Human-centered Interactions with Text Classifiers: Fusing Concept-based Knowledge with Local Surrogate Explanation Models

Sebastian Manuel Kiefer né Bruckert

Submitted in partial fulfilment of the requirements for the degree of Doctor of Natural Sciences (Dr. rer. nat.) of the University of Bamberg

Committee Members:

Prof. Dr. Ute Schmid (Advisor & Reviewer)
Prof. Dr. Andreas Henrich (Reviewer)
Prof. Dr. Oliver Posegga (Head of Committee)

Submitted: 24.05.2023

Day of Defense: 06.07.2023

URN: urn:nbn:de:bvb:473-irb-897715
DOI: https://doi.org/10.20378/irbo-89771

Abstract

Human interactions often take place with the aim of exchanging understanding between individuals. Thus, there is a human need to develop and communicate explanations on the one hand and to receive explanations on the other hand in order to further develop one's own understanding. Explanations are an attempt to trace events back to their causes, for example, to provide answers to the question of "why".

The concept of bidirectional explanations can also be applied to the interaction between humans and machines, for example, in the context of machine learning (ML), which is a sub-area of artificial intelligence (AI). Whenever machines are used to support human decision-making or even to provide recommendations for action, there is a need to be able to understand how the results were obtained and to influence them if necessary. Human decision-makers should develop trust in machine learners, especially in high-stakes application domains, like in medical diagnostics, but also in application domains that are strongly influenced by regulations, like the domain of financial auditing. The comprehension of concrete decisions and the incorporation of expert knowledge can contribute to the development of trust.

Modern machine learning methods have recently been improved in terms of their predictive accuracy, which has led to them already surpassing human performance in some tasks. On the other hand, such powerful methods are usually so-called *black-box*-methods, what makes it difficult or even impossible for humans to develop an understanding of the general decision logic or the specific model behavior. Lack of comprehensibility not only impairs the formation of trust in machine companions, but also impairs the ability to interact in the sense of correctability. In comparison to global explanations that explain *how* a system works, local explanations that provide information about correlations of specific inputs and outputs and thus justify *why* a particular output was produced are often considered more reliable and purposeful. This fact especially holds true for non-expert AI users.

Researchers in the field of explainable artificial intelligence have developed approaches that enable post hoc model-agnostic and mostly local explanations of supervised machine learners. The goal is to explain individual ML results retrospectively, i.e., after predictions have been made for previously unobserved instances, and independently of the ML model used, i.e., without knowledge about its inner workings. Especially in the case of surrogate explanation models that locally approximate the model to be explained, there is a risk that, when contextual information, such as dependencies between features, is neglected, this will result in data points that are described as "out-of-distribution". Explanations based on such unrealistic data points can easily be misinterpreted. Furthermore, explanation approaches are often developed exclusively from a technical point of view. Insights from psychology or the social sciences, according to which explanations that are understandable for humans should, for example, have a coherent structure, often are not taken into account.

In this dissertation, a new approach is proposed and validated that enables human-centered interactivity with text classifiers by using bidirectional model-agnostic explanations. To this end, a framework is introduced that defines comprehensible and interactive artificial intelligence in an interdisciplinary manner using cognitive concepts such as explainability, interpretability, transparency, and interactivity. It elaborates that semantic and contextual information available within a certain textual application domain should be taken into account during the generation and representation of explanations such that contextual explanations result. In order to fill this identified research gap and obtain coherent explanations, a new technique is presented that generates model-agnostic concept-based explanations whose explanatory features consist of semantically related words. The consideration of context enables more realistic and meaningful local perturbation distributions, which form the basis of many model-agnostic and local explanation approaches, such as LIME. A technical evaluation of the new explanation methodology called *topicLIME*, in particular of the underlying surrogate models and the resulting explanations, analyzes its local fidelity related to the text classifier to be explained. Besides a technical evaluation, the obtained results are empirically analyzed by means of two user studies. It is investigated how concept-based contextual explanations in comparison to contextless explanations are perceived by humans interacting with a text classifier that predicts to which content category a presented document belongs.

As another contribution of this dissertation, a method is presented that extends state-of-the-art for explanatory interactive learning, especially for text classification. The method intends to enable humans to engage in broader interactions, such as correcting predictions and explanations in a constructive and contextual manner. Previous model-agnostic methods for explanatory interactive learning, such as the CAIPI approach, only allow correcting a classifier regarding its functioning via explanations in case a correct prediction has been made for the wrong reasons. The newly developed Semantic Push approach qualifies humans to perform corrections via concept-based explanations and integrate them into the learner using non-extrapolating training documents, across different error types of a classifier. The evaluation of the approach shows that the newly developed method outperforms the baseline CAIPI method in terms of learning performance and local explanation quality.

In summary, the fusion of concept-based knowledge with local surrogate explanation models is a promising research direction. The results of this dissertation further show that more interdisciplinary research is needed to address the challenges in the field of human-centered machine learning.

Zusammenfassung

Menschliche Interaktionen finden häufig mit dem Ziel statt, ein Verständnis zwischen Individuen auszutauschen. Es besteht somit ein menschliches Bedürfnis, einerseits Erklärungen zu erarbeiten und zu kommunizieren, andererseits Erklärungen entgegenzunehmen, um eigenes Verständnis weiterzuentwickeln. Bei Erklärungen handelt es sich um einen Versuch, Vorgänge auf ihre Ursachen zurückzuführen, also Antworten auf die Frage nach dem "Weswegen" zu liefern.

Das Konzept von gegenseitigen Erklärungen lässt sich auch auf die Interaktion von Mensch und Maschine, zum Beispiel im Rahmen des maschinellen Lernens (ML), welches einen Teilbereich der künstlichen Intelligenz (KI) darstellt, übertragen. Überall dort, wo maschinelle Verfahren herangezogen werden, um Menschen Entscheidungsunterstützung zu leisten oder gar Handlungsempfehlungen zur Verfügung zu stellen, besteht die Notwendigkeit, das Zustandekommen der Ergebnisse nachvollziehen und bei Bedarf beeinflussen zu können. Insbesondere bei Entscheidungen von hoher Tragweite, beispielsweise in der medizinischen Diagnostik, aber auch bei Entscheidungen innerhalb von Anwendungsdomänen, die stark regulatorisch beeinflusst sind, wie beispielsweise die Domäne der jährlichen Wirtschaftsprüfung, sollten menschliche Entscheidungsträger Vertrauen in maschinelle Verfahren entwickeln. Das Nachvollziehen konkreter Entscheidungen sowie das Einbringen von Expertenwissen kann dabei zur Vertrauensbildung beitragen.

Moderne maschinelle Lernverfahren wurden in jüngster Vergangenheit stark in Bezug auf ihre Vorhersagegenauigkeit verbessert, was dazu geführt hat, dass sie bei einigen Aufgaben menschliche Leistungsfähigkeit bereits übertreffen. Auf der anderen Seite handelt es sich bei solchen leistungsstarken Verfahren in der Regel um sogenannte "Black-Box"-Verfahren, die es Menschen erschweren oder gar unmöglich machen, ein Verständnis der generellen Entscheidungslogik oder des spezifischen Modellverhaltens zu entwickeln. Mangelnde Nachvollziehbarkeit beeinträchtigt nicht nur die Bildung von Vertrauen in maschinelle Partner, sondern beeinträchtigt auch die Interaktionsfähigkeit im Sinne einer Korrigierbarkeit. Inbesondere lokale Erklärungen, die Informationen über Korrelationen von Ein- und Ausgaben bieten und damit rechtfertigen, *warum* eine bestimmte Ausgabe produziert wurde, werden im Vergleich zu globalen Erklärungen, die darlegen, *wie* ein System funktioniert, vor allem von KI-Laien als zuverlässiger und zielführender erachtet.

Forscher aus dem Bereich der erklärbaren künstlichen Intelligenz haben Ansätze entwickelt, die es ermöglichen, post hoc modell-agnostische und meist lokale Erklärungen von überwachten maschinellen Lernverahren zu erstellen. Zielsetzung ist es, individuelle ML-Ergebnisse nachträglich, also nachdem Vorhersagen für bisher unbeobachtete Instanzen getätigt wurden, und unabhängig vom eingesetzten ML-Modell, also ohne Wissen über dessen innere Funktionsweise, zu erklären. Insbesondere bei Surrogat-Erklärungsmodellen, die das zu erklärende Modell lokal approximieren, besteht die Gefahr, dass bei der Erstellung einer lokalen Nachbarschaft ohne Berücksichtigung kontextueller Informationen, wie

beispielweise Abhängigkeiten zwischen Merkmalen, Datenpunkte resultieren, die als "out-of-distribution" beschrieben werden können. Erklärungen, die auf solchen unrealistischen Datenpunkten basieren, können leicht fehlinterpretiert werden. Weiterhin werden Erklärungsansätze häufig ausschließlich unter technischen Gesichtspunkten entwickelt. Erkenntnisse aus der Psychologie oder aus den Sozialwissenschaften, wonach für Menschen verständliche Erklärungen beispielsweise eine kohärente Struktur aufweisen sollten, finden häufig keine Berücksichtigung.

In dieser Dissertation wird ein neuer Ansatz vorgeschlagen und validiert, der mensch-zentrierte Interaktivität mit Textklassifikatoren durch die Verwendung von bidirektionalen modell-agnostischen Erklärungen ermöglicht. Dafür wird zunächst ein Framework eingeführt, das nachvollziehbare und interaktive künstliche Intelligenz interdisziplinär unter Verwendung kognitiver Konzepte wie Erklärbarkeit, Interpretierbarkeit, Transparenz und Interaktivität definiert. Es wird erarbeitet, dass semantische und kontextbezogene Informationen, die innerhalb einer gewissen textuellen Anwendungsdomäne verfügbar sind, während der Generierung und Darstellung von Erklärungen berücksichtigt werden sollten, sodass kontextbezogene Erklärungen resultieren. Um diese identifizierte Forschungslücke zu schließen und kohärente Erklärungen zu erhalten, wird eine neue Technik vorgestellt, die modell-agnostisch konzept-basierte Erklärungen generiert, deren Erklärungsbestandteile aus semantisch zusammenhängenden Worten bestehen. Durch die damit verbundene Berücksichtigung von Kontext werden realistischere und bedeutungsvollere lokale Perturbationsverteilungen ermöglicht, die die Grundlage vieler modell-agnostischer und lokaler Erklärungsansätze wie LIME darstellen. Eine technische Evaluation der neuen Erklärungsmethodik namens topicLIME, insbesondere der zugrundeliegenden Surrogatmodelle sowie der resultierenden Erklärungen, analysiert unter anderem die lokale Wiedergabetreue bezogen auf den zu erklärenden Textklassifikator. Neben einer technischen Evaluation werden die erzielten Ergebnisse empirisch anhand zweier Nutzerstudien analysiert. Dabei wird untersucht, wie konzept-basierte kontextuelle Erklärungen im Vergleich mit kontextlosen Erklärungen von Menschen wahrgenommen werden, die mit einem Textklassifikator interagieren, welcher vorhersagt, zu welcher Inhaltskategorie ein präsentiertes Dokument gehört.

Als weiterer Beitrag dieser Dissertation wird eine Methode vorgestellt, die den aktuellen Stand der Technik für erklärendes interaktives Lernen, insbesondere für die Textklassifizierung, erweitert. Die Methode beabsichtigt, Menschen umfassendere Interaktionen zu ermöglichen, beispielsweise die Korrektur von Vorhersagen und Erklärungen auf konstruktive und kontextbezogene Weise. Bisherige modell-agnostische Verfahren für erklärendes interaktives Lernen wie der CAIPI-Ansatz erlauben lediglich, einen Klassifikator hinsichtlich dessen Funktionsweise über Erklärungen zu korrigieren, wenn eine korrekte Vorhersage getätigt wurde, jedoch aus den falschen Gründen. Der im Rahmen dieser Arbeit neu entwickelte Ansatz Semantic Push qualifiziert Menschen dazu, auf Basis lokalgetreuer und konzept-basierter Erklärungen Korrekturen vorzunehmen und diese mit Hilfe nicht-extrapolierender Trainingsdokumente in einen Textklassifikator zu integrie-

ren, und zwar über unterschiedliche Fehlerarten eines Klassifikators hinweg. Die Evaluation des Ansatzes zeigt, dass die neu entwickelte Methode die Basismethode CAIPI in Bezug auf Lernperformanz sowie lokale Erklärungsqualität übertrifft.

Zusammenfassend lässt sich festhalten, dass die Verschmelzung konzeptbasierten Wissens mit lokalen Surrogat-Erklärungsmodellen eine vielversprechende zukünftige Forschungsrichtung darstellt. Die Ergebnisse dieser Dissertation zeigen weiterhin, dass verstärkt interdisziplinäre Forschung erforderlich ist, um die Herausforderungen auf dem Gebiet des mensch-zentrierten maschinellen Lernens anzugehen.

Publications and Further Scientific Contributions

I. The following publications contain some of the main scientific contributions elaborated during this doctoral research:

- **Sebastian Kiefer (né Bruckert)**, Bettina Finzel, and Ute Schmid. "The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions." In *frontiers in Artificial Intelligence, 3* (2020). DOI: 10.3389/frai.2020.507973.
- **Sebastian Kiefer**. "CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge." In *Information Fusion, 77* (2022), pp. 184-195. DOI: 10.1016/j.inffus.2021.07.014.
- **Sebastian Kiefer** and Günter Pesch. "Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations." In *Edelkamp, S., Möller, R., Rueckert, E. (eds) KI 2021: Advances in Artificial Intelligence. KI 2021. Lecture Notes in Computer Science()*, vol 12873. Springer, Cham. DOI: 10.1007/978-3-030-87626-5_22.
- **Sebastian Kiefer**, Mareike Hoffmann, and Ute Schmid. "Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions." In *Machine Learning and Knowledge Extraction, 4* (2022), pp. 994-1010. DOI: 10.3390/make4040050.

The written scientific contents contributed by me are listed in detail in the appendix A.

II. The following empirical evaluation that is part of the synopsis is planned for publication:

- **Sebastian Kiefer**, Alisa Münsterberg, and Ute Schmid. "CaSE-Study: An Empirical Evaluation of Topic-based Explanations". (2022).

III. The following published book chapters also contain some of the ideas that emerged during this doctoral research:

- Ute Schmid and **Sebastian Kiefer (né Bruckert)**. "Künstliche Intelligenz in Unternehmen - Zielgruppenspezifische KI-Kompetenzen identifizieren und vermitteln". In *Philipp Ramin (Ed.). (2021). Handbuch Digitale Kompetenzentwicklung: Wie sich Unternehmen auf die digitale Zukunft vorbereiten.* Carl Hanser Verlag GmbH Co KG.
- Ute Schmid and **Sebastian Kiefer (né Bruckert)**. "Artificial Intelligence in Business and Industry - Identifying and Imparting Competencies for Different Target Groups". In *Philipp Ramin (Ed.). (2022). Digital Competence and Future Skills: How companies prepare themselves for the digital future.* Carl Hanser Verlag GmbH Co KG.

IV. The following bachelor's and master's theses that I advised and the scientific talks that I held contributed to some of the results achieved during this doctoral research:

- **Sebastian Kiefer (né Bruckert)**, Gregor Fischer, and Jeffrey Ahmad. "Explainable Artificial Intelligence". DigiCamp 2019. DATEV eG.
- Jonas Amling. "Explaining Text Classification Decisions with LIME – Topic-based versus Random Input Manipulation". Bachelor's Thesis. 2020.
- Mareike Hoffman. "Explanatory interactive machine learning for text classification – A model agnostic approach for semantic corrections". Master's Thesis. 2022.
- Günter Pesch and **Sebastian Kiefer**. "Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations". 17. Deggendorfer Forum für digitale Datenanalyse. 2023.

Acknowledgment

This dissertation would not have been possible without the encouragement and support of some special people I was privileged to learn from.

First and foremost, I would like to thank my adviser Prof. Dr. Ute Schmid for her constant guidance, patience, and personal support. Thanks for accepting me as an external doctoral candidate, for drawing my attention to DATEV eG and inspiring my professional career.

Many thanks also go to the other members of the committee Prof. Dr. Andreas Henrich and Prof. Dr. Oliver Posegga, who both supported me during my exploration of research directions.

Also, thank you Alisa Münsterberg for your continuous support during my empirical research and for all the inspiring discussions.

Additionally, I would like to express my sincere gratitude to DATEV eG, especially to my colleagues Dr. Thilo Edinger and Gregor Fischer, who made it possible for me to do my doctoral research in cooperation with DATEV eG.

I am also very grateful to my parents who enabled me a life with freedom of choice. I extend my gratitude to my best friends Philip and Max who continuously motivated me during this research journey. Furthermore, thank you Mathias and Amira for making me smile even in the early morning. Last but not least, a very special thanks, with all my heart, to Silvana, who always stood by my side, encouraged me to continue, and patiently covered my ass during my research.

Table of Contents

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence.

CAI Comprehensible Artificial Intelligence.

CIAI Comprehensible and Interactive Artificial Intelligence.

GDPR General Data Protection Regulation.

HCI Human-Computer Interaction.

HCML Human-centered Machine Learning.

LDA Latent Dirichlet Allocation.

LIME Local Interpretable Model-agnostic Explanations.

ML Machine Learning.

NLP Natural Language Processing.

NPMI Normalized Pointwise Mutual Information.

SHAP SHapley Additive exPlanations.

XAI Explainable Artificial Intelligence.

Part I

Synopsis of Thesis

1. Introduction

> *"The computer is incredibly fast, accurate, and stupid. Man is incredibly slow, inaccurate, and brilliant. The marriage of the two is a force beyond calculation."*
>
> – Presumably by Leo M. Cherne (1912–1999)

1.1 Motivation and Background

As social beings, humans - both experts and laypersons - engage in interactions, often attempting to communicate an understanding between individuals. Therefore, humans are naturally driven to acquire and provide explanations, but also to receive explanations in order to expand their understanding (Keil 2006). An explanation is defined as the answer to a "why"-question (Dennet 1989; Overton 2011; David Lewis 1986; Lipton 1990; T. Miller 2019). Even young children start to ask "why"-questions early in their life to uncover the reasons of an observation. As human explanations are often framed by stances or modes of construal and are therefore interpretative and diverse in nature, humans need to perform mental calculations in order to understand such explanations (Dennet 1989). Often, the human capabilites to flexibly use contextual and background information and to use intuition and feeling are consulted to distinguish "brilliant" and "real" intelligence (Bergstein 2017) from Artificial Intelligence (AI), as computers generally are deemed "stupid" with regard to such tasks.

As a common characteristic, both human and artificial intelligence lack the ability to truthfully and introspectively analyze and explicitly communicate their inner reasoning (Nosek, Hawkins, and Frazier 2011). For example, humans are generally unaware of their implicit cognitions and can hardly describe their brain's conscious thoughts. Also the products of human thoughts cannot reveal the brain's inner workings despite pretending to do so (Kahneman and Klein 2009).

A similar inability applies to AI as well. Machine Learning (ML), which for many tasks manages to achieve or even exceed human performance with regard to prediction accuracy, mostly is not capable of providing an explicit knowledge representation and hides the underlying explanatory structure (Holzinger, Biemann, et al. 2017). Such approaches are therefore considered black-boxes as their inner approach stays unknown, concealing the way the results have been produced. This leads to ML models that cannot be interpreted easily by humans (Petch, Di, and Nelson 2022; Patel et al. 2008).

In contrast to human intelligence, AI systems shine in terms of signal speed, direct connectivity, updatability, scalability, and especially data processing and storage capacity (Bostrom 2014). All of this leads to very different qualities and limitations of human and artificial intelligence. Moravec's Paradox puts it in a nutshell by stating: "It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (Moravec 1988).

The research area Human-centered Machine Learning (HCML) (Gillies et al. 2016) was initiated to foster integrated research that lies at the intersection of human and artificial intelligence. It aims at linking the capabilities of human perception with the capacity and performance of computers. Striving for better consideration of human goals, contexts, and ways of working when designing ML algorithms and corresponding interfaces, HCML shall lead to more useful and usable artificially intelligent systems (Gillies et al. 2016).

HCML builds on concepts and techniques elaborated in Human-Computer Interaction (HCI). HCI in general deals with the creation of appropriate user interfaces that enable an efficient interaction between humans and computers. Its focus is especially on user interfaces which rely on human behaviors that can hardly be reasoned about explicitly, for example, training virtual video game characters to use body language when interacting with humams (Kleinsmith and Gillies 2013). Applying HCI methods to ML-related interaction tasks leads to interactive machine learning (Fails and Olsen 2003) where humans improve an ML model by adding individual training examples in an iterative manner until its quality is sufficient.

However, in order to allow users to co-work effectively and efficiently with ML systems, to make ML accessible to non-experts, and to take advantage of the richness of human expertise, ML results need to be interpretable and therefore transparent and comprehensible for humans (Patel et al. 2008; Akata et al. 2020). To achieve this kind of ex post understanding of an ML model's specific behavior, HCML further includes techniques elaborated within the research field of Explainable Artificial Intelligence (XAI). XAI strives for generating explanations for individual ML results providing justifications why a certain output was produced. Often, so-called *local surrogate explanation models*, which learn an interpretable model locally around a prediction to be explained (Marco Tulio Ribeiro, Singh, and Guestrin 2016), are used for the generation of explanations.

In summary, three major desiderata for a human-AI-partnership are formulated in the context of HCML. The first is a more general desideratum, the others feature human-centered interactivity in particular:

1. Humans shall benefit from machines' outstanding performance and accuracy. This requirement is often already met by progress in "classical" AI research, like using Convolutional Neural Networks (CNNs) to identify faces (Parkhi, Vedaldi, and Zisserman 2015), detect objects (Krizhevsky, Sutskever, and

Hinton 2017), or assist self-driving cars (Bojarski et al. 2016). Additionally, great success was achieved in the context of Natural Language Processing (NLP), which is a field of artificial intelligence and linguistics, devoted to make computers understand words or phrases written in human languages (Khurana et al. 2022). As an example, encoder-decoder-based neural machine translation systems were successfully applied to bilingual translation tasks (Sutskever, Vinyals, and Le 2014; Cho et al. 2014; Bahdanau, Cho, and Bengio 2014; Gehring et al. 2017; Vaswani et al. 2017). In many cases, such systems already exceed human performance.

2. As a basis for efficient interaction, machines shall be able to justify why they produced a certain prediction in a way that can be considered human-friendly (refer to section 3.3 for details on human-friendly textual explanations).

3. Machines shall allow the integration of conceptual human knowledge into their inner reasoning, for example, in case of wrong predictions.

Overall, this thesis aims at proceeding one step further towards HCML by combining techniques and insights from research disciplines like artificial intelligence, human-computer interaction (especially interactive machine learning), and psychology of explanation (especially explanatory understanding). It proposes a framework for Comprehensible and Interactive Artificial Intelligence (CIAI) and, in the following chapters, instantiates it mainly for the domain of text classification, which involves assigning text documents to a set of predefined classes using ML techniques (Dalal and Zaveri 2011).

To be more concrete, the focus of this thesis is to facilitate effective and efficient collaboration between humans and artificially intelligent text classifiers via bidirectional explanations used as an attempt to communicate an understanding or to enforce a desired reasoning. It is elaborated how fusing concept-based knowledge[1], which is inherent in text corpora or provided by human annotators, with local surrogate explanation models allows for faithful explanations that enable contextual interpretation of and intervention on a text classifier without knowing its internal structure. Technically, a local surrogate explanation model called *Local Interpretable Model-agnostic Explanations (LIME)* is combined with a topic modeling technique called *Latent Dirichlet Allocation (LDA)*. LDA belongs to the field of text mining, where the goal is to discover knowledge from text data, which are unstructured or semi-structured (Cai and Sun 2009).

[1] In this thesis, the term "concept-based knowledge" is used. It is derived from the term "conceptual knowledge", which has been defined as "understanding of the principles and relationships that underlie a domain" or "knowledge of concepts and relations which are fundamental in a certain domain" (Byrnes 1992; Crooks and Martha W Alibali 2014; Hiebert 1986; Rittle-Johnson, Siegler, and Martha Wagner Alibali 2001). Concepts technically are approximated by topics in the further course of this thesis.

1.2 Application Domain

This doctoral research in parts has been conducted in cooperation with the company DATEV eG. DATEV eG is a software company and IT service provider for tax consultants, auditors, lawyers, and their clients. The company mainly provides software solutions for the automation of processes in financial accounting that are often highly regulated. Offering more than 200 software products and IT services for more than 500.000 customers, DATEV eG is the third largest provider of business software in Germany and among the largest European IT service providers[2].

Some of the research results presented in this thesis have been achieved during work on projects of the company DATEV eG. For example, one publication of this thesis (Kiefer and Pesch 2021), which is contained in appendix A.2.2, presents work on unsupervised anomaly detection for financial auditing with model-agnostic explanations. Financial auditing refers to the process of an annual audit that covers all transactions of a client with all business partners in one fiscal year. Also for another publication (Kiefer 2022), which is attached in appendix A.2.1, some preliminary work has been done while working on a DATEV project that involves the semi-automatic classification of incoming invoices into defined ledger accounts. The task to categorize scans of invoices using supervised ML techniques can be redefined as a text classification task by representing incoming invoices as text documents by applying Optical Character Recognition (OCR).

As the work on DATEV-internal projects involves working with confidential data, the remainder of this thesis presents the achieved results harnessing public text classification datasets that are further detailed in the related publications (Kiefer 2022; Kiefer, Hoffmann, and Schmid 2022), both contained in appendix A.2.1 and A.3.1.

Text classification, also known as text categorization, is the most fundamental task in NLP (Li et al. 2022). Text classification comprises a variety of different tasks, like sentiment analysis, topic labeling, news classification, question answering, natural language inference, named entity recognition, and syntactic parsing (Gasparetto et al. 2022). Obtaining the relevant categories can be performed on document-level, paragraph-level, sentence-level, or sub-sentence-level. A typical text classification pipeline comprises steps like text preprocessing (inter alia, text cleaning, tokenization, removing of stop words, capitalization, stemming, lemmatization, or using N-grams), feature extraction (like applying embedding techniques), dimensionality reduction, classification harnessing ML-based text classifiers, and evaluation (Kowsari et al. 2019).

The main contributions (refer to sections 3.4 and 4.4) of this doctoral research are related to multi-class news classification tasks on document-level that are supposed to assign categories to news articles.

Given a corpus $D = \{D_1, D_2, ..., D_N\}$, D_i refers to a document comprising a

[2]https://www.datev.de/web/de/m/ueber-datev/das-unternehmen/kurzprofil/, last accessed on 10.12.2022.

number s of sentences such that each sentence is made of w_s words with l_w letters. In addition, a fixed set of predefined classes $C = \{C_1, C_2, ..., C_M\}$ is given. Based on a training set comprising training documents with their associated classes, a learning algorithm strives for learning a classification function $f : D \rightarrow C$ that maps documents to classes.

1.3 Research Questions

This thesis primarily deals with the question how human-centered interactivity with text classifiers can be achieved using bidirectional model-agnostic explanations. From the perspective of XAI, it is investigated how conceptual explanations can be generated that are considered locally faithful and human-friendly. As an excursus, it is furthermore elaborated how model-agnostic explanation techniques like LIME can be applied to unsupervised ML tasks. From the perspective of interactive ML, this work contributes to the question of how human users can interact with ML learners by providing concept-based feedback based on conceptual explanations such that the learner's reasoning is pushed towards a desired behavior. Specifically, the following research questions are addressed in this thesis:

1. How to conceptually define comprehensible and interactive AI using cognitive concepts and interdisciplinary research areas?

2. How to achieve concept-based explainable machine learning for the domain of text classification?

   (a) How to approximate a conceptual representation of a textual input domain that humans and ML systems can harness for interaction? How to ensure that coherent facts are taken into account during the exchange of explanations?

   (b) How to generate concept-based and locally faithful local model-agnostic explanations? How to obtain realistic and meaningful local perturbation distributions in order to avoid extrapolation?

   (c) How can ML-users compare concept-based a-priori domain knowledge about input-output relations with the reasoning of text classifiers?

   (d) Do humans prefer concept-based explanations consisting of topics that are made of coherent words compared to explanations whose explanatory features lack an interrelated structure?

3. How to apply model-agnostic explanation techniques to unsupervised machine learning? How to post-process explanations such that they can be considered selective and receiver-dependent?

4. How to develop an explanatory interactive machine learning approach for text classification that offers concept-based means for performing corrections and providing hints?

   (a) How to integrate concept-based human corrections into text classifiers while avoiding counterexamples that are considered "out-of-distribution"?

   (b) How does the elaborated interactive system that supports contextual interpretation and intervention compare to state-of-the-art methods in terms of predictive performance of downstream multi-class classification tasks?

   (c) How do explanations of the concept-based interactive system compare to explanations generated by state-of-the-art methods with regard to local explanation quality?

1.4 Structure of Thesis

Part I provides a synopsis of the research and development performed as part of this doctoral work. The synopsis is intended to (a) define the overall area of research, (b) contextualize the individual scientific publications and supply the reader with a reading assistance, and (c) include further related research results that have not been published yet. To address the research questions stated in section 1.3, the synopsis is organized as follows:

Chapters 2 to 4 explain the main scientific contributions. While chapter 2 mainly comprises conceptual contributions to the field of comprehensible and interactive artificial intelligence, chapters 3 and 4 describe the technical, empirical and applied contributions of this research.

Chapter 2 presents a distilled version of a framework for comprehensible and interactive AI that is used as a structuring aid for this doctoral thesis. In its original form, it was proposed first as conceptual contribution in (Bruckert, Finzel, and Schmid 2020). The framework unites interdisciplinary concepts, approaches and measures related to comprehensible and interactive AI. Section 2.1 describes the need for CIAI, mainly from regulatory and legal point of view. Next, sections 2.2 to 2.3 show how the concepts explainability, interpretability, transparency and interactivity relate and contribute to the development of CIAI.

Chapter 3 describes a way to generate concept-based and selective explanations in a model-agnostic way, mainly for text classifiers. After introducing XAI and a local surrogate explanation model named LIME in sections 3.1 to 3.2, insights from psychology on properties of human-friendly textual explanations are presented in section 3.3. Next, section 3.4 introduces and evaluates an architecture that is capable of improving model-agnostic local explanations for the domain of text classification. This approach builds locally faithful explanations such that its explanatory features are made of coherent words. Section 3.5 presents and discusses

the results of an empirical evaluation of the contextual explanations generated by the newly developed explanation method.

As an excursus, section 3.6 introduces an architecture that allows the generation of model-agnostic explanations for unsupervised ML tasks, like clustering or association problems (e.g., anomaly detection, customer segmentation, recommender systems), which has not been possible ante hoc. The developed approach is exemplary showcased for the task of anomaly detection for financial auditing and a technique to tailor the explanations to the needs of different target groups, like data scientists or financial auditors, is proposed.

Chapter 4 first introduces interactive machine learning and explanatory interactive machine learning in section 4.1. Subsequently, section 4.2 describes CAIPI as a state-of-the-art and model-agnostic method for explanatory interactive machine learning. In a next step, section 4.3 characterizes requirements for effective and efficient co-work between humans and ML systems based on insights from psychology. Based on the results from chapter 3, section 4.4 describes the development of a novel interaction framework. It is capable of incorporating concept-based user feedback via topics into a text classification system based on transparent queries.

Finally, chapter 5 concludes this thesis and provides an outlook for future research.

Part II extends the synopsis - contained in part I - in the form of an appendix. Appendix A includes all the publications that resulted from the doctoral research. Along with each publication, the related research questions from section 1.3 and the types of contribution are addressed. A full reference of each paper is given and the scientific contributions and written contents of the author of this doctoral thesis are detailed. In appendix B, some additional results produced during this work are presented.

2. A Framework for Comprehensible and Interactive Artificial Intelligence (AI)

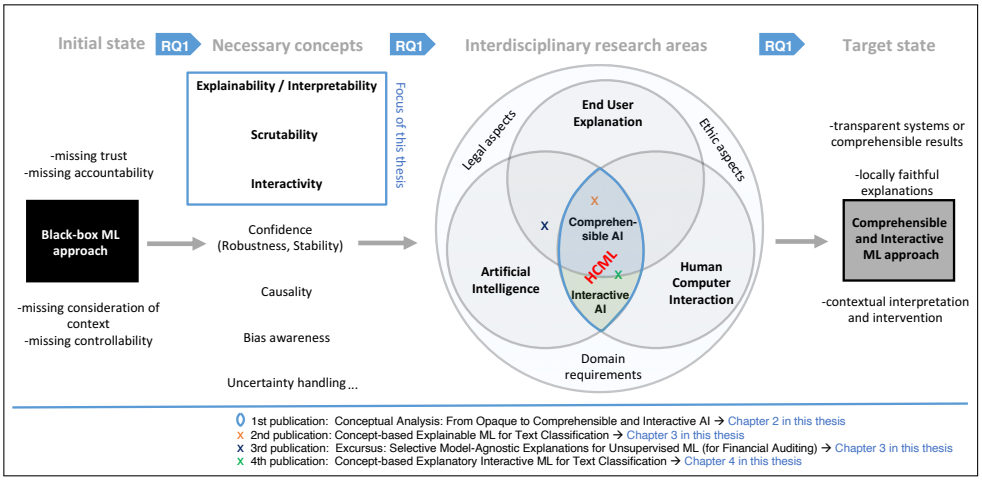


Figure 2.1: Comprehensible and interactive artificial intelligence framework (distilled version): the framework is used as a structuring aid for this doctoral thesis, locates the associated scientific publications, and, in combination with the publication (Bruckert, Finzel, and Schmid 2020) contained in appendix A.1.1, contributes to research question 1 from section 1.3.

Figure 2.1 presents a distilled and slightly modified version of the Comprehensible Artificial Intelligence (CAI) transition framework first suggested in figure 3 of the conceptual analysis paper (Bruckert, Finzel, and Schmid 2020), contained in appendix A.1.1. It was originally intended to work as a broader roadmap framework showing how to conceptually achieve the transition from black-box ML approaches to comprehensible ML approaches using interdisciplinary, yet integrated research. The main goal was to develop AI-infused systems that can act as interpretable and therefore transparent and comprehensible expert companions. At this point in time, the reader is referred to the first publication associated with this dissertation, included in appendix A.1.1.

The slightly modified Comprehensible and Interactive Artificial Intelligence (CIAI) framework in this thesis locates the research area HCML as the integration of comprehensible and interactive AI and introduces underlying cognitive concepts like explanation and interpretation (refer to section 2.2 for further details).

Additionally, it addresses typical drawbacks of black-box systems, like missing accountability, missing consideration of context and missing controllability, that hinder human-centered interactivity. Furthermore, the implications, especially for highly regulated or high-stakes application domains, are addressed in section 2.1.

Figure 2.1 is not only intended to provide a general overview of CIAI, but also to frame the four research publications of this cumulative dissertation and how they jointly contribute to the development of human-centered interactivity with AI approaches, especially with text classifiers. The first publication (Bruckert, Finzel, and Schmid 2020), denoted as blue frame in figure 2.1, performs a conceptual analysis and discusses the elaborated concepts with regard to the task of medical diagnosis in order to make the considerations more tangible. Furthermore, it addresses the prospect of interactive machine learning, especially human-in-the-loop-learning. Additionally, it introduces semantic alignment between ML classifiers and its human users as a prerequisite for bidirectional semantically meaningful explanations.

