Bachelor thesis, Institute of Computer Science, Freie Universität Berlin

Human-Centered Computing (HCC), AG NBI

<Titel der Arbeit>

– Exposé –

Tim Korjakow

tim.korjakow@campus.tu-berlin.de

Supervisor: Prof. Dr. C. Müller-Birn

Berlin, April 23, 2019

1 Motivation of the thesis

Research on Explainable Artificial Intelligence, often called XAI, is currently wildly distributed and characterized by several competing ideas and approaches from a vast number of fields including computer science, mathematics, social sciences and philosophy[**TK:** citation needed]. Research is mostly focused on explainability in computer vision since the inner workings of a model can directly be translated into a graphic representation. In contrast to that development stands the fact that most of humanity's knowledge is encoded in text and XAI algorithms and methods from computer vision most often cannot be directly applied to results of NLP algorithms. Therefore there is an urgent need to research these methods in the context of NLP.

In Project IKON, we are developing a data-driven application at a major natural history research institution in order to make potentials for knowledge transfer between research projects and society actionable. To this end, it features a data visualization of semantic relations between research projects supplemented by links to infrastructures (e.g., collections or labs) and knowledge transfer activities (e.g., workshops or lectures). To generate the semantic relations, we have developed a Topic Modelling Pipeline with a Singular Value Decomposition at its heart. It is thought that this method shares a limited amount of expressiveness with different other linear methods due to its purely linear nature [AHM+17]. Additionally, when considering how we can make the results of our pipeline more interpretable, we encountered significant doubts over the meaningfulness of approaches such as parameter manipulation or choosing between algorithms for dimensionality reduction [BMG18] in our use case.

In short, the fundamental challenge of interpretability in Project IKON is: which model can we use that is potentially more interpretable, and how precisely can what aspect be made interpretable for humans in order to support the context of identifying potentials for knowledge transfer at the research institution? The latter illustrates a significant gap in the related work, as contextuality (both considering the way in which the output of a machine learning algorithm is operationalized in an algorithmic system as well as the situated context of use) is a sorely neglected aspect of interpretability [Mil17].

2 Thematische Einordnung der Arbeit

- Welche Artikel/Literatur sind/ist relevant für diese Arbeit?
- Bitte geben Sie die relevanten Inhalte der Artikel kurz wieder.
- Das Ausarbeiten von ausgewählter Literatur bzw. verwandten Arbeiten hilft Ihnen, Ihre Ziele im folgenden Abschnitt zu definieren. Daher ist eine Auseinandersetzung mit der Literatur von Beginn an notwendig, wenn es zu diesem Zeitpunkt noch nicht erschöpfend sein muss.

[**TK:** Same as before?] As algorithmic systems increasingly regulate, evaluate and adapt to individual as well as geopolitical human activity, there is a particular ethical

urgency to make their operation and results interpretable for people with different backgrounds (ML experts, non-technical experts etc.). This has been especially apparent in recent years, with a growing discourse on algorithmic influence in political events and lawmakers seeking a right to explanation [8] for people using algorithmic systems. As a result, interpretability has become a frequent if ill-defined [Lip16] objective for research and development in machine learning (ML) algorithms. For example, ML algorithms have been showcased as 'interpretable' in the form of network graphs of neural network layers [9], as well as feature visualisations [7] or textual explanations [3] in image classification.

3 Goal setting

This thesis tries to answer the following two questions:

- Which techniques to enhance interpretability of models are out there and how are they characterized?
- How can the existing topic modeling pipeline be augmented or changed in order to enhance interpretability?
- Does the usage of these methods result in an enhanced understanding on the end users side? [**TK:** Needs to be discussed!]

4 Procedure and methods

In order to gain a reproducible overview over the status of XAI research in the field of NLP a literature mapping study according to Petersen et al. [PFMM] is going to be conducted. This should result in a number of papers which are, according to the process, good representatives of the literature base and therefore also of current research efforts.

These papers are going to be analyzed quantitatively and qualitatively for occurring XAI methods and categorized according to generally accepted criteria e.g. Miller's "Properties of Interpretable Models" [Lip16] or Robnik-Sikonja's criteria [RB18].

Based on that analysis a number of techniques are selected and going to be applied to one of the state-of-art NLP topic modelling techniques - e.g. Doc2Vec [LM14].

One of the most crucial parts of the existing topic modelling pipeline is the document embedding. Currently a simple information retrieval method - Tf-Idf - is used, but the main drawback of the technique is that it, as all Bag-of-Words method, completely disregards word order and semantic dependencies in texts. Since that may be a integral part especially in scientific literature a more complex model is needed which is able to capture this dependencies. In order to do this a short summary of the state of research of document embeddings is presented and an applicable model will be chosen.

This model will be augmented by interpretability techniques and their impact will be studied.

5 Technische Umsetzung

- Mit welchen softwaretechnischen Hilfsmitteln soll die Arbeit realisiert werden?
- Selbstverständlich können Sie an der Stelle noch nicht alles wissen, aber Sie sollen sich hier bereits einen guten Überblick verschaffen.

One of the main technical challenges and parts of the implementational work will be the augmentation of the current topic modelling pipeline by a document embedding technique. Since the performance of the model greatly depends on this step, it is crucial to have well learned vector representations of the document base. Currently there is a corpus of circa 114000 scientific documents available in order to train the model. If that is not enough to gain expressive document embeddings, one may include pretrained word embeddings via e.g. BERT to introduce external information into the model and enhance the semantic coherence of the learned embeddings. This path should be taken with cation since it is connected to an hardly determinable amount of complexity. Research in the field of transfer learning for document embeddings is still in its infancy.

In order to adhere to the current research and industry standards, the implementation of this thesis is going to be done in Python, more precisely Tensorflow. Based on the chosen model which is going to be augmented with explainability mechanisms further packages and technologies are going to be selected. Additionally to the results of the previous literature analysis further features for the decision will be the ease of implementation and applicability to the chosen model.

6 Time calculation

[**TK:** Needs to be done!]

- Wie ist der generelle Zeitplan der Arbeit?
- Sie sollten bereits wissen, wann Sie fertig sein wollen und von dort mit der Rückwärtsterminierung starten.
- Ihre Arbeit ist ein Projekt, daher planen Sie es auch wie eines. Nutzen Sie zur Visualisierung ein Gantt-Chart.

References

[AHM+17] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):e0181142, August 2017.

[BMG18] Jesse Benjamin, Claudia Müller-Birn, and Rony Ginosar. Transparency and the Mediation of Meaning in Algorithmic Systems. October 2018.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, June 2016.

[LM14] Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*, May 2014.

[Mil17] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*, June 2017.

[PFMM] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattson. Systematic Mapping Studies in Software Engineering.

[RB18] Marko Robnik-Sikonja and Marko Bohanec. Perturbation-Based Explanations of Prediction Models. pages 159–175. June 2018.