Maturing Artificial Intelligence – Data Science for Real-World Applications

"AI-DSRWA" Special Track at the
Fourteenth International Conference on Pervasive Patterns and Applications
– PATTERNS 2022 –
April 24, 2022 to April 28, 2022 – Barcelona, Spain

Joschka Kersting
Paderborn University
Warburger Str. 100, 33098 Paderborn, Germany
joschka.kersting@uni-paderborn.de

***Abstract*—The world's data is growing enormously, and great advancements have been made in Artificial Intelligence (AI) recently. This applies to machine learning and data science in general and Natural Language Processing (NLP) in particular. However, AI is often developed for research purposes instead of real-world applications. This special track aims at collecting studies laying an application focus on submitted studies. Hence, the state-of-the-art is enhanced by applicable research results that do not focus on percentage gains in evaluation scores, but on solving problems for usage in the real world. Organizations may use them as a blueprint for own developments. This paper aims at summarizing the results of the track. The track deals especially with applications of recent data science methods for small and large corporations. Hence, production topics as well as data tracking and data acquisition are treated. Furthermore, information extraction of implicit and yet not investigated data types are discussed, because most research papers focus on standard datasets and tasks.**

***Index Terms*—Data Science; Natural language processing; Industry Applications; AI for SMEs.**

I. Introduction

For small and large firms as well as for a society as a whole, AI is the next thriving technology that is understood by few. To enable broader ranges of people and organizations to understand machine learning, commonly named AI, it must be applicable in everyday tasks. However, most research has focused on common problems that are said to provide a sufficient element of reality. For example, most research in Aspect-based Sentiment Analysis (ABSA) dealt with nouns and adjectives and hence directly mentioned aspects [1]–[3]. Furthermore, most studies focused on commercial domains such as products, hotels and restaurants [4], [5]. Phrases that rate inter-personal behavior such as services were neglected. Among the studies published in this special track is one that tackles this issue [6].

Approaches like the above mentioned help dealing with non-standard areas of interest and less clear topics. Other areas of interest concern non-sandbox applications of data science tools like in production. Firms nowadays plan large manufacturing plants, but not everything can be calculated before the plant is built and used. That is, wear part and chip abrasion cannot be predicted, because they may form certain entities that block machines. This causes production line shutdowns that cost millions per day. Hence, investigating this issue is key.

Even common analysis solutions such as text classification, information extraction, and connected engineering solutions are for most practitioners the topics in need. More and more researchers are keen on developing algorithms for industrial problems. With this special track, we want to provide several AI solutions as blue prints to help practitioners cope with the issues they face.

II. Submissions

The first paper by Kersting and Bäumer [6] is concerned with "Implicit Statements in Healthcare Reviews: A Challenge for Sentiment Analysis". The study builds on related work [4] that was set in the domain of physician reviews. While the related literature aimed at rating phrases, the corresponding paper aims to find a definition of implicit rating phrases in general. The authors present a working paper that comes with an annotated text dataset. The contents are implicit and explicit aspect phrases that that can be found using machine learning algorithms. Among others, the authors use XLM-RoBERTa [7] that performs best. They manage to find a basis for discussion in their working paper that is supported by the experiments.

That is, they manage to shift the focus away from specific aspects to general considerations and thus follow the overall flow in recent NLP research. Their work may serve well to find broader but also detail-richer analyses based on text data in the future.

The second presented study by Bäumer, Denisov, Sirvend, and Weber [8] deals with the challenge of having only few data in many machine learning contexts: "Tackling the "We have no Data" Challenge: Domain-Specific Machine Translation in SMEs". The idea is here to help small and medium sized enterprises (SMEs) with the data collection and the data

usage. The authors provide a distinguished contribution to this topic while focusing on translation software. Professional domains have their own vocabulary, universities and private companies offer language courses focusing, e.g., on English vocabulary for mechanical engineering. This is also an issue when it comes to translation software that can help free human employees from enduring and tedious tasks such as translating long instruction manuals. In a globalized world, this is of special importance.

However, Bäumer et al. [8] help companies finding and processing data they have at hand, because firms often do not know that they have data at hand. Most translation data is already present in companies. In this sense, the study provide a great contribution to applied data science practices.