The second publication (Kiefer 2022), a single-author publication denoted as orange cross and attached in appendix A.2.1, identifies the research gap of missing contextual information in local surrogate explanation models and the resulting explanations. Consequently, it proposes an architecture motivated by insights from the psychological field of "explanatory understanding". Its instantiation called *topicLIME* is capable of generating coherent and semantically meaningful explanations for text classifiers independently of the underlying model. It improves the standard LIME method and generates explanations with higher local fidelity. As an addition, the publication presents the prospect of concept-based interactivity based on contextual explanations by introducing a new algorithm for semantic interrogations. It enables humans to explicate their conceptual domain knowledge, represented as topic distributions as a proxy, translate it to computational representations and compare it to the classification results of an ML learner. As the proposed explanation generation and interrogation approaches work independently of the harnessed ML models, the contribution of the publication is located more towards end user explanation than towards AI in figure 2.1.

The third publication's scope (Kiefer and Pesch 2021), denoted as black cross and attached in appendix A.2.2, is located slightly outside the main scope of HCML as it leaves the interactive part aside. It rather identifies the research gap of missing techniques to achieve model-agnostic explainability for unsupervised ML tasks, such as clustering or association problems. Exemplary for the task of unsupervised anomaly detection for financial auditing, an architecture is developed that is capable of applying perturbation-based local explanation generators, like LIME or SHapley Additive exPlanations (SHAP) (Lundberg and S.-I. Lee 2017), to unsupervised ML tasks. The key contribution is to integrate an intermediate supervised classification task as global approximation of the unsupervised model that is to be explained. Additionally, publication 3 elaborates on an explanation-post-processing-technique in order to generate selective explanations tailored to

the needs of the explainees[3] (refer to section 3.6 for further details). Accordingly, the contribution of this application-oriented research is located close to the center of AI in figure 2.1 while integrating parts of end user explanation.

The fourth publication (Kiefer, Hoffmann, and Schmid 2022), denoted as green cross and attached in appendix A.3.1, enables humans to incorporate conceptual corrections into a text classifier as a direct answer to the concept-based explanations generated by topicLIME. It describes the development of a new framework which is instantiated by a method called *Semantic Push*. Semantic Push improves a state-of-the-art method for explanatory interactive learning, called *CAIPI*, by allowing users the integration of constructive and continuous conceptual feedback based on contextual explanations. The human corrections are in turn translated to adequate computational representations in the form of counterexamples that subsequently push an ML learner towards the desired behavior. It is shown that Semantic Push outperforms several baselines with regard to learning performance and local explanation quality. The proposed interaction strategy operates model-agnostically with regard to the classifier, builds on concept-based explanations developed during research in the field of comprehensible AI and further integrates an active learning paradigm. Therefore, the contribution of the fourth publication is located at the intersection of comprehensible AI and interactive AI in figure 2.1.

2.1 The Need for Comprehensibility and Interactivity

The need for comprehensible and interactive AI is manifold. Among others, the main drivers towards transparent, comprehensible, and interactive AI systems are requirements related to general areas like law, ethics, or society or requirements related to specific application domains. While ethical and social aspects demand a broad discourse and domain-specific aspects need to be analyzed task-specifically, some legal aspects or regulatory requests from authorities like the European Union can be concisely summarized. As parts of the research results presented in this thesis have been achieved during work within highly regulated application domains, like semi-automatic classification of incoming invoices for tax consultants or financial auditing, the remainder of this subsection focuses on legal and regulatory requirements for comprehensible and interactive AI. In addition, some psychological insights are included that relate concepts like explainability and interactivity to trust that users should develop in AI systems.

Generally, the relationship between AI and applicable law contains tremendous potential for clarification (Rodrigues 2020). While some legal aspects pertaining to AI broadly cover a variety of risks, others focus on specific issues or relate to specific domains, like healthcare (Price 2017), defense, or transport. Legal and human rights issues can comprise topics like intellectual property of AI (Schönberger

[3] In the remainder of this thesis, the term "explainee" is used to describe a human on the receiving end of an explanation, while the term "explainer" is used to describe a technical method that produces explanations for an ML system.

2018), fairness, and algorithmic bias in autonomous systems (Hacker 2018; Danks and London 2017; Courtland 2018), privacy and data protection (Wachter and Mittelstadt 2019), access to justice (Raymond and Shackelford 2013), algorithmic transparency (Coglianese and Lehr 2019; Bodo et al. 2018), liability for harms (Vladeck 2014), and accountability (Liu, Lin, and Y.-J. Chen 2019). Technical, regulatory as well as legal considerations across different application domains are required for an overall analysis of the relationship between AI and law.

A significant issue in terms of legal discussions regarding AI is the lack of algorithmic transparency (Bodo et al. 2018). Especially in areas involving high-stakes decisions, it is inevitable for AI systems to be accountable, fair, and transparent (Cath 2018). As an example, people who were denied jobs or refused loans due to an algorithmic decision might want to understand the algorithm's functionality. To open the algorithmic black box, the EU General Data Protection Regulation (GDPR) suggests the "right to an explanation" as a basis for accountability (Edwards and Veale 2017). Nevertheless, achieving algorithmic transparency only constitutes a necessary, not a sufficient condition for a legally satisfactory interaction between humans and machines as decision-subjects might not agree with the explained outcome (Ananny and Crawford 2018).

Another issue in legal discussions regarding AI is the lack of accountability for harm. The question arises of who should be accountable for the development, deployment and use of AI systems. Even if a potential harm can be identified due to the algorithms being transparent, it still can be difficult to ensure legal accountability for violations (Rodrigues 2020). The "right to an explanation" as a tool for AI accountability might reveal some major challenges as well. Often, it is not practical or even possible to explain all decisions made by an algorithm (Edwards and Veale 2017), which might even lead to a transparency fallacy. Additionally, some AI explanation techniques that computer scientist have elaborated so far might only partially match the requirements of the legal conception of explanations as "meaningful information about the logic of processing" (Edwards and Veale 2017).

From a legal perspective, identifying those accountable for harm caused by an AI system is not enough. The question arises, how legal liability should be established if something goes wrong. The question for liability for harm arises if third parties suffer damages that are caused by recommendations or decisions made by AI approaches. For example, driverless cars might run over pedestrians or harmful medical treatment might be advised by an AI system. Since the specific behavior and the individual actions of such systems often are not fully predictable due to lack of algorithmic transparency (Lepri et al. 2018), this poses a big challenge for justice systems. Additionally, many parties are involved in the development of an AI system, like data providers, manufacturers, programmers, developers, users as well as the technical AI component itself (Rodrigues 2020). In such cases, liability is difficult to establish as there are many factors that need to be considered in case of faulty behavior. According to the latest jurisdiction, software architects, software developers, and users are only liable for their actions and artifacts if

a certain behavior of the system would have been predictable. In case that a specific incorrect behavior of a system, be it unexpected, damaging, unfair, or discriminatory (Edwards and Veale 2017), or its effects was not foreseeable, strict liability is considered a poor solution (Bathaee 2017).

Furthermore, liability issues with regard to AI potentially could be addressed by civil or criminal liability. It is doubtful whether the question of criminal liability could be posed at all, and if so, who or what the accused would be (Kingston 2016). Additionally, it is debatable whether an AI entity itself could be criminally liable beyond the criminal liability of the manufacturer and beyond the corresponding civil liability.

As a consequence, a legal gap results regarding the question of liability, which has to be reduced or even closed by means of legal considerations and technical measures in the near future. From a legal standpoint, there are several efforts being made, for example, by the European Union. According to (GDPR 2016/679), individuals must have the right to challenge and request a review for automated decision-making if they are affected in their rights or legitimate interests by automatic decisions (Rodrigues 2020). Furthermore, it is argued that individuals need to have a right (a) to obtain an explanation of the decision reached, (b) to obtain human intervention, and (c) to challenge the decision (Roig 2017). In a nutshell, the authors of (Hildebrandt and Janssens 2016) and (Edwards and Veale 2017) underline that overcoming algorithmic opacity and establishing means to challenge the algorithmic systems will be a prerequisite for accountability and liability of AI owners in a legal sense.

Thus, from a technical standpoint, an improvement of interpretability of recommendations and decisions of AI systems could address some open points regarding accountability and liability. In addition to predictability of the AI behavior (ante hoc), especially the generation of meaningful and human-friendly post hoc explanations could contribute to many questions arising during the interplay of AI and law.

Besides legal considerations, there are other major drivers towards explainable, interpretable, and interactive AI. According to (Pu, L. Chen, and Hu 2011; Prahl and Van Swol 2017; T. Miller 2019), explanations generated by a system are capable of providing a valuable basis for transparency and comprehensibility regarding systems' decisions and therefore can lead to increased trust of users. The authors of (Schaefer et al. 2016) refer to a high level of user-trust as well as interaction and influencing possibilities as future acceptance criteria for the usage of such systems. Work from (Madhavan and Wiegmann 2007) also states that initial trust in AI partners might decrease rapidly in case of incorrect or unexpected reactions, which is why future AI-infused systems would benefit from being comprehensible and allowing interventions in the form of user-specified corrections.

The considerations from this section constitute the motivation for this thesis that strives for overcoming algorithmic opacity by improving local surrogate explanation models (refer to sections 3.1 and 3.2 for further details). Furthermore, human interactions via concept-based explanations and corrections shall be en-

abled in order to allow comprehension of AI decisions and the possibility for human intervention on those.

2.2 Comprehensible Artificial Intelligence

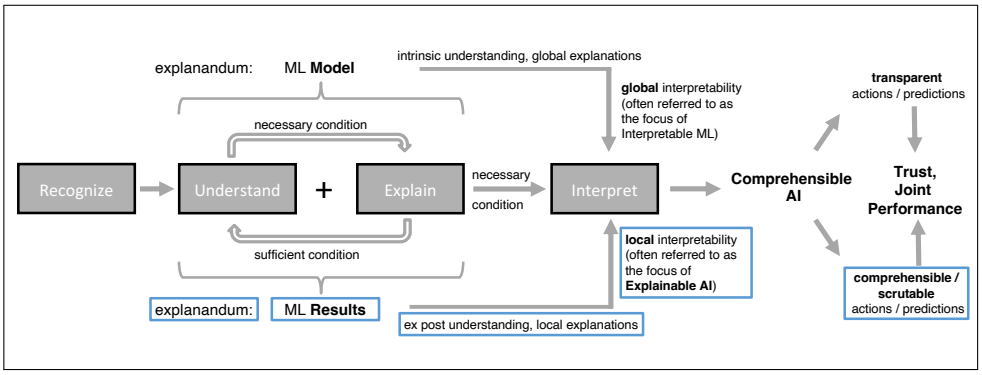


Figure 2.2: Derivation of comprehensible artificial intelligence considering basic cognitive concepts. The emphasized parts in blue denote the main scope of this thesis. This figure is a slightly modified version of figure 3 of (Bruckert, Finzel, and Schmid 2020).

This section in some parts summarizes work published in (Bruckert, Finzel, and Schmid 2020). Figure 2.2 depicts how comprehensible artificial intelligence can be described and derived using basic cognitive concepts known from cognitive psychology. This thesis refers to the overall objective of CAI as developing intrinsically transparent AI systems or systems that produce comprehensible, both in the sense of reasonable and, as a basis for further interactions, scrutable results. Humans shall develop trust into such systems and a high joint performance between a system and its users is aimed for. While global interpretability can be seen as a property related to transparent AI systems, local interpretability refers to making reasonable justifications why a system produced a certain single output.

This thesis primarily deals with local surrogate explanation models that shall help users develop an ex post understanding. Such models do not require any knowledge of or assumptions about the internal structure of the model to be explained. They are referred to as model-agnostic explainers, contrasting model-specific explainers that use the model's internal structure to offer explanations. Generally, enabling local interpretability by generating local explanations is often referred to as the focus of XAI. For further details on local surrogate explanation models, please refer to sections 3.1 and 3.2.

Figure 2.2 further breaks down the objective of the research underlying this dissertation. The main scope is to develop or improve local surrogate explanation models such that locally faithful explanations for individual results are generated for end users who shall be able to comprehend the results. For a detailed description

of local faithfulness, please refer to subsection 3.4.3. Based on comprehensible and scrutable results, users shall be able to perform contextual interpretation and intervention.

The terms "explainability" and "interpretability" are often used interchangeably in the context of XAI. As an example, the authors of (Kulesza et al. 2015) refer to explainability as the capability of an ML system to accurately explain the reasons for its predictions to an end user. In a similiar way, work from (Doshi-Velez and Kim 2017) describes interpretability as a model's "ability to explain or to present in understandable terms to a human". This thesis closely sticks to the definitions made by (Tomsett et al. 2018): explanation in the context of XAI is defined as "the information provided by a system to outline the [...] reason for a decision or output". Accordingly, explainability describes the degree to which an AI system is capable of providing explanations for the underlying reasons. Interpretation in the context of XAI, in contrast, is referred to as the understanding of a user about the underlying reasons and is therefore related to a user's mental model of the system to be explained. Finally, interpretability describes the degree of understanding a user gains based on explanation.

This thesis argues that comprehensibility and its associated concepts explainability and interpretability constitute a necessary, but not a sufficient condition for paving the way towards human-centered machine learning. As depicted in figure 2.1, scrutability and interactivity in the context of AI are required additionally and therefore constitute other necessary conditions for HCML.

2.3 Scrutable and Interactive Artificial Intelligence

Besides explainability and interpretability, there are further concepts related to comprehensible and interactive artificial intelligence. In general, scrutability can be defined as a concept that is about "allowing users to tell the system if it is wrong" (Tintarev and Masthoff 2015). Although the concept originally stems from research on social recommender systems (Balog, Radlinski, and Arakelyan 2019), it fits seamlessly into a framework for comprehensible and interactive artificial intelligence. One aspect of scrutability relates to interpretability, as the authors of (Cheverst et al. 2005) define the concept as "the ability of a user to interrogate her user model in order to understand the system's behavior". Another facet of scrutability more relates to the concept of interaction, as it provides users "with a direct and meaningful way to revise their mental model [through interactions]" (Smith-Renner et al. 2020).

The concepts of explainability, interpretability, and scrutability can be seen as conceptual prerequisites for explanatory interactive machine learning. Generally, interactive ML describes an iterative learning process that tightly couples human input with machine learners (Fails and Olsen 2003), however detached from the concepts explainability and interpretability. The overall process can be characterized as a training–feedback–correction cycle.

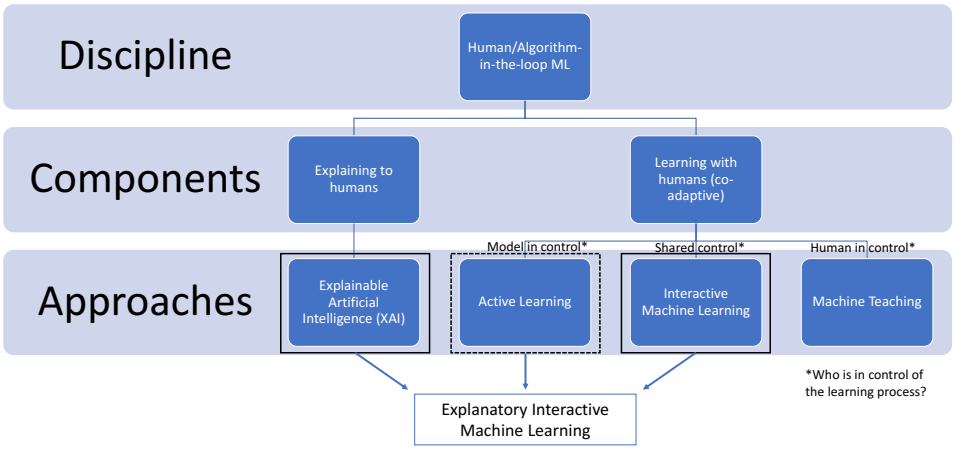


Figure 2.3: Derivation of explanatory interactive machine learning: own representation based on (Mosqueira-Rey et al. 2022), the black framed approaches for human-in-the-loop machine learning are combined within explanatory interactive machine learning.

Explanatory interactive ML goes one step further. It is intended to allow humans to correct predictions based on explanations (Teso and Kersting 2019). Users shall be able to iteratively integrate corrective feedback into an ML learner (Amershi et al. 2014) after having analyzed its decisions. As depicted in figure 2.3, current approaches to model-agnostic explanatory interactive learning integrate an *active learning* paradigm with local explainers like LIME. In active learning, human experts label query instances selected by a learner according to some preference mechanism (Bùrr Settles 2011). A more detailed description of explanatory interactive ML is provided in section 4.1.

The combination of these approaches allows for a "back-and-forth dialogue between a user and a system" (Chromik 2021), where both parties directly influence each other's behavior and therefore act in a co-adaptive way (Dudley and Kristensson 2018) while including the concepts explainability and interpretability. In a nutshell, the concept of *scrutability* can be referred to as an intermediate concept between *explainability/interpretability*, which is the focus of comprehensible AI, and *interactivity*, which is the focus of interactive AI. The latter is defined as "the ability [of users] to tweak and interact with the models" (Arrieta et al. 2020). In the further course of this thesis, some contributions to comprehensible AI are described in chapter 3. Chapter 4 is related to the heart of HCML as it integrates the concepts *explainability*, *interpretability*, *scrutability* and *interactivity* in order to allow users to perform informed corrections of a system and to integrate concept-based expert knowledge.

3. Concept-based Explainable Machine Learning for Text Classification

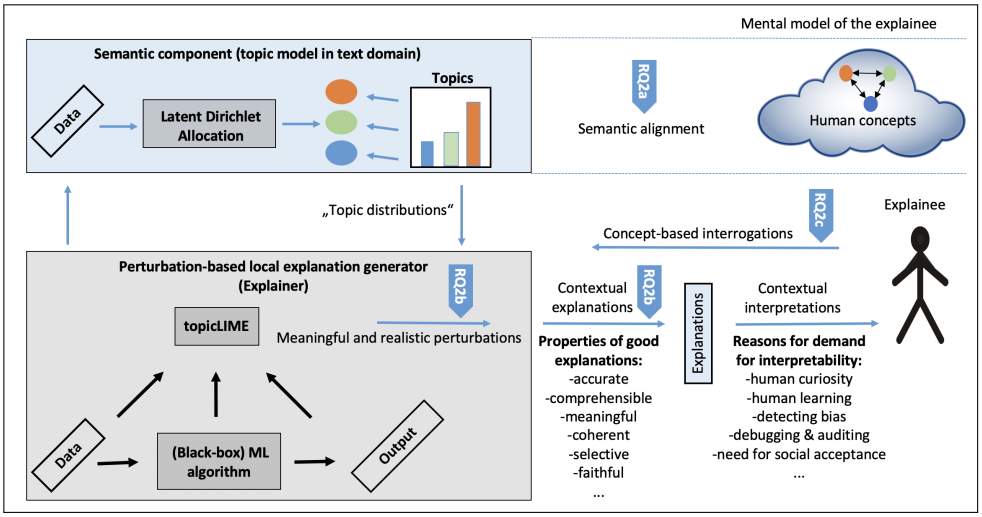


Figure 3.1: Fusion of local surrogate explanation models with contextual and semantic knowledge: the blue colored parts of this figure indicate the explicit main contributions of this chapter. The properties of good explanations and the reasons for the demand for interpretability are taken from (Doshi-Velez and Kim 2017; Robnik-Šikonja and Bohanec 2018). The proposed architecture and its instantiation constitute the main contributions to research question 2 from section 1.3.

The following chapter is primarily intended to describe an architecture that is capable of generating concept-based and coherent explanations for text classifiers. It mainly integrates and contextualizes the methods developed in the second (Kiefer 2022) and third (Kiefer and Pesch 2021) publication associated with this dissertation in a tangible and more intuitive manner. First, some background information on the overall research area explainable artificial intelligence, which is motivated as the basis for human-centered machine learning, is provided. Next, the theoretical foundations and drawbacks of LIME that the newly developed topic-based approach builds on and compares to as a baseline are introduced. Subsequently, this chapter provides information on desiderata for human-friendly explanations known from disciplines like psychology or the social sciences. Additionally, the central methodological connections between the different components

of the proposed architecture and its integration are described. The new algorithm for generating topic-based explanations called *topicLIME* is introduced in an illustrative way by using analogies to the domain of computer vision, depicting its perturbation-based sampling in a visual manner and comparing the resulting explanations to its baseline. For in-depth technical details, the reader is referred to the according publication in appendix A.2.1. Lastly, topic-based explanations are compared to the word-based explanations generated by LIME, both in terms of a technical and an empirical evaluation.

3.1 Explainable Artificial Intelligence

A prominent definition was provided by (Arrieta et al. 2020): "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand". This definition implies that considering the characteristics of an audience, like an explainee's mental model, during the explanation-process is of great importance, even though considering related insights from cognitive psychology, social sciences, or human-computer interaction has often been neglected by XAI research (Abdul et al. 2018; T. Miller 2019; Inkpen et al. 2019). Based on these considerations, several XAI goals have been formulated over time. The following enumeration focuses on the four most important ones when trying to achieve a step towards human-centered machine learning:

1. **Trustworthiness**: Trustworthiness is often referred to as the primary aim of an explainable AI (Marco Tulio Ribeiro, Singh, and Guestrin 2016) and can be considered as "the confidence of whether a model will act as intended when facing a given problem" (Arrieta et al. 2020). The main target audience for trustworthy AI comprises domain experts and users that are affected by decisions of AI-infused systems. As an example, financial auditors might only accept a system supporting the process of an annual audit if they have an idea of how the system works and why it arrives at certain decisions.

2. **Informativeness**: Especially ML models are typically used for supporting decision making performed by its users (Huysmans et al. 2011), which is why a great deal of information is required to relate decisions of users to results of an ML model (Arrieta et al. 2020). Therefore, models should give insights about the problem at hand, for example, by extracting some information about the inner representations and relations of a model.

3. **Accessibility**: Accessibility can be regarded as a predecessor of end user interactivity which is introduced below. Some researchers (Chander et al. 2018; T. Miller, Howe, and Sonenberg 2017) argue that accessibility via explainability is a prerequisite to get non-technical or non-expert users involved in developing and improving ML models.

4. **Interactivity**: Some contributions to XAI (Harbers, Bosch, and Meyer 2010; Langley et al. 2017) state that explainable ML models should be capable of being interactive with its users. This thesis argues that interactivity is not a direct goal of XAI. Instead, scrutability, which constitutes a concept located between explainability and interactivity, can result as a valuable by-product of explainable models.

Technically, explainability methods in the context of XAI can be differentiated regarding the *scoop of interpretability*, whether they shall enable global or local interpretability, and whether they need access to the internals of the model to be explained. Local explainers are focused on providing individual explanations for single predictions or a group of predictions and therefore shall establish trust in model outcomes (Adadi and Berrada 2018). Global explainers, on the other hand, focus on providing an understanding of the model's mechanism in terms of its decision process, answering the question of "*how* does the system work?". Generally, in terms of trustworthiness, local explanations are often considered more faithful compared to global explanations (Adadi and Berrada 2018) as they offer information on the correlations of inputs and outputs and thus provide justifications *why* a certain output was produced.

As model-specific techniques use knowledge on the model's internal structure, the resulting explanations are typically considered more accurate compared to model-agnostic methods. The latter often harness approximations of the model to be explained, like surrogate models do, and are therefore regarded as more flexible.

In this chapter, an approach is introduced that shall enable human users contextual interpretation of text classifiers without knowing its internal structure. Later in this thesis, contextual intervention shall be possible based on interpretable results. The associated architecture, as depicted in figure 3.1, is based on a local surrogate model called *LIME* for generating explanations. It allows highest flexibility with regard to the types of models to be explained. If the explanation-generation-process is further enriched with human-like concept-based knowledge, as learned by an LDA model that is described in detail in subsection 3.4.1, then such a combination is assumed to be well-suited as a basis for human-centered interactions.

3.2 Local Interpretable Model-agnostic Explanations

As LIME is of central importance and the basis for explanations for all approaches introduced in the remainder of this thesis, this section replicates and extends the formal description of LIME that is published in (Kiefer 2022) and (Kiefer, Hoffmann, and Schmid 2022). The authors of (Marco Tulio Ribeiro, Singh, and Guestrin 2016) developed LIME, a method explaining a prediction by locally approximating a classifier's decision boundary in the neighborhood of an instance to be explained. LIME harnesses a local linear surrogate explanation model and therefore can be characterized as an additive feature attribution method (Lundberg and S.-I.

Lee 2017). Given the original representation $x \epsilon \mathbb{R}^d$ of an instance that shall be explained, $x' \epsilon \{0,1\}^{d'}$ denotes a binary vector for its interpretable input representation. Furthermore, let an explanation be represented as a model $g \epsilon G$, where G is a class of potentially interpretable models, such as linear models or decision trees. Additionally, let $\Omega(g)$ be a measure of complexity of the explanation $g \epsilon G$, for example, the number of non-zero weights of a linear model. The original model for which explanations are searched is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A measure $\pi_x(z)$ defining the locality around x is used to capture the proximity between an instance z and x. The final objective is to minimize a measure $\mathcal{L}(f, g, \pi_x(z))$ that evaluates how unfaithful g (the local explanation model) is in approximating f (the model to be explained) in the locality defined by $\pi_x(z)$. Striving for both interpretability and local fidelity, an explanation is obtained by minimizing $\mathcal{L}(f, g, \pi_x(z))$ as well as keeping $\Omega(g)$ low enough to be an interpretable model.

$$\xi(x) = \underset{g \epsilon G}{\arg\min} \ \mathcal{L}(f, g, \pi_x(z)) + \Omega(g). \tag{3.1}$$

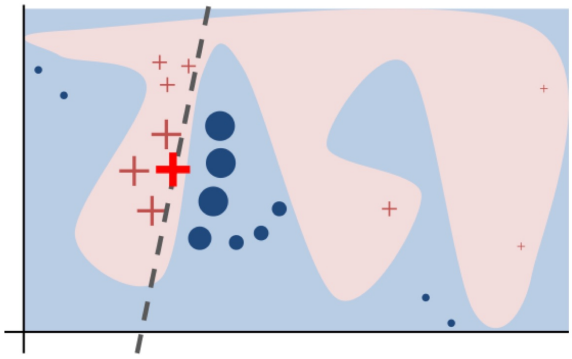


Figure 3.2: Intuition of LIME: the background colored in light blue and light pink represents a non-linear decision function (not known to LIME) of a model to be explained. The bold red cross is the instance that shall be explained locally. Therefore, LIME generates a local neighborhood based on independently drawing samples from a local perturbation distribution and then obtains predictions by applying the prediction function of the explained model. Subsequently, the samples are weighted by their proximity to the original instance. Finally, a local linear surrogate model (the dashed line) is learned from which the feature attributions are extracted and used as a, desirably locally faithful, explanation. This illustration is taken from (Marco Tulio Ribeiro, Singh, and Guestrin 2016), figure 3.

To be a model-agnostic explainer, the local behavior of f must be learned without making any assumptions about f. Therefore, $\mathcal{L}(f, g, \pi_x(z))$ needs to be approximated by drawing random samples weighted by $\pi_x(z)$. Instances around x', representing a binary vector for the interpretable input representation of x, are sampled from a local perturbation distribution by drawing nonzero elements of x' uniformly at random. Finally, a perturbed sample z' is obtained.

Recovering z from z' and applying $f(z)$ then yields a label that is used as label for the explanation model. The last step consists of optimizing equation (3.1) by making use of dataset $\mathcal{Z}$, which includes all perturbed samples with the associated labels. Figure 3.2 provides an intuition of how LIME works.

3.3 Desiderata for Human-friendly Explanations

After developing algorithms that explain the predictions of ML models, the explanation methods and the individual explanations need to be evaluated with regard to some common criteria. Therefore, the authors of (Robnik-Šikonja and Bohanec 2018) developed a set of properties for judging the quality of individual explanations from a technical point of view. One of the most important properties of technically sound explanations is fidelity. It is even considered more important than high accuracy that measures the degree to which an explanation predicts unseen data (Molnar 2022). The main question with regard to fidelity is how well an explanation approximates the prediction performed by a model to be explained. Especially for local model-agnostic explanation methods, like LIME or SHAP, which can not rely on any knowledge on the model's internal structure and therefore need to approximate the model locally, high fidelity of explanations is of utmost importance. Local explainers can only achieve high local fidelity, meaning that a local explanation only approximates the model's prediction for a group of data or even an individual data point (Molnar 2022). The newly developed topic-based approach for explanation generation presented next in section 3.4 focuses on improving LIME. It strives for enabling its sampling to harness more realistic and non-extrapolating local perturbation distributions such that local explanations with higher local fidelity are obtained.

Generally, the process of generating and communicating explanations is far more than just a technical process. In the social sciences, an explanation process is defined as a social process between an explainer and an explainee, where the goal is to transfer knowledge about a cognitive process (T. Miller 2019). Therefore, it is not enough to develop and evaluate explanation systems solely based on technical properties. There is some prominent research (T. Miller, Howe, and Sonenberg 2017) arguing that many XAI approaches might reveal a phenomenon referred to as the "inmates running the asylum". This phrase coined by (Cooper 2004) describes the situation in which programmers design software for themselves instead of for their target audience. As a solution, the authors of (T. Miller, Howe, and Sonenberg 2017) call upon XAI researchers and developers to "understand, adopt, implement, and improve models from the vast and valuable bodies of research

in philosophy, psychology, and cognitive science" and "to focus more on people than on technology". Based on that, (T. Miller 2019) provides a summary of what characteristics make explanations "good" for humans. Some of the properties that the concept-based XAI approach introduced next tries to address are now enumerated in condensed form while including prominent work from (Keil 2006):

1. Explanations shall be **coherent**: Explanatory features in category learning that are the most causally interdependent on others are preferred by human users (W.-k. Ahn et al. 2000; Sloman, Love, and W.-K. Ahn 1998). Furthermore, explanatory features should not contradict themselves. In addition, explanations should be built in a way that they can be characterized as an internally consistent package whose elements form an interconnected, mutually supporting relational structure (Thagard 2002; Gentner and Toupin 1986).

2. Explanations are **selective**: Good explanations focus on the main causes of a decision (Arrieta et al. 2020). Humans are typically used to select one or two causes from a bunch of possible causes. In general, humans are capable of perceiving, processing, and remembering a limited number of pieces of information. This capacity is somewhere around seven plus or minus two pieces of information according to Miller's law (G. A. Miller 1956). Additionally, the *inference to the best explanation* process (Harman 1965) suggests that explainees might favor explanations made of a simpler internal structure despite being less predictive compared to explanations with higher predictive power revealing a more complex structure.

3. Explanations are **social** and therefore **receiver-dependent**: As explanations are part of a social interaction, the social context matters when defining the content and the way in which an explanation is presented. Typically, domain experts require a different type of explanation, like the level of abstraction or the mode of presentation, than non-experts, like end users of an ML system. In general, there are different abstraction levels of an explanation's information units, like raw features, derived features, semantic features, or abstract semantic features (Schwalbe and Finzel 2021). An approach could be to use an explanation facility that "moderates the social process of explanation between the user and the XAI system" (Chromik 2021). An effective explanation facility should meet technical requirements, like fidelity, but also softer requirements, like naturalness, responsiveness, flexibility, and sensitivity (Moore and Paris 1991). As a consequence, explanations especially designed for end users should (a) for the domain of text be organized in coherent natural language, (b) allow for follow-up questions and explanations that take prior explanations into consideration, and (c) account for a user's prior knowledge. To be a human-centered explanation facility, the latter must possess or create a model of the mental model of the explainee (Hoffman, Clancey, and Mueller 2020).

In this thesis, explanations that aim at fulfilling desiderata as mentioned above are referred to as human-friendly explanations. It is to be noted that some of those desiderata assume non-ML-experts or people with little time as the recipients of the explanations.

3.4 Topic-based Approach for Contextual Explanations

This section introduces a topic-based approach for generating model-agnostic contextual explanations for text classifications.

Typically, local model-agnostic explanation systems reveal some common drawbacks. Perturbation-based methods, which sample the local neighborhood from a local perturbation distribution by drawing nonzero elements of the instance to be explained uniformly at random, in case of LIME from a Gaussian distribution (Molnar 2022), ignore feature dependence. Such a behavior can have some severe negative consequences. First, when LIME's text explainer perturbs a certain document by deleting its words independently from each other, the links between the words representing some sort of context are ignored (refer to figure 3.6 for an intuition). In this manner, semantic information is lost and cannot be included in the resulting explanation. Second, the individual explanatory words that an explanation comprises are presented in a seemingly unrelated form that is far away from an interconnected, mutually supporting relational structure. Third, when ignoring the effect of correlated features during neighborhood generation, unrealistic, in the sense of unlikely, datapoints are taken as a basis for subsequent explanations. This phenomenon called extrapolation might lead to explanations that are characterized by low local fidelity and are therefore likely to be misinterpreted. Please refer to section 3.4.3 for a formalization of and further technical details on local fidelity.

Another problem becoming relevant in the future might be that generated explanations are not adjustable with regard to the semantic abstraction level. Despite the fact that more and more research, like research directed by Kristian Kersting[4] on semantic, symbolic, and interpretable machine learning, strives for developing ML approaches "which operate at the human abstraction level, where the world is described by entities, concepts, and their mutual relationships", many explainers for text classifications work solely at the level of raw features. In such a case, the explanatory features are words of the document to be explained. This thesis argues that flexibility in terms of the abstraction level of explanations will become more important, especially when dealing with heterogeneous groups of explainees and striving for the next step towards HCML.

In the progress of this section, an approach is introduced that is built on top of LIME's standard text explainer. It is intended to generate locally faithful, coherent, and selective explanations that offer the explainee additional contextual

[4] https://ellis.eu/programs/semantic-symbolic-and-interpretable-machine-learning, last accessed on 19.12.2022.

information. Technically, it strives for obtaining realistic and meaningful local perturbation distributions by avoiding extrapolation, which is a line of research that, according to the authors of Ribeiro et al., "would benefit multiple explanation methods"[5] (Marco Tulío Ribeiro, Singh, and Guestrin 2018).

In figure 1 of the second publication, attached in appendix A.2.1, the outline of generating semantically meaningful explanations that operate at a higher level of abstraction compared to explanations generated by state-of-the-art explainers is presented. A detailed view of the proposed approach is given in figure 3.1, showing the integration of a black-box ML algorithm, a perturbation-based local explanation generator, and a semantic approach that extracts statistical and meaningful regularities from the input domain's data. The individual components of the proposed approach are described in the following subsections.

3.4.1 Latent Dirichlet Allocation for Semantic Alignment

LDA represents the basis for the extraction of concept-based knowledge that is later used for enriching explanations and for incorporating constructive and contextual feedback into a learner. Due to its central importance, this section replicates and extends the formal description of LDA that is published in (Kiefer 2022) and (Kiefer, Hoffmann, and Schmid 2022). LDA can be summarized as an unsupervised generative probabilistic three-level hierarchical Bayesian model for collections of discrete data (Blei, Ng, and Jordan 2003). In text modeling, its goal is to find short representations of the documents within a corpus while preserving essential statistical relationships, like inter- or intra-document statistical structure. Therefore, LDA models documents from a corpus as an infinite mixture over an underlying set of topics. Each word is modeled as an infinite mixture over an underlying set of topic probabilities. In essence, documents are represented as random mixtures over latent topics and each topic is defined by a distribution over words. For each document w in a corpus D, a generative process from which the according documents have been created is assumed:

1. Choose N (the number of words) $\sim$ Poisson(ξ).
2. Choose θ (a topic mixture) $\sim$ Dir(α).
3. For each of the N words w_{n}:
   (a) Choose a topic $z_{\mathrm{n}} \sim$ Multinomial(θ).
   (b) Choose a word w_{n} from $p(\mathrm{w_n}|z_{\mathrm{n}}, \beta)$, a multinomial probability conditioned on the topic z_{n}.

