<Bachelor-/Masterarbeit> am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC)

<Titel der Arbeit>

<Ihr Vor- und Nachname>

Matrikelnummer: <IhreMatrikelnummer>

<ihreemail@adresse.de>

Betreuerin und Erstgutachterin: Prof. Dr. C. Müller-Birn

Zweitgutachter: <Name des Zweitgutachters>

Berlin, <Datum>

Eidesstattliche Erklärung

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den June 23, 2019

<Name>

Abstract

<Please summarize your thesis in a brief but meaningful way (about one page). Include in your abstract the topic of this thesis, important contents, results of your research and an evaluation of your results.>

Zusammenfassung

<Hier sollten Sie eine kurze, aussagekräftige Zusammenfassung (ca. eine Seite) Ihrer Arbeit geben, welche das Thema der Arbeit, die wichtigsten Inhalte, die Arbeitsergebnisse und die Bewertung der Ergebnisse umfasst.>

Contents

List of Figures

List of Tables

Vorwort

Allgemeine Hinweise zur Erstellung einer Abschlussarbeit

- Beachten Sie, dass diese Vorlage für einen zweiseitigen Ausdruck angelegt wurde.
- Über die Papierqualität können Sie entscheiden, aber wir empfehlen aber Seiten mit wichtigen, farbigen Grafiken auch in Farbe auszudrucken und dabei ein höherwertiges Papier zu verwenden.
- Bitte stimmen Sie mit dem Betreuer Ihrer Arbeit auch den Zweitgutachter ab. Die Anfrage des Zweitgutachters erfolgt von Ihnen. Es ist an dieser Stelle sinnvoll, die Anfrage mit einer kurzen Zusammenfassung der Arbeit zu stellen.
- Bitte beachten Sie, dass Sie Ihre Abschlussarbeit mit einer Klebebindung versehen, eine Ringbindung ist nicht erwünscht.

1 Einleitung

Im folgenden werden Ihnen Hinweise zur Strukturierung und zum Inhalt des ersten Kapitels gegeben.

1.1 Thema und Kontext

- Wo setze ich an? (Problemstellung / Ausgangslage)
- Identifikation der signifikanten Problemen im betrachteten Forschungsbereich
- Ein kurzer Überblick über den aktuellen Forschungsstand in dem Bereich inklusive vorhandener Lösungen (ausführlicher dann in den Folgeabschnitten)

1.2 Zielsetzung der Arbeit

- Was sind die mit dieser Arbeit verfolgten Ziele? Welches Problem soll gelöst werden?
- Eine Beschreibung der ersten Ideen, der vorgeschlagene Ansatz und die aktuell erreichten Resultate
- Eine Beschreibung, welchen Beitrag die Arbeit leistet, um das vorgestellte Problem zu lösen
- Eine Diskussion, wie die vorgeschlagene Lösung sich von bestehenden unterscheidet, was ist neu oder besser?

1.3 Vorgehen bei der Umsetzung

- Wie will ich meine Ziele erreichen? (Methodische Überlegungen)
- Darstellung zum Forschungsdesign.
- Insbesondere bei Master: Wie kann die Zielerreichung "gemessen" werden?

1.4 Aufbau der Arbeit

- Welche Schritte werden durchlaufen, um die Ziele zu erreichen?
- An dieser Stelle ist beispielsweise eine Grafik hilfreich, um den Aufbau der Arbeit und welche Ergebnisse/Erkenntnisse wo genutzt werden, zu visualisieren.
- Ebenfalls sollten noch Anmerkungen zur Gestaltung der Arbeit gegeben werden, vor allem, da in vielen deutschen Arbeiten englische Fachbegriffe verwendet werden. Ein solcher Text könnte folgendermaßen lauten:
  - "Abschließend sind hier noch eine Anmerkungen zur Gestaltung der vorliegenden Arbeit. Für die im Folgenden verwendeten personenbezogene Ausdrücke wurde, um die Lesbarkeit der Arbeit zu erhöhen, die männliche Schreibweise gewählt. Des Weiteren werden eine Reihe von englischen Bezeichnungen verwendet, um einerseits dem interessierten Leser das Studium der häufig vorliegenden englischen Originalliteratur zu erleichtern oder andererseits bestehende Fachbegriffe nicht durch die Übersetzung zu verfälschen. Diese Begriffe sind vom herkömmlichen Text in kursiver Schrift unterschieden."

