of improvement and potential feature gaps. By targeting misclassified samples, we uncover patterns that may elude conventional models, enabling us to propose targeted adjustments in model architecture and feature engineering. We leverage SHAP (SHapley Additive exPlanations) analysis together with unsupervised learning techniques, such as PCA or t-SNE, to reveal nonlinear relationships and natural data groupings based on feature vectors. Furthermore, we employ K-Nearest Neighbors (KNN) and cluster analysis to detect annotation errors by identifying homogeneous feature vector clusters and to enhance data integrity by flagging potential misannotations for review. We applied the proposed framework to an entity matching project, where text-based features are compared between different documents to identify matching pairs. This approach allowed us to identify the limitations of our models and guide the creation of new features specifically designed to distinguish between samples with very similar feature vectors but different annotations. Clustering analysis also helped identify and correct erroneous annotations in the dataset, resulting in a significant improvement in model performance. Our framework not only identifies and corrects model weaknesses, but also proposes strategies to build more robust, accurate, and interpretable models, ultimately advancing their applicability in real-world scenarios. Although tailored for NLP challenges, the framework is also applicable beyond NLP for any feature-based ML model. This study serves as a guide for data scientists and machine learning practitioners seeking to optimize model performance through comprehensive feature analysis and enhancement techniques.

16 "Radikale Diskurse lichten" Automated Telegram monitoring for analysis & research

Lars Schmid

The RaDisli ("Radikale Diskurse Lichten") project introduces an automated, dynamic monitoring tool that systematically collects and analyzes extremist content from Telegram channels. The prototype leverages advanced NLP techniques to provide real-time analytical insights into radical discourse, speci cally supporting monitoring and analytical e orts within social work. Key features include individual ltering by channels, time range, and search terms. Each message undergoes automated classi cation into categories: "hate speech," "toxicity," "threat," and "extremism." The Streamlit-based web application visualizes activity patterns through heat maps, highlighting peak communication times. Word clouds summarize frequently used terms per channel or group, and topic modeling via Latent Dirichlet Allocation (LDA)

provides insights into prevalent themes within the discourse. Additionally, a network graph visualizes interconnections between channels based on forwarded messages, highlighting in uential hubs and dissemination pathways. Evaluations of the prototype indicate that the application signi cantly enhances analytical capabilities. Users report that the streamlined, image-free interface reduces emotional stress and allows for a more objective, neutral assessment of extremist content compared to direct interaction within Telegram. To date, the system has processed and analyzed over 3.1 million messages from more than 180 channels, demonstrating robust scalability and performance.