[5] For a short introduction of further model-agnostic explanation methods, please refer to section 2 of (Kiefer 2022).

The joint distribution of a topic mixture θ, a set of topics z and a set of words w given the hyper-parameters α and β is characterized by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_{\mathrm{n}}|\theta) p(w_{\mathrm{n}}|z_{\mathrm{n}}, \beta). \tag{3.2}$$

The *concentration* hyper-parameters α and β for the Dirichlet distribution enable the injection of prior beliefs about topic and word sparsity. For a symmetric Dirichlet distribution, high α-values lead to documents that, with a high probability, contain a mixture of many or most of the topics, whereas high β-values lead to topics that are likely to be made of many of the words from the vocabulary. Another hyper-parameter that has to be chosen and should be optimized by the user with regard to human topic-interpretability is the number of topics *K*.

The major inferential problem, which consists of computing the posterior distribution of the hidden variables given a document, is described as follows:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \tag{3.3}$$

As the posterior distribution from equation 3.3 is intractable to compute in general, LDA makes use of approximate inference algorithms, like Laplace approximation, variational approximation, or Markov-Chain-Monte-Carlo.

On the one hand, LDA can be used as dimension reduction technique, on the other hand, it is also capable of making sense of the input data due to its generative probabilistic semantics properties. Referring to the latent multinomial variables as topics enables LDA to capture text-oriented intuitions, global statistics in a corpus as well as synonymy and polysemy (Blei, Ng, and Jordan 2003). For evaluating learned topic models, topic coherence measures are used that score a topic by measuring semantic similarity between highly contributing words within the according topic (Stevens et al. 2012). An important coherence measure that approximates human ratings regarding semantic topics best (Röder, Both, and Hinneburg 2015), namely $\mathrm{C_v}$ coherence, is introduced next.

Röder et al. found a combination of already existing coherence approaches, which they called C_v coherence, to be the best in terms of its correlation with respect to all available human topic ranking data (Röder, Both, and Hinneburg 2015). Topic ranking data are a state-of-the-art proxy for human topic-interpretability (Syed and Spruit 2017). Therefore, $\mathrm{C_v}$ coherence is often used to choose the best hyper-parameter "number of topics *K*" within an LDA approach and to measure the quality of identified topics with regard to human interpretability.

This coherence measure combines an indirect cosine measure with a Normalized Pointwise Mutual Information (NPMI) measure and a boolean sliding window (Röder, Both, and Hinneburg 2015). $\mathrm{C_v}$ is obtained by combining four parts (Syed and Spruit 2017):

1. Data segmentation: first, the top-words within each topic are paired with each other. Let W be a set of a topic's top-M most probable words.

2. Calculation of word probabilities: with Boolean document calculation the probabilities of single words or the joint probability of words from a word pair from step (1) are calculated. In order to partially consider word frequencies and distances, a Boolean sliding window is used that slides over one word of a document per step.

3. Calculation of confirmation measure: For each segmented word pair, a semantic confirmation measure is calculated by representing words as context vectors (see eq. 3.4) that are obtained using NPMI measure (see eq. 3.5).

$$\overrightarrow{v}(W') = \left\{ \sum_{w_{\mathrm{i}} \in W'} NPMI(w_{\mathrm{i}}, w_{\mathrm{j}})^{\gamma} \right\}_{j=1,\ldots,|W|}. \tag{3.4}$$

   γ is a free hyper-parameter used to put higher weights on higher NPMI values.

$$NPMI(w_{\mathrm{i}}, w_{\mathrm{j}})^{\gamma} = \left(\frac{\log(\frac{P(w_{\mathrm{i}}, w_{\mathrm{j}}) + \epsilon}{P(w_{\mathrm{i}}) * P(w_{\mathrm{j}})})}{-\log(P(w_{\mathrm{i}}, w_{\mathrm{j}}) + \epsilon)} \right)^{\gamma}. \tag{3.5}$$

   $0 < \epsilon \leq 1$ is a free hyper-parameter that prevents the logarithm from being zero. Ultimately, the cosine vector similarity of all context vectors is used to obtain the final confirmation measure.

4. The last step consists of applying the arithmetic mean of all confirmation measures in order to retrieve the final coherence score.

The use of a measure like C_{v} coherence as a quality criterion for human topic-interpretability can be justified by the distributional hypothesis of linguistics (Syed and Spruit 2017). It states that a difference of meaning correlates with a difference of distribution (Harris 1954). Expressed differently, words with similar meanings tend to occur in similar contexts. In the further course of this work, C_{v} coherence is used in order to find an adequate number of topics *K* of an LDA topic model such that topic coherence is maximized and high correspondence with human topic-interpretability is achieved. The overall goal is to infer semantically cohesive topics as an interpretable latent space that correspond to natural groupings for humans (Chang et al. 2009). Table 3.1 depicts the results of an exemplary application of the combined use of LDA and C_{v} coherence for the publicly available AG News dataset (X. Zhang, Zhao, and LeCun 2015).

Several LDA models were trained on the corpus with different values for the "number of topics" hyper-parameter *K*. A final selection was made by determining the optimal number K^* of topics $t = 1, \ldots, K$ by solving $\arg\max_K \frac{1}{K} \sum_{t=1}^{K} C_{\mathrm{v}}(t)$,

Table 3.1: Learned LDA topics and most representative words for the *AG News* dataset.

Topic	Representative Words
0	Iraq, Baghdad, Nuclear, Iran, Force, Military
1	Microsoft, Company, Software, IBM, System
2	European, United, Bank, Million, Trade, Deal
3	Bush, President, Press, Washington, John, Kerry
4	Internet, Search, Service, Phone, Online, Google
5	Oil, Price, Percent, Sale, Profit, Rate
6	Court, Company, Charge, Million, Trial, Drug
7	World, Cup, Win, Gold, Final, Champion
8	Game, Season, Team, League, Coach, Sport
9	New, York, Stock, Dollar, Share, Investor
10	Game, India, Australia, Fan, Video, Cricket
11	Police, People, Killed, Attack, Palestinian, Bomb
12	Minister, Election, Leader, President, Vote, Party

where C_{v} is the $\mathrm{C_v}$ coherence as introduced above. K was set to 30 and an optimal number of $K^* = 13$ topics was determined.

Combining LDA with a $\mathrm{C_v}$ coherence measure and harnessing the structural topic-based knowledge for subsequent explanation generation contributes to research question 2, especially 2a, from section 1.3. Research conducted by (Chang et al. 2009) proves that topic models, especially LDA, infer topics that are characterized by a semantic coherence that humans appreciate. Furthermore, the authors show that humans can associate the same documents with a topic as a topic model, like LDA, does.

Using $\mathrm{C_v}$ coherence combined with LDA to infer human-interpretable topics, to associate those topics with documents in a human-like manner and to harness the resulting knowledge in XAI leads to a concept named *semantic alignment*. The concept has been invented during this research and defines an alignment with regard to topic-encoded concepts between humans and a semantic component, to which a perturbation-based explanation system has access. Striving for a common understanding of the textual input domain based on topics that correspond to a higher abstraction level and include more contextual information than a plethora of independent facts might support explanation systems. For example, such systems might be better suited for taking the mental model of an explainee into account during explanation generation.

3.4.2 TopicLIME

The main idea of topicLIME is to use higher-level concepts, represented by topics in contrast to single words, of the textual input domain for subsequent explanation generation. Therefore, an approach is introduced that takes LIME's explainer for the visual domain as an analogy. When explaining images with LIME, a local neighborhood of an image to be explained is not sampled by perturbing individual pixels of an image, but using a higher-level representation based on so-called superpixels. Figure 3.3 depicts different superpixel algorithms. In general, superpixels represent image patches that have uniform pixel intensity and are aligned with intensity edges (PAN, LI, and ZHOU 2014). Superpixel algorithms are often harnessed during image preprocessing, like the computation of local image features.

Figure 3.3: In the visual domain, superpixels of images often are used by explanation systems, like LIME, as explanatory features. This illustration has been obtained by applying different superpixel algorithms to an image of the astronaut Eileen Collins, it is included in python's *skimage* library.

According to figure 3.3, the outputs of applying different superpixel algorithms heavily vary in terms of granularity of the contours identified, like the astronaut's eyes, ears, nose, mouth, or hair as higher-level concepts. When generating local explanations with LIME, the identified superpixels are used as the interpretable input representations for perturbation and in the local surrogate explanation models whose feature attributions are extracted. As individual pixels would lack the ability to include further contextual information, an identification of expressive input representations is of utmost importance for an explainer like LIME.

This thesis suggests to use an analogy of superpixels from the visual domain for explaining text classifications with topicLIME at a higher level of abstraction.

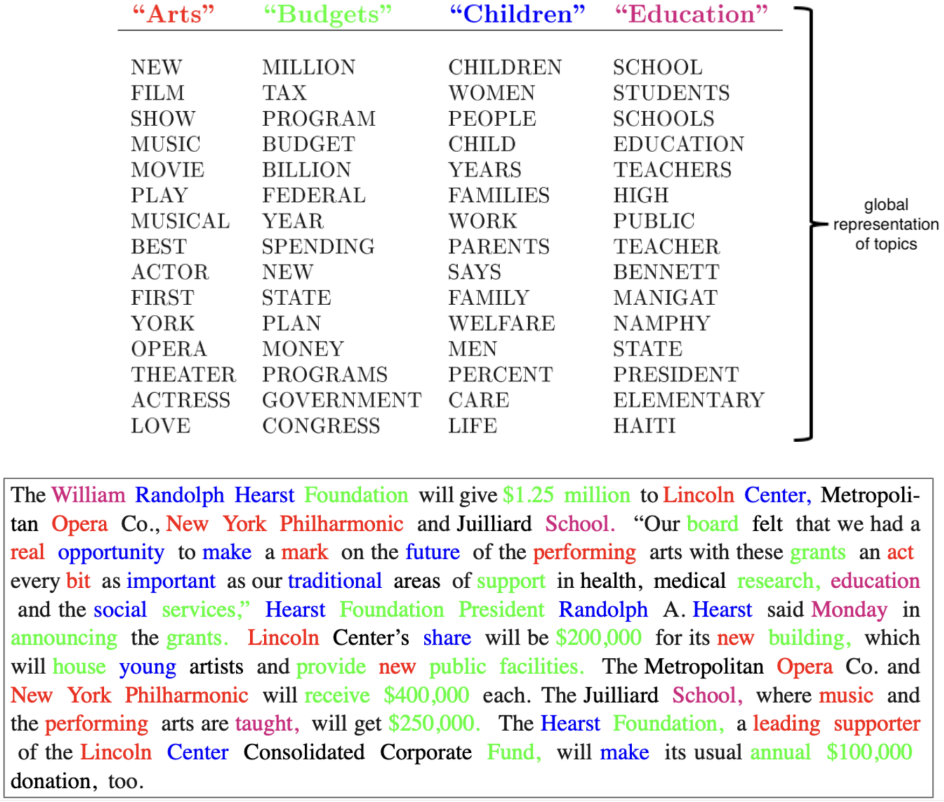


Figure 3.4: In the textual domain, local topic attributions of documents could, in analogy to superpixels in the visual domain, be used by explanation systems as explanatory features. In the upper part of this figure, the global factors, a.k.a. LDA topics, are represented. In the lower part, the words are assigned with color codes to the different factors that generated a word. This illustration is taken from (Blei, Ng, and Jordan 2003), figure 8.

The lower part of figure 3.4 depicts a single document from a text corpus called *TREC-AP* (Dàvid Lewis 1996). The individual words are assigned to different factors that generated the words. Those factors, also called topics, depicted in the

upper part of figure 3.4, that an LDA model identified globally, are shown using the most representative words of the whole corpus. Harnessing the multinomial distributions learned by LDA, each word of a document can be assigned to the topic that most likely was the generating factor of the word. For an example, please refer to figure 3.5.

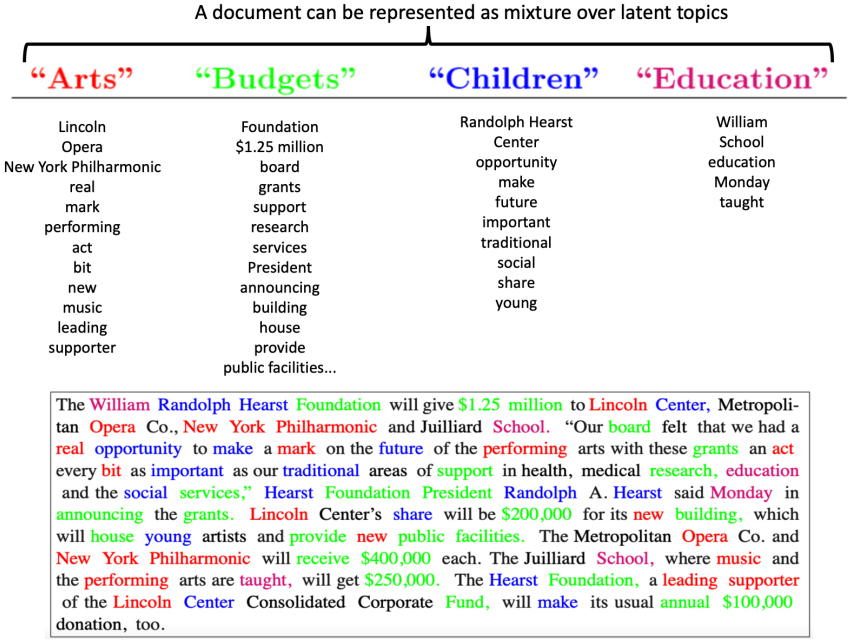


Figure 3.5: Representation of a single document as a mixture over a set of latent topics. Each word of the document is assigned to the topic it most probably belongs to.

In analogy to locally identified superpixels of an image, a specific document can be partitioned into "local topics", where each local topic consists of a subset of the words from the document at hand. This thesis argues that for generating explanations, those local topics can be used as interpretable input representations for perturbation and in a local surrogate explanation model. In the analogy (superpixels $\rightarrow$ pixels and topics $\rightarrow$ words) there is one difference to be noted. While superpixels are identified locally, meaning directly in a certain image, topics are identified globally for a corpus of documents. As such, there is an additional step involved in order to generate local topics that can be compared to superpixels, namely applying the global corpus-based knowledge learned by an LDA model locally to a document. The technical details are described in (Kiefer 2022).

Figure 3.6 depicts how word-based perturbation performed by standard LIME's text explainer differs from topic-based perturbation using local topics as performed by topicLIME. Having generated a local neighborhood of a document to be explained by applying topic-based perturbations, topicLIME outputs explanations as local topic attributions each characterized by a set of coherent words from

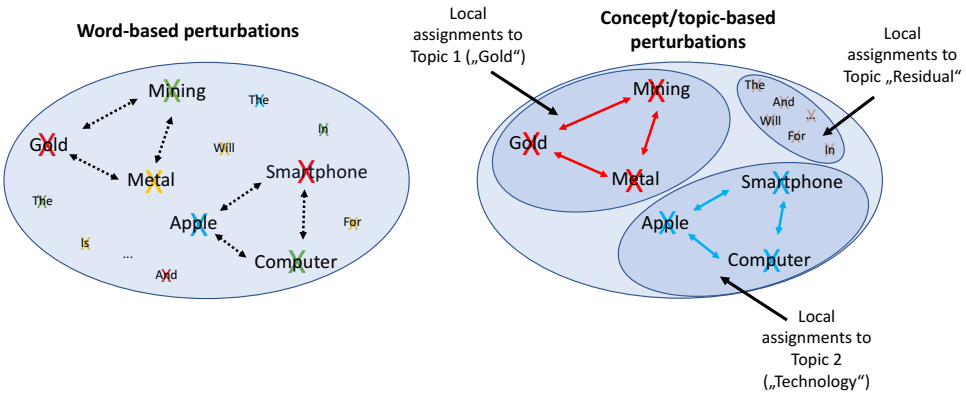


Figure 3.6: Comparison of word-based and topic-based perturbations of a document to be explained: for word-based perturbation, one or more most probably independent words are removed at a time not considering its distributional and semantic relations with the other words. In this example, first the red, then the green, then the yellow, and finally the blue crossed words are removed. A contextual interpretation of the word-explanations is complicated as the semantic "links" of a word are not reflected in the explanations. For topic-based perturbation, coherent and most likely, at least distributionally, but often semantically related words are considered at once. In this example, all red crossed words belonging to topic 1, then all blue crossed words belonging to topic 2, and finally all words from the residual topic including words that could not be assigned or were excluded, like stopwords, are removed. By simultaneously removing coherent words, the semantic "links" that in turn provide the context in the explanations that are based on these perturbations are included. Furthermore, as feature dependence between words is considered during neighborhood generation, the common problem of interpreting explanations resulting from unlikely datapoints, as a consequence of extrapolation to feature areas with low data density (Molnar et al. 2020), is mitigated.

the document. In order to additionally offer some sort of global context in the explanations, the labels of the globally identified LDA topics can be included. The global topics can be labeled either with the help of human annotators, which is called *eyeballing*, or by applying automated topic labeling approaches. For further details on automatic topic labeling, please refer to section 3.4 of the publication contained in appendix A.2.1.

Figure 3.7 compares word-based explanations generated by standard LIME with topic-based explanations generated by topicLIME.

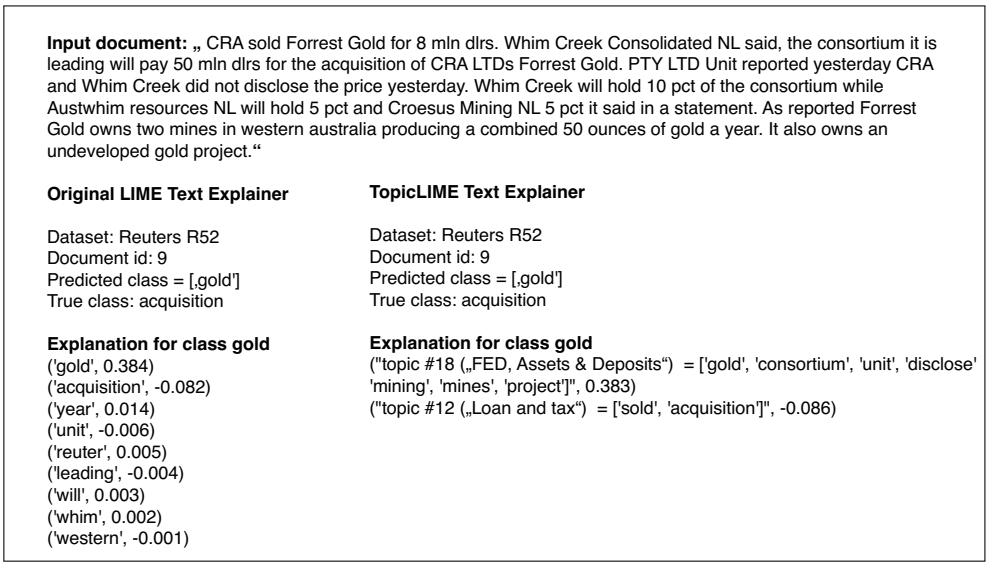


Figure 3.7: Exemplary explanations generated by LIME and topicLIME for a document of the Reuters R52 dataset (Davìd Lewis 1993). This illustration is taken from (Kiefer 2022), figure 6.

The two explanation modalities primarily differ with regard to the following aspects:

- **Level of abstraction**: In general, but still depending on the number of topics K identified by LDA, topicLIME generates explanations at a higher level of semantic detail as it uses topics as explanatory features instead of words. Doing so, topicLIME explanations include fewer explanatory features, represented as the number of topics, compared to LIME explanations, represented as the number of words. The level of semantic detail can be individually adapted by incorporating prior knowledge into the LDA process, what makes topicLIME explanations adaptable to the needs of receivers.
- **Consideration of context**: While LIME treats explanatory words independently of each other, topicLIME provides contextual information by considering word dependencies. Therefore, topicLIME explanations reveal an interdependent and coherent internal structure.
- **Mode of presentation**: While LIME presents words as the explanatory features in a flat list-format, topicLIME explanations can be regarded as more hierarchical as they include topic-represented context and additionally drill down with regard to granularity by including the words that have been assigned to the respective topics.

- **Quality of the explanatory features**: The two explanation modalities do not only differ in terms of presentation of the explanations, but also in terms of the included explanatory features. Due to a more realistic local perturbation distribution, from which the local neighborhood is sampled, often more representative words with regard to the true inductive local behavior of the classifier to be explained are included in explanatory topics generated by topicLIME compared to explanatory words identified by LIME. For this reason, topicLIME is less prone to extrapolation as the underlying local neighborhood of the document to be explained is more realistic with regard to the generative process - as estimated by LDA - from which the corpus' documents originally have been created. As a consequence, both the surrogate explanation models itself and the identified words included in topicLIME's explanatory topics reveal a higher local fidelity to the classifier compared to LIME, as the next subsection shows.

At this point in time, the reader is referred to the second publication associated with this dissertation for the technical details of topicLIME (Kiefer 2022). The publication is included in appendix A.2.1. Section 3.5 of the second publication can be referred to as an intermediary between concept-based explainable ML, described in this chapter of the synopsis, and explanatory interactive ML, as introduced in chapter 4 of the synopsis. It strives for answering research question 2c from section 1.3. It introduces an algorithm called *Semantic Interrogations* that enables human ML users to explicate their conceptual domain knowledge, represented as topic distributions as a proxy, translate it to computational representations and compare it with the results of a classifier. The exploitation of the generative process of the textual input domain to create documents revealing a specific semantic content and semantic structure provides the foundation for incorporating conceptual user feedback into a text classification system, based on comprehensible and scrutable predictions.

3.4.3 Local Fidelity of Surrogate Explanation Models

This subsection shortly summarizes results achieved during the experiments conducted in the publication attached in appendix A.2.1. It also includes some further results for the dataset *AG News* that have not been published in that publication. In order to evaluate topicLIME with regard to the question whether the learned local surrogate explanation models constitute a good approximation of the classifier's prediction locally, different measures for analyzing local fidelity have been adapted and developed.

Local fidelity is said to be achieved if an explanation model $g \epsilon G$ is found such that $f(z) \approx g(z')$ for $z, z' \epsilon Z$, where Z constitutes the vicinity of x and f is the model to be explained. In this thesis, Mean Local Approximation Error, see equation (3.6), and MeanR^2, see equation (3.7), are used as a proxy to measure local fidelity of the explanation models to be compared.

$$MLAE = \frac{\sum_{i=1}^{N} |f(x_i) - g_i(x_i)|}{N}. \tag{3.6}$$

$$MeanR^2 = \frac{\sum_{i=1}^{N} R^2(g_i)}{N}, R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} (f(z_i) - g(z_i'))^2}{\frac{1}{n}\sum_{i=1}^{n} (f(z_i) - f_{mean})^2}. \tag{3.7}$$

In both cases, N is the number of instances in the associated test dataset.

In order to also measure local fidelity of individual explanations, a new measure called *Combined Removal Impact* has been designed. It constitutes a measure inspired by the *Area Over The Perturbation Curve* (AOPC)[6] (Nguyen 2018). CRI is defined as follows:

$$CRI = \frac{1}{N}\sum_{i=1}^{N} p(\hat{y}|x_i) - p(\hat{y}|\tilde{x}_i^{(k)}), \tag{3.8}$$

where the top *k%* explanatory features are removed from x_i to yield $\tilde{x}_i^{(k)}$, $\hat{y}$ denotes the predicted label for x_i, and N is the number of instances in the associated test dataset.

Overall, three different text classification datasets, namely Reuters R52 dataset (Davìd Lewis 1993), 20 Newsgroups dataset (20NG 1995), and AG News dataset (X. Zhang, Zhao, and LeCun 2015), have been evaluated with regard to those measures in order to compare LIME and topicLIME in terms of local fidelity. The results are summarized in the tables 3.2, 3.3, 3.4 and 3.5. For details on how the individual measures have been applied within the experimental settings, please refer to section 4 of the publication included in appendix A.2.1.

Table 3.2: Comparison of LIME and topicLIME with regard to local fidelity via Mean Local Approximation Error and $MeanR^2$ for Reuters R52 dataset.

	MLAE			**MeanR^2**		
	Lime	**topicLime**	**Delta**	**Lime**	**topicLime**	**Delta**
XGBoost	0.0195	0.0076	-61%	0.864	0.951	+10%
Log. Regr.	0.0545	0.0409	-25%	0.697	0.847	+21%
SVM	0.0371	0.0277	-25%	0.733	0.862	+17%

[6] In the second publication associated with this doctoral thesis, the measure was referred to as the Area Over The Perturbation Curve. Although it is strongly related, Combined Removal Impact is a term better describing the underlying functionality. Therefore, the measure is referred to as CRI in the progress of this thesis.

Table 3.3: Comparison of LIME and topicLIME with regard to local fidelity via Mean Local Approximation Error and $MeanR^2$ for 20 Newsgroups dataset.

	MLAE			**MeanR^2**		
	Lime	**topicLime**	**Delta**	**Lime**	**topicLime**	**Delta**
XGBoost	0.0071	0.0032	-55%	0.755	0.798	+6%
Log. Regr.	0.0321	0.0169	-47%	0.692	0.738	+7%
SVM	0.0240	0.0156	-35%	0.733	0.765	+4%

Table 3.4: Comparison of LIME and topicLIME with regard to local fidelity via Mean Local Approximation Error and $MeanR^2$ for AG News dataset.

	MLAE			**MeanR^2**		
	Lime	**topicLime**	**Delta**	**Lime**	**topicLime**	**Delta**
XGBoost	0.0394	0.0342	-13%	0.863	0.884	+2.5%
Log. Regr.	0.0736	0.0678	-8%	0.732	0.761	+4%
SVM	0.06342	0.0602	-5%	0.785	0.808	+3%

The evaluation reveals that topicLIME consistently performs better in terms of local fidelity of the surrogate explanation models. On the one hand, a lower Mean Local Approximation Error of topicLIME indicates that the classifiers are better approximated in the locality of the instances to be explained by using a more realistic local neighborhood. On the other hand, higher $MeanR^2$ values achieved by topicLIME show that topic-encoded features harnessed within the locally approximated interpretable linear models can be considered as more predictive of the dependent class.

Table 3.5: Comparison of LIME and topicLIME with regard to local fidelity via Combined Removal Impact for the three datasets. As classifier, XGBoost is used throughout the experiments.

	CRI		
	Lime	**topicLime**	**Delta**
Reuters R52	0.271	0.302	+11%
20 Newsgroups	0.097	0.112	+15%
AG News	0.229	0.277	+21%

Besides higher fidelity of the surrogate models, higher CRI values indicate that the removed explanatory features are considered more important by the classifier for the class decision at hand. Therefore, topicLIME explanations itself reveal higher local fidelity compared to LIME explanations.

3.5 Empirical Evaluation of Topic-based Explanations

The following empirical evaluation that addresses research question 2d from section 1.3 has been conducted as part of this doctoral research. It has not been published yet, but is planned for publication in the near future.

For the validation of the capabilities of concept-based contextual explanations, two user experiments in within-subjects design have been conducted that compare word-based and topic-based explanations. The overall question was to analyze how the different explanation modalities are perceived by humans that interact with a text classifier that predicts which class, e.g., content category, a presented document belongs to. First, the different explanation modalities were compared with regard to perceived helpfulness for understanding a certain prediction performed by a text classifier. The according task is referred to as preference selection task in the remainder of this section. In a second study, participants had to choose the class they assumed the classifier to predict most likely for a given text based on different explanatory features generated by the two modalities. In addition, the attendees were asked to indicate their level of confidence towards the question whether the chosen class equals the class the classifier would predict based on the provided explanations. The according task is referred to as forward prediction task in the remainder of this section. In both experiments, the classifier was instantiated with a logistic regression model. The according LDA model that was used in the studies comprised K=19 topics, which constituted the global optimum for C_v coherence.

3.5.1 Preference Selection Task

In the preference selection task, participants were presented ten different text documents from the Reuters R52 dataset along with the class of the documents as determined by a text classifier. Additionally, two different kinds of local explanations for why the classifier chose a certain class for a specific document were shown. On the one hand, standard LIME explanations made of single independent word attributions were presented. On the other hand, contextual explanations that are made of higher-level concepts represented by topics and its assigned topic-related words were offered. The participants of the first experiment were then asked, per presented document, to select the explanation modality they perceived most helpful for understanding why the classifier made the according prediction, which is referred to as comprehensibility of individual predictions in the context of XAI. Finally, the attendees should indicate their general preference with regard to the different explanation modalities at the end of the study. Precisely, it was asked, whether they preferred independent words or groups of related words as explanations and whether they preferred fewer or more bars in the visualization of the explanations.

Hypothesis and Experimental Design

According to insights from psychology, it was assumed that more human-like explanations, like explanations whose elements reveal a coherent structure, will be preferred by human explainees. The baseline to which topicLIME explanations are compared are standard LIME explanations whose explanatory features lack an interrelated structure.

To test this hypothesis, ten documents, together with their predicted classes and the according explanations for those classes, were selected from the Reuters R52 dataset. All documents were labeled with one class out of the following set of five classes:

- Consumer Price Index (economical)
- Acquisition (economical)
- Ship (economical and political)
- Grain (agricultural)
- Sugar (agricultural)

Five different sequences, in which the documents are presented, were created for randomization of the stimuli. By that, it shall be ensured that the order of document presentation does not bias the participants' assessment of the explanations. Only true positives, such documents whose predicted class equals the true class, were part of the study. Along with each document, an attention check in the form of a yes-or-no question with regard to the text's content was conducted in order to ensure that the participants carefully read the text.

The study was distributed online via the University of Bamberg. 27 participants took part, out of which 14 were females, twelve were males, and one was diverse. The participants were between 15 and 54 years old and differed in their fields of study. Overall, ten attendees studied subjects related to computer science, like applied computer science, software engineering, computing in the humanities or information systems. 13 participants were students of the overall subject psychology, four of the participants were no students at all. After the complete study was finished, those participants' answers to such documents, where the attention check was not passed, were excluded.

Figures B.1, B.2, B.3, and B.4 of appendix B.1 depict an excerpt of the conducted study. Summing up, a within-subjects study design was applied. The number of explanation units (the number of word attributions for LIME explanations and the number of topic attributions for topicLIME explanations) was varied between different text documents, but the number of words was kept the same across the conditions for a single stimulus.

- **Word-based LIME explanations**: In this condition, the participants were presented with standard LIME explanations. Thus, the top n ($8 \leq n \leq 16$)

most important explanation units, which are the words with the highest attribution w.r.t. a single classification's confidence, including their LIME scores were shown to the participants. Only positive evidences, which are the words with positive attribution for the given class, were included. The explanations were presented visually with the help of bar charts.

- **Topic-based explanations**: This condition comprised explanations whose explanation units were each made of a set of coherent words together with the according positive topicLIME scores. The total number of words distributed over topics that were included in the explanation was kept equal to the number of words from the word-based condition per stimulus. Also, this explanation modality was presented visually to the participants using bar charts.

After the ten document-related questions, the participants were asked for their general preferences with regard to the different explanation modalities (refer to figure B.6 of appendix B.1). The participants could answer with the help of a 5-point Likert scale.

Analysis and Results

In order to analyze the preferences of the attendees with regard to the type of explanations per presented document, a descriptive analysis was performed. Please refer to figure B.5 of appendix B.1 for the results. The results suggest ambiguous preferences as, on average, around 53.3% of the participants preferred the concept-based explanations, with a high standard deviation of 0.28. As further analysis, the relationship between the document-related questions and the general questions regarding the explanation-preference was studied. It was analyzed whether participants who preferred topic-based explanations in the document-specific preference selection task also indicated a general preference towards groups of related words as explanations in contrast to single independent words. With the help of a linear regression model, it turned out that participants who found topic-based explanations more helpful were significantly more likely to prefer explanations in the form of related words in groups ($R^2 = .11$, $F(1, 213) = 25.37$, $\beta = 1.09$, $p < 0.001$, 95% CI [0.66, 1.52]).

Additionally, it was analyzed whether participants who preferred topic-based explanations in the document-specific preference selection task also indicated a general preference towards fewer explanation units in contrast to more individual explanation units. With the help of a linear regression model, it turned out that participants who found topic-based explanations more helpful were significantly less likely to prefer explanations with more individual explanation units ($R^2 = .02$, $F(1, 213) = 3.91$, $\beta = -0.46$, $p = 0.049$, 95% CI [-0.909, -0.002]).

In summary, people who in general stated to prefer explanations consisting of fewer explanation units made of related words tended to prefer the topic-based explanations generated by topicLIME.

3.5.2 Forward Prediction Task

In the forward prediction task, participants were presented short text documents from the Reuters R52 dataset. For some of those, explanations of the classifier's local reasoning were provided. Some explanations comprised the words of standard LIME presented in a flat list. Other explanations were made of the words extracted from the contextual explanations generated by topicLIME, also flattened into a list format. Please refer to figure B.7 of appendix B.1 for more information on the way the explanations were transformed and presented. During the study, the attendees had to select the class the classifier would predict based on the explanation, along with the level of confidence that this is the prediction the classifier would make based on the presented explanation.

Hypothesis and Experimental Design

Due to the modified and therefore improved process of neighborhood generation resulting from the consideration of feature dependencies during sampling, topicLIME often identifies a different set of top words included in the topics compared to the top words identified by standard LIME (refer to subsection 3.4.2). It could be shown quantitatively that the top n words of topicLIME explanations itself are jointly capable of better describing the actual behavior underlying a text classifier's local reasoning, independently of the way the words have been grouped. As such, topicLIME explanations reveal higher local fidelity compared to LIME explanations as the identified top words cause a larger drop in the classifier's confidence towards the predicted class when being removed.

As hypothesis for this experiment, it was assumed that humans receiving topicLIME explanations are enabled to better anticipate the local predictive behavior of a classifier while being more confident towards their estimation.

As the first experiment could not identify a general human preference towards a single explanation modality, primarily with regard to the way of presenting explanations, the second experiment constitutes an extension of the first one and confronts attendees with a forward prediction task. It is designed to work as a proxy for evaluating whether the explanatory features itself lead to a better human understanding of why the classifier made a certain prediction, independently of the way they are presented. It is assumed that explanations allowing participants to more effectively and with higher confidence solve such a kind of forward prediction task in turn lead to better local comprehension of the predictions made. In such a case, the higher local fidelity of the topicLIME explanation model and its explanatory features would push participants to develop a better local understanding of the classifier.

To implement the forward prediction task, fifteen documents from the Reuters R52 dataset belonging to nine classes were selected. The set of possible classes is described below:

- Acquisition (economical; used as true positive class during the active phase of the forward prediction task)
- Cocoa (agricultural; used as true positive class during the active phase of the forward prediction task)
- Consumer Price Index (economical; used as true positive class during the training phase and as distractor during the active phase of the forward prediction task)
- Gold (economical and political; used as true positive class during the active phase of the forward prediction task)
- Grain (agricultural; used as true positive class during the training phase and as distractor during the active phase of the forward prediction task)
- Interest (financial; used as true positive class during the active phase of the forward prediction task)
- Job (economical and political; used as true positive class during the active phase of the forward prediction task)
- Ship (economical and political; used as true positive class during the training phase and as distractor during the active phase of the forward prediction task)
- Sugar (agricultural; used as true positive class during the training phase and as distractor during the active phase of the forward prediction task)

Again, only true positives and only positive evidences were part of the study and along with each document, an attention check in the form of a yes-or-no question with regard to the text's content was conducted in order to ensure that the participants carefully read the text.

The experiment started with a short training phase in which the participants should get familiar with the way the classifier predicts classes for documents of the Reuters R52 corpus. The participants first received information on the set of possible classes for the documents contained in the study. Next, five documents together with the classifier's prediction and an attention check were presented.