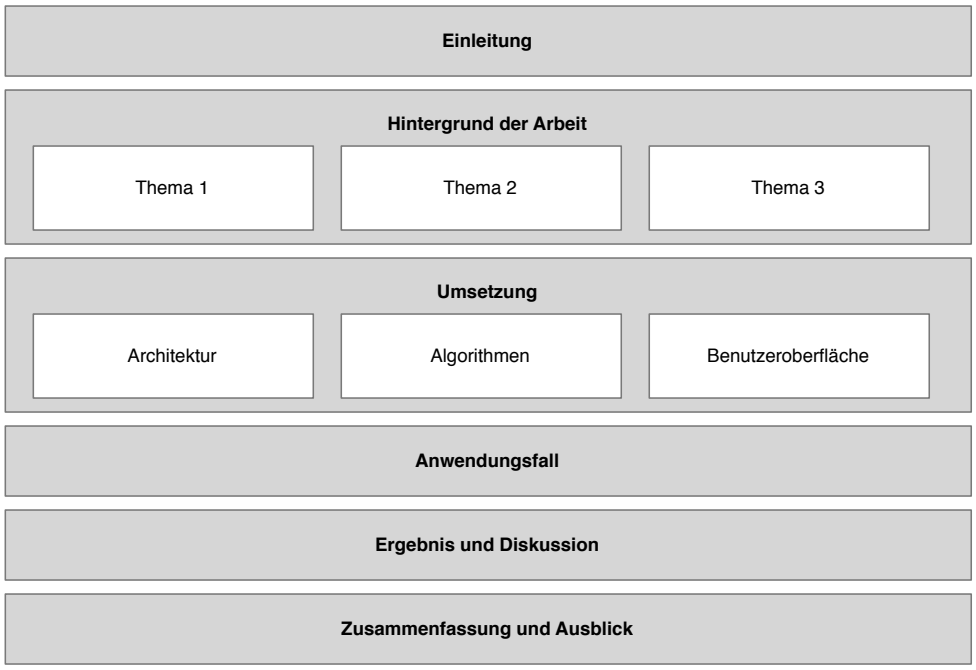

Figure 1.1: Beispiel einer möglichen Darstellung zum Aufbau der Arbeit (vgl. Beschreibung Abschnitt 3).

2 Introduction

2.1 Interpretability

2.2 Project IKON

3 Kapitel

- Abhängig vom Ziel der Arbeit und dem verwendeten Forschungsdesign unterscheidet sich dieser Hauptteil der Arbeit erheblich.

- Eine sehr allgemeine Struktur ist die folgende:
  - Hintergrund der Arbeit (Theoretische Einordnung der Arbeit)
    - Hier sollte enthalten sein, welche Anwendungen in diesem Bereich bereits existieren und warum bei diesen ein Defizit besteht.
    - Falls genutzt, sollten hier die entsprechenden Algorithmen erläutert werden.
    - Es sollten die Ziele der Anwendungsentwicklung, d.h. die Anforderungen herausgearbeitet werden. Dabei sollte die bestehende Literatur geeignet integriert werden.
  - Umsetzung (Praktischer Anteil der Arbeit)
    - Zunächst sollte die Softwarearchitektur und die genutzten Anwendungen, APIs etc. erläutert werden. Ebenfalls gehört dazu das Datenbankschema.
    - Es sollten die zentralen Elemente der Software (abhängig von der Aufgabenstellung) beschrieben werden, wie implementierte Algorithmen oder das Oberflächendesign.
    - Zentraler Quellcode sollte entsprechend aufgelistet werden:

```
1  public class Main {
2      public static void main(String[] args) {
3          System.out.println("Hello World!");
4      }
5  }
```

  - Evaluation (zumeist nur für Masterarbeiten relevant)
    - Jede Software muss auch getestet werden. Dieses Tests werden entweder mit einem vorgegebenen Datensatz erfolgen oder aber die Evaluation erfolgt auf Basis von Experimenten. In diesem Kapitel sollte daher entweder der genutzte Datensatz oder der experimentelle Aufbau beschrieben werden.
  - Ergebnis und Diskussion

* Die Ergebnisse der Anwendung werden in diesem Kapitel vorgestellt und anschließend diskutiert. Wenn möglich sollte die Ergebnisse in Relation zu bestehenden Arbeiten in dem Bereich erörtert werden.

