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 16, 2019

1 Motivation of the thesis

- In welchem Bereich/Themenfeld bewegt sich Ihre geplante Arbeit?
- Erläutern Sie kurz, in welchem Themenbereich Ihre Arbeit angesiedelt ist. Wo werden Sie einen Beitrag leisten?
- Nutzen Sie bei den Erläuterungen die Ihnen bereitgestellte Kurzaufgabenstellung.

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.

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

- Welche Ziele werden mit der Arbeit verfolgt? Und welche zentralen Fragen lassen sich daraus ableiten?

- Die Ziele sollten so spezifisch wie möglich sein. Das hilft Ihnen im Verlauf der Umsetzung zu prüfen, ob Sie Ihre Ziele erreichen konnten.

4 Procedure and methods

- Welche einzelnen Aktivitäten müssen umgesetzt werden, um die Fragen zu beantworten und das Ziel der Arbeit zu erreichen?

- Aus den Fragen (vorheriger Abschnitt) können Sie dann gut Aktivitäten ableiten, die Ihnen helfen, Ihre weitere Arbeit zu strukturieren.

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 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 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 are going to be applied to one of the state-of-art NLP topic modelling techniques - e.g. Doc2Vec [LM14].

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.

The implementation of this thesis is going to be done in Python since most research in machine learning is conducted in this language. Based on the chosen model which is going to be augmented with explainability mechanisms further packages and technologies are going to be selected. Features for the decision will be ease of implementation and applicability additionally to the results of the previous literature analysis.

6 Time calculation

- 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.

7 Wie geht es nach dem Exposé weiter?

Nachdem die Phase der Exposé-Erstellung abgeschlossen ist (das kann bis zu drei Iterationen dauern), können Sie mit der Erstellung der eigentlichen Abschlussarbeit beginnen. Bitte nutzen Sie die Inhalte des Exposés gleich als inhaltlichen Rahmen für die Arbeit (vor allem in Kapitel 1). Ihnen wird wieder eine LaTeXVorlage zur Verfügung gestellt. In dieser Vorlage finden Sie wieder viele Informationen und Hilfestellungen zur Erstellung der Arbeit. Sie sollten nun Ihre Arbeit anmelden. Das entsprechende Formular finden Sie auf den Institutsseiten (Link).Bitte bringen Sie das ausgefüllte Formular zu einer unserer Sitzungen mit. Ich unterschreibe es und leite es weiter. Nun sollten Sie auch bald darüber nachdenken, wer der Zweitgutachter Ihrer Arbeit sein könnte. Ich berate Sie dabei gern.

Viele weitere, nützliche Informationen finden Sie in der Prüfungsordnung Ihres Studiengangs. Bitte lesen Sie den Sie betreffenden Absatz im Anhang (**??**).

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.