The study was again distributed online via the University of Bamberg. 29 participants took part, out of which 16 were females and 13 were males. The participants were between 15 and 54 years old and differed in their fields of study. Overall, six attendees studied applied computer science, 13 participants were students of the overall subject psychology, and ten of the participants were no students at all or could not clearly be categorized. Six persons already took part in the preference selection task study.

Figures B.8, B.9, B.10, B.11, B.12, and B.13 of appendix B.1 depict an excerpt of the conducted study. Participants were asked to select one class, out of four,

the classifier would most probably predict for a given document based on one of three conditions (LIME explanation, topicLIME explanation, no explanation provided). Additionally, the attendees should indicate their confidence that this is the prediction the classifier would make based on the presented explanation. They could answer with the help of a 5-point Likert scale, ranging from 0 (not confident at all) to 5 (very confident). In the forward prediction task, five documents were accompanied with LIME explanations, five with topicLIME explanations, and five did not include any explanations at all. The number of explanation units *n*, representing the number of words contained in the flattened list, was varied between different text documents ($6 \leq n \leq 18$).

Analysis and Results

A descriptive analysis was performed visualizing how many participants predicted the correct label based on the different types of explanations (refer to figure B.14 of appendix B.1). According to the results, no clear trend could be identified towards a certain type of explanation. It is noticeable that participants were good at predicting the correct classes in cases where no explanations were provided at all. To statistically test whether different types of explanations lead to performance differences in predicting the label the classifier would choose, a mixed effects logistic regression model with random intercepts for each subject and item was computed. Adding the type of explanation as a fixed effect did not significantly increase the goodness-of-fit ($\chi^2(2) = 0.59$, $p = 0.75$). Therefore, it can be concluded that no differences in forward prediction performance could be found for different types of explanations.

Additionally, a descriptive analysis was performed visualizing the mean confidence of the participants towards the fact that their selected class equals the prediction the classifier would make based on the different types of explanations (refer to figure B.15 of appendix B.1). As before, no clear trend can be identified towards a certain type of explanation. To statistically test whether different types of explanations lead to differences in the confidence estimation, a mixed effects linear regression model with random intercepts was computed for each subject and item. Adding the type of explanation as a fixed effect did not significantly increase the goodness-of-fit ($\chi^2(2) = 1.62$, $p = 0.45$). Therefore, it can be concluded that no differences in confidence estimation could be found for different types of explanations.

To sum it up, the different types of explanations do not significantly differ in helping attendees to solve a forward prediction task. However, it is to be noted that huge effects should not be expected due to the high baseline modality (no explanation provided) to which the different types of explanations were compared. Obviously, participants could already perform the forward prediction task with high performance by just using the input texts and their ex ante knowledge on the reasoning of the classifier gained during the training phase.

3.5.3 Implications and Discussion

According to the results of both studies, the original hypotheses have to be rejected. On the one hand, the first study could not identify a general preference of human explainees towards concept-based explanations consisting of higher-level topics made of coherent words. On the other hand, the second study could not show that human receivers of topicLIME explanations are enabled to better anticipate the actual local predictive behavior of a text classifier while being more confident towards their estimation.

Nevertheless, some further interesting insights - in addition to the ones gained from the technical evaluation (refer to subsection 3.4.3) - regarding the newly developed explanation method called *topicLIME* can be gained from the synopsis of the two studies. It can be concluded that topicLIME does not significantly differ from LIME, meaning it is not inherently better or worse in terms of participants' performance achieved during the reference tasks. The results of the document-specific task suggest that preferences towards the different types of explanations are rather ambiguous. On average, around 53.3% of the participants preferred the concept-based explanations generated by topicLIME. The answers to the questions for general preferences were related to the behavior of the attendees in the document-specific task. Jointly considering these facts, a possible interpretation could be that individuals' preferences regarding the kind of explanations provided cannot be considered obvious. Some participants stated they preferred fewer and more coherent explanation units. Those were inherently satisfied with the explanations provided by topicLIME. Thus, the generation of topicLIME explanations as an alternative to LIME explanations can be considered useful, especially when striving for the consideration of the different needs of explanation receivers. Summing up, topicLIME can be considered as an explanation method that should be offered individually. Its explanations fulfill some of the requirements from psychology on human-friendly explanations and generate explanations with higher local fidelity (refer to subsection 3.4.3). On average, topicLIME explanations allow human explainees to perform the proxy tasks of this study as well as LIME explanations do.

For identifying more nuanced differences in the two modalities, larger sample sizes would be required in follow-up experiments. Additionally, topicLIME explanations might perform even better when dealing with longer text documents whose contents are structured by the composition of more complex sub-concepts.

3.6 Excursus: Model-agnostic and Receiver-dependent Explanations for Unsupervised Anomaly Detection

This section strives for answering research question 3 from section 1.3 framing it with the application context of anomaly detection for financial auditing. It shall be elaborated, how model-agnostic explanation techniques, like LIME, can be applied to unsupervised ML use cases, like anomaly detection or customer segmentation.

In addition, a suggestion is made on how to post-process explanations such that they are considered more selective and receiver-dependent. Examples from the domain of financial auditing are given in order to showcase the elaborated concepts more intuitively.

Financial auditing refers to the annual process of reassessing all transactions of a client with all business partners. Theoretically, a complete examination must be performed by an auditor in order to be able to make a reliable statement. For economic reasons, only a selective approach to an audit is feasible, where often 95 percent or even less is referred to as reasonable assurance. For this reason, an auditor conducts a selection test in the form of a sample which leads to an incomplete search for anomalies. Typically, such anomalies result from incorrect, probably in the sense of negligent, accounting or from violations in the sense of deliberate fraud.

During this doctoral research, an ML architecture has been elaborated for detecting anomalous datapoints in the absence of ground truth for the domain of financial auditing. In the third publication (Kiefer and Pesch 2021), an ensemble-based approach is introduced that shall support financial auditors by detecting anomalous datapoints, so-called punctual anomalies, in the absence of labeled data for training. The task to identify such anomalies can therefore be characterized as an unsupervised clustering task that involves dealing with unbalanced data, as anomalies are defined as subsets of data whose characteristics differ from expected values and as such from the remainder of the data (Chandola, Banerjee, and Kumar 2009; Mehrotra, Mohan, and Huang 2017). The overall architecture of the approach is depicted in figure 3.8. At this point in time, the reader is referred to the third publication for the technical details on how the anomaly detection approach has been designed and implemented (Kiefer and Pesch 2021), which is included in appendix A.2.2.

Having identified anomalous datapoints, end users, like financial auditors, might ask the question why the unsupervised system made certain decisions and what input parameters influenced the model's predictions.

Out of the box, model-agnostic explainers, like the perturbation-based approaches LIME or SHAP, only work for classifiers trained in a supervised manner. For being able to apply those approaches to unsupervised ML tasks, like clustering or association tasks, with a variety of use cases, like recommending, anomaly detection, or customer segmentation, this doctoral research proposes to approximate the original unsupervised clustering with a supervised classification. Taking the predictions of the initial clustering process as labels for the input data and synthetically oversampling the minority class comprising all datapoints identified as outliers, traditional supervised classification techniques, like random forests, support vector machines or artificial neural networks, can be applied. Those act as a global approximation of the unsupervised model to be explained. Doing so, standard model-agnostic explainers, like *tabularLIME*, can be harnessed to extract local feature attributions as explanations. For the technical details regarding oversampling and supervised approximation, refer to (Kiefer and Pesch 2021).

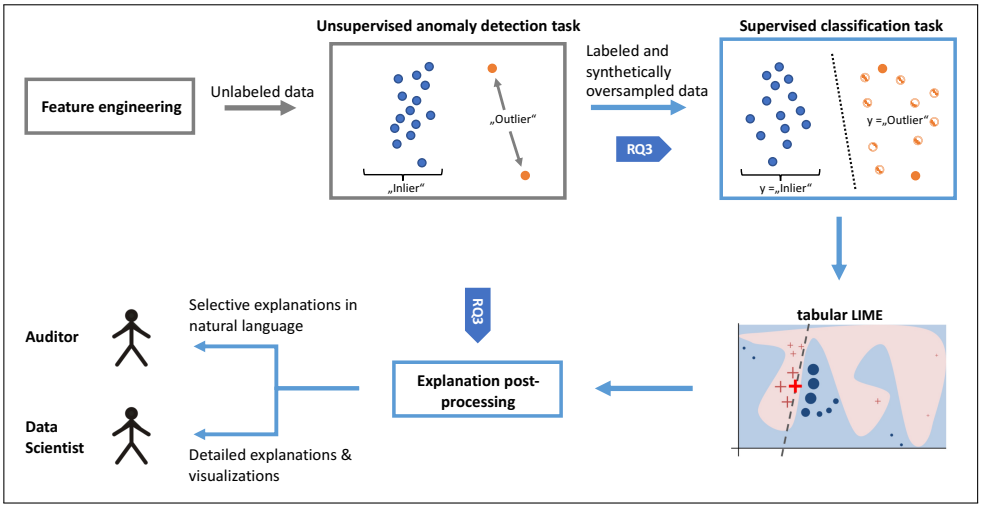


Figure 3.8: Unsupervised anomaly detection for the domain of financial auditing combined with Local Interpretable Model-agnostic Explanations (LIME) and explanation post-processing. The blue colored parts of this figure indicate the explicit contributions of the research underlying this chapter's excursus. The proposed architecture and its instantiation constitute the main contributions to research question 3 from section 1.3.

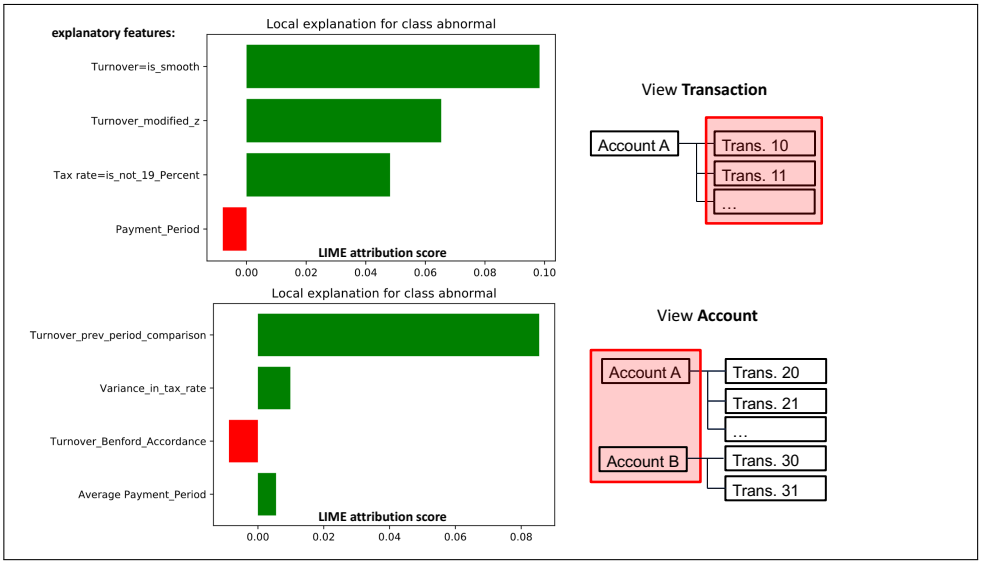


Figure 3.9: Detailed explanations for expert AI users.

Figure 3.9 visualizes the raw output when applying *tabularLIME* to the data underlying a DATEV use case called *anomaly detection for financial auditing*.

It comprises explanations for detected outliers for two different views, namely *transaction* and *account*. The use case itself, the according views for detection of outliers as well as the used features are further detailed in (Kiefer and Pesch 2021). For each explanatory feature that was identified as positive evidence for the decision towards the class *anomaly*, tabularLIME outputs the type of attribution. Green bars indicate that a feature attributes as positive evidence in the form of high values and red bars indicate feature attribution as positive evidence in the form of low values. Work from (Tomsett et al. 2018) argues that such explanations are well-suited for explainees like developers (e.g., data scientists) as they are typically interested in understanding the underlying models and data in terms of technical details.

As described earlier in this thesis in section 2.2, explainability and interpretability are similar, but not identical concepts. Thus, it is not enough to only focus on how to generate explanations when striving for human-friendly and receiver-dependent explanations, as interpretability depends on the explainee, the recipient of explanations, while explainability is only related to the XAI system as the technical explainer. Human-friendly explanations are typically selective and as such focus on the main causes of a decision with n, representing the number of causes, preferably being $1 \leq n \leq 7 \pm 2$ as stated by Miller's law (G. A. Miller 1956). Standard LIME, according to (Molnar 2022), partially accounts for human-friendly explanations. When using a Lasso model within LIME, the interpretable surrogate model from which the feature attributions are later extracted as explanation units can be forced to perform training with exactly N features. As N can only be determined a priori in a static manner, the identification of the most important causes is not possible. As an example, refer to the left part of figure 3.7, where $N = 9$, but only either one or two explanatory words, namely "gold" as positive evidence and maybe "acquisition" as negative evidence for the predicted class "gold", constitute major causes of the prediction.

As a solution, this research suggests to use explanation-post-processing for identifying the major explanatory features of an explanation dynamically. It proposes a concept called *relative explanation selectivity* that allows to dynamically choose N in a truly selective manner. Figures 3.10 and 3.11 depict two possible realizations of relative explanation selectivity. Both have in common, that the absolute feature attributions as produced by LIME first need to be sorted in descending order. As first option, this thesis suggests to use a relative weight threshold (see figure 3.10), which constitutes a parameterized relative threshold. All features that have a higher relative attribution, compared to the first one that is the most important one, than the relative weight threshold are considered relevant. Therefore, all features whose feature attributions are above the threshold, indicated by the blue colored area in figure 3.10, are included in the selective explanation.

Another option is to use methods that, given a set of x and y values, return knee points, for concave functions, or elbow points, for convex functions, of the respective function. One method that is based on regression lines through n adjacent points, has been proposed in (Kiefer and Pesch 2021). Another, more

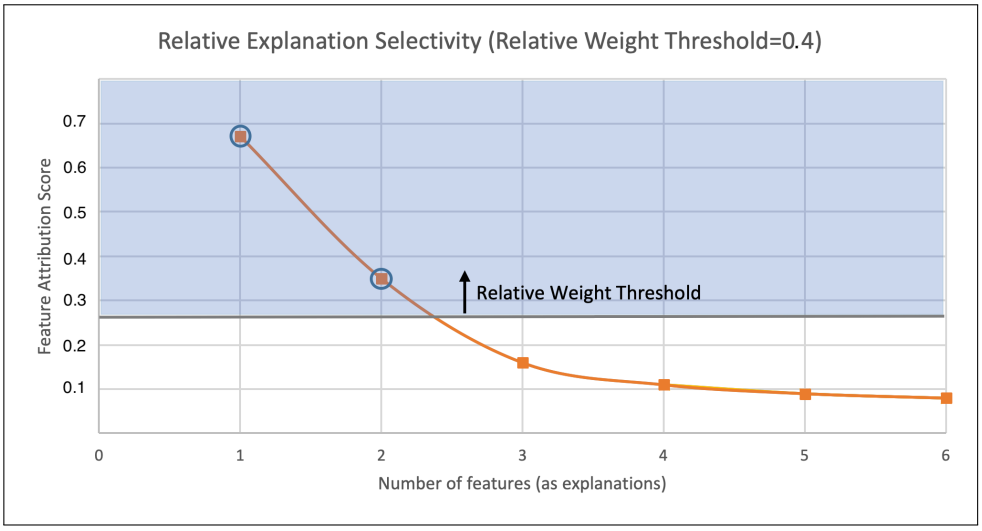


Figure 3.10: Relative explanation selectivity using relative weight threshold.

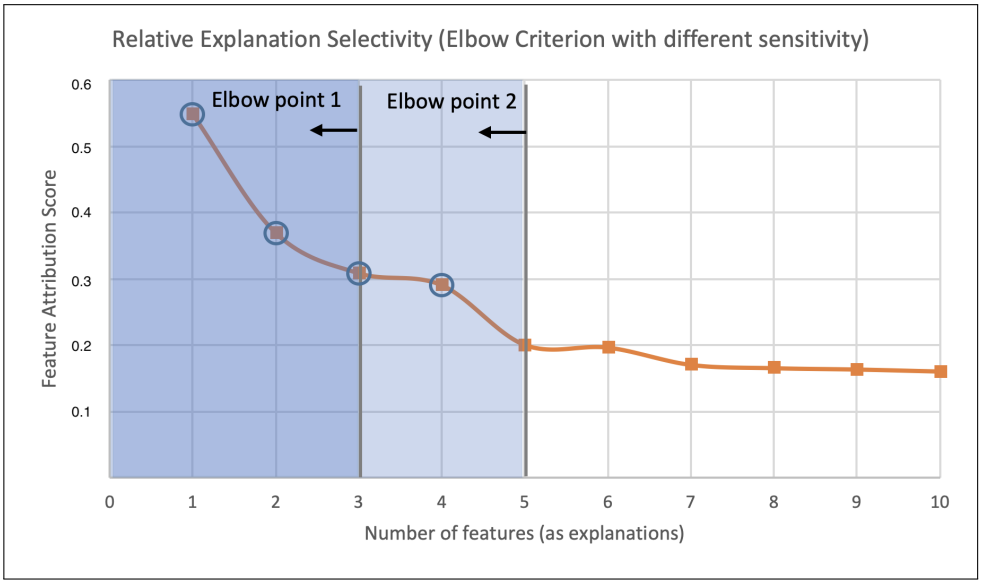


Figure 3.11: Relative explanation selectivity using elbow points.

sophisticated method is to use the *kneedle* algorithm for identifying knee or elbow points (Satopaa et al. 2011). After specifying several hyper-parameters, like the direction of the curve and sensitivity, kneedle automatically detects knee or elbow points by identifying points with maximum curvature. Figure 3.11 depicts the identification of elbow points using two different levels of sensitivity, where smaller values for sensitivity detect elbows at an earlier stage. Finally, all features located left of the elbow points are included in the selective explanation.

In contrast to developers, end users, which can be further differentiated into end user decision makers, like financial auditors, or affected end users (Hind et al. 2019), either use explanations to inform their decisions or to evaluate whether certain predictions can be trusted (Belle and Papantonis 2021). As such, those groups commonly can be characterized as non-experts with regard to ML knowledge and therefore prefer selective explanations, at best presented in coherent natural language. To achieve this, the raw explanations are post-processed in two ways. First, the explanation is reduced to the main causes in order to make it more selective. Therefore, either a relative weight threshold or methods for detecting elbow points are applied as described in figures 3.10 and 3.11. Second, the raw explanations are transferred to natural language explanations as this can lead to various advantages, like higher efficiency (Alonso et al. 2017) and coverage (Sokol and Flach 2018) regarding the explainees. For demonstration purposes, a simple template-based approach for verbalizing the raw explanation units has been applied in this research. By now, a bunch of more sophisticated methods for generating natural language explanations is available as reviewed by (Cambria et al. 2023). The resulting explanations designed for financial auditors are shown below. The number of explanatory features has been reduced from four to three for the *transaction*-view and from four to one for the *account*-view by applying the elbow method.

- **Human-friendly explanation for view *transaction***: "This transaction might differ from the remainder. It comprises a smooth and comparably high turnover and a tax rate which is not 19 percent!"
- **Human-friendly explanation for view *account***: "This account might differ from the remainder. It comprises a high turnover compared to the previous period!"

4. Concept-based Explanatory Interactive Machine Learning for Text Classification

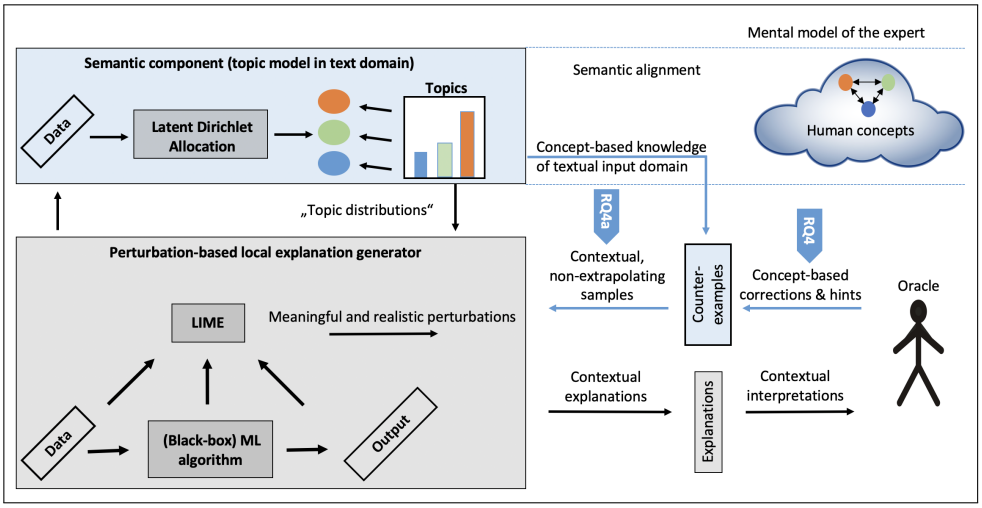


Figure 4.1: Semantic explanatory interactive learning: the blue colored parts of this figure indicate the explicit contributions of this chapter. The proposed architecture and its instantiation constitute the main contributions to research question 4 from section 1.3.

The following chapter describes a novel interaction framework for the domain of text classification that further places the human in the loop and allows for constructive and concept-based human corrections. First, some background information on the overall research area explanatory interactive machine learning is provided. Next, the theoretical foundations and drawbacks of CAIPI that the newly developed topic-based approach for constructive corrections builds on and compares to are introduced. Subsequently, this chapter provides information on desiderata for efficient interactivity between humans and machines known from disciplines like psychology or the social sciences. The new algorithm called *Semantic Push* is introduced in an illustrative way by depicting its functionality using a graphical model and a visual conceptualization. For technical details, the reader is referred to the publication (Kiefer, Hoffmann, and Schmid 2022). In addition, topic-based corrections are compared to corrections provided by CAIPI, both in terms of its implications on predictive performance and local explanation quality.

4.1 Explanatory Interactive Machine Learning

Figure 2.3 in section 2.3 depicts the derivation of explanatory interactive machine learning. It strives for integrating the two components related to the overall discipline human-in-the-loop ML, namely providing humans with explanations of ML decisions and having humans participate in the learning process in a co-adaptive way. Approaches to explanatory interactive machine learning therefore integrate techniques from XAI, active learning and interactive learning. Active learning constitutes an attempt to maximize an ML model's performance while requiring a minimum of human-annotated samples (Ren et al. 2021). In contrast to active learning, where the ML system remains in control of the learning process and uses humans as oracles for annotating specific samples (Mosqueira-Rey et al. 2022), interactive ML goes a step further and provides humans with more control. It harnesses active learning as a basis, integrates techniques from HCI, and allows a human designer to train, correct, and teach a model until a sufficient performance is met (Fails and Olsen 2003), albeit independently from the concepts explainability and interpretability. The integration of all the approaches mentioned above leads to explanatory interactive machine learning, a term coined by (Teso and Kersting 2019). The authors combine the local and model-agnostic explanation method LIME with an (inter-)active learning setting and propose a method called *CAIPI*. Please refer to section 4.2 for an introduction of CAIPI. Explanatory interactive ML shall enable users to iteratively integrate corrective feedback into a model after having analyzed its decisions, both in a model-agnostic way.

Recently, a term called "algorithm-in-the-loop" has been introduced in this context (Green and Y. Chen 2020). In contrast to classical human-in-the-loop-learning approaches, like active learning, more interactive ML approaches allow users to modify data or features in a more dynamic manner (Fails and Olsen 2003; Gillies et al. 2016). Algorithm-in-the-loop-learning targets non-technical experts, like end user decision makers or affected end users, treats the learning algorithm as "part of a human design process" (Gillies et al. 2016) and therefore puts humans into the central focus of interactive decision making (Green and Y. Chen 2020).

Freely altering or generating training data is proven to have several advantages compared to solely providing data labels (R. A. Fiebrink 2011). On the one hand, users can provide corrective examples in the case of system errors. On the other hand, it is possible to perform changes in a system's behavior as domain knowledge evolves (Gillies et al. 2016). As an example, the reasoning of a supervised classification system to predict a certain class can be adapted over time by altering the decision boundary of the classifier using user-specified training data. Doing so, users are enabled to define a desired model behavior in an efficient way requiring only few data (R. A. Fiebrink 2011). Overall, interactive ML approaches that support interpretation and produce scrutable results enable users to critique learner outputs on the basis of a deeper understanding (Amershi et al. 2014).

In general, learning from explanations has first been used in concept learning (Mitchell, Keller, and Kedar-Cabelli 1986; DeJong and Mooney 1986) and proba-

bilistic logic programming (Kimmig, De Raedt, and Toivonen 2007), mostly in a model-specific way as logic-based models have been considered. More recently, similar methods have been proposed, such as approaching interactive learning with mutual explanations, especially in relational domains. Research conducted by (Schmid 2021) suggests using Inductive Logic Programming (ILP) in combination with verbal explanations generated from Prolog rules in order to be expressive enough to capture information on relations between different aspects or sub-concepts. Based on mutual explanations, a learner's reasoning shall be correctable not only in terms of the classes it predicts, but also with regard to its explanations.

By now, there is a small amount of approaches that in part contribute to the field of explanatory interactive ML in the context of text classification, albeit in a poorly integrated manner. Work from (Heimerl et al. 2012) shows the development of three active learning approaches with various degrees of human involvement for the task of text document retrieval. Besides a basic active learning approach, this work additionally integrates a visual method that allows users to explore the classification context by projecting documents to low-dimensional space. Furthermore, a user-driven method is proposed where the users have full control over the selection of documents that shall be labeled. Although this work constitutes a first step towards interactive ML, it lacks the ability to provide the user with explanations of individual decisions that explainable interactive ML requires for further informed interactions.

The authors of (Savelka, Trivedi, and Ashley 2015) apply an interactive machine learning approach to allow statutory analysis, which is an important component of research on legal issues. In their approach, a support vector machine with a linear kernel is used for classification. Together with a graphical user interface, the system provides users with information on its classification confidence, highlights globally prominent features, and allows the user to suggest relevant features. In this manner, the approach constitutes a further step towards explainable interactive ML, albeit it is solely working in a model-specific way applicable only to support vector machines. Additionally, it is not capable of generating local explanations, which are preferably used when dealing with end users with limited technical experience. Other examples of that category are systems such as EluciDebug (Kulesza et al. 2015) or Crayon (Fails and Olsen 2003) that use feedback based on explanations to adapt a learner, albeit model-specifically.

There is further research that is more focused on identifying interpretable data representations for high-dimensional textual data as a basis for efficient and human-friendly interactivity (Kim et al. 2015). The authors primarily harness LDA as feature compression in order to prove statistically significant high human interpretability of a textual input domain, both in terms of conventional metrics and human subject experiments. Although they do not explicitly use those intermediate concepts as means for enabling interactivity between humans and ML models, they provide the outlook that the intermediate topic layer can offer semantically meaningful information to humans. Those could be used for aiding "human reasoning about data characteristics in connection with compressed topic space"

when interacting with ML models regarding tasks like interpretation, evaluation and debugging (Kim et al. 2015).

There is a bunch of research that proposes feature supervision (Raghavan, Madani, and Jones 2006; Raghavan and Allan 2007; Druck, Mann, and McCallum 2008; Druck, Burr Settles, and McCallum 2009; Burr Settles 2011; Attenberg, Melville, and Provost 2010) and rationales (Zaidan, Eisner, and Piatko 2007; Zaidan and Eisner 2008; Sharma, Zhuang, and Bilgic 2015), both for label- and feature-level supervision, in order to improve learning efficiency. Although it is shown that providing rationales can be efficiently performed by human annotators, most of those approaches reveal a black-box nature, do not provide any explanations, and force the learner and the user to interact by using the same features, which is often hard to accomplish.

Summing up, for the domain of text classification, most research focuses on standard active learning approaches to achieve interactivity with humans while omitting interpretability and therefore scrutable results. If explanations are integrated in an interactive setting, they are mostly generated model-specifically and on a global level. Only a single work addresses the importance of generating human-interpretable concepts as intermediate layer, but misses to incorporate the achieved results into an interactive learning setting. Another method, called *CAIPI*, has been proposed (Teso and Kersting 2019) that combines model-agnostic local explanations with explanation corrections. However, it misses to include human-interpretable and contextual representations in the bidirectional explanations. As a consequence, there is a research gap that needs to be addressed. In order to achieve human-centered explanatory interactive ML, methods for enabling contextually meaningful and model-agnostic interventions based on scrutable results to correct learner mistakes or adapt a learner's local reasoning need to be elaborated.

As a solution, Semantic Push, a concept-based explanatory interactive ML approach for text classification, is introduced later in this chapter. It describes a method that is effective for translating concept-based corrections of humans, based on local model-agnostic and contextual explanations, to non-extrapolating training examples such that the learner's reasoning is pushed towards the desired behavior. Semantic Push in parts is build on the basis of CAIPI and is later evaluated with CAIPI as a baseline. As Semantic Push enables semantic alignment between humans and machines while generating explanations and at the same time allows the generation of user-defined counterexamples based on meaningful corrections, this combination is assumed to be well-suited as a basis for HCML.

4.2 CAIPI

As CAIPI constitutes the basis for concept-based explanatory interactive ML, which is instantiated by Semantic Push in this doctoral work, this section extends the description of CAIPI published in (Kiefer, Hoffmann, and Schmid 2022). CAIPI is an approach that integrates active learning with any model-agnostic local explainer of choice (Teso and Kersting 2019). In each iteration of interaction, the learner predicts

and explains a query instance to the user, who is able to respond by correcting the prediction and the according explanations in a model- and explainer-agnostic way. CAIPI enables users to correct a learner when its predictions are right for the wrong reasons by adding counterexamples in a "destructive" manner. In the domain of text classification, words that are falsely identified as relevant for a class decision are masked from the original document, then the resulting counterexamples recur as additional training documents. Algorithm 1 introduces CAIPI's basic functionality.

Algorithm 1 The CAIPI algorithm (Teso and Kersting 2019).

Require: A set of labeled examples L, a set of unlabeled instances U, and an iteration budget T.
  $f \leftarrow FIT(L)$
  repeat
    $x \leftarrow \mathbf{S}elect\,\mathbf{Q}uery\ (f, U)$
    $\hat{y} \leftarrow f(x)$
    $\hat{z} \leftarrow \mathbf{E}xplain\ (f, x, \hat{y})$
    Present $x, \hat{y}$, and $\hat{z}$ to the user
    Obtain y and explanation correction C
    $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow \mathbf{T}o\,\mathbf{C}ounterexamples(C)$
    $L \leftarrow L \cup \{(x, y)\} \cup \{\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$
    $U \leftarrow U \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^{c})$
    $f \leftarrow FIT(L)$
  until budget T is exhausted or f is good enough
  return f

As common in active learning, a set of initially labeled examples L and a set of unlabeled instances U as well as an iteration budget T are given. It is assumed that $|U| \gg |L|$. First, a learner f is fit on the set L before interaction with the user begins. At each interaction step, CAIPI selects a query instance from the set of unlabeled instances U based on a certain sampling strategy, like maximum classification uncertainty. Classification uncertainty is defined as $U(x) = 1 - P_\theta(\hat{y}|x)$, where x is the instance to be predicted and $\hat{y}$ is the most likely prediction. Intuitively, a query instance is selected for which the classifier f is maximum uncertain about its correct class, i.e., an instance lying near the decision boundary the classifier has learned so far. Next, the CAIPI algorithm presents the query to the learner f, which predicts the most probable class $\hat{y}$. Subsequently, the learner provides a local explanation $\hat{z}$ for its prediction using a model-agnostic local explainer, like LIME, and presents the triple $(x, \hat{y}, \hat{z})$ to the interacting user. Based on that, the user responds in a corrective manner by providing, if necessary, the true class y for the instance x or, in case of $\hat{y} = y$, an explanation correction C after having analyzed $\hat{z}$. At that point, it becomes clear that CAIPI is only capable of including corrective explanations in the case of predictions that are right for the wrong reasons. In that case, the human annotator is asked to "indicate the

components that have been wrongly identified by the explanation as relevant" (Teso and Kersting 2019). More formally, the correction set C is defined as $C = \{j : |w_j| > 0 \wedge \text{the user believes the } j\text{th component to be irrelevant}\}$. In the domain of text classification, C comprises words that the user judges to be irrelevant for the classification, but the explainer identifies those to be relevant for the learner. In order to feed the user-specified corrections C back to the learner, a method called "ToCounterExamples" converts C to a set of additional training examples $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$ such that the learner "unlearns" to depend on the irrelevant components. When classifying documents, counterexamples would simply be generated by masking irrelevant words from the original document. As the labels $\bar{y}_i$ of the counterexamples are kept identical to the prediction $\hat{y}$ which equals y in the "right for the wrong reasons" case, the learner "unlearns" the correlations between the irrelevant components and the label. The last steps of CAIPI involve adding the generated counterexamples to L, updating the set of unlabeled instances U as well as refitting the learner f on the updated set L.

4.3 Desiderata for Human-friendly and Efficient Corrections

In order to allow effective and efficient co-work between humans and ML systems by taking advantage of the richness of human expertise, the explanation-process must not be treated unidirectional. The social sciences explore the transactional nature of explanations between individuals and refer to explanations as an attempt to communicate understanding between social and interacting agents (Keil 2006; Chromik 2021). According to theories of interpersonal relationships developed in psychology (Simpson 2007), trust in an interaction, among others, depends on understandability and directability. The latter is defined as "the degree to which the trustor can rapidly assert control or influence when something goes wrong" (Hoffman, Johnson, et al. 2013). As a consequence, correctability of an ML system based on a previously developed understanding is required for modulating trust (Kulesza et al. 2015). According to insights from the social sciences, psychology and explicit research on interactive ML, some desiderata for human-friendly and efficient corrections based on bidirectional explanations can be derived and summarized as follows:

1. Human explanations are typically **contrastive**. Often, human explanations rely on causal factors explaining why a certain event occurred instead of an alternative event (T. Miller 2019; Lipton 1990). As a consequence, this thesis argues that human explanations used to perform meaningful corrections of a text classifier should be capable of including information on both the predicted class and the alternative classes, like the true class if the classifier made an incorrect prediction.

2. Humans tend to prefer interpretation of an input domain using **high-level concepts**, like topic-representations for the textual domain (Kim et al. 2015).

Research from (Holzinger, Malle, et al. 2021) states as a central hypothesis that "using conceptual knowledge as a guiding model of reality will help to train more explainable, more robust and less biased machine learning models, ideally able to learn from fewer data". Therefore, interaction strategies between humans and machine learners should not force both partners to interact using the same, often low-level features, although many state-of-the-art methods do so (Teso and Kersting 2019).

3. In research on interactive ML conducted by (Amershi et al. 2014) and (Odom and Natarajan 2018), it was elaborated that humans typically want to teach learners **by demonstration**, not solely by feedback. Therefore, a step from the "human-in-the-loop" towards the "algorithm-in-the-loop" should be done when developing new model-agnostic interaction strategies. In this manner, humans can teach learners by providing examples of a concept or by providing natural feedback, such as suggesting alternative or new features (Stumpf et al. 2007). As a consequence, this thesis argues that such constructive feedback is required instead of pure destructive feedback that prevents a learner from learning concepts based on an incorrect reasoning.

4. User input in interactive ML shall be **focused** and **incremental** (Amershi et al. 2014). As such, user input shall only affect a certain part or aspect of the model and shall only result in a small change. Therefore, locally weighted and continuous corrections based on local explanations could provide a promising basis for efficient interactivity.