4 Literature mapping study

4.1 Motivation

4.2 Methodology

In order to generate a reproducible and current overview over the fast-moving field of interpretability research in machine learning a rigorous methodology by Peterson et al. [**?**] is used.

The recommended process is augmented by further steps in order to tailor it to the existing use case and consists of the following seven procedures:

1. Definition of research questions: The overall process starts by defining clear questions which should guide the development of the whole literature study and subsequently the result as well. Since I am interested in the gaining an overview over the existing interpretability techniques, I chose the following questions:

   a) What kind of explainability techniques are mentioned in the corpus?

   b) In which domain and for which applications are they most commonly used?

   c) Which techniques are applicable to results produced by the pipeline or the pipeline itself? [**TK:** Is that right here already?]

2. Construction of a search string:

   Based on the questions one is able to gather a set of key words which are most relevant to the field which is analyzed. Each word is augmented by synonyms which are concatenated with boolean OR operators and several of these synonymous groups are again connected via logical ANDs. Applying this method to the previously found questions yields the following search string:

   ("explainability" OR "explainable" OR "explanation" OR "explaining" OR "interpretability" OR "interpretable" OR "interpretation" OR "interpret" OR "understanding") AND ("machine learning" OR "neural network" OR "neural networks" OR "AI" OR "XAI" OR "artificial intelligence" OR "model") AND ("text" OR "document" OR "NLP" OR "natural language programming" OR "review" OR "method" OR "technique" OR "visualization")

3. Analysis of the main publishers using a meta search and the search string:

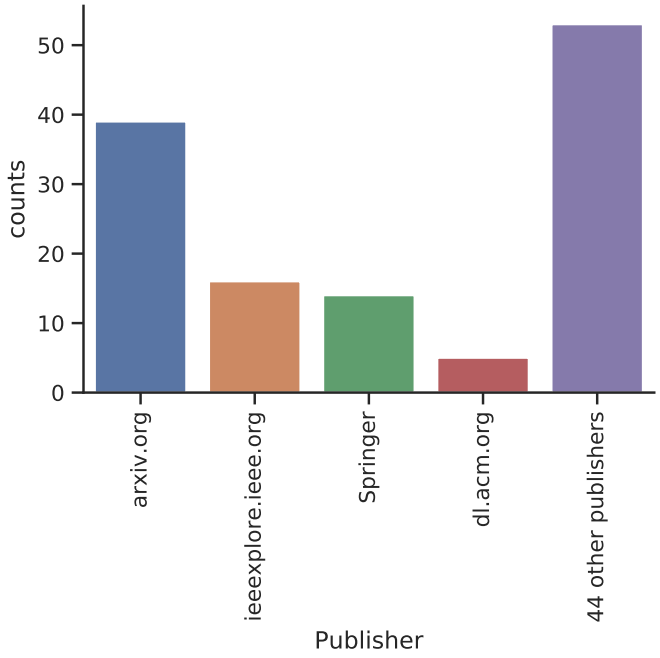

Figure 4.1: Barplot displaying the distribution of publishers occurring in the meta search results

Due to the presumed distributed nature of interpretability research it is not easy to pinpoint the main publishers of scientific articles. In order to mitigate this a pre-search in the meta-search engine 'Google Scholar' is conducted. It should be noted at this point that any biases which are apparent in the meta search engine therefore apply to this analysis as well. One can see in Figure 4.1 that the main publishers are respectivly Arxiv, IEEE, Springer and ACM. Since all of these publishers are mainly focused on publications in computer science, mathematics and engineering, this speaks in favor of the hypothesis that most of the research is still very technical and research from social sciences rarely influences it. Even though Arxiv is not a credible publisher per se, it seems like the research community uses it as the first place to publish ones work and therefore it should not be excluded in this analysis.