The study by Denisov, Bäumer, and Geierhos [9] deals with an utterly different topic that is nevertheless relevant to applied data science: "Track Me If You Can: Insights into Profile Interlinking on Social Networks". The topic here are profile interlinks across social media platforms. The idea is to model Digital Twins that were created to harm the users behind original profiles. The authors therefore analyze profile contents and links between profiles to warn users of potential threats. The study deals with data collected from YouTube[1] and Twitter[2] and identifies links referring to other domains.

Denisov et al. [9] herewith provide a first step for systematize profile data and identifying Digital Twins. In the time of dis-information, this is a fruitful direction that will help when being used for real-world applications.

Spoor, Weber, Hagemann, and Bäumer [10] present an industry-related study that deals with production issues: "Concept of an Inference Procedure for Fault Detection in Production Planning". They basically work on a structured feedback loop that runs between production and planning. The idea is to help planning plants by learning from running factories. As mentioned earlier, not everything in modern plants can be foreseen or planned, which is why the scholars investigate the issue.

This is a complex process that, as an example, also requires an ontology and knowledge base, but also the definition of a metric. The topic is therefore an integral part of applied data science, even if, or rather because it is not dealing with standard tasks that are commonly investigated in research.

III. Conclusion

As a result, it can be seen that the corresponding studies have contributed to a maturing AI and thus to data science for real-world applications. There are several contributions presented, such as the definition and method to process implicit aspect phrases with machines. Furthermore, Digital Twins and issues of the globalized social media networks are investigated in a practical manner. SMEs can benefit from the solved 'no data challenge' and lastly, we have authors who help detect faults in corporate plants.

[1]YouTube is available at https://youtube.com, accessed 2022-04-14.

[2]Twitter can be found at https://twitter.com, accessed 2022-04-14.

There is still work to be done, e.g., Spoor et al. [10] see it as future work to deviate a useful metric that measures similarity. All in all, this special track has helped provide example cases and practical solutions for data science scenarios from the real world.

Acknowledgements

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center On-The-Fly Computing (SFB 901).

References

[1] J. Kersting and M. Geierhos, "Human Language Comprehension in Aspect Phrase Extraction with Importance Weighting," in *Natural Language Processing and Information Systems*, ser. LNCS, E. Métais, F. Meziane, H. Horacek, and E. Kapetanios, Eds., vol. 12801. Saarbrücken, Germany: Springer Nature, 2021, pp. 231–242.

[2] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, CA, USA: ACL, 2016, pp. 19–30.

[3] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "SemEval 2016 Task 5: Aspect Based Sentiment Analysis (ABSA-16) Annotation Guidelines," pp. 1–20, 2016.

[4] J. Kersting and M. Geierhos, "Towards Aspect Extraction and Classification for Opinion Mining with Deep Sequence Networks," in *Natural Language Processing in Artificial Intelligence – NLPinAI 2020*, ser. Studies in Computational Intelligence (SCI), R. Loukanova, Ed. Cham, Switzerland: Springer, 2021, vol. 939, pp. 163–189.

[5] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306456

[6] J. Kersting and F. S. Baeumer, "Implicit statements in healthcare reviews: A challenge for sentiment analysis," in *AI-DSRWA Special Track: Maturing Artificial Intelligence – Data Science for Real-World Applications at the Fourteenth International Conference on Pervasive Patterns and Applications, PATTERNS 2022*, J. Kersting, Ed. IARIA.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, 2020, pp. 8440–8451.

[8] F. S. Bäumer, S. Denisov, B. Sirvend, and J. Weber, "Tackling the "we have no data" challenge: Domain-specific machine translation in smes," in *AI-DSRWA Special Track: Maturing Artificial Intelligence – Data Science for Real-World Applications at the Fourteenth International Conference on Pervasive Patterns and Applications, PATTERNS 2022*, J. Kersting, Ed. IARIA.

[9] S. Denisov, F. S. Bäumer, and M. Geierhos, "Track me if you can: Insights into profile interlinking on social networks," in *AI-DSRWA Special Track: Maturing Artificial Intelligence – Data Science for Real-World Applications at the Fourteenth International Conference on Pervasive Patterns and Applications, PATTERNS 2022*, J. Kersting, Ed. IARIA.

[10] J. M. Spoor, J. Weber, S. Hagemann, and F. S. Baeumer, "Concept of an inference procedure for fault detection in production planning," in *AI-DSRWA Special Track: Maturing Artificial Intelligence – Data Science for Real-World Applications at the Fourteenth International Conference on Pervasive Patterns and Applications, PATTERNS 2022*, J. Kersting, Ed. IARIA.