5. This thesis states that interactions should be characterized by high **data efficiency** and high **data faithfulness** towards the statistical characteristics of the input domain. As in active learning, where instances to be labeled are efficiently selected by a preference mechanism in order to avoid the need for intense human supervision, a single human correction should be translated to $1 \leq n$ non-extrapolating training examples, where n constitutes a domain- and task-specific hyperparameter. In this way, human annotators would just need to specify the characteristics of the corrections, which would then automatically be transferred to several in-distribution counterexamples.

Summing up, efficient interactions with ML learners should enable humans, based on a local and contextual understanding of the prediction at hand, to iteratively formulate, based on background knowledge, and locally integrate concept-based corrections in a constructive manner. Corrections should be possible for both the scenarios, where the learner is right for the wrong reasons or completely wrong. Additionally, corrections should be capable of transporting corrective information for alternative classes of the downstream classification task. Lastly, corrections should be efficiently and faithfully incorporated into the classifier in a model-agnostic way.

4.4 Topic-based Approach for Contextual and Constructive Corrections

This section and its subsections introduce a topic-based approach for generating constructive corrections for text classifications in a model-agnostic way. The elaborated architecture naturally continues the architecture developed for concept-based XAI. As many of the techniques elaborated in chapter 3, like topicLIME or LDA, are reused in this approach, the central aspects of the proposed methodology are explained rather shortly.

The newly developed interaction method addresses the following common drawbacks that most current approaches to interactive ML for the domain of text reveal:

1. Typically, explanatory interaction methods enable humans to perform corrections based on **contextless explanations** that lack semantically meaningful information between the explanation units. The topic-based approach, in contrast, harnesses contextual and hierarchical explanations consisting of different higher-level topics that in turn are made of coherent words. Using locally faithful topic-based explanations, generated by topicLIME, as a basis for further interactions, humans shall be encouraged to respond to those explanations in a similarly meaningful way.

2. In analogy to explanations transferred from ML systems to humans, current interactive systems, like CAIPI, also lack the ability to include **contextual information in the corrections**. As such, CAIPI solely allows the generation of counterexamples by deleting certain words independently of the overall context of the associated document. Doing so, a text classifier might be trained on "out-of-distribution"-corrections stemming from counterexamples sampled from unrealistic local perturbation distributions, especially in cases of documents made of highly dependent words. This circumstance might lead to generalization errors. The topic-based approach overcomes this by allowing users to incorporate concept-based information into the corrections. Experts are now capable of gradually manipulating a certain document's concept composition by analogy to their conceptual knowledge. To achieve this, topic-based corrections rely on the capabilities of an LDA model that helps to constrain the corrections such that the statistical characteristics of the input domain are maintained.

3. As a central drawback, CAIPI solely allows for **destructive feedback**. It operates by deleting irrelevant explanatory words, i.e., those that have been incorrectly used for a certain prediction. As a solution, the newly developed approach enables constructive feedback by generating words from certain topics that should be informative for a specific class. Technically, this is possible by harnessing the generative process learned by the included LDA model.

4. Many interactive systems are limited to certain types of a **learner's reasoning and prediction errors**. As an example, CAIPI is designed to only address the situation in which a learner is "right for the wrong reasons". In contrast, the topic-based approach allows for local corrections of a learner's reasoning used to arrive at completely false predictions in the domain of text classification. In the computer vision domain, the authors of (Slany et al. 2022) also suggest improving CAIPI's interaction possibilities such that it is capable of dealing with wrong predictions.

4.4.1 Semantic Push

This subsection introduces the new interaction method called *Semantic Push* in an illustrative way. First, a graphical model is shown illustrating the objective of the approach. Next, a visual conceptualization is provided that depicts how Semantic Push is intended to work locally for class decisions under high uncertainty in the domain of text classification. Finally, an exemplary application to a text document of the Reuters R 52 dataset is presented that portrays the extensions that Semantic Push offers compared to CAIPI. In some parts, work already published in (Kiefer, Hoffmann, and Schmid 2022) is reused for the introduction of Semantic Push.

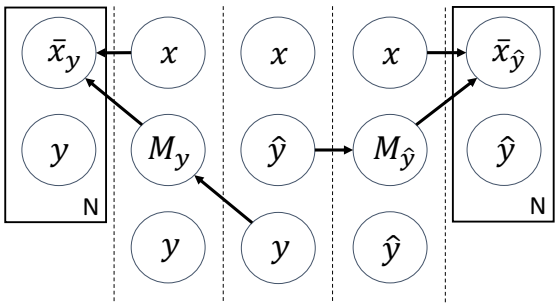


Figure 4.2: Graphical model of Semantic Push. This illustration is taken from (Kiefer, Hoffmann, and Schmid 2022).

Figure 4.2 introduces Semantic Push as graphical model. Let X and Y be the input and output space for a binary classification[7], where $x \in X$ represents a query instance, $y \in Y$ is the accompanying true label, and $\hat{y} \in Y$ is the predicted label. As an answer to a local explanation, the overall goal is to find a matrix M depending on the labels y or $\hat{y}$ that adequately incorporates human feedback into the classifier's reasoning in a model-agnostic way by generating counterexamples $\bar{x}$ based on x. Thus, we seek a set of L input manipulations $M = \{m_1, ..., m_L\}$

[7] This does not imply that Semantic Push is only applicable to binary classification scenarios. It is by design able to work in multiclass classification scenarios. For an individual prediction, however, the algorithm just needs to consider the true and predicted class as possible classes.

and a manipulation function $q : M \times X \to \bar{X}$. Here, $q(m, x)$ is a local function such that it only affects a part of the input x.

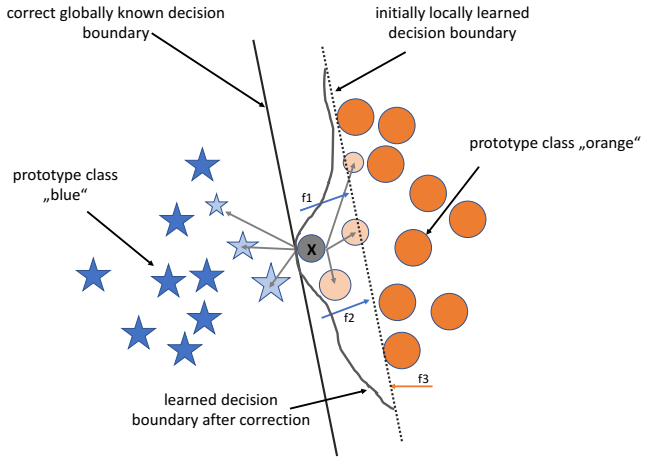


Figure 4.3: Conceptualization of Semantic Push: the grey query document in the middle is predicted as class "blue", but should be "orange" instead according to ground truth. Local explanatory features f1 and f2 (i.e., words or topics created by LIME or topicLIME) are used by the classifier locally to assign the query document to class "blue". According to expert knowledge (either real human expert knowledge or simulated knowledge for evaluation purposes), those features push the learned local decision boundary too far towards the class "orange". Feature f3, according to expert knowledge, is, among others, significantly used globally by the classifier to assign documents to class "orange". Semantic Push incorporates this expert knowledge by generating new documents, shown in light color, for both classes and eventually weighs them by their distance to the query document. The degree of locality of applying the potentially rather global expert knowledge to the query document can be customized via hyperparametrization to decide how focused the knowledge shall affect the classifier. Sampling new documents solely based on rather global expert knowledge might result in prototypical documents, potentially located in dense regions, which might not lead to a great benefit for the classifier. After transferring the expert knowledge to counterexamples in the form of further training documents accompanied by their true classes as labels, the classifier is retrained on those examples and the decision boundary is adapted accordingly. This illustration is taken from (Kiefer, Hoffmann, and Schmid 2022) and has been slightly adapted.

The setting and high-level functionality of Semantic Push can be depicted more illustratively by visualizing an individual local incorrect prediction as starting point for interaction. Please refer to figure 4.3 for a conceptualization of Semantic

Push. How interaction via Semantic Push works more practically for a specific document, is shown in figure 4.4.

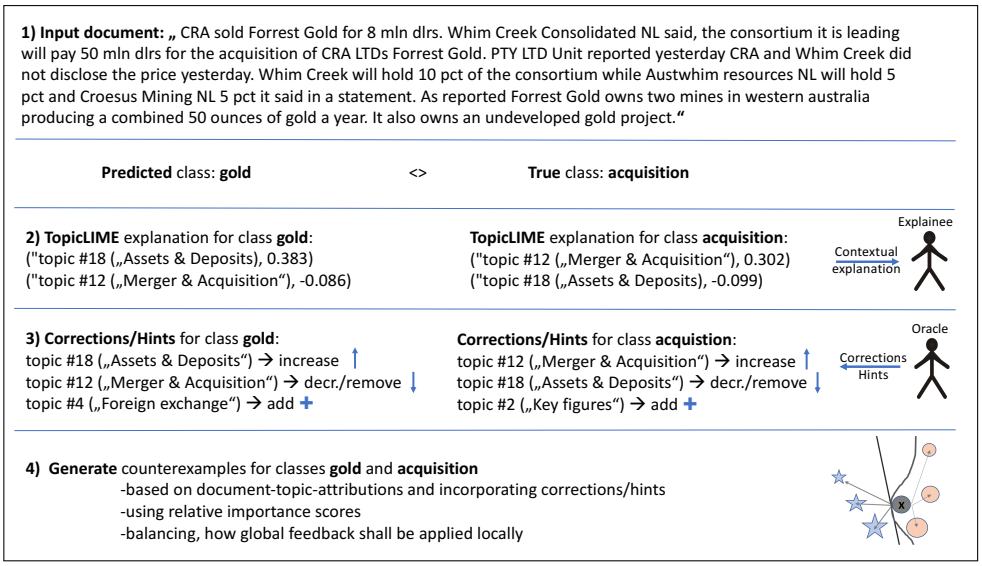


Figure 4.4: An exemplary application of Semantic Push to document ID 9 of the Reuters R 52 dataset. This illustration is taken from (Kiefer, Hoffmann, and Schmid 2022), it has been further elaborated in this thesis.

In comparison to CAIPI, Semantic Push not only enables human annotators to indicate and address (a) components that a learner wrongly identified as relevant -as CAIPI does -, but also (b) components that the learner has forgotten to learn, and (c) relevant components that have been used incorrectly.

At this point in time, the reader is referred to the fourth publication associated with this dissertation for the technical details of Semantic Push (Kiefer, Hoffmann, and Schmid 2022). The publication is included in appendix A.3.1.

4.4.2 Predictive Performance and Local Explanation Quality

This subsection shortly provides evaluation details and further evaluation results that mostly have not been part in the respective publication.

In order to efficiently evaluate Semantic Push, a simulated oracle that can be replaced by a human expert in a practical real-life scenario is used. Therefore, Semantic Push is based on a newly developed conceptual gold standard that works as a proxy for an expert's knowledge. Specifically, the gold standard contains concepts in the form of LDA-retrieved topics that should be informative for a specific class. Please refer to section 4.3 of publication (Kiefer, Hoffmann, and Schmid 2022) for further information on the used gold standard. The performance of Semantic Push is evaluated two-fold: first with regard to the predictive performance of downstream classification tasks using the macro-averaged F1 score and

second with regard to local explanation quality[8].

The F1 score is defined as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{4.1}$$

where

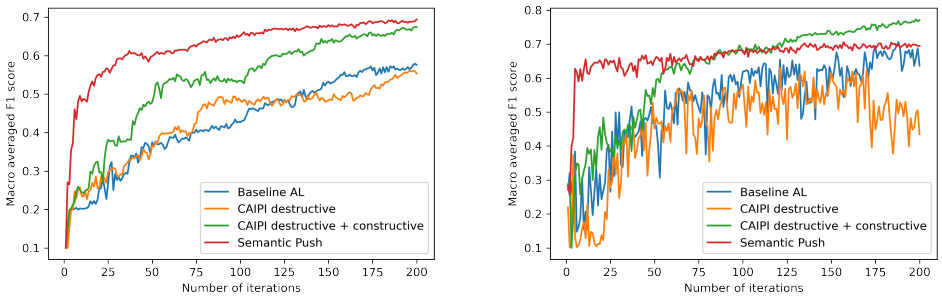
$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

and

$$Recall = \frac{TP}{TP + FN}. \tag{4.3}$$

In the context of precision and recall, TP represents the amount of true positives, FP the amount of false positives, and FN the amount of false negatives in a classification scenario. The F1 score in turn is defined as the harmonic mean of precision and recall. In multiclass classification, the F1 score is first calculated for each class in an one-vs-rest approach and then averaged using the arithmetic mean in the case of macro-averaging.

Figures 4.5 to 4.7 compare four different interaction strategies in terms of its macro-averaged F1 score.

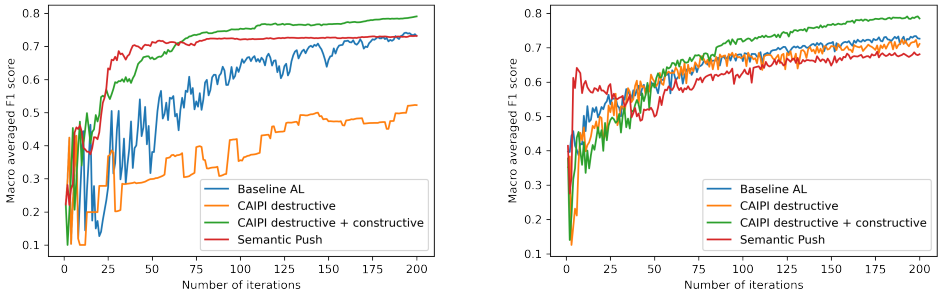


(a) (b)

Figure 4.5: Learning performance of different interactive learning strategies for *AG News* dataset: (**a**) XGBoost as learner, (**b**) Random Forest as learner.

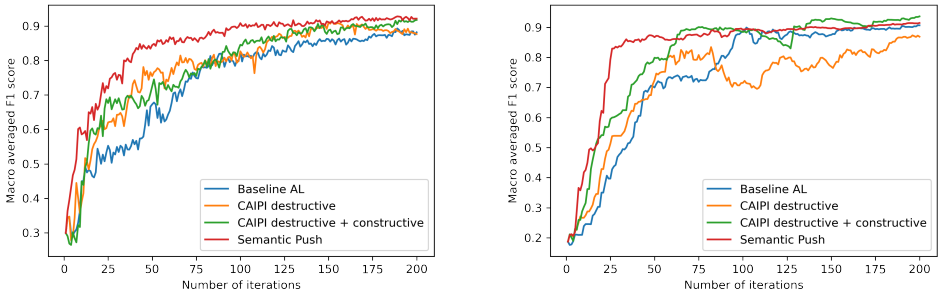
The according downstream tasks comprised multiclass classification scenarios for two different text corpora, where overall five different classifiers have been used. The first interaction strategy called *Baseline AL* is represented by a standard active learner harnessing maximum classification uncertainty as query strategy. It is compared against the standard CAIPI strategy, against a constructive variant

[8]Please note that local explanation quality can be analyzed in both directions of interaction with a human user. As Semantic Push integrates topicLIME, which has already been evaluated in chapter 3, this evaluation only considers local explanation quality from the opposite direction - from the human to the learner.



(a) (b)

Figure 4.6: Learning performance of different interactive learning strategies for *AG News* dataset: (**a**) Multinomial naive Bayes as learner, (**b**) multilayer perceptron as learner.



(a) (b)

Figure 4.7: Learning performance of different interactive learning strategies for Reuters R10 dataset: (**a**) XGBoost as learner, (**b**) SVM as learner.

developed during this doctoral research as well as against Semantic Push. According to the results, both constructive strategies - "CAIPI destructive + constructive" and Semantic Push - outperform the standard active learner and standard CAIPI across most types of learners and datasets. Semantic Push especially performs well at early stages of interactions, despite a simulated gold standard that is around ten percent worse than that used for CAIPI. The standard CAIPI algorithm is not able to consistently beat the active learner's baseline, in some cases it performs clearly worse. More generally, Semantic Push shows high data efficiency with respect to the queried documents as it incorporates the oracle's expert knowledge efficiently at a much earlier stage compared to the other approaches. At some point, Semantic Push's performance suffers from the worse gold standard as most of the correct knowledge has already been applied. Partially, the strategy even starts to incorporate "incorrect corrections". That is the case when the performance of the

classifier exceeds the performance of the simulated conceptual gold standard. It seems that especially the multilayer perceptron suffers from "incorrect corrections", presumably due to overfitting as a result of a comparably low inductive bias. Such a low inductive bias might especially have a strong negative impact in a low data setting as it is the case in an active learning setting.

For a more realistic comparison of the different strategies, the underlying gold standards should either be equal in terms of their predictive performance when being simulated or a real human oracle should be integrated into the loop. Regardless of that, it can be concluded that offering the possibility for constructive corrections comes with clear gains in terms of learning performance. Incorporating concept-based knowledge in the corrections further enhances performance, especially at early interaction stages, as long as the performance of the gold standard is not exceeded.

In addition to learning performance, Semantic Push is evaluated by analyzing local explanation quality. It measures the extend to which expert knowledge is adequately adopted by the learner such that its reasoning is pushed towards the desired behavior. The average explanatory accuracy is defined as follows:

$$ExplanatoryAccuracy_{AVG} = \frac{1}{N}\sum_{i=1}^{N}\frac{|GS_{local}(x_i) \cap \epsilon(x_i)|}{|GS_{local}(x_i)|}, \tag{4.4}$$

with x being a test document, $\epsilon(x)$ a local explanation, $GS_{local}(x)$ a local gold standard, and N the number of documents in a test dataset.

Table 4.1 summarizes the results of analyzing the different interaction strategies applied to different datasets and learners with regard to the average explanatory accuracy. It is striking that only Semantic Push is capable of clearly transferring the expert knowledge, included in the corrective explanations, in a way that it is adopted by the learner. The two versions of CAIPI do not reveal better results than the standard active learner that solely allows the correction of labels. To sum up, Semantic Push not only improves learning performance, especially in early stages of interactions, but also pushes the reasoning of the learner towards the behavior desired by a human. The proposed approach therefore takes a step towards human-centered machine learning by offering contextual interpretation and intervention in an interactive setting. Effective and efficient co-work between users and an ML learner is enabled, allowing the learner to take advantage of the richness of human expertise.

Table 4.1: Comparison of different interactive ML strategies regarding explanatory accuracy across datasets and learners: AL represents the baseline active learner, CAIPI_d the standard CAIPI algorithm and $\text{CAIPI}_{d/c}$ the constructive extension of CAIPI.

	Explanatory Accuracy$_{AVG}$			
	AL	**CAIPI$_d$**	**CAIPI$_{d/c}$**	**Semantic Push**
AG News (XGBoost)	0.690	0.683	0.685	**0.711**
AG News (Random Forest)	0.708	0.708	0.716	**0.726**
AG News (Naive Bayes)	0.722	0.715	0.723	**0.731**
AG News (Multilayer Perceptron)	0.726	**0.727**	**0.727**	0.726
Reuters R10 (XGBoost)	0.741	0.739	0.742	**0.768**
Reuters R10 (SVC)	0.786	0.785	0.788	**0.796**

5. Conclusion and Outlook

In this dissertation, a new approach has been proposed and validated that enables human-centered interactivity with text classifiers by using bidirectional model-agnostic explanations. To achieve this, the doctoral research has been split into three main parts that contribute to the development of human-centered machine learning.

As a first contribution, a framework has been introduced that conceptually defines comprehensible and interactive artificial intelligence in an interdisciplinary way using cognitive concepts like explainability, interpretability, transparency and interactivity. By describing and putting the basic cognitive concepts in relation, this thesis has been able to address and discuss its research questions in an interdisciplinary manner. Following along the framework, semantic alignment between ML classifiers and human users has been suggested as a prerequisite for comprehensibility and interactivity. As a major claim it has been stated that contextual information provided by the input domain must be taken into account during explanation generation and presentation such that more human-friendly and reliable explanations are obtained.

As a second contribution, a novel proposal for obtaining coherent and semantically meaningful explanations has been worked out. The explanation method called *topicLIME* builds explanations in a way, in which its explanatory features are made of coherent words providing explainees with contextual information. It offers a realistic and meaningful local perturbation distribution by avoiding extrapolation, which is a line of research that, according to the authors from Ribeiro et al., "would benefit multiple explanation methods" (Marco Tulío Ribeiro, Singh, and Guestrin 2018). As a consequence, topicLIME reveals higher local fidelity compared to LIME, which results in more reliable explanations regarding the classifier's actual local reasoning. A higher local fidelity of topicLIME has been identified from two different perspectives. On the one hand, the newly developed approach generated local surrogate explanation models that, as a whole, were characterized by higher local fidelity compared to the surrogate explanation models approximated by standard LIME. On the other hand, it has been shown that also individual explanations offered by topicLIME revealed a higher local fidelity in terms of explaining the local behavior of the classifier. Such explanations characterized by higher local fidelity are less likely to be misinterpreted by human explainees due to a lower degree of extrapolation.

Besides the technical evaluation of the capabilities of the new explanation method, topicLIME has further been evaluated empirically with the help of two user studies. It has been analyzed how different explanation modalities are perceived by humans that interact with a text classifier that predicts which content category a presented document belongs to. As a main result, it has turned out that the generation of contextual explanations as an alternative to LIME explanations can be considered useful, especially when striving for receiver-dependent explanations. The new method should be offered individually as it provides explanations that especially satisfy a certain group of humans that overall prefer explanations made of fewer but coherent and concept-based explanation units.

Latest research has mainly focused on the generation of model-agnostic explanations for supervised ML, which is why this doctoral research furthermore has proposed a way to provide model-agnostic and receiver-dependent explanations despite the lack of class labels. By approximating any unsupervised ML approach globally with a supervised classification approach, it is now possible to explain the decisions of an unsupervised model by applying model-agnostic explainers, like LIME or SHAP. When additionally using post-processing techniques, like an elbow criterion or a relative weight threshold, the resulting explanations can be forced to be more selective. In this way, explanations are obtained that, according to insights from psychology, can be characterized as more human-friendly, especially for non-ML-experts.

As a third main contribution, this doctoral research has elaborated on improving state-of-the-art approaches to explanatory interactive learning, especially for text classification. A new interaction framework has been proposed with the intention of further closing the loop by allowing humans richer interactions, like correcting predictions and explanations of a query in a constructive and contextual way. Its instantiation called *Semantic Push* is among the first model-agnostic interactive learning strategies that enable semantically meaningful corrections of a learner, also for completely false predictions. Based on locally faithful and contextual explanations, it qualifies humans to provide concept-based corrections that in turn are integrated into a learner's reasoning via non-extrapolating additional training documents. As a consequence of combining richer explanations with more extensive semantic corrections, the proposed interaction paradigm outperforms its baselines, like different variants of CAIPI, with regard to learning performance and local explanation quality in several downstream classification tasks.

Summing up, this thesis has contributed to the overall question of how human-centered interactivity with text classifiers can be achieved using bidirectional model-agnostic explanations. Concretely, two state-of-the-art methods, namely LIME and CAIPI, have been improved and extended for the domain of text such that they are now able to include concept-based knowledge within the generated explanations or corrections. In order to approximate the relevant conceptual representations of textual input domains that model-agnostic explainers and human users can access for more aligned interactions, Latent Dirichlet Allocation models have been integrated into an interactive ML architecture.

Besides the achievements listed so far, there are also some prerequisites that should be mentioned. As the entire semantic functionality of the proposed approach is based on Latent Dirichlet Allocation, some expertise in topic modeling is required, for instance, to implement a suitable data preprocessing or to find an adequate hyper-parametrization. The latter needs to be addressed from different perspectives. On the one hand, topics shall be identified globally for a given text corpus such that they correlate well with human topic-interpretability. On the other hand, the learned topics also need to be suited for locally generating groups of coherent words as explanatory features for single documents to be explained. In addition, the approximated generative process must preserve the statistical properties of the input domain such that it is capable of generating realistic "in-distribution" documents based on human corrections. Lastly, it is to be noted that LDA generates documents represented as bag-of-words. Therefore, future research could include an encoder-decoder language model into the interactive architecture to ensure that generated counterexamples are meaningful both semantically and linguistically, and especially that they are syntactically correct. Masked language modeling could be harnessed to check for linguistically sensible counterexamples, while autoencoders could be used to identify "out-of-distribution" counterexamples by analyzing the reconstruction error.

To conclude, this thesis advocates for conducting further research in order to facilitate non-expert human involvement in interactive classification scenarios. A promising approach could be the combination of visual analytics and explanatory interactive learning. The integration of exploratory visual data analysis into the framework of topic-based interactions could help non-expert users that interact with a text-classifier via topics to explore and understand the statistical characteristics of the input domain and to evaluate and optimize the quality of the topics learned by a topic model. In this way, humans could analyze meaning and prevalence of topics in a given corpus and identify how topics and documents relate to each other. Based on this conceptual domain understanding, the explication of human knowledge might be simplified.

To put it in a nutshell, the fusion of concept-based knowledge with local surrogate explanation models turned out to be a promising research direction when striving for linking the capabilities of human perception with the capacity and performance of computers. To further converge towards the target state of human-centered machine learning, more interdisciplinary research in this area is required, especially integrating findings from psychology or the social sciences.

> *"We must design for the way people behave, not for how we would wish them to behave."*
>
> – By Donald A. Norman, born in 1935

References

20NG (1995). *20 Newsgroups Dataset.* https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html.

Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli (2018). "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda". In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–18.

Adadi, Amina and Mohammed Berrada (2018). "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6, pp. 52138–52160.

Ahn, Woo-kyoung, Nancy S Kim, Mary E Lassaline, and Martin J Dennis (2000). "Causal status as a determinant of feature centrality". In: *Cognitive Psychology* 41.4, pp. 361–416.

Akata, Zeynep, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling (2020). "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence". In: *Computer* 53.8, pp. 18–28. DOI: 10.1109/MC.2020.2996587.

Alonso, Jose M, Alejandro Ramos-Soto, Ehud Reiter, and Kees van Deemter (2017). "An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems". In: *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, pp. 1–6.

Amershi, Saleema, Maya Cakmak, William Bradley Knox, and Todd Kulesza (2014). "Power to the people: The role of humans in interactive machine learning". In: *Ai Magazine* 35.4, pp. 105–120.

Ananny, Mike and Kate Crawford (2018). "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". In: *New media & society* 20.3, pp. 973–989.

Arrieta, Alejandro Barredo, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58, pp. 82–115.

Attenberg, Josh, Prem Melville, and Foster Provost (2010). "A unified approach to active dual supervision for labeling features and examples". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010*. Springer, pp. 40–55.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Balog, Krisztian, Filip Radlinski, and Shushan Arakelyan (2019). "Transparent, scrutable and explainable user models for personalized recommendation". In: *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 265–274.

Bathaee, Yavar (2017). "The artificial intelligence black box and the failure of intent and causation". In: *Harvard Journal of Law and Technology* 31, p. 889.

Belle, Vaishak and Ioannis Papantonis (2021). "Principles and practice of explainable machine learning". In: *Frontiers in big Data*, p. 39.

Bergstein, Brian (2017). "AI isn't very smart yet. But we need to get moving to make sure automation works for more people". In: *MIT Technology*.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3, pp. 993–1022.

Bodo, Balazs, Natali Helberger, Kristina Irion, Frederik Zuiderveen Borgesius, Judith Moller, Bob van de Velde, Nadine Bol, Bram van Es, and Claes de Vreese (2018). "Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents". In: *Yale Journal of Law and Technology* 19, p. 133.

Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. (2016). "End to end learning for self-driving cars". In: *arXiv preprint arXiv:1604.07316*.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bruckert, Sebastian, Bettina Finzel, and Ute Schmid (2020). "The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions". In: *Frontiers in Artificial Intelligence* 3. ISSN: 2624-8212. DOI: 10.3389/frai.2020.507973.

Byrnes, James P (1992). "The conceptual basis of procedural learning". In: *Cognitive Development* 7.2, pp. 235–257.

Cai, Yanli and Jian-Tao Sun (2009). "Text Mining". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Springer US, pp. 3061–3065. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_418.

Cambria, Erik, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani (2023). "A survey on XAI and natural language explanations". In: *Information Processing & Management* 60.1, p. 103111.

Cath, Corinne (2018). "Governing artificial intelligence: ethical, legal and technical opportunities and challenges". In: *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences* 376.2133.

Chander, Ajay, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino (2018). "Working with beliefs: AI transparency in the enterprise". In: *IUI Workshops*.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM computing surveys* 41.3, pp. 1–58.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei (2009). "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems* 22.

Cheverst, Keith, Hee Eon Byun, Dan Fitton, Corina Sas, Chris Kray, and Nicolas Villar (2005). "Exploring issues of user model transparency and proactive

behaviour in an office environment control system". In: *User Modeling and User-Adapted Interaction* 15.3, pp. 235–273.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078*.

Chromik, Michael (2021). "Human-centric Explanation Facilities: Explainable AI for the Pragmatic Understanding of Non-expert End Users". PhD thesis. Ludwig Maximilian University of Munich, Germany.

Coglianese, Cary and David Lehr (2019). "Transparency and algorithmic governance". In: *Administrative Law Review* 71, p. 1.

Cooper, Alan (2004). *The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity*. Pearson Education.

Courtland, Rachel (2018). "Bias detectives: the researchers striving to make algorithms fair". In: *Nature* 558, pp. 357–360.

Crooks, Noelle M and Martha W Alibali (2014). "Defining and measuring conceptual knowledge in mathematics". In: *Developmental review* 34.4, pp. 344–377.

Dalal, Mita K and Mukesh A Zaveri (2011). "Automatic text classification: a technical review". In: *International Journal of Computer Applications* 28.2, pp. 37–40.

Danks, David and Alex John London (2017). "Algorithmic Bias in Autonomous Systems." In: *IJCAI*. Vol. 17, pp. 4691–4697.

DeJong, Gerald and Raymond Mooney (1986). "Explanation-based learning: An alternative view". In: *Machine learning* 1, pp. 145–176.

Dennet, Daniel (1989). "The Intentional Stance". In: *The MIT Press*.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.

Druck, Gregory, Gideon Mann, and Andrew McCallum (2008). "Learning from labeled features using generalized expectation criteria". In: *Proceedings of the*

31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 595–602.

Druck, Gregory, Burr Settles, and Andrew McCallum (2009). "Active learning by labeling features". In: *Proceedings of the 2009 conference on Empirical methods in natural language processing*, pp. 81–90.

Dudley, John J and Per Ola Kristensson (2018). "A review of user interface design for interactive machine learning". In: *ACM Transactions on Interactive Intelligent Systems* 8.2, pp. 1–37.

Edwards, Lilian and Michael Veale (2017). "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for". In: *Duke L. & Tech. Rev.* 16, p. 18.

Fails, Jerry Alan and Dan R. Olsen (2003). "Interactive Machine Learning". In: *Proceedings of the 8th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, pp. 39–45. ISBN: 1581135866. DOI: 10.1145/604045.604056.

Fiebrink, Rebecca Anne (2011). *Real-time human interaction with supervised learning algorithms for music composition and performance*. Princeton University.

Gasparetto, Andrea, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli (2022). "A Survey on Text Classification Algorithms: From Text to Predictions". In: *Information* 13.2, p. 83.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin (2017). "Convolutional sequence to sequence learning". In: *International conference on machine learning*, pp. 1243–1252.

Gentner, Dedre and Cecile Toupin (1986). "Systematicity and surface similarity in the development of analogy". In: *Cognitive science* 10.3, pp. 277–300.

Gillies, Marco, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. (2016). "Human-centred machine learning". In: *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pp. 3558–3565.

Green, Ben and Yiling Chen (2020). "Algorithm-in-the-loop decision making". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 13663–13664.

Hacker, Philipp (2018). "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law". In: *Common Market Law Review* 55.4.

Harbers, Maaike, Karel van den Bosch, and John-Jules Meyer (2010). "Design and evaluation of explainable BDI agents". In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 2. IEEE, pp. 125–132.

Harman, Gilbert H (1965). "The inference to the best explanation". In: *The philosophical review* 74.1, pp. 88–95.

Harris, Z (1954). "Distributional hypothesis". In: *Word* 10.23, pp. 146–162.

Heimerl, Florian, Steffen Koch, Harald Bosch, and Thomas Ertl (2012). "Visual classifier training for text document retrieval". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12, pp. 2839–2848.

Hiebert, J (1986). "Conceptual and procedural knowledge: The case of mathematics". In: *Hillsdale, NJ*, pp. 1–27.

Hildebrandt, M. and L. Janssens (2016). *The New Imbroglio – Living with Machine Algorithms*. Philipps-Universität Marburg.

Hind, Michael, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney (2019). "TED: Teaching AI to explain its decisions". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123–129.

Hoffman, Robert R, William J Clancey, and Shane T Mueller (2020). "Explaining AI as an exploratory process: The peircean abduction model". In: *arXiv preprint arXiv:2009.14795*.

Hoffman, Robert R, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink (2013). "Trust in automation". In: *IEEE Intelligent Systems* 28.1, pp. 84–88.

Holzinger, Andreas, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell (2017). "What do we need to build explainable AI systems for the medical domain?" In: *arXiv preprint arXiv:1712.09923*.

Holzinger, Andreas, Bernd Malle, Anna Saranti, and Bastian Pfeifer (2021). "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI". In: *Information Fusion* 71, pp. 28–37.

Huysmans, Johan, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens (2011). "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models". In: *Decision Support Systems* 51.1, pp. 141–154.

Inkpen, Kori, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric PS Baumer (2019). "Where is the human? Bridging the gap between AI and HCI". In: *Extended abstracts of the 2019 chi conference on human factors in computing systems*, pp. 1–9.

Kahneman, Daniel and Gary Klein (2009). "Conditions for intuitive expertise: a failure to disagree." In: *American psychologist* 64.6, p. 515.

Keil, Frank C (2006). "Explanation and Understanding". In: *Annual Review of Pschology* 57, pp. 227–254.

Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh (2022). "Natural language processing: State of the art, current trends and challenges". In: *Multimedia Tools and Applications*, pp. 1–32.

Kiefer, Sebastian (2022). "CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge". In: *Information Fusion* 77, pp. 184–195. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.07.014.

Kiefer, Sebastian, Mareike Hoffmann, and Ute Schmid (Nov. 2022). "Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions". In: *Machine Learning and Knowledge Extraction* 4.4, pp. 994–1010. ISSN: 2504-4990. DOI: 10.3390/make4040050.

Kiefer, Sebastian and Günter Pesch (2021). "Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations". In: *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, pp. 291–308.

Kim, Been, Kayur Patel, Afshin Rostamizadeh, and Julie Shah (2015). "Scalable and interpretable data representation for high-dimensional, complex data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.

Kimmig, Angelika, Luc De Raedt, and Hannu Toivonen (2007). "Probabilistic explanation based learning". In: *Machine Learning: ECML 2007: 18th European Conference on Machine Learning*. Springer, pp. 176–187.

Kingston, John KC (2016). "Artificial intelligence and legal liability". In: *International conference on innovative techniques and applications of artificial intelligence*. Springer, pp. 269–279.

Kleinsmith, Andrea and Marco Gillies (2013). "Customizing by doing for responsive video game characters". In: *International Journal of Human-Computer Studies* 71.7, pp. 775–784. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2013.03.005.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown (2019). "Text classification algorithms: A survey". In: *Information* 10.4, p. 150.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf (2015). "Principles of explanatory debugging to personalize interactive machine learning". In: *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137.