4. Sourcing of publications in scientific databases:

   Based on the insights from the previous step each of the main publisher's databases is scraped using the search string. Since most searches result in more than 1000 publications only the top 100 results ordered by the relevance scoring of the database are taken into account. These publications then form the corpus which is the basis for further analysis.

5. Filtering of these publications by keywording their abstracts

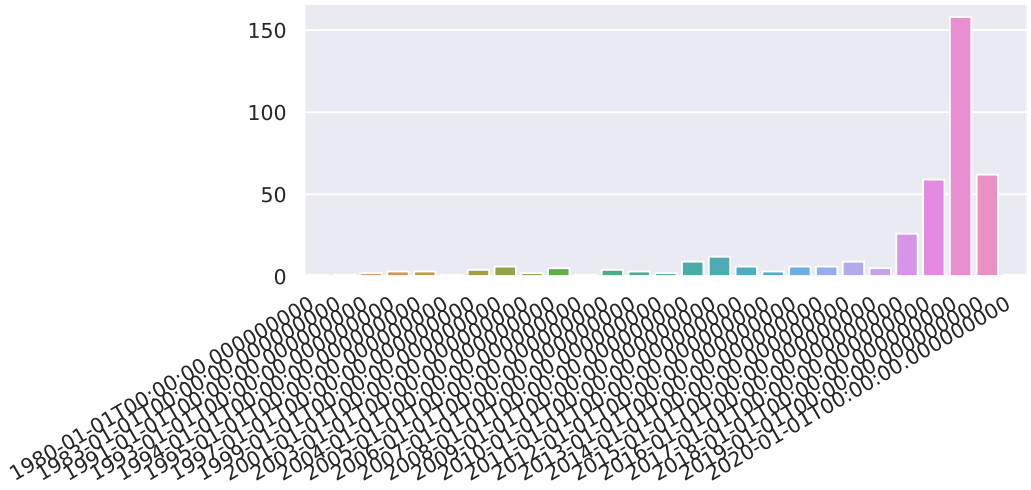

Figure 4.2: Barplot displaying the distribution of publishers occurring in the meta search results

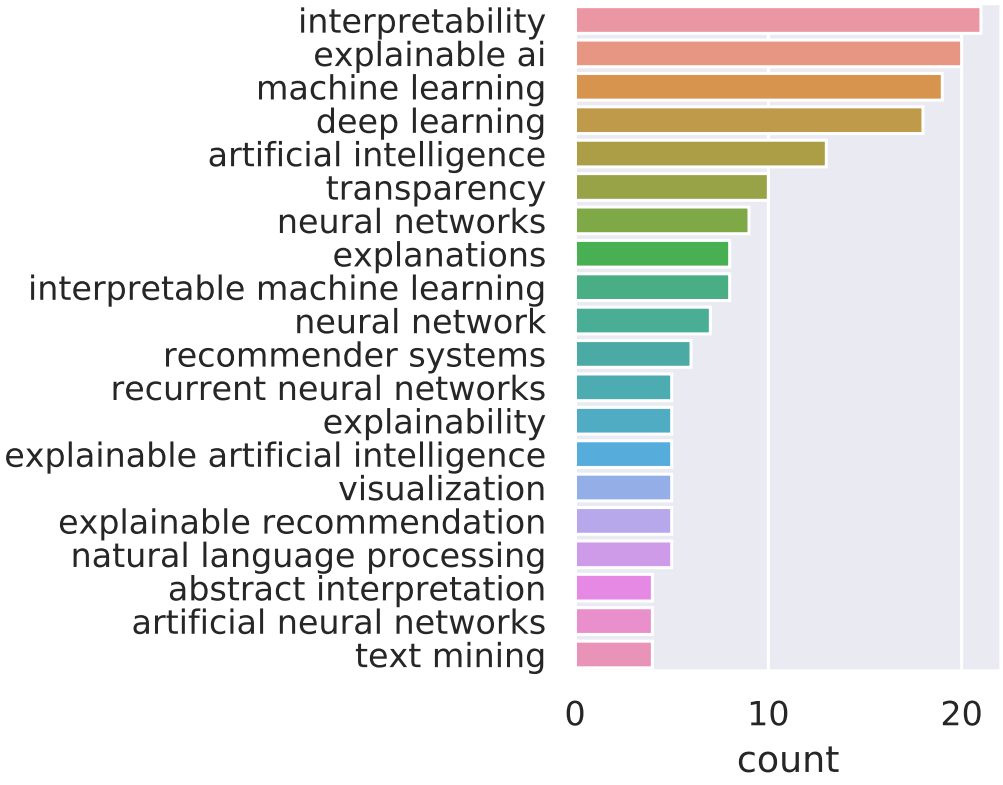

Figure 4.3: List of the 20 most used tags and their absolute frequency

Most scientific databases do a full text search on publications and possibly find the supplied keywords from the search string in sections of the paper which are not relevant e.g. the bibliography or in the outlook. Therefore another filtering step is necessary which searches for the search string in the abstracts of the papers of the corpus.

In order to enhance the quality of the filtering process, the search string is enhances with key words generated by an analysis of the current corpus. As seen in Figure 4.3 the previous search string already contains most of the relevant keywords.

6. Definition and application of inclusion and exclusion criteria to narrow down the pool of publications further

Inclusion criteria	Exclusion criteria
• Reviews the current state of explainability research • Presents a specific method for enhancing explainability for models	• Is not scientific literature • Does not describe the used explainability method • The publication does not focus on explainability • The described method is neither general, nor focused on NLP

gg

7. Quantitative and qualitative assessment of the resulting corpus

   hghbhbhb

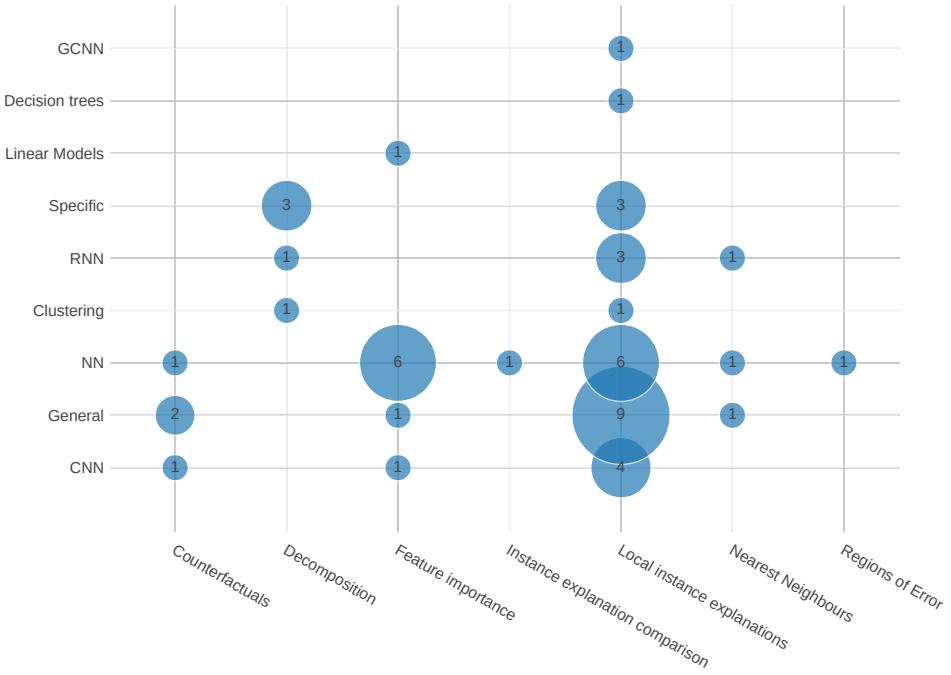

Figure 4.4: Mapping of applicability and gamut classification

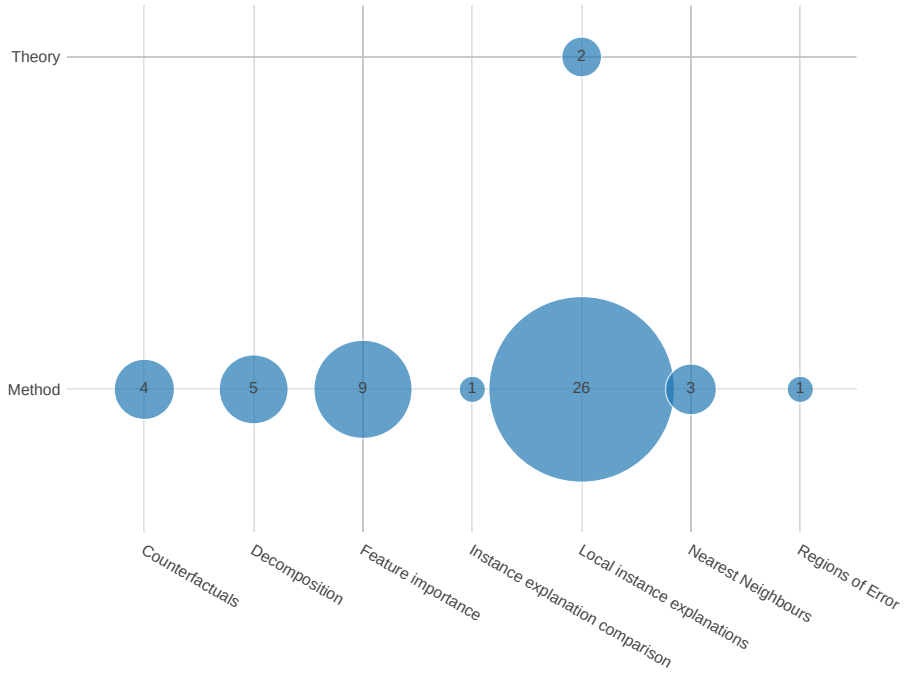

Figure 4.5: Mapping of applicability and gamut classification

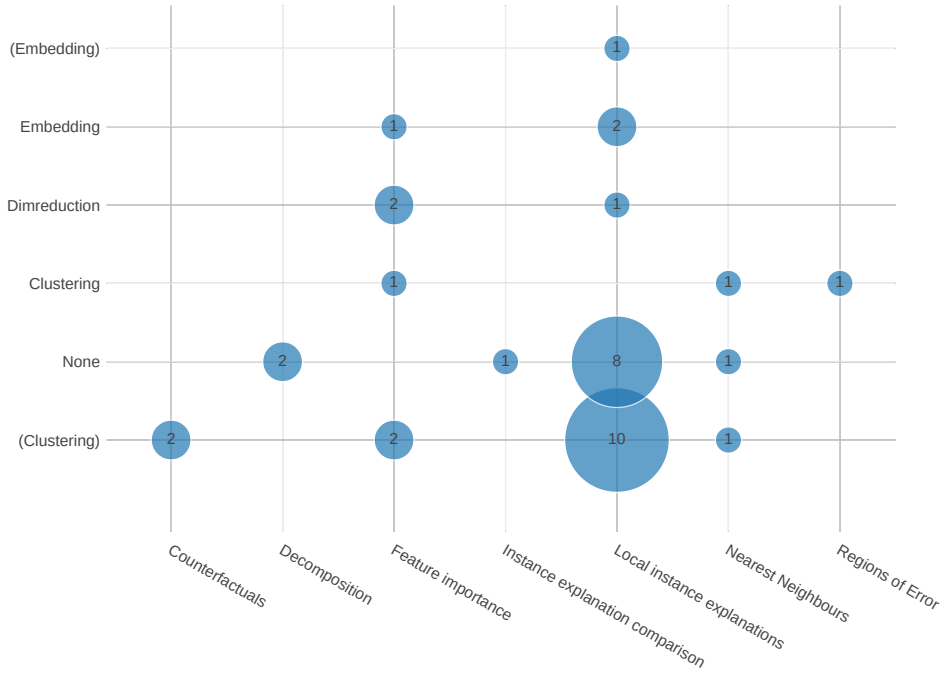

Figure 4.6: Mapping of applicability and gamut classification

5 Zusammenfassung und Ausblick

- Die Zusammenfassung sollte das Ziel der Arbeit und die zentralen Ergebnisse beschreiben. Des Weiteren sollten auch bestehende Probleme bei der Arbeit aufgezählt werden und Vorschläge herausgearbeitet werden, die helfen, diese Probleme zukünftig zu umgehen. Mögliche Erweiterungen für die umgesetzte Anwendung sollten hier auch beschrieben werden.

Bibliography