Langley, Pat, Ben Meadows, Mohan Sridharan, and Dongkyu Choi (2017). "Explainable agency for intelligent autonomous systems". In: *Twenty-Ninth IAAI Conference*.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouze, Alex Pentland, and Patrickl Vinck (2018). "Fair, Transparent, and Accountable Algorithmic Decision-making Processes". In: *Philosophy and Technology* 31, pp. 611–627.

Lewis, Davìd (1993). *REUTERS-21578*. https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.

Lewis, Dàvid (1996). *TREC-AP*. http://www.daviddlewis.com/resources/testcollections/trecap/.

Lewis, David (1986). "Causal Explanation". In: *Philosophical Papers*. Ed. by David Lewis. Oxford University Press, pp. 214–240.

Li, Qian, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He (Apr. 2022). "A Survey on Text Classification: From Traditional to Deep Learning". In: *ACM Trans. Intell. Syst. Technol.* 13.2. ISSN: 2157-6904. DOI: 10.1145/3495162.

Lipton, Peter (1990). "Contrastive Explanation". In: *Royal Institute of Philosophy Supplement* 27, pp. 247–266. DOI: 10.1017/S1358246100005130.

Liu, Han-Wei, Ching-Fu Lin, and Yu-Jie Chen (2019). "Beyond State v Loomis: artificial intelligence, government algorithmization and accountability". In: *International journal of law and information technology* 27.2, pp. 122–141.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Madhavan, Poornima and Douglas A Wiegmann (2007). "Effects of information source, pedigree, and reliability on operator interaction with decision support systems". In: *Human factors* 49.5, pp. 773–785.

Mehrotra, Kishan G, Chilukuri K Mohan, and HuaMing Huang (2017). *Anomaly detection principles and algorithms: Terrorism, Security, and Computation*. Vol. 1. Springer International Publishing.

Miller, George A (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2, p. 81.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267, pp. 1–38. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007.

Miller, Tim, Piers Howe, and Liz Sonenberg (2017). "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences". In: *International Joint Conference on Artificial Intelligence, Workshop on Explainable AI (XAI)* 36, pp. 36–40.

Mitchell, Tom M, Richard M Keller, and Smadar T Kedar-Cabelli (1986). "Explanation-based generalization: A unifying view". In: *Machine learning* 1, pp. 47–80.

Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed.

Molnar, Christoph, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio (2020). "Model-agnostic Feature Importance and Effects with Dependent Features–A Conditional Subgroup Approach". In: *arXiv preprint arXiv:2006.04628*.

Moore, Johanna D and Cécile L Paris (1991). "Requirements for an expert system explanation facility". In: *Computational Intelligence* 7.4, pp. 367–370.

Moravec, Hans (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Mosqueira-Rey, Eduardo, Elena Hernandez-Pereira, David Alonso-Rios, Jose Bobes-Bascaran, and Angel Fernandez-Leal (2022). "Human-in-the-loop machine learning: A state of the art". In: *Artificial Intelligence Review*, pp. 1–50.

Nguyen, Dong (June 2018). "Comparing Automatic and Human Evaluation of Local Explanations for Text Classification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, pp. 1069–1078. DOI: 10.18653/v1/N18-1097.

Nosek, Brian A, Carlee Beth Hawkins, and Rebecca S Frazier (2011). "Implicit social cognition: From measures to mechanisms". In: *Trends in cognitive sciences* 15.4, pp. 152–159.

Odom, Phillip and Sriraam Natarajan (2018). "Human-guided learning for probabilistic logic models". In: *Frontiers in Robotics and AI* 5, p. 56.

Overton, James A. (2011). "Scientific Explanation and Computation". In: *Proceedings of the 6th International Explanation-Aware Computing workshop*, pp. 41–50.

PAN, Xiao, Yun-liang LI, and Yuanfeng ZHOU (2014). "A Research Review of Superpixels Generation Algorithms". In: *Computer Aided Drafting, Design and Manufacturing*.

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman (2015). "Deep face recognition". In: *Proceedings of the British Machine Vision Conference (BMVC)*.

Patel, Kayur, James Fogarty, James A. Landay, and Beverly Harrison (2008). "Investigating Statistical Machine Learning as a Tool for Software Development". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Association for Computing Machinery, pp. 667–676. ISBN: 9781605580111. DOI: 10.1145/1357054.1357160.

Petch, Jeremy, Shuang Di, and Walter Nelson (2022). "Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology". In: *Canadian Journal of Cardiology* 38.2, pp. 204–213. ISSN: 0828-282X. DOI: https://doi.org/10.1016/j.cjca.2021.09.004.

Prahl, Andrew and Lyn Van Swol (2017). "Understanding algorithm aversion: When is advice from automation discounted?" In: *Journal of Forecasting* 36.6, pp. 691–702.

Price, William Nicholson (2017). "Artificial Intelligence in Health Care: Applications and Legal Issues". In: *The SciTech Lawyer* 10.

Pu, Pearl, Li Chen, and Rong Hu (2011). "A user-centric evaluation framework for recommender systems". In: *Proceedings of the fifth ACM conference on Recommender systems*, pp. 157–164.

Raghavan, Hema and James Allan (2007). "An interactive algorithm for asking and incorporating feature feedback into support vector machines". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 79–86.

Raghavan, Hema, Omid Madani, and Rosie Jones (2006). "Active learning with feedback on features and instances". In: *The Journal of Machine Learning Research* 7, pp. 1655–1686.

Raymond, Anjanette H and Scott J Shackelford (2013). "Technology, ethics, and access to justice: should an algorithm be deciding your case". In: *Michigan Journal of International Law* 35, p. 485.

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang (Oct. 2021). "A Survey of Deep Active Learning". In: *ACM Comput. Surv.* 54.9. ISSN: 0360-0300. DOI: 10.1145/3472291.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining. KDD '16, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778.

Ribeiro, Marco Tulío, Sameer Singh, and Carlos Guestrin (2018). "Anchors: High-precision model-agnostic explanations". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Rittle-Johnson, Bethany, Robert S Siegler, and Martha Wagner Alibali (2001). "Developing conceptual understanding and procedural skill in mathematics: An iterative process." In: *Journal of educational psychology* 93.2, p. 346.

Robnik-Šikonja, Marko and Marko Bohanec (2018). "Perturbation-Based Explanations of Prediction Models". In: *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Ed. by Jianlong Zhou and Fang Chen. Cham: Springer International Publishing, pp. 159–175. ISBN: 978-3-319-90403-0. DOI: 10.1007/978-3-319-90403-0_9.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.

Rodrigues, Rowena (2020). "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities". In: *Journal of Responsible Technology* 4, p. 100005. ISSN: 2666-6596. DOI: https://doi.org/10.1016/j.jrt.2020.100005.

Roig, Antoni (2017). "Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)". In: *European Journal of Law and Technology* 8.3.

Satopaa, Ville, Jeannie Albrecht, David Irwin, and Barath Raghavan (2011). "Finding a" kneedle" in a haystack: Detecting knee points in system behavior". In: *2011 31st international conference on distributed computing systems workshops*. IEEE, pp. 166–171.

Savelka, Jaromir, Gaurav Trivedi, and Kevin D. Ashley (2015). "Applying an Interactive Machine Learning Approach to Statutory Analysis". In: *International Conference on Legal Knowledge and Information Systems*.

Schaefer, Kristin E, Jessie YC Chen, James L Szalma, and Peter A Hancock (2016). "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems". In: *Human factors* 58.3, pp. 377–400.

Schmid, Ute (2021). "Interactive Learning with Mutual Explanations in Relational Domains". In: *Human-Like Machine Intelligence*. Oxford University Press. ISBN: 9780198862536. DOI: 10.1093/oso/9780198862536.003.0017. eprint: https://academic.oup.com/book/0/chapter/350717106/chapter-pdf/43431156/oso-9780198862536-chapter-17.pdf.

Schönberger, Daniel (2018). "Deep Copyright: Up- and Downstream Questions Related to Artificial Intelligence (AI) and Machine Learning (ML)". In: *Zeitschrift für geistiges Eigentum (ZGE)* 10, pp. 35–58.

Schwalbe, Gesina and Bettina Finzel (2021). "XAI method properties: A (meta-) study". In: *arXiv preprint arXiv:2105.07190*.

Settles, Bùrr (2011). "From theories to queries: Active learning in practice". In: *Active learning and experimental design workshop in conjunction with AISTATS 2010*. JMLR Workshop and Conference Proceedings, pp. 1–18.

Settles, Burr (2011). "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1467–1478.

Sharma, Manali, Di Zhuang, and Mustafa Bilgic (2015). "Active learning with rationales for text classification". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 441–451.

Simpson, Jeffry A (2007). "Psychological foundations of trust". In: *Current directions in psychological science* 16.5, pp. 264–268.

Slany, Emanuel, Yannik Ott, Stephan Scheele, Jan Paulus, and Ute Schmid (2022). "Caipi in practice: Towards explainable interactive medical image classification". In: *Artificial Intelligence Applications and Innovations. AIAI 2022 International Workshops*. Springer, pp. 389–400.

Sloman, Steven A, Bradley C Love, and Woo-Kyoung Ahn (1998). "Feature centrality and conceptual coherence". In: *Cognitive Science* 22.2, pp. 189–228.

Smith-Renner, Alison, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater (2020). "No explainability without accountability: An empirical study of explanations and feedback in interactive ml". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Sokol, Kacper and Peter A Flach (2018). "Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements." In: *IJCAI*, pp. 5785–5786.

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski, and David Buttler (2012). "Exploring topic coherence over many models and many topics". In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 952–961.

Stumpf, Simone, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker (2007). "Toward harnessing user feedback for machine learning". In: *Proceedings of the 12th international conference on Intelligent user interfaces*, pp. 82–91.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27.

Syed, Shaheen and Marco Spruit (2017). "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation". In: *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, pp. 165–174.

Teso, Stefano and Kristian Kersting (2019). "Explanatory interactive machine learning". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 239–245.

Thagard, Paul (2002). *Coherence in thought and action*. MIT press.

Tintarev, Nava and Judith Masthoff (2015). "Explaining recommendations: Design and evaluation". In: *Recommender systems handbook*. Springer, pp. 353–382.

Tomsett, Richard, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty (2018). "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems". In: *arXiv preprint arXiv:1806.07552*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vladeck, David C (2014). "Machines without principals: liability rules and artificial intelligence". In: *Washington Law Review* 89, p. 117.

Wachter, Sandra and Brent Mittelstadt (2019). "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI". In: *Columbia Business Law Review*, p. 494.

Zaidan, Omar and Jason Eisner (2008). "Modeling annotators: A generative approach to learning from annotator rationales". In: *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pp. 31–40.

Zaidan, Omar, Jason Eisner, and Christine Piatko (2007). "Using "annotator rationales" to improve machine learning for text categorization". In: *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pp. 260–267.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level Convolutional Networks for Text Classification". In: *Advances in Neural Information Processing Systems*. Vol. 28.

Part II

Appendix

A. Publications and Details on Written Scientific Contributions

A.1 A Framework for Comprehensible and Interactive Artificial Intelligence

A.1.1 Kiefer et al. "The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions." In: frontiers in Artificial Intelligence 2020

Full Reference of Paper

Sebastian Kiefer (né Bruckert), Bettina Finzel, and Ute Schmid. "The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions." In *frontiers in Artificial Intelligence, 3* (2020). DOI: 10.3389/frai.2020.507973.

My Scientific Contributions

Related research question(s): **RQ1**.
Type of contribution: **conceptual**.

During the research underlying this dissertation, I made the following scientific contributions published in this paper:

- Addressing the nature, drawbacks and implications of black-box-ML approaches for human decision-makers in a conceptual manner.
- Deriving and describing the field of *comprehensible artificial intelligence* using cognitive concepts provided by psychology.
- Structuring and classifying the research disciplines *explainable artificial intelligence*, *interpretable machine learning*, and *interactive machine learning* within *comprehensible artificial intelligence* (together with Bettina Finzel).
- Systematizing desiderata, necessary concepts, and related interdisciplinary research areas of comprehensible AI by introducing an integrated transition framework that can be used by practitioners as a roadmap towards transparent expert companions (together with Bettina Finzel).

- Showcasing *comprehensible artificial intelligence* fundamentals applied to medicine (together with Bettina Finzel and Ute Schmid) focusing on missing semantic alignment between humans and machines and proposing bidirectional explanations considering the mental model of humans.
- Outlining future research, like enabling *semantic alignment* and improving perturbation-based explanation systems by using realistic perturbation distributions in order to generate meaningful and contextual explanations.

My Written Contents

I structured the paper together with Bettina Finzel and contributed to all parts except to subsections 3.2 and 3.3 (those were contributed by Bettina Finzel and Ute Schmid). My written contribution to the paper is more than 60%.

frontiers in Artificial Intelligence

CONCEPTUAL ANALYSIS
published: 24 September 2020
doi: 10.3389/frai.2020.507973

The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions

Sebastian Bruckert, Bettina Finzel and Ute Schmid*

Cognitive Systems, University of Bamberg, Bamberg, Germany

Increasing quality and performance of artificial intelligence (AI) in general and machine learning (ML) in particular is followed by a wider use of these approaches in everyday life. As part of this development, ML classifiers have also gained more importance for diagnosing diseases within biomedical engineering and medical sciences. However, many of those ubiquitous high-performing ML algorithms reveal a black-box-nature, leading to opaque and incomprehensible systems that complicate human interpretations of single predictions or the whole prediction process. This puts up a serious challenge on human decision makers to develop trust, which is much needed in life-changing decision tasks. This paper is designed to answer the question how expert companion systems for decision support can be designed to be interpretable and therefore transparent and comprehensible for humans. On the other hand, an approach for interactive ML as well as human-in-the-loop-learning is demonstrated in order to integrate human expert knowledge into ML models so that humans and machines act as companions within a critical decision task. We especially address the problem of *Semantic Alignment* between ML classifiers and its human users as a prerequisite for semantically relevant and useful explanations as well as interactions. Our roadmap paper presents and discusses an interdisciplinary yet integrated Comprehensible Artificial Intelligence (cAI)-transition-framework with regard to the task of medical diagnosis. We explain and integrate relevant concepts and research areas to provide the reader with a *hands-on-cookbook* for achieving the transition from opaque black-box models to interactive, transparent, comprehensible and trustworthy systems. To make our approach tangible, we present suitable state of the art methods with regard to the medical domain and include a realization concept of our framework. The emphasis is on the concept of Mutual Explanations (ME) that we introduce as a dialog-based, incremental process in order to provide human ML users with trust, but also with stronger participation within the learning process.

Keywords: explainable artificial intelligence, interactive ML, interpretability, trust, medical diagnosis, medical decision support, companion

OPEN ACCESS

Edited by:
David Benrimoh, McGill University, Canada

Reviewed by:
Usman Qamar, National University of Sciences and Technology (NUST), Pakistan
Shivanand Sharanappa Gornale, Rani Channamma University, Belagavi, India

****Correspondence:***
Sebastian Bruckert
sebastian-manuel.bruckert@stud.uni-bamberg.de

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *28 October 2019*
Accepted: *17 August 2020*
Published: *24 September 2020*

Citation:
Bruckert S, Finzel B and Schmid U (2020) The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. Front. Artif. Intell. 3:507973. doi: 10.3389/frai.2020.507973

1. INTRODUCTION

Although modern ML approaches improved tremendously in terms of quality (prediction accuracy) and are able to even exceed human performance in many cases, they currently lack the ability to provide an explicit declarative knowledge representation and therefore hide the underlying explanatory structure (Holzinger et al., 2017). Due to this inability, modern ML approaches often result in black-box approaches—models and techniques, whose internal approach stays unknown and that just connect observable input- and output information without allowing an understanding nor an explanation of the way results have been produced (see **Figure 1**). Exactly that missing transparency makes it difficult for users of ML techniques to develop an understanding of the recommendations and decisions, which mostly constitutes an inherent risk (Sliwinski et al., 2017).
In a legal sense the question of legal security and liability security arises. Since the European General Data Protection Regulation (GDPR and ISO/IEC 27001) has entered into force in May, 2018, the relationship between AI and applicable law contains tremendous potential for clarification (Holzinger et al., 2017). As an example, the question of liability arises, especially if third parties suffer damages that are caused by recommendations or decisions made by ML approaches. According to latest jurisprudence, software architects, software developers as well as users are only liable for their actions and artifacts if a certain behavior of the system would have been predictable (Burri, 2016).

Most current architectures as described in **Figure 1** often lead to several problems. On the one hand, the system of internal rules itself often is not interpretable by humans. On the other hand, the ML results in terms of classification, regression or policy outputs are not comprehensible nor explainable due to biases and uncertainty introduced by the used model, the data or other factors. In addition, human experts have difficulties in integrating their expert knowledge into the learning process. All of the just mentioned points of criticism have led to a steadily increasing importance of the research areas Explainable Artificial Intelligence (xAI), Interpretable Machine Learning (iML) and Interactive ML that we summarize and refer to as cAI. These primarily aim at developing approaches that in addition to a precise prediction accuracy fulfill concepts like interpretability, explainability, confidence including stability and robustness, causality, interactivity, liability and liability security in a legal sense, socio-technical and domain aspects, bias awareness as well as uncertainty handling. The intention of cAI can be characterized by either achieving interpretability regarding the models or by making at least the results itself understandable and explainable and therefore interpretable (see **Figure 2**). We develop and present our cAI framework with regard to the application of ML for medical diagnosis. Since medical diagnosis comprises a complex process relevant for many succeeding medical sub-disciplines with high human involvement, diagnostic decisions not only need to be done accurately and precisely, but also in a comprehensible and trustworthy manner. Convolutional neural networks can be used to demonstrate the current trade-off between ML performance and interpretability. Such deep learning approaches often used for image-based medical diagnosis perform well in terms of prediction accuracy, but the models as well as their decisions cannot be interpreted easily without further investigations.

2. METHOD/DESIGN

In order to address the shortcomings mentioned above, we first provide an overview of the cognitive concepts that are used in the course of this paper to differentiate between different research

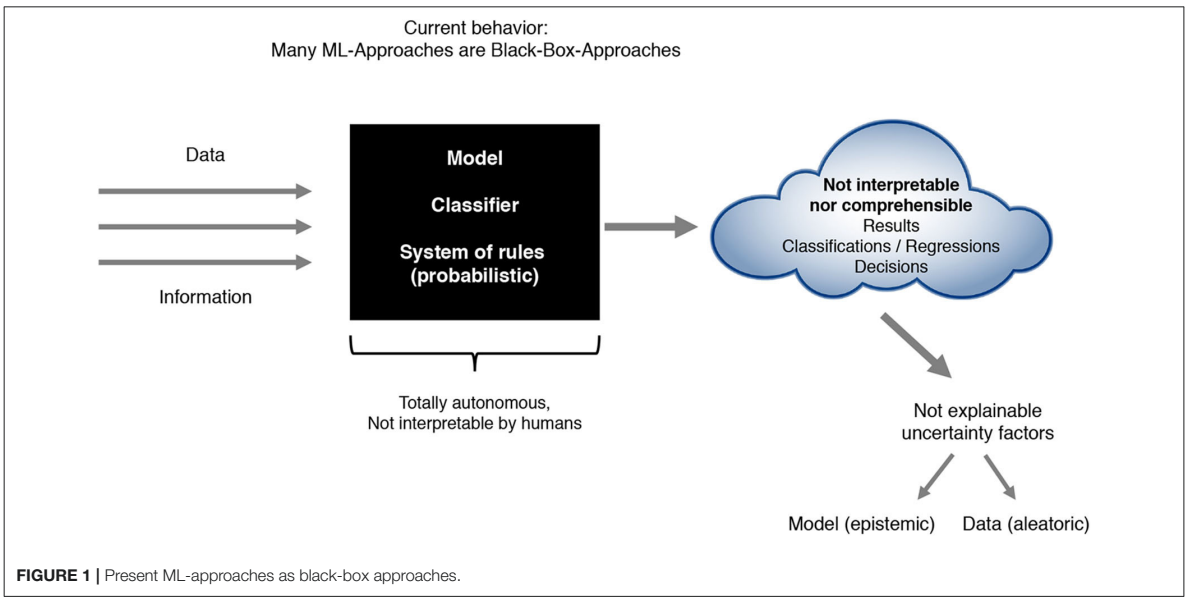


FIGURE 1 | Present ML-approaches as black-box approaches.

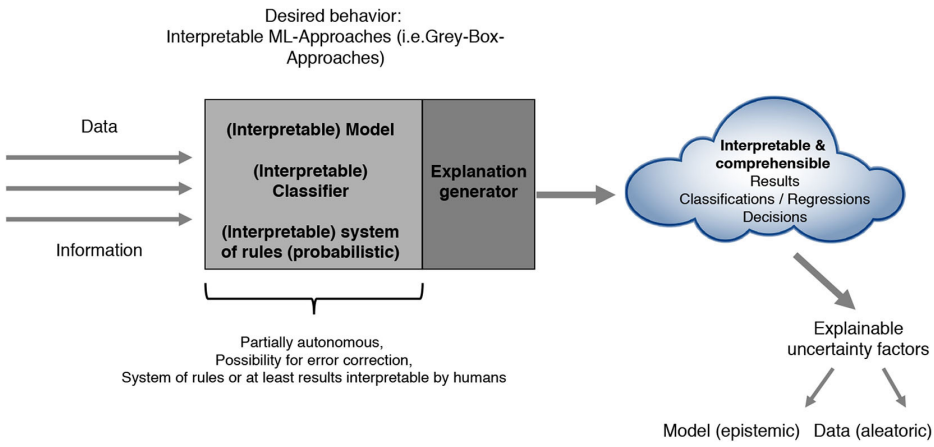


FIGURE 2 | Future comprehensible ML-approaches (i.e., gray-box approaches).

branches (see **Figure 3**). The cognitive concept of interpretation can be seen as the key concept, whose different shapings can be used as a criterion for differentiation of iML and xAI. From a philosophical and hermeneutical perspective, understanding and explaining are correlated terms and sometimes considered as symmetric cognitive concepts (Schurz, 2002). Having recognized and understood an issue therefore leads to having an explanation for it, and, reaching the state of understanding comes with having generated explanations. Thus the concept of understanding can be seen as necessary and sufficient condition for explaining and explaining represents a sufficient condition for understanding. Both concepts, understanding and explaining, in combination constitute a necessary condition for interpretation. iML and xAI differ in the explanandum as well as in the nature of desired interpretability, which the authors from Adadi and Berrada (2018) call the *scoop of interpretability*.

The task of making classifications, regressions or derived policies of an ML approach interpretable, contains sub-tasks like understanding and explaining as described in **Figure 3**. *Understanding*, which means recognizing correlations (context) in an intellectual way, can be seen as the bridge between human recognition and decision and is therefore the basis of explanation. Humans are performing really good in understanding a context and based on this generalizing from observations, whereas there is a long way still to go for AI especially in terms of contextualizing. On the basis of understanding a context, the explanation task, in addition, includes making the reasons of observed facts by stating logical and causal correlations comprehensible for humans (Holzinger, 2018). We draw a distinction between the attributes explainable and explicable within the AI context in stating that making facts explicable is a sub-task of the explanation task, meaning that purely explicating facts is not enough for humans to build an understanding. In terms of our cAI terminology (see **Figure 3**), ML models and results need to be explicable so that they are transparent to human users, but they need to be explainable for being comprehensible, too. We therefore refer to explicability as a property, which forms the basis for explainability and states that something potentially can(!) be explained, but it doesn't necessarily correspond to the concrete explanation for a certain set of facts in rationale terms. The focus of explaining can be differentiated regarding the explanation of the reasoning, the model or the evidence for the result (Biran and Cotton, 2017). However, in all cases, the goal of the explanation task can be seen as updating the humans' mental models (Chakraborti et al., 2018), where good explanations must be relevant to a, potentially implicit, human question as well as relevant to the mental model of the explainee (Miller, 2019).

Explanations can provide a valuable basis for providing transparency and comprehensibility regarding systems' decisions and therefore can lead to increased trust of ML users (Pu et al., 2011; Prahl and Swol, 2017; Miller, 2019). A high level of initial trust in ML systems, which often decreases rapidly in case of erroneous or unexpected reactions (Madhavan and Wiegmann, 2007), as well as interaction and influencing possibilities might be an acceptance criterion for the usage of such systems (Schaefer et al., 2016). As illustrated in **Figure 3**, we distinguish between two different shapings of the cognitive concept of interpretation—namely iML and xAI, which differ in the kind of understanding as well as in the way explanations are revealed. In our opinion, iML focuses on using or generating global interpretability by providing intrinsic—*ex ante*—understanding of the whole logic of the corresponding models (Adadi and Berrada, 2018). Global explanations therefore relate to the inner functioning of models, meaning the entire and general behavior in terms of the entire reasoning describing HOW the systems work internally. Hence, the scoop of this type of interpretability is to inform about the global effects giving some indication on the real concepts that a system has learned. The explanandum is therefore the ML model itself where we consider the *rules*

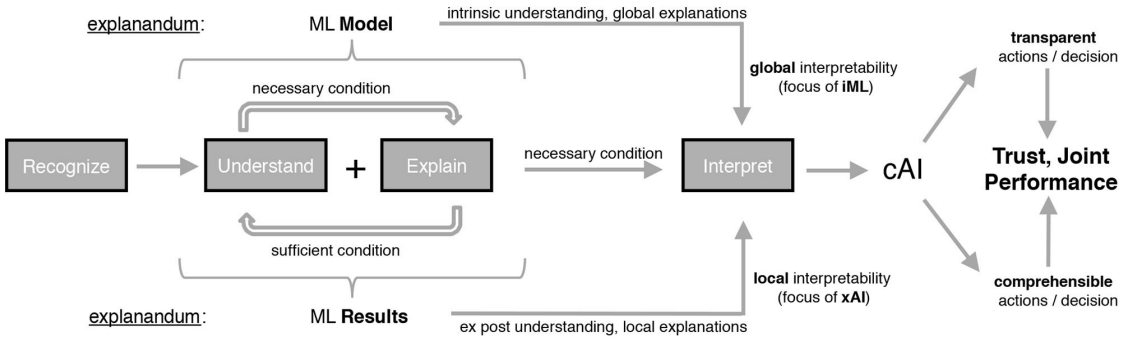


FIGURE 3 | Derivation of cAI from iML and xAI considering the underlying cognitive concepts.

of reasoning as the explanans giving information about how all of the different possible outcomes are connected to the inputs. On the other hand, we see xAI's focus more on enabling local interpretability by providing an *ex post*-understanding of the model's specific behavior. Local explanations for individual decisions or single predictions strive for making the input-output-correlations clear to the user without the need for knowing the model's internal structure (Adadi and Berrada, 2018). Thus, the scoop of this type of interpretability is to make justifications WHY a model produced its output in the way it did. The explanandum is therefore an individual ML result or a group of results where we see the occurrences, importances, and correlations of input features as the explanans giving information about the logical and causal correlations of inputs and outputs. The two dimensions spanned by cAI, in our understanding of interpretability, namely transparency and comprehensibility (see **Figure 3**), might aim at different requirements of different kind of users. Therefore, we refer to transparency as a property especially relevant to domain or ML experts that are not solely interested in why a certain output was made but also trying to explore the nature and characteristics of the underlying concepts and its context. In contrast, we refer to comprehensibility as a requirement raised particularly by humans that are directly affected by the outputs and the correlated consequences trying to understand why a specific decision was made. We define the overall objective of cAI as developing transparent and comprehensible AI systems that humans can trust in as well as improving the systems' "joint performance," both by means of global interpretability (iML) in combination with local interpretability (xAI). Depending on the domain and the ML problem to be solved an adaptive combination of white-box approaches and black-box approaches with connected explanation generators and interfaces (gray-box approach) will be necessary in order to reach cAI.

Figure 4 illustrates our suggestion for a possible transition framework, which includes interdisciplinary concepts, approaches and measures to reach cAI and thus the next level of transparent and interactive companions for decision support. As discussed, current ML approaches lack conceptual properties like interpretability of the model as well as the results. Additionally, missing reproducibility of ML predictions and the according explanations imposes requirements on a concept called *confidence*, which the authors from Arrieta et al. (2020) refer to as a generalization of robustness and stability of ML approaches. Furthermore and due to missing interpretability, state of the art ML systems often do not provide any possibility for human interaction, since humans are not able to understand the rules the system has learned. Therefore, any correction of erroneous rules or any inclusion of domain-specific knowledge through human experts (i.e., physicians) is not possible. In addition, the points of criticism mentioned so far also lead to tremendous potential for clarification in terms of the relationship between AI and applicable law. Legal security and liability security will play a crucial role in the near future. As an example, in the medical domain the question of liability arises, especially if a patient suffers damages that are caused by a medical treatment of a physician who acted on the recommendation of an ML approach. Additionally, we consider socio-technical and domain aspects as other important conceptual properties, since in most cases ML pipelines need to be adapted to the according context of the problem to be solved. In the same way, explanation and interpretation techniques need to be in accordance with the individual domain and social as well as ethical requirements. Causality is another necessary concept (Pearl, 2009) and refers to making underlying mechanisms transparent beyond computing correlations (Holzinger et al., 2019) to derive the *true* reasons that lead to a particular outcome. Therefore, causality depends on available interpretability and explainability of models. This requirement as precondition to causality can be referred to as causability and is currently examined in the context of explanation evaluation, especially for the medical domain (Holzinger et al., 2019). Analogously to our differentiation between explicability and explainability, we strongly agree with the authors from Holzinger et al. (2019) that results gained from explainable and interpretable models should not only be usable but also useful to humans. In this regard they refer to Karl Popper's hypothetical deductive model in order to derive facts from laws and conditions in a deductive manner by causal explanations. Bias awareness as further concept focuses on avoiding ML-related biases in predictive modeling like sample

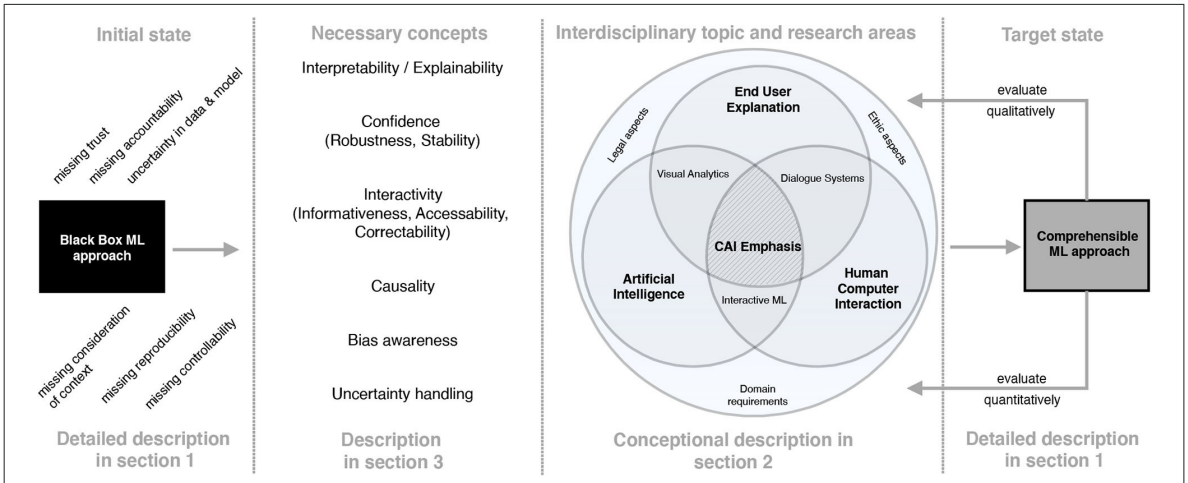


FIGURE 4 | cAI transition framework using interdisciplinary concepts, approaches and measures to reach next level of AI.

bias, exclusion bias, label bias, bias in ground truth as well as other more general biases like observer bias, prejudice bias and measurement bias. A remedy can be to use techniques such as FairML, which is a toolbox for diagnosing bias in predictive modeling (Sgaard et al., 2014; Adebayo, 2016). Uncertainty is another concept that should be taken into account. In ML two types of uncertainty are distinguished (Kendall and Gal, 2017). Uncertainty that originates from noise in observations, meaning for example missing measurements, irrelevant data or mislabeled examples, is called *aleatoric* uncertainty. The other type of uncertainty is called *epistemic* uncertainty. It refers to uncertainty that results from the model. In particular in image classification, approaches such as Bayesian deep learning can be applied and extended to handle and explicate uncertainties.

For enabling such conceptual properties, an integration of concepts, approaches, techniques and measures from a variety of disciplines is necessary as depicted in **Figure 4**. We refer to and extend a proposal from the Defense Advanced Research Projects Agency (DARPA) to elaborate cAI emphasis by showing relevant research disciplines and its relationships to AI (Gunning, 2016). In this context, the emphasis of cAI is defined as an overlapping of the disciplines AI, Human Computer Interaction (HCI) and End User Explanation with its interdisciplinary techniques and approaches like visual analytics, interactive ML and dialog systems. Furthermore, domain requirements, legal as well as ethic aspects participate and contribute to an overall understanding of cAI.

3. FUNDAMENTALS OF cAI TRANSITION APPLIED TO MEDICINE

The relevance of cAI becomes clear when ML is applied to medicine. In common, medical sub-disciplines rely on high sensitivity and specificity of diagnostic decisions. In order to choose the right therapy and to avoid delays in treatment caused by initial misdiagnosis, neither false alarms nor miss outs are desirable. Several recent studies show that ML can help to increase the accuracy of diagnosis (Weng et al., 2017; Haenssle et al., 2018; Hu et al., 2019). Applying ML therefore has the potential to save lives and resources. Especially sub-disciplines that are based on image processing and classification, like histology, could benefit from high performing approaches such as convolutional neural networks (Buetti-Dinh et al., 2019). However, since these approaches remain a black-box, medical experts cannot comprehend why a certain classification was performed and thus convolutional neural networks should not be applied in decision-critical tasks unless their predictions are made comprehensible and robust. Even though an ML approach shows a high classification accuracy, it still might be biased (Gianfrancesco et al., 2018). In the following sections we present the cornerstones as well as some specific approaches for improving comprehensibility of expert companions for the medical domain.

3.1. Explanation Generation and Visual Analytics

Visual analytics techniques, which in our transition framework from **Figure 4** are located at the intersection of AI and End User Explanation, can be used to provide visualizations that are helpful for humans to interpret according models or its results. Therefore, human comprehensible End User Explanations need to be built on top of formal explanations by considering and using knowledge from psychological and philosophical investigations. These, inter alia, strive for the generation of explanations understandable for humans and for an efficient communication by conveying the causal history of the events to be explained (Lewis, 1986). As a consequence, most state of the art explanation generators try to use visualization techniques in order to generate explanations that are relevant both to the implicit questions of the explainees as well as to their mental models (Miller, 2019).

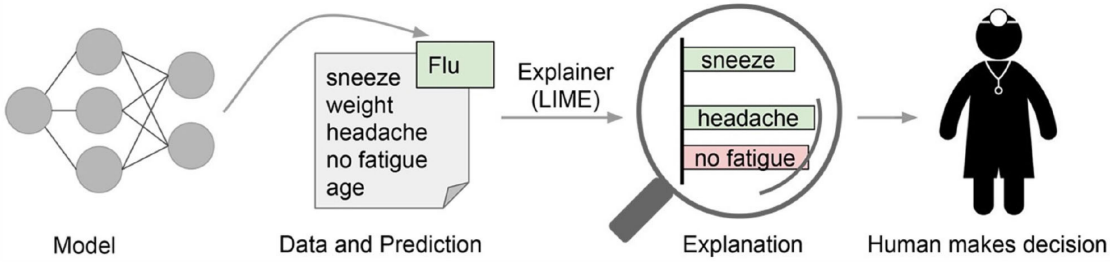


FIGURE 5 | LIME: Explaining individual predictions in the medical context (Ribeiro et al., 2016).

A prominent xAI technique, which allows for local, model-agnostic and *post-hoc* interpretations by approximating black-box models locally in the neighborhood of predictions of interest, was proposed by Ribeiro et al. (2016). LIME uses a local linear explanation model and can thus be characterized as an additive feature attribution method (Lundberg and Lee, 2017). Given the original representation $x \in \mathbb{R}^d$ of an instance to be explained, $x' \in \{0,1\}^{d'}$ denotes a binary vector for its interpretable input representation. Furthermore, let an explanation be represented as a model $g \in G$, where G is a class of potentially interpretable models like linear models or decision trees. Additionally, let $\Omega(g)$ be a measure of complexity of the explanation $g \in G$, for example the number of non-zero weights of a linear model. The original model that we are searching explanations for is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A measure $\pi_x(z)$ defining the locality around x is used that captures proximity between an instance z to x. The final objective of LIME is to minimize a measure $\mathcal{L}(f, g, \pi_x(z))$ that evaluates how unfaithful g is in approximating f in the locality defined by $\pi_x(z)$. Striving for both interpretability and local fidelity, a LIME explanation is obtained by minimizing $\mathcal{L}(f, g, \pi_x(z))$ as well as keeping $\Omega(g)$ low enough to be an interpretable model:

$$\xi(x) = \underset{g \in G}{\arg\min}\ \mathcal{L}(f, g, \pi_x(z)) + \Omega(g) \tag{1}$$

For being a model-agnostic explainer, the local behavior of f must be learned without making any assumptions about f. This is achieved by approximating $\mathcal{L}(f, g, \pi_x(z))$, drawing random samples weighted by $\pi_x(z)$. Having drawn non-zero elements of x' uniformly at random, a perturbed sample $z' \in \{0,1\}^{d'}$ is obtained. Recovering z from z' and applying $f(z)$ then yields a label, which is used as label for the explanation model. The last step consists of optimizing Equation (1), making use of dataset $\mathcal{Z}$ that includes all perturbed samples with the associated labels. **Figure 5** depicts an exemplary explanation process of LIME in the medical domain that explains why a patient was classified as having the flu by portraying the features *sneeze* and *headache* as positive contributions to having the flu, while *no fatigue* was considered as evidence against the flu. Other techniques for generating explanations, especially for concrete predictions of neural networks, comprise Layer-wise Relevance Propagation (LRP), which identifies properties pivotal for a certain prediction, as well as neural network rule extraction techniques like Neurorule, Trepan, and Nefclass (Beasens et al., 2003; Lapuschkin, 2019). All of these approaches share in common that they either provide explanations in terms of visualizations by showing the most important features relevant for a single prediction or by providing rules that are represented as decision table. As an example, Binder et al. (2018) developed an approach for predictive learning of morphological and molecular tumor profiles. In addition to purely focusing on prediction accuracy, the authors applied LRP in order to analyze the non-linear properties of the learning machine by mapping the results of a prediction onto a heatmap that reveals the morphological particularities of the studied pathological properties. Hägele et al. (2019) analyzed histopathological images and applied LRP for visual and quantitative verification of features used for prediction as well as for detection of various latent but crucial biases using heatmapping.

Out of such explanations and visualizations, experts might get valuable interpretations, but to even improve interpretability especially for lay humans it could be helpful to include other explanation modalities. As an example, combining visual explanations with natural language explanations as well as allowing for more interactivity between ML systems and users could further improve trust in the system. Additionally, in our opinion the process of transferring and presenting generated explanations should be made up in a way such that semantic level of detail as well as semantic context are aligned between the ML system, the explanation system and the human user. Therefore, our transition framework includes an interdisciplinary, psychologically motivated research area that deals with *End User Explanations*. Psychological insights into the process of generating and communicating explanations can be derived from explanatory understanding (Keil, 2011). According to that, explanations reveal a transactional nature and communicate an understanding between individuals. Additionally, as humans adapt stances or modes of construal (Dennett, 1987) that frame explanations, the latter ones reveal an interpretative nature and require humans to perform mental calculations in order to understand explanations. Therefore, the authors from Sloman et al. (1998) and Ahn et al. (2000) name circularity, relevance and especially coherence as

further important dimensions that guide systematic evaluation of explanations. Coherence in the domain of explanations describes the fact that humans prefer explanatory features within induction, which are most causally interdependent on others and therefore coherent. Furthermore, explanations are deemed relevant and informative when being presented to humans at the correct level of semantic detail. In essence, high quality explanations stick together and represent an internally consistent package, whose elements form an interconnected, mutually supporting relational structure (Gentner and Toupin, 1986; Thagard, 2000).

Many state of the art explanation systems, especially those based on perturbations, reveal some significant drawbacks. One of them is the fact that they sample instances around the instance to be explained by drawing samples uniformly at random. Doing so they ignore feature dependence when sampling from a marginal distribution (Molnar, 2019). Thus, there is a high chance that subsequent explanation strategies put too much weight on unlikely data points and are therefore susceptible for extrapolation. In such a case, explanations can then easily be misinterpreted. As a further consequence, context between the explanation features is not considered, yielding explanations, where humans have to perform many mental calculation steps in order to interpret and understand the explanations properly. Another potential problem is described by the authors (Alvarez-Melis and Jaakkola, 2018), namely potential instabilities of explanations manifesting in great variances for explanations of two close data points. Due to the random-sampling-step, one of the necessary concepts from our transition framework, namely confidence, is often violated. The authors from Arrieta et al. (2020) refer to confidence as a generalization of robustness and stability, which are themselves also motivated by the problem of missing reproducibility of the ML predictions as well as the according explanations. Finally, missing context between explanation features can lead to a lack of semantic interactivity between ML system and human users, since humans think and explain via *semantic coherent concepts* that the explanation systems are often not able to deal with.

As LIME is a representative of perturbation-based explanation systems and constitutes state of the art within xAI for image as well as for text classification (both of which are highly relevant within the medical domain), we propose an architecture to overcome some of the drawbacks mentioned above especially for text classification combined with LIME. Therefore, we propose the integration of (a) a ML classification algorithm, (b) an explanation system like LIME as well as (c) a *semantic* approach. In text domain, the latter is represented as a text modeling approach, in specific a topic modeling approach like Latent Dirichlet Allocation (LDA) that captures semantic and contextual information of the input domain. The goal of this integrated architecture (as illustrated in **Figure 6**) is to provide the basis for *coherent* and therefore *human-interpretable*, *contextual* explanations and to enable insights into the classifier's behavior from conceptual point of view. Harnessing semantic and contextual meta-information of the input domain by learning human-interpretable latent topics with LDA enables a Perturbation-based Local Explanation Generator like LIME to sample from a realistic local distribution via topic-based perturbations. As a result, topic-encoded explanations are obtained, which allow humans to recognize correlations

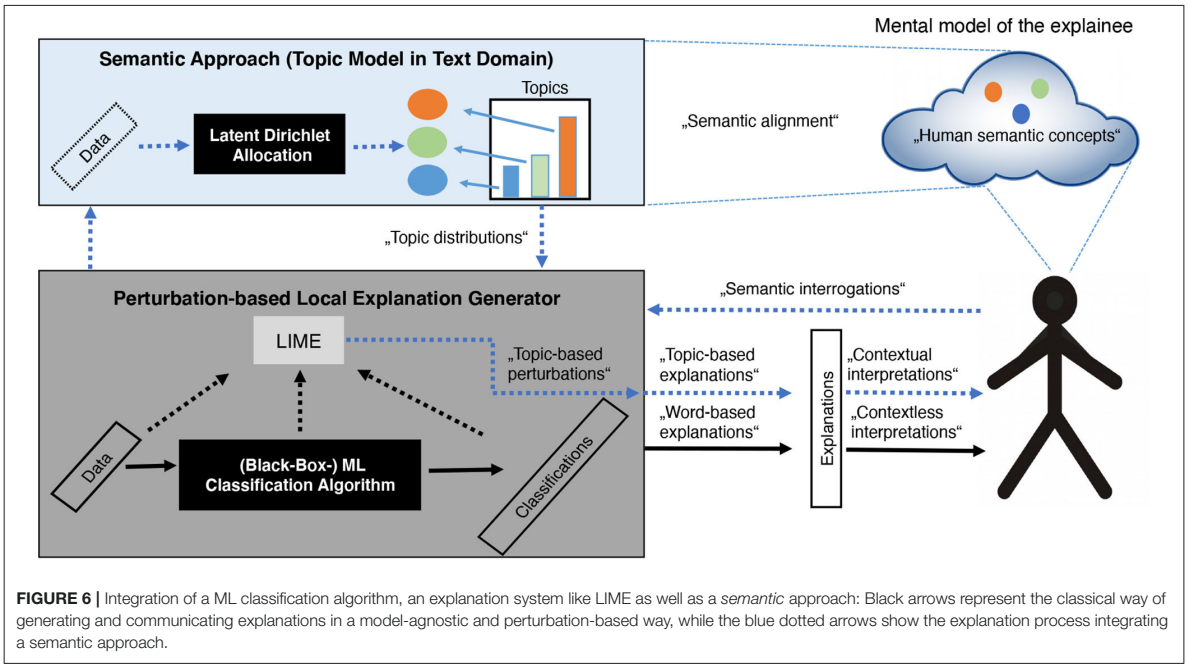


FIGURE 6 | Integration of a ML classification algorithm, an explanation system like LIME as well as a *semantic* approach: Black arrows represent the classical way of generating and communicating explanations in a model-agnostic and perturbation-based way, while the blue dotted arrows show the explanation process integrating a semantic approach.

(context) and to perform interpretations more intuitively by aligning the encoded semantic concepts with their mental model. Another interesting property of a combination of an explanation system combined with a semantic approach is its *semantic interrogation ability*. When it comes to the point, whether to trust a ML classifier, potential questions to be answered could be: "Does a classifier behave in a manner that is expected by humans?" or "How much does a classifier resemble human intuition". A semantically enriched architecture enables humans to generate documents that reveal a specific semantic content (represented as a mixture of certain topics and according words) as well as semantic structure. Presenting those user-specified documents to the classifier and receiving the according classifications, human users can interact with the classifier through an explanation system via *semantic interrogations*. These will be answered by the classification system with topic-based explanations allowing the user to interpret them in terms of human semantic concepts.

Applying such an architecture to the medical domain can help with improving and explaining automatic recognition of medical concepts in (un-)structured text (i.e., patient records), which is a complicated task due to the broad use of synonyms and non-standard terms in medical documents (Arbabi et al., 2019). In essence, better reproducibility of explanations can be achieved by reducing randomness during perturbation by the integration of *semantic sampling* that also allows to generate contextual explanations, which in turn can be interpreted by humans more intuitively.

3.2. Verbal Explanations

As described in the previous subsection, providing visual explanations and semantics helps to increase the interpretability of opaque classifiers. In addition, natural language explanations constitute an important explanation modality, since, for their expressiveness, they capture complex relationships better than visualizations (Finzel et al., 2019; Rabold et al., 2019; Schmid and Finzel, 2020) and increase comprehensibility (Muggleton et al., 2018).

In our transition framework (see **Figure 4**) we include verbal explanations at the intersection of End User Explanation and HCI. We show in the following paragraphs that natural language plays a key role in enhancing the comprehensibility of classifier results and that it is an important modality to allow for meaningful interactions between the classifier and a medical expert. Medical diagnosis often relies on the visual inspection of image- or video-based data, such as microscopy images, cardiograms or behavioral data from videos (Schmid and Finzel, 2020). In many cases, diagnostic decisions are not made solely based on the mere occurrence or absence of symptoms and abnormalities. The analysis of images and videos often takes into account spatial information and spatial relationships between the entities of interest. Visual explanations are limited with respect to representing relations. Visualizations, such as heatmaps and superpixel-based highlights are restricted to presenting conjunctions of information, i.e., (co-)occurrence of entities of interest. Although negation can be encoded with the help of the color space (e.g., in LRP-based heatmaps, where highlights in a color opposite to positive relevance indicate that some important property is missing), interpreting and semantically embedding which property is negated in comparison to the properties of contrasting classes, remains the task of the human expert. Therefore, enhancing understanding by visual explanations is limited, since the latter can only be interpreted with respect to positions of entities and given conjunctions of highlights encoded by the color space. They lack to express more complex relationships, such as spatial relations between two or more entities. Arbitrary relationships and special cases of relational concepts, for example recursion, can be better represented in natural language. Therefore, verbal explanations better qualify for giving insights into causal chains behind classification and thus diagnostic problems. This is especially important, since expert knowledge is often implicit and making it explicit can be hard or even impossible for experts. Particularly interesting are therefore systems that are capable of learning relational rules, which can then be translated into natural language expressions for generating verbal explanations. As presented for example in Schmid and Finzel (2020), spatial relationships are considered in the analysis of microscopy images to verbally explain the classification of the depth of invasion for colon tumors. In this use case, not only the occurrence of tissues, but also the complex spatial relationships between different types of tissue must be taken into account. For example, if tumor tissue has grown passed muscle tissue and already invades fat, the tumor class is more critical compared to a tumor that resides within tissue of the mucosa (Wittekind, 2016). As further pointed out in Schmid and Finzel (2020), ML approaches should therefore be able to reveal which relationships lead to a certain classification. Furthermore, relationships should be communicated in a comprehensible way to medical experts and this can be achieved with the help of natural language explanations. In their project the authors utilize Inductive logic programming (ILP) to implement a comprehensible explanation interface for a *Transparent Medical Expert Companion*, a system that explains classification outcomes of black-box and white-box classifiers and allows for interaction with the medical expert. ILP is an ML approach that produces output that can be transformed into verbal explanations for classification outcomes. In the *Transparent Medical Expert Companion*, microscopy scans are classified either by human experts or by an end-to-end black-box ML system. In the given example (see **Figure 7**), target class is tumor class *pT3*. Scans that are classified as *pT3* are positive examples, scans with different classification are negative. Learning can be realized by a one-against-all-strategy or separated in different sub-problems, such as discriminating one target class from the most similar alternative classes. An ILP system can now be used to learn over the given examples. In **Figure 7**, an illustration for one learned rule is given. A new scan is classified as *pT3* if it fulfills all components of the rule. In order to transform such rules into verbal explanations, methods similar to those introduced in the context of expert systems can be utilized (Schmid and Finzel, 2020).

In addition, experts can still provide their knowledge to the algorithm, as illustrated in **Figure 8**, where an exemplary spatial relationship *touches* is defined in the background knowledge

```
positive examples for diagnostic class pT3
---------------------------------------
scan123 is classified as pT3. The scan is composed of areas
of different tissues such as fat and tumor
which are in specific spatial relations.

pt3(scan123).
contains_tissue(scan123,t1).
contains_tissue(scan123,f1).
contains_tissue(scan123,f2).
is_tumor(t1).
is_fat(f1).
is_fat(f2)
touches(t1,f1).
disjoint(f1,t1).

negative examples for diagnostic class pT3 (e.g. pT2, pT4)
-----------------------------------------
...


Induced Rules:
A scan is classified as pT3
if a scan A contains a tissue B and B is a tumor and B touches C
and C is fat.

pT3(A) :-
   contains_tissue(A,B), is_tumor(B), touches(B,C), is_fat(C).

further rules ...
```

FIGURE 7 | Training examples and learned rules for a hypothetical diagnostic domain of colon cancer (Schmid and Finzel, 2020).

```
Background Theory for Spatial Relations
---------------------------------------
Area X touches area Y if holds that they have at least one boundary
point in common, but no interior points.

touches(X,Y) :-
   I is intersection(X,Y), not(empty(I)),
   InteriorX is interior(X), InteriorY is interior(Y),
   J is intersection(InteriorX,InteriorY), empty(J).

disjoint(X,Y) :- ...
includes (X,Y) :- ...
...
```

FIGURE 8 | Background theory with domain rules for a hypothetical diagnostic domain of colon cancer (Schmid and Finzel, 2020).

and can thus be found by the algorithm in the data if relevant to the classification of *pT3*. It has been shown that due to the implicitness of expert knowledge and variants in how health symptoms manifest, it is easier for an expert to determine why a certain example belongs to a diagnostic class rather than describing the class in its entirety (Možina, 2018). Rules learned by ILP can be traced, meaning that they can be applied to the background knowledge, which contains the data from examples like a data base. This way, the learned program, consisting of the learned rules and the data base, can explain its reasoning to the human expert. This is done by showing the output from the chain of reasoning steps, as it has been implemented for example in the diagnostic system MYCIN (Clancey, 1983). Traces can be translated into natural language expressions and then used in explanatory interaction in the form of a dialog between the system and the human expert, where the expert can ask for clarification in a step-wise manner. Research on how these dialogs could be implemented are concerned with

rule-based argumentation (Možina et al., 2007), argumentation schemes and the form of argumentative input [e.g., free-form, structured, or survey-based (Krening et al., 2017)]. Finally, natural language is the basis for more expressive correction of classification decisions, which will be discussed in the next subsection.

3.3. Interactive Machine Learning

Since medical knowledge changes steadily (which can lead to a bias in models learned on outdated data), ML approaches are needed that are able to adapt or that can be adapted easily by medical experts. This is where interactive ML comes into play. The main motivation of *human-in-the-loop*-based interactive ML is to build systems that improve their learning outcome with the help of a human expert who interacts with them (Holzinger, 2016). The human expert interacts for example with the data to improve the prediction outcomes and helps to reduce the search space through her expertise. The vision behind interactive ML is to "enable what neither a human nor a computer could do on their own" (Holzinger, 2016). Still, in the context of comprehensible interfaces for machine learned classifiers, mostly explanations are unidirectional—from the AI system to the human (Adadi and Berrada, 2018). Therefore, there exists a big potential for the medical domain to improve diagnosis with the help of developing new interactive ML approaches. State of the art approaches include systems where the human expert labels an example that was chosen by the algorithm according to some preference mechanism. In adherence to the so-called *active learning* paradigm, the system learns from the interaction with the user and may produce better prediction outcomes afterwards. Likewise experts can change labels of incorrectly classified examples or may add new examples with new labels in an incremental way. Furthermore, there exist approaches where the user has the possibility to indicate which features are relevant or irrelevant to a certain classification. An exemplary system is the EluciDebug prototype (Kulesza et al., 2015) for categorizing emails. After putting an incoming email into a certain folder, the system lists the words that have been considered as relevant. The user can adjust weights, for example to decrease the importance of words in order to remove them from decision rules. With the help of active learning particularly the data bias can be controlled by the human expert. There exist approaches and proposals for systems that offer explainable classification and allow user feedback in form of corrections beyond re-labeling and feature weighting that in a next step are used to adapt the machine learned model. One of the first approaches is the interactive learning system Crayon that enables the user to correct a classification of objects in an image by simply re-coloring some of the misclassified pixels (Fails and Olsen, 2003) to retrain the model. A second approach is named CAIPI (Teso and Kersting, 2019). It combines querying an example image, making a local prediction with a black-box learner and explaining the classification with an xAI approach, allowing the user to give feedback in the form of pixel re-coloring and re-labeling of false positives. Although both approaches offer promising ways of user interaction, they only take into account pixel-based visual information, omitting textual or relational information that might be relevant for expert decision making.

Interaction can be taken a step further. In domains where class decisions are based on complex relationships, interaction that allows for correction of relational models can improve the human-AI partnership (Schmid and Finzel, 2020). It has been shown also in other domains of AI that explanations can be used to revise current models (Falappa et al., 2002). A bi-directional exchange between an ILP system and a human expert is realized in the exemplary system *LearnWithME* (Schmid and Finzel, 2020) that integrates the principle of ME.

The aim behind the application *LearnWithME* is to provide medical experts a companion system for improved diagnosis. Companion systems serve as assistants to support humans in their daily or work routine. Adaptive machine learning, which incorporates interaction with the human and incremental learning, is suitable to enhance such companions (Siebers and Schmid, 2019). Furthermore, cognitive conditions imposed by the context

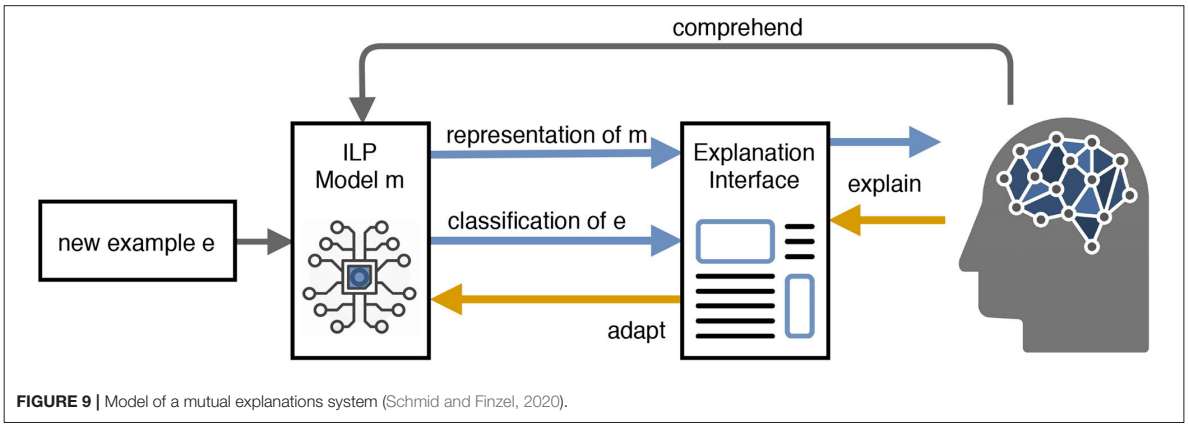


FIGURE 9 | Model of a mutual explanations system (Schmid and Finzel, 2020).

of use and the user should be considered (Cawsey, 1991, 1993).

Accordingly, the concept of *Mutual Explanations* is a cooperative, interactive and incremental act of information exchange between humans and machines with the goal to improve the joint performance of the involved partners in classification problems. The process of explanation refers (1) to providing arguments that make simple and complex relations, which apply to the domain of interest, explicit and (2) to integrating corrective explanations into existing internal models in order to adapt these (Schmid and Finzel, 2020). A model of such a ME system, which allows for bidirectional communication via explanations as well as interactive ML (corrections for model adaptation), is given in **Figure 9**: Starting with an initial ILP model, a new instance *e* is classified. The class decision for *e* is presented to the human who can accept the label or ask for an explanation. The explanation can be accepted or corrected with the help of defining constraints over the verbalized model at class level or at the level of the instance explanation.

Learning expressive, explicit rules rather than a black-box classifier has the advantage that generation of verbal explanations is quite straight-forward. However, in image-based medical diagnosis, it is clearly desirable to indicate a system's decision directly in the image. Often, only a combination of visual highlighting and verbal relational explanations allows to convey all information relevant to evaluate a decision. We believe that our cAI framework therefore provides a guideline to the development of interpretable systems for the medical domain by integrating visual and verbal explanations as well as interactive machine learning at the level of model adaptation through corrective feedback.

4. CONCLUSION

In the course of this paper we described why comprehensibility and interactivity will be crucial properties of modern ML systems in many application domains and especially for the task of designing transparent expert companions for the medical domain. Since thoughts on improved interpretability started to get considerable attention and many related concepts and terms have not been clearly defined yet, we introduced the term and concept of *Comprehensible Artificial Intelligence*. By describing and putting the basic cognitive concepts for cAI research and practice in relation, we were able to assign and discuss many current related research questions in an integrated manner from conceptual point of view. Furthermore, we gave a brief summary of connected interdisciplinary research areas and their overlappings, jointly being able to address many of the shortcomings mentioned in current literature. An integrated cAI transition framework was introduced revealing the guiding principles for exploring and implementing ML approaches that humans have trust in and can interact with. Our framework can be considered by developers and practitioners as a guideline to identify necessary concepts and possible solutions for their individual medical context. To the best of our knowledge, this has not been done yet beyond the scope of a literature review. We based our transition framework on theoretical foundation, derived practical implications and gave examples for possible solutions.

Following along our framework during some prototypical use cases, we identified *Semantic Alignment* between ML classifiers and human users, which is often overlooked in current approaches, as necessary prerequisites for comprehensibility as well as interactivity. Considering psychological insights from explanatory understanding, we proposed to properly account for the individual mental models of the explainees by integrating a semantic approach into a classification pipeline and presenting explanations at an appropriate level of semantic details. Especially when using black-box-algorithms and perturbation-based explanation systems, such an architecture can be used to enable realistic perturbations that reflect the underlying joint distribution of the input features and to generate meaningful, useful and more reproducible explanations. Our claim is that semantic and contextual information provided by the input domain must be taken into account during explanation generation and presentation, such that coherent and human-interpretable explanations are obtained bringing to light logical as well as causal correlations. For the task of classifying and explaining text documents being made of medical concepts, we describe a process that allows to find local topic-based explanations using topic models like Latent Dirichlet Allocation together with LIME. To even increase comprehensibility of explanations in terms of expressiveness, we suggest to include other explanation modalities as well. In addition to visual inspection as often conducted in medical diagnosis, verbal explanations and according methods to directly obtain them from classification systems are analyzed and shown exemplary with the help of Inductive Logic Programming. Furthermore, we provide the prospect of *Semantic Interrogations* to compare a classifier's semantic classification ability with human semantic concepts. As a kind of overall realization concept this paper introduces ME that in our opinion can provide a valuable basis for providing bidirectional information exchange between humans and machines. Summarizing and integrating all mentioned concepts in a single framework shall guide practitioners when attempting to create interactive, transparent and comprehensible ML systems that even laymen can interpret and build trust in.

Although many topics have been discussed with regard to the medical domain, the main points remain valid across different application domains. Adapting these approaches to the context of the individual problem as well as assessing explanations' quality quantitatively as well as qualitatively in a pragmatic way, these are the points that in our opinion constitute main future demands on cAI. Trying to anticipate ML's future in research and practice, we request for a stronger interdisciplinary thinking on cAI. This implies not just researching for formal explanations for ML systems and decisions, but trying to allow for an efficient generation and

transportation of interpretation artifacts to human users considering disciplines like explanatory understanding. It shall allow humans to gain a deeper understanding leading to improved interpretations forming the basis for transparent and comprehensible AI that we refer to as cAI.

AUTHOR CONTRIBUTIONS

SB, BF, and US made substantial contributions to conception and design of their approach. All authors involved in drafting the manuscript or revising it critically for important intellectual content, gave final approval of the version to be published, read, and approved the final manuscript.

FUNDING

Part of the work presented in this paper was funded by BMBF grant FKZ 01IS18056 B (ML-3 project Transparent Medical Expert Companion).

ACKNOWLEDGMENTS

We say many thanks to our project partners from the Fraunhofer IIS (Volker Bruns, Dr. Michaela Benz) and the University Hospital Erlangen (Dr. med. Carol Geppert, Dr. med. Markus Eckstein, and Prof. Dr. Arndt Hartmann, head of institute of pathology) who provided us the data and the knowledge about colon cancer microscopy scans.

REFERENCES

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Adebayo, J. A. (2016). *FairML: toolbox for diagnosing bias in predictive modeling* (Master's thesis), Massachusetts Institute of Technology, Institute for Data, Systems, and Society, Department of Electrical Engineering and Computer Science, Cambridge, MA, United States.

Ahn, W.-K., Kim, N. S., Lassaline, M. E., and Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cogn. Psychol.* 41, 361–416. doi: 10.1006/cogp.2000.0741

Alvarez-Melis, D., and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv* 1806.08049.

Arbabi, A., Adams, D. R., Fidler, S., and Brudno, M. (2019). Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Med. Inform.* 7:e12596. doi: 10.2196/12596

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Beasens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Manage. Sci.* 49, 312–329. doi: 10.1287/mnsc.49.3.312.12739

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., et al. (2018). Towards computational fluorescence microscopy: machine learning-based integrated prediction of morpho- logical and molecular tumor profiles. *arXiv* 1805.11178.

Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: a survey," in *IJCAI-17 Workshop on Explainable AI (XAI) Proceedings* (Melbourne). Available online at: https://ijcai-17.org/workshop-program.html.

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., et al. (2019). Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnol. Rep.* 22:e00321. doi: 10.1016/j.btre.2019.e00321

Burri, T. (2016). "Machine learning and the law: five theses," in *Conference on Neural Information Processing Systems (NeurIPS) (Short Paper)* (Barcelona). Available online at: https://nips.cc/Conferences/2016

Cawsey, A. (1991). "Generating interactive explanations," in *Proceedings of the Ninth National Conference on Artificial Intelligence*, Vol. 1 (Anaheim), 86–91. Available online at: https://dl.acm.org/doi/proceedings/10.5555/1865675

Cawsey, A. (1993). Planning interactive explanations. *Int. J. Man Mach. Stud.* 38, 169–199. doi: 10.1006/imms.1993.1009

Chakraborti, T., Sreedharan, S., and Kambhampati, S. (2018). "Balancing explicability and explanations emergent behaviors in human-aware planning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (Macao). doi: 10.24963/ijcai.2019/185

Clancey, W. J. (1983). The epistemology of a rule-based expert system—a framework for explanation. *Artif. Intell.* 20, 215–251. doi: 10.1016/0004-3702(83)90008-5

Dennett, D. (1987). *The Intentional Stance*. Cambridge MA: MIT Press.

Fails, J. A., and Olsen, D. R. Jr. (2003). "Interactive machine learning," in *International Conference on Intelligent User Interfaces*, Vol. 8 (Miami, FL), 39–45. doi: 10.1145/604045.604056

Falappa, M. A., Kern-Isberner, G., and Simari, G. R. (2002). Explanations, belief revision and defeasible reasoning. *Artif. Intell.* 141, 1–28. doi: 10.1016/S0004-3702(02)00258-8

Finzel, B., Rabold, J., and Schmid, U. (2019). "Explaining relational concepts: when visualizationand visual interpretation of a deep neural network's decision are not enough," in *European Conference on Data Analysis, Book of Abstracts* (Bayreuth), 60–61.

Gentner, D., and Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cogn. Sci.* 10, 277–300. doi: 10.1207/s15516709cog1003_2

Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* 178, 1544–1547. doi: 10.1001/jamainternmed.2018.3763

Gunning, D. (2016). *Explainable Artificial Intelligence (XAI)–Proposers Day*. Defense Advanced Research Projects Agency (DARPA).

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* 29, 1836–1842. doi: 10.1093/annonc/mdy166

Hägele, M., Seegerer, P., Lapuschkin, S., Backmayr, M., Samek, W., Klauschen, F., et al. (2019). Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Sci. Rep.* 10:6423. doi: 10.1038/s41598-020-62724-2

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* 3, 119–131. doi: 10.1007/s40708-016-0042-6

Holzinger, A. (2018). Explainable AI (ex-AI). *Inform. Spektrum* 41, 138–143. doi: 10.1007/s00287-018-1102-5

Holzinger, A., Biemann, C., Pattichis, C. B., and Kell, D. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv* 1712.09923.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining Knowl. Discov.* 9:e1312. doi: 10.1002/widm.1312

Hu, L., Bell, D., Antani, S., Xue, Z., Yu, K., Horning, M. P., et al. (2019). An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J. Natl. Cancer Inst.* 111, 923–932. doi: 10.1093/jnci/djy225

Keil, F. C. (2011). Explanation and understanding. *Annu. Rev. Psychol.* 57, 227–254. doi: 10.1146/annurev.psych.57.102904.190100

Kendall, A., and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision. *arXiv* 1703.04977. doi: 10.5555/3295222.3295309

Krening, S., Harrison, B., Feigh, K. M., Isbell, C. L., Riedl, M., and Thomaz, A. (2017). Learning from explanations using sentiment and advice in Rl. *IEEE Trans. Cogn. Dev. Syst.* 9, 44–55. doi: 10.1109/TCDS.2016.2628365

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, Vol. 15 (Atlanta, GA), 126–137. doi: 10.1145/2678025.2701399

Lapuschkin, S. (2019). *Opening the machine learning black box with layer-wise relevance propagation* (Dissertation). Fraunhofer Heinrich Hertz Institute Berlin.

Lewis, D. K. (1986). "Causal explanation," in *Oxford Scholarship Online: Philosophical Papers 2* (Oxford). doi: 10.1093/0195036468.003.0007

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing* (Red Hook, NY), 4768–4777.

Madhavan, P., and Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Hum. Factors* 49, 773–785. doi: 10.1518/001872007X230154

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Molnar, C. (2019). *Interpretable Machine Learning*. Christoph Molnar. Available online at: https://christophm.github.io/interpretable-ml-book/.

Možina, M. (2018). Arguments in interactive machine learning. *Informatica* 42, 53–59.

Možina, M., Žabkar, J., and Bratko, I. (2007). Argument based machine learning. *Artif. Intell.* 171, 922–937. doi: 10.1016/j.artint.2007.04.007

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., and Besold, T. (2018). Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Mach. Learn.* 107, 1119–1140. doi: 10.1007/s10994-018-5707-3

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Prahl, A., and Swol, L. V. (2017). Understanding algorithm aversion: when is advice from automation discounted? *J. Forecast.* 36, 691–702. doi: 10.1002/for.2464

Pu, P., Chen, L., and Hu, R. (2011). "A user-centric evaluation framework for recommender systems," in *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, IL), 157–164. doi: 10.1145/2043932.2043962

Rabold, J., Deininger, H., Siebers, M., and Schmid, U. (2019). "Enriching visual with verbal explanations for relational concepts-combining lime with aleph," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Würzburg), 180–192. doi: 10.1007/978-3-030-43823-4_16

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 1135–1144. doi: 10.1145/2939672.2939778

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 58, 377–400. doi: 10.1177/0018720816634228

Schmid, U., and Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. *Künstliche Intell.* 34, 227–233. doi: 10.1007/s13218-020-00633-2

Schurz, G. (2002). *Erklären und verstehen: Tradition, transformation und aktualität einer klassischen kontroverse*. Erfurt: Philosophical Prepublication Series at the University of Erfurt.

Sgaard, A., Plank, B., and Hovy, D. (2014). "Selection bias, label bias, and bias in ground truth," in *Proceedings of COLING 2014, The 25th International Conference on Computational Linguistics: Tutorial Abstracts* (Dublin), 11–13.

Siebers, M., and Schmid, U. (2019). Please delete that! Why should I? Explaining learned irrelevance classifications of digital objects. *Künstliche Intell.* 33, 35–44. doi: 10.1007/s13218-018-0565-5

Sliwinski, J., Strobel, M., and Zick, Y. (2017). "A characterization of monotone influence measures for data classification," in *IJCAI-17 Workshop on Explainable AI (XAI) Proceedings* (Melbourne).

Sloman, S. A., Love, B. C., and Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cogn. Sci.* 22, 189–228. doi: 10.1207/s15516709cog2202_2

Teso, S., and Kersting, K. (2019). "Explanatory interactive machine learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI), 239–245. doi: 10.1145/3306618.3314293

Thagard, P. (2000). *Coherence in Thought and Action*. Cambridge MA: MIT Press.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12:e0174944. doi: 10.1371/journal.pone.0174944

Wittekind, C. (2016). *TNM: Klassifikation maligner Tumoren*. Weinheim: John Wiley & Sons.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A.2 Concept-based Explainable Machine Learning for Text Classification

A.2.1 Kiefer, S. "CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge." In: Information Fusion 2022

Full Reference of Paper

Sebastian Kiefer. "CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge." In *Information Fusion, 77* (2022), pp. 184-195. DOI: 10.1016/j.inffus.2021.07.014.

My Scientific Contributions

Related research question(s): **RQ2a - RQ2c**.
Type of contribution: **technical**.

During the research underlying this dissertation, I made the following scientific contributions published in this paper:

- Identification of the research gap of missing contextual and semantic information in model-agnostic explanations generated by LIME text explainer.
- Collection and structuring of literature on state-of-the-art explanation techniques, especially on model-agnostic explainers, like LIME or SHAP.
- Identification of major drawbacks of current perturbation-based explanation methods and its implications on human interpretability.
- Proposition of the CaSE architecture that is motivated by insights from the psychological field of "explanatory understanding".
- Implementation of topicLIME as an instantiation of the CaSE framework.
- Definition of the evaluation methodology comprising new evaluation criteria called "consistency property" and "combined removal impact".
- Outlining future research, like semantic explanatory interactive ML by providing the prospect of "semantic interrogations".

My Written Contents

I wrote all of the contents as single author. My written contribution to the paper is 100%.

Information Fusion 77 (2022) 184–195

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge

Sebastian Kiefer

Cognitive Systems, University of Bamberg, Bamberg, Germany

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence
Interpretable Machine Learning
Explanatory understanding
Human-like explanations
Contextual and semantic interrogations
Topic modeling

ABSTRACT

Generating explanations within a local and model-agnostic explanation scenario for text classification is often accompanied by a local approximation task. In order to create a local neighborhood for a document, whose classification shall be explained, sampling techniques are used that most often treat the according features at least semantically independent from each other. Hence, contextual as well as semantic information is lost and therefore cannot be used to update a human's mental model within the according explanation task. In case of dependent features, such explanation techniques are prone to extrapolation to feature areas with low data density, therefore causing misleading interpretations. Additionally, the "the whole is greater than the sum of its parts" phenomenon is disregarded when using explanations that treat the according words independently from each other. In this paper, an architecture named CaSE is proposed that either uses Semantic Feature Arrangements or Semantic Interrogations to overcome these drawbacks. Combined with a modified version of Local interpretable model-agnostic explanations (LIME), a state of the art local explanation framework, it is capable of generating meaningful and coherent explanations. The approach utilizes contextual and semantic knowledge from unsupervised topic models in order to enable realistic and semantic sampling and based on that generate understandable explanations for any text classifier. The key concepts of CaSE that are deemed essential for providing humans with high quality explanations are derived from findings of psychology. In a nutshell, CaSE shall enable *Semantic Alignment* between humans and machines and thus further improve the basis for Interactive Machine Learning. An extensive experimental validation of CaSE is conducted, showing its effectiveness by generating reliable and meaningful explanations whose elements are made of contextually coherent words and therefore are suitable to update human mental models in an appropriate way. In the course of a quantitative analysis, the proposed architecture is evaluated w.r.t. a consistency property and to Local Fidelity of the resulting explanation models. According to that, CaSE generates more realistic explanation models leading to higher Local Fidelity compared to LIME.

1. Introduction

Although modern ML approaches improved tremendously in terms of prediction accuracy and are able to even exceed human performance in many cases, they currently lack the ability to provide an explicit declarative knowledge representation and therefore hide the underlying explanatory structure [1]. Exactly that missing transparency makes it difficult for users of ML techniques to develop an understanding of the recommendations and decisions, a fact that often leads to an inherent risk [2].

A scientific field named Interpretable Machine Learning focuses on using or generating global interpretability by providing intrinsic – ex ante – understanding of the whole logic of the corresponding models [3]. In contrast, methods from the research area Explainable Artificial Intelligence strive for encountering these problems by enabling local interpretability that shall provide an ex post understanding of the ML model's specific behavior [3]. Combining both results in Comprehensible Artificial Intelligence, which strives for generating results that are transparent and comprehensible [4]. Those two properties form the basis of trust in AI. A prominent additive feature attribution method, which allows for local, model-agnostic and post hoc interpretations by approximating black-box models locally in the neighborhood of predictions of interest, was proposed by Ribeiro et al. [5]. LIME tries to minimize a locality-aware loss by an approximation technique that in the NLP domain samples instances by randomly removing features yielding perturbed instances, for which an interpretable model is learned [5]. Although current explanation frameworks often enable humans to gain insights into the classifier's behavior, most of them

E-mail address: sebastian.kiefer@uni-bamberg.de.

https://doi.org/10.1016/j.inffus.2021.07.014
Received 19 November 2020; Received in revised form 21 May 2021; Accepted 25 July 2021
Available online 2 August 2021

consider features, at least semantically, independently from each other. Therefore, they cannot provide any contextual nor semantic information during the explanation process. This paper contributes in three ways.

First, a general architecture named *Contextual and Semantic Explanations* (CaSE) is proposed that allows a combination of semantic knowledge with any text classification- and according local explanation approach. In particular it is shown, how topic models trained in an unsupervised manner can be leveraged to provide structural information in terms of semantics and context of text documents for subsequent explanations. Such an extended architecture is able to guide and improve an explanation process in a way that it generates coherent explanations. These shall be relevant to a, potentially implicit, human question as well as relevant to the mental model of the explainee [6].

The second contribution is to describe in detail, how explanations for a text classifier can be generated such that they reveal some desirable properties like coherence as demanded within scientific philosophy. This paper states that explanations should be built as a statement of coherent facts such that they support each other. A coherent fact set then can be interpreted in a context that covers all or most of the facts and therefore improves human understandability [7,8]. In the course of this work, Latent Dirichlet Allocation (LDA)-topics are leveraged to assign words to coherent word groups that capture semantic and contextual information. For harnessing those semantic information during the explanation generation, *topicLIME*, which uses *Semantic Feature Arrangements*, conducts topic-based perturbations and outputs topic attributions together with the assigned coherent words as explanations, is introduced.

Third, an algorithm named *Semantic Interrogations* is proposed. It embeds CaSE into an interactive process, in which humans can compare their inherent mental models of the classification domain with the functionality of the ML classifier.

Finally, the properties of topic-encoded explanations depending on the base classifier are analyzed and compared to word-encoded explanations within a quantitative experimental setting. It turns out that Local Fidelity is increased substantially when using topic-based perturbations that explicitly consider feature dependence.

2. Related work

This section provides a brief overview, including strengths and weaknesses, of current state of the art explanation strategies and, in particular, introduces LIME.

Ribeiro et al. proposed LIME, a method that explains an individual model's prediction by locally approximating the model's decision boundary in the neighborhood of the given instance [5]. LIME uses a local linear explanation model and can thus be characterized as an additive feature attribution method [9]. Given the original representation $x \in \mathbb{R}^d$ of an instance to be explained, $x' \in \{0,1\}^{d'}$ denotes a binary vector for its interpretable input representation. Furthermore, let an explanation be represented as a model $g \in G$, where G is a class of potentially interpretable models like linear models or decision trees. Additionally, let $\Omega(g)$ be a measure of complexity of the explanation $g \in G$, for example the number of non-zero weights of a linear model. The original model that we are searching explanations for is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A measure $\pi_x(z)$ defining the locality around x is used that captures proximity between an instance z to x. The final objective of LIME is to minimize a measure $\mathcal{L}(f, g, \pi_x(z))$ that evaluates how unfaithful g is in approximating f in the locality defined by $\pi_x(z)$. Striving for both interpretability and local fidelity, a LIME explanation is obtained by minimizing $\mathcal{L}(f, g, \pi_x(z))$ as well as keeping $\Omega(g)$ low enough to be an interpretable model:

$$\xi(x) = \underset{g \in G}{\arg\min}\ \mathcal{L}(f, g, \pi_x(z)) + \Omega(g) \tag{1}$$

For being a model-agnostic explainer, the local behavior of f must be learned without making any assumptions about f. This is achieved by approximating $\mathcal{L}(f, g, \pi_x(z))$, drawing random samples weighted by $\pi_x(z)$. Having sampled instances around x' by drawing nonzero elements of x' uniformly at random, a perturbed sample $z' \in \{0,1\}^{d'}$ is obtained. Recovering z from z' and applying $f(z)$ then yields a label, which is used as label for the explanation model. The last step consists of optimizing Eq. (1), making use of dataset $\mathcal{Z}$ that includes all perturbed samples with the associated labels. For even more expressive explanations LIME can be combined with Inductive Logic Programming (ILP) in order to generate first-order rules as explanations [10].

Ribeiro et al. also proposed another novel model-agnostic explanation system that uses high-precision "if-then-rules" called *anchors*, which represent local, sufficient conditions for predictions of interest [11]. Hence, the authors claim to contribute with intuitive, faithful and easy to comprehend explanation rules that clearly state their generalizability in terms of a coverage measure and therefore solve the *unclear coverage problem*. In this context, unclear coverage describes the fact that for local explanation techniques like LIME, it is not clear, whether the found local explanations also apply to unseen instances [11]. In other words, the degree of "global correctness" of the given explanations is unclear (i.e. in which regions these explanations apply). As a consequence, humans might draw wrong conclusions when applying insights from an explanation to other instances.

SHAP (SHapley Additive exPlanations) has been presented by Lundberg et al. as a unified framework for interpreting predictions. The resulting SHAP values can be interpreted as a unified measure of additive feature importance [9]. Although Shapley values fulfill desirable properties like local accuracy (between the explanation model g and the original model f), missingness (such that missing features in the original input get a Shapley value of zero and thus have no attributed impact), and consistency (meaning consistent behavior between the prediction scores of two different models when removing a certain feature and the according attribution scores of that feature per model), its model-agnostic approximation technique *Kernel SHAP* is slow with regard to computational time complexity on the one hand. On the other hand, it is not capable of dealing with feature dependence as it involves sampling from a marginal distribution [12].

Local Rule-Based Explanations (LORE) [13] is another local approach for explaining the reasons for a decision of a certain instance. It comprises a model-agnostic method only applicable for relational tabular data that learns a local interpretable predictor from a neighborhood, which has been generated by a genetic algorithm synthetically. Just like LIME, the next step of LORE consists of extracting the reasons for the individual decision from the local surrogate. In order to account for meaningful derived explanations, LORE on the one hand uses a decision rule for explaining the logic behind the decision. On the other hand, it is capable of producing a set of counterfactual rules, where changes of the according features necessary to achieve another prediction are explicated.

Many of those model-agnostic approaches reveal some significant drawbacks. First, most perturbation-based interpretation methods ignore feature dependence due to the fact that they replace feature values with values from random instances, as it is commonly easier to randomly sample from a marginal distribution. Thus, in case of dependent, for instance correlated features, such techniques are likely to put too much weight on unlikely data points due to extrapolation. In that case according explanations could potentially be misinterpreted as they are based on unrealistic samples. Unlikely data points harnessed for explanation generation can also be a common problem when using LIME, since samples are drawn from a Gaussian distribution not considering any dependencies like correlations between features [12]. Another major problem related to LIME is discussed by the authors Alvarez-Melis et al. namely potential instabilities of explanations manifesting in great variances for explanations of two close data points [14]. Consequently, finding realistic local perturbation distributions, which are expressive with regard to the original model's behavior while being interpretable and maintaining Local Fidelity is a major challenge [11]. The proposed

approach CaSE aims at encountering the disadvantages like treating features independently and therefore ignoring context. Its goal is to improve Local Fidelity as well as human comprehensibility by reducing the amount of mental calculations that humans have to perform in order to interpret an explanation by constraining explanations to offer a "contextually coherent frame".

3. CaSE

The CaSE architecture (see Fig. 1) is characterized by the integration of (a) an ML classification algorithm, (b) an explanation system like LIME as well as (c) a *semantic* approach. In a text domain the latter is represented as a text modeling approach, in specific a topic modeling approach like LDA that captures semantic and contextual information of the input domain. Providing the basis for *coherent* and therefore *comprehensible* and *contextual* as well as realistic explanations with high Local Fidelity is the major goal of CaSE. As a side note, it is worth mentioning that the proposed architecture is not limited to the domain of text. It can especially be extended to multimodal settings, where an LDA model fuses both text as well as image features (i.e. represented as *Bag of Visual Words* [15]) into a joint multimodal feature representation space. According to Holzinger et al. [16], harnessing conceptual knowledge "as a guiding model of reality" might help to develop more explainable and robust ML models, which are less biased. As CaSE captures semantic and conceptual information, it might pave the way for better interactive explainability and human-in-the-loop-learning.

This section is divided into five subsections. First, a reference is made to the psychological field of explanatory understanding, from which some essential concepts are utilized to build an adequate architecture for CaSE. The second subsection contains some theoretical foundations of LDA (see Section 3.2) that forms the semantic basis of the integrated CaSE approach. As the "number of topics k" represents an important LDA hyper-parameter, it is shortly explained, how to find an adequate number of such latent dimensions with the help of a C_v coherence measure that is introduced within Section 3.3. Sections 3.4 and 3.5 serve as an introduction to two variants of the semantic explanation framework CaSE, namely *Semantic Feature Arrangements* together with *topicLIME* and *Semantic Interrogations*. The first variant arranges all the features of a document to be classified in a contextual way using the LDA topic structure of the input domain and then proceeds with a revised version of LIME (topicLIME) that generates explanations in the form of topic attributions including the according coherent words. The second variant addresses the problem of finding a user-specified realistic (w.r.t. topic distributions of documents as well as resulting coherent and interdependent words) local perturbation distribution. It can be useful for human interaction with a classifier through perturbation explanation systems like LIME, SHAP or anchors. The way of creating a semantic neighborhood of a classified document that shall be explained can thus be user-parameterized. In so doing, *Semantic Interrogations* enable humans to ask the classifier, which semantic aspects it has learned and how it separates specific topics of interest from each other. This method allows humans to compare their expectations on how a specific classifier works within a specific domain with the actual behavior of the classifier.

3.1. Insights into explanatory understanding

To motivate and justify the concepts that CaSE builds upon, this subsection provides a brief overview of explanatory understanding, which deals with the process of creating and transferring explanations [17]. Explanations therefore reveal a transactional nature and reflect an attempt to communicate an understanding between individuals. Furthermore, as humans often adopt stances [18] or modes of construal that frame explanations, the latter ones are often interpretative in nature and therefore require humans to perform mental calculations in order to understand an explanation.

As for most natural as well as artificial phenomena, the full set of relations that shall be explained is enormous [19] and often overwhelming, explanations often contain compressed information that are outcomes of interpretative tricks. Beneath some pivotal properties of individual good explanations like accuracy, fidelity, consistency, stability, representativeness and comprehensibility [12], the authors from Ahn et al. as well as Sloman et al. identified softer attributes that humans put a demand on. Thus, explanatory features in category learning and induction that are early in a causal chain or are the most causally interdependent on others (centrality effect with causal relations) and therefore coherent are preferred by human users [20,21]. As an example, the concept of "boomerang" can be considered. It is connected with other concepts such as "throwing", "air" and "speed", which all are intricately connected. Such features that play an important role in commonsense are considered more essential in categorization than others that do not [20]. In general, circularity, relevance and coherence constitute three important dimensions that guide systematic evaluation of explanations [17]. As an example, a child might judge an explanation as satisfactory if it is relevant (to a "how" or "why" question), is not contradicting itself (i.e., being coherent) and is providing new information (not being circular) [22]. Explanations are further deemed relevant if they are presented to humans at the correct level of semantic detail. Additionally, explanations should stick together and represent an internally consistent package whose elements form an interconnected, mutually supporting relational structure [23,24].

In essence, human individuals prefer to provide and communicate their understanding as explanations, which are made of coherent elements and thus frame explanations with a context. CaSE, as described in the following, is designed to work analogously and to build upon concepts that reveal a close relation to the insights of explanatory understanding.

3.2. Latent Dirichlet allocation

LDA can be summarized as an unsupervised generative probabilistic three-level hierarchical Bayesian model for collections of discrete data [25]. In text modeling, its goal is to find short representations of the documents within a corpus while preserving essential statistical relationships like inter- or intra-document statistical structure. Therefore, LDA models documents from a corpus as an infinite mixture over an underlying set of topics. Each word is modeled as an infinite mixture over an underlying set of topic probabilities. In essence, within text modeling documents are represented as random mixtures over latent topics and each topic is defined by a distribution over words. For each document $\mathbf{w}$ in a corpus $\mathbf{D}$ a generative process from which the according documents have been created, is assumed:

1. Choose N (the number of words) $\sim$ Poisson(ξ).
2. Choose θ (a topic mixture) $\sim$ Dir(α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ).
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

The joint distribution of a topic mixture θ, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ given the hyper-parameters α and β is characterized by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta). \quad (2)$$

The *concentration* hyper-parameters α and β for the Dirichlet distribution enable the injection of prior beliefs about topic and word sparsity. For a symmetric Dirichlet distribution, high α-values lead to documents that with a high probability contain a mixture of many or most of the topics, whereas high β-values lead to topics that are likely

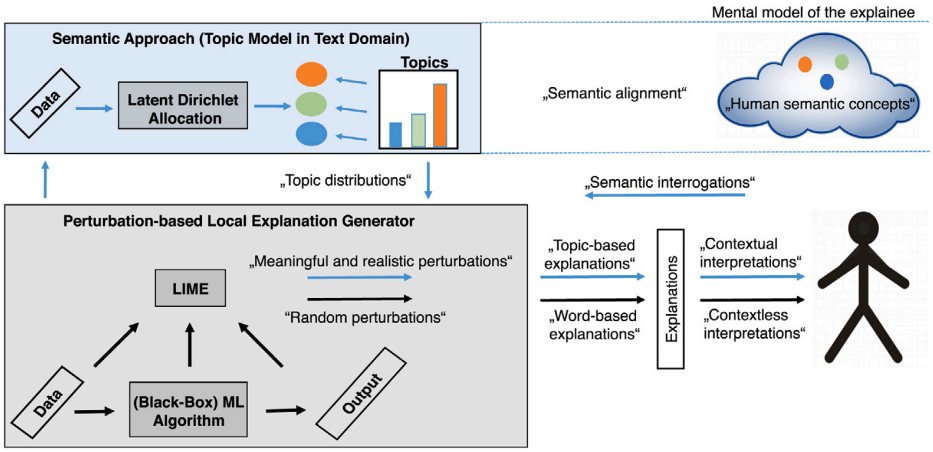


Fig. 1. Integration of an ML classification algorithm, an explanation system (LIME) as well as a semantic approach: Black arrows represent the classical way of generating and communicating explanations in a model-agnostic and perturbation-based way, while the blue arrows show the explanation process of CaSE integrating a semantic approach [4]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to be made of most of the words from the vocabulary. Another hyper-parameter, which has to be chosen and should be optimized by the user with regard to human topic-interpretability is the “number of topics k”.

The major inferential problem, which consists of computing the posterior distribution of the hidden variables given a document, is described as follows:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \tag{3}$$

As the posterior distribution from Eq. (3) is intractable to compute in general, LDA makes use of approximate inference algorithms like Laplace approximation, variational approximation or Markov chain Monte Carlo.

On the one hand, LDA can be used as dimension reduction technique, on the other hand, it is also capable of making sense of the input data due to its generative probabilistic semantics properties. Referring to the latent multinomial variables as topics enables LDA to capture text-oriented intuitions, global statistics in a corpus as well as synonymy and polysemy [25]. For evaluation, topic coherence measures are used that score a topic by measuring semantic similarity between highly contributing words within the according topic [26]. An important coherence measure, which approximates human ratings regarding semantic topics best [7], namely C_v coherence, is introduced in Section 3.3.

3.3. C_v coherence

Röder et al. found a combination of already existing coherence approaches, which they called C_v coherence, to be the best in terms of its correlation with respect to all available human topic ranking data [7]. Topic ranking data are a state of the art proxy for human topic-interpretability [8]. Therefore, C_v coherence is often used to choose the best hyper-parameter “number of topics k” within an LDA approach and to measure the quality of identified topics with regard to human interpretability.

This coherence measure combines an indirect cosine measure with *normalized pointwise mutual information* (NPMI) measure and a boolean sliding window [7]. C_v is obtained by combining four parts [8]:

1. Data segmentation: First, the top-words within each topic are paired with each other. Let W be a set of a topic‘s top-M most probable words.
2. Calculation of word probabilities: With Boolean document calculation the probabilities of single words or the joint probability of words from a word pair from step (1) are calculated. In order to at least partially consider word frequencies and distances a Boolean sliding window is used that slides over one word from a document per step.
3. Calculation of confirmation measure: For each segmented word pair (a pair of each word W′ combined with all other words) a semantic confirmation measure is calculated by representing words as context vectors (see Eq. (4)) that are obtained using NPMI measure (see Eq. (5)).

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|}. \tag{4}$$

γ is a free hyper-parameter used to put higher weights on higher NPMI values.

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log(\frac{P(w_i, w_j)+\epsilon}{P(w_i) * P(w_j)})}{-\log(P(w_i, w_j)+\epsilon)} \right)^{\gamma}. \tag{5}$$

$0 < \epsilon \leq 1$ is a free hyper-parameter that prevents the logarithm from being zero. Ultimately the cosine vector similarity of all context vectors is used to obtain the final confirmation measure.
4. The last step consists of applying the arithmetic mean of all confirmation measures in order to retrieve the final coherence score.

The use of a measure like C_v coherence as a quality criterion for human topic-interpretability is justified by the distributional hypothesis of linguistics [8]. It states that a difference of meaning correlates with a difference of distribution [27]. Expressed differently, words with similar meanings tend to occur in similar contexts. In the further course of this work, C_v coherence is used in order to find an adequate “number of topics k” of an LDA topic model such that topic coherence is maximized and thus high correspondence with human topic-interpretability is achieved.

3.4. Semantic feature arrangements for contextual explanations

Enabling an explanation context by offering coherent facts that support each other is a major goal of CaSE. For arranging input features from the input domain in a semantic, coherent and interpretable way an LDA topic model is trained and applied. With the main focus of identifying human-interpretable topics, the “number of topics k” is set in a way that it maximizes C_v coherence over all topics. Another possibility to automatically determine a suitable “number of topics k” is to use the Hierarchical Dirichlet Process [29].

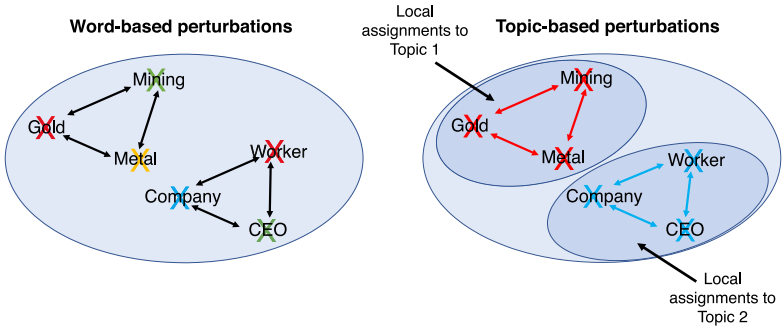


Fig. 2. Comparison of word-based and topic-based perturbation of a document: For word-based perturbation, one or more most probably independent words are removed at a time not considering its distributional as well as semantic relations with the other words (in this example: first both red, then both green, then single yellow, and finally the single blue crossed word). A contextual interpretation of the word-explanations is complicated as the semantic "links" of a word are not reflected in the explanations. For topic-based perturbation, coherent and most likely, at least semantically, related words are considered at once (in this example: all red crossed words belonging to topic 1 first, then all blue crossed words belonging to topic 2 next) including the semantic "links" that in turn provide the context in the explanations that are based on these perturbations. Furthermore, as feature dependence between words is considered during neighborhood generation, the common problem of interpreting explanations resulting from unlikely datapoints (as a consequence of extrapolation to feature areas with low data density [28]) is mitigated. By also allowing the explainee to leverage the LDA-topic-distribution of the document to be explained (refer to Section 3.5), prior knowledge in the form of domain knowledge can be integrated into the process of perturbation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For further using the interpretable components within an explanation scenario, these semantic topics on the one hand can be used without providing additional information and just harnessing the internal relational structure to offer contextual explanations. On the other hand, a semantic label can be assigned to each topic. One possibility is to conduct human labeling of topics via *eyeballing*. Therefore humans assign meaning to topics by supplying each topic with a title based upon the top n keywords of a topic [30]. As alternative, automated topic labeling approaches that either use unsupervised graph-based methods [31], top-ranking topic terms, titles and subphrases from Wikipedia [32] or sibling and parent–child relations [33], enrich interpretable topics with further summarized semantics.

Algorithm 1 Semantic Feature Arrangements t_{map} (by threshold)

Require: Topic Model lda, word w
Require: Minimum probability p_{min}
  $t_w \leftarrow \{\}$ ▷ the set of topics that w is assigned to
  for $topic \in lda$ **do**
    if $(w \in lda\_vocabulary \wedge p(w|topic) \geq p_{min})$ **then**
      $t_w \leftarrow t_w \cup topic$
    end if
  end for
  if $t_w \in \{\}$ **then**
    $t_w \leftarrow \{'UNKNOWN'\}$ ▷ Return topic named 'UNKNOWN'
  end if
  return t_w

Algorithm 2 Semantic Feature Arrangements t_{map} (maximum assignment)

Require: Topic Model lda, word w
  if $w \in lda\_vocabulary$ **then**
    return $\arg\max_{topic \in lda} p(w|topic)$
  else
    return $\{'UNKNOWN'\}$ ▷ Return topic named 'UNKNOWN'
  end if

Having retrieved the latent coherent topics from LDA, a *Semantic-Feature-Arrangement*-Algorithm assigns one or more topics to each word of the input document. Algorithms 1 and 2 show two possible variants how words can be assigned to semantically coherent topics. Algorithm 3 shows the overall explanation process, which is referred to as *topicLIME* in the course of this paper, as pseudo-code.

Algorithm 3 topicLIME

Require: Classifier f, Topic Model lda
Require: Text document x
Require: Number of topics T_k, Number of samples N
Require: t_{map}: function that produces Semantic Feature Arrangements
Require: Similarity kernel π_x, Length of explanation K
  $Z \leftarrow \{\}$
  for i ∈ {1,2,3,...,N} **do**
    $x'_i \leftarrow generate\_interpretable\_components(x, T_k, t_{map})$
    $z'_i \leftarrow randomly\_draw\_components(x'_i)$ ▷ topics are drawn randomly
    $z_i \leftarrow mask\_words(x, z'_i)$ ▷ mask all words (topic elements) in x that are not included in z'_i
    $Z \leftarrow Z \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $t \leftarrow$ K-Lasso(Z,K) ▷ with z'_i as features, f(z) as target
  return t ▷ returns as explanations a set of topics; each topic includes the assigned words

  function GENERATE_INTERPRETABLE_COMPONENTS(x,T_k,t_{map})
    $x' \leftarrow \{\}$ ▷ The interpretable version of x
    for $t \in \{1, 2, 3, ..., T_k\}$ **do**
      $w_t \leftarrow \{\}$
      for $w \in x$ **do** ▷ find all words in t given x
        if $t \in t_{map}(w)$ **then**
          $w_t \leftarrow w_t \cup w$
        end if
      end for
      $x' \leftarrow x' \cup w_t$
    end for
    return x'
  end function

Semantic Feature Arrangements in the text domain, where single words are grouped by their semantic similarity and their common context, can be compared to Superpixels within the Computer Vision domain. Superpixels conduct a grouping of perceptually similar pixels with the purpose of creating visually meaningful entities while reducing the number of basic units, for instance the number of features within a classification task [34]. Lifting textual explanations to a higher semantic level harnessing topics as explanation units by introducing a word-topic-assignment can be seen analogously to the Pixel-Superpixel-relationship for images.

101

Input document: „The Federal Home Loan Bank Board adjusted the rates on its short term discount notes as follows: (maturity new rate) (old rate) (maturity days) (7 per cent 5 per cent 3 days).“

Original LIME Text Explainer	**TopicLIME Text Explainer**
Dataset: Reuters R52	Dataset: Reuters R52
Document id: 645	Document id: 645
Predicted class = ['interest']	Predicted class = ['interest']
True class: interest	True class: interest
Explanation for class interest	**Explanation for class interest**
('rate', 0.157)	("topic #7 („Financial rates“) = ['discount', 'rates', 'rate']", 0.288)
('rates', 0.113)	("topic #18 („FED, Assets & Deposits“)= ['federal', 'bank']", 0.035)
('discount', 0.035)	("topic #4 („Foreign exchange“) = ['short', 'term']", 0.030)
('bank', 0.026)	("topic #12 („Loan and tax“) = ['loan']", 0.004)
('term', 0.014)	
('federal', 0.004)	
('short', 0.003)	
('follows', 0.002)	

Fig. 3. Textual comparison of original LIME text explainer (left) and topicLIME text explainer (right).

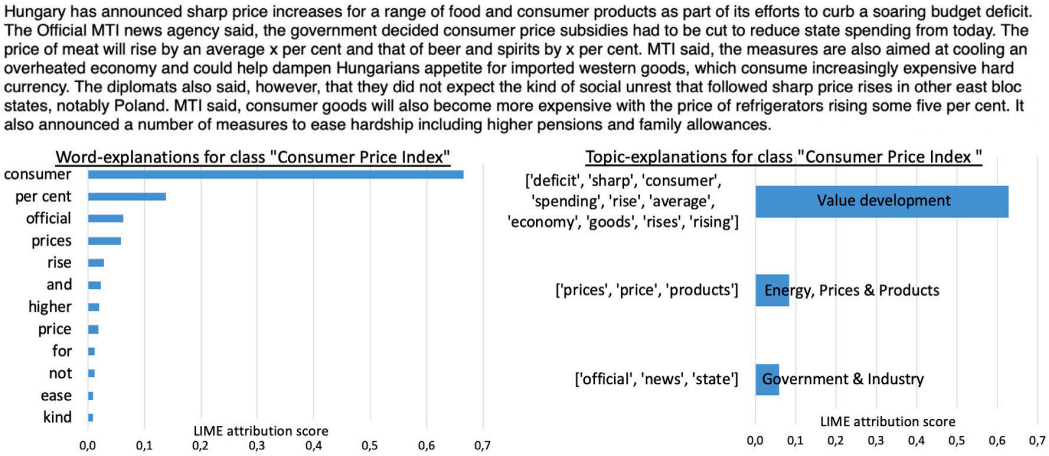


Fig. 4. Visual comparison of original LIME text explainer (left) and topicLIME text explainer (right).

Having arranged all words in a semantic way, a revised topic-based version of LIME is applied. Fig. 2 compares word- and topic-based perturbations for neighborhood generation. *topicLIME* generates a local neighborhood of the document to be explained by removing coherent words. In this case, a single topic or more topics are chosen, all words assigned to these topics are removed at once and similarity of the generated documents to the original document is determined. While LIME measures similarity between documents using similarity kernels that are based on distance measures like *cosine distance*, topicLIME can additionally measure similarity between documents using Kullback–Leibler divergence regarding the topic distributions offered by LDA, which enables a notion of “semantic proximity” of two documents. Hence, context as well as coherent semantic facts that should reveal a similar meaning are analyzed as basic explanation units regarding their effect on the classification when being removed. In the next step, in analogy to Ribeiro et al. [5], a K-Lasso-Algorithm is applied, which selects K topic-represented features with Lasso and learns the weights via least squares afterwards. Finally, the most relevant topics with its assigned words are presented as explanations and the ‘UNKNOWN’-topic is included in case not all words have been assigned to topics during algorithm 1 or 2.

Considering all semantically coherent words explicitly at the same time during a single topic-sampling-iteration, distributional as well as semantic knowledge is included in the topic-explanations in an intuitively interpretable way. Furthermore, the “the whole is greater than the sum of its parts” phenomenon, in ML context potentially resulting from feature interaction effects [12], can be expressed by topic-encoded explanations. In a case where features interact with each other in a prediction model, the prediction cannot be expressed as the sum of individual feature effects as individual feature effects depend on the other feature values [12].

Figs. 3 to 6 illustrate textual and visual comparisons between explanations that are generated with the original LIME text explainer and such that are produced by topicLIME. Both algorithms have been applied to specific documents about “Financial Interest”, “Consumer Price Index”, “Earning” and “Acquisition” of the *Reuters R52*-dataset. While LIME just outputs single words locally most relevant to the specific classification, topicLIME groups contextually coherent words together, outputs according explanations as topic-attributions (including the coherent words) and optionally labels these topics, either via the help of humans by *eyeballing* or by applying automated topic labeling approaches as described in Section 3.4. Thus, contextual effects can be captured and measured that put coherent semantic facts in relation to the classification result. Consequently, explanations are presented at a higher level of information granularity and are made of coherent facts, what, according to insights from psychology (refer to Section 3.1), aligns well with the human conceptual way of thinking. As an example, Fig. 3 shows how topicLIME enhances LIME by performing a contextual

Input document: „Royal Gold and Silver Corp stated to six months: Net share profit two cents vs loss two cents (net profit vs loss), revenues 8 mln vs nil note period that includes 8 dlr extraordinary gain.“

Original LIME Text Explainer	**TopicLIME Text Explainer**
Dataset: Reuters R52	Dataset: Reuters R52
Document id: 1842	Document id: 1842
Predicted class = [‚earn']	Predicted class = [‚earn']
True class: earn	True class: earn
Explanation for class earn	**Explanation for class earn**
(‚cents', 0.198)	("topic #7 („Financial rates“) = ['six', 'net', 'nil', 'note', 'extraordinary', 'gain']",
('net', 0.120)	0.031)
('profit', 0.028)	("topic #18 („FED, Assets & Deposits“) = ['gold', 'silver']", -0.023)
('loss', 0.025)	("topic #2 („Key figures“) = ['profit', 'loss']", 0.022)
('gold', -0.024)	
('mln', 0.009)	
('corp', -0.008)	
('dlr', -0.006)	
('gain', 0.004)	

Fig. 5. Textual comparison of original LIME text explainer (left) and topicLIME text explainer (right) where explanations include both positive as well as negative attributions to the predicted class.

Input document: „ CRA sold Forrest Gold for 8 mln dlrs. Whim Creek Consolidated NL said, the consortium it is leading will pay 50 mln dlrs for the acquisition of CRA LTDs Forrest Gold. PTY LTD Unit reported yesterday CRA and Whim Creek did not disclose the price yesterday. Whim Creek will hold 10 pct of the consortium while Austwhim resources NL will hold 5 pct and Croesus Mining NL 5 pct it said in a statement. As reported Forrest Gold owns two mines in western australia producing a combined 50 ounces of gold a year. It also owns an undeveloped gold project.“

Original LIME Text Explainer	**TopicLIME Text Explainer**
Dataset: Reuters R52	Dataset: Reuters R52
Document id: 9	Document id: 9
Predicted class = [‚gold']	Predicted class = [‚gold']
True class: acquisition	True class: acquisition
Explanation for class gold	**Explanation for class gold**
('gold', 0.384)	("topic #18 („FED, Assets & Deposits“) = ['gold', 'consortium', 'unit', 'disclose'
('acquisition', -0.082)	'mining', 'mines', 'project']", 0.383)
('year', 0.014)	("topic #12 („Loan and tax“) = ['sold', 'acquisition']", -0.086)
('unit', -0.006)	
('reuter', 0.005)	
('leading', -0.004)	
('will', 0.003)	
('whim', 0.002)	
('western', -0.001)	

Fig. 6. Textual comparison of original LIME text explainer (left) and topicLIME text explainer (right) w.r.t. a wrong classification.

assignment of words, that also LIME offers as explanation units, to semantic concepts represented as topics. Fig. 4 visualizes how topicLIME in addition relates an increase of prices in the context of products to Consumer Price Index and inflation via the first two topic explanation units and how it separates these concepts from government and industry matters (as shown by the last topic attribution). In Fig. 5, topicLIME contrasts the concept ‘‘Assets’’ (represented by the words ‘‘gold’’ and ‘‘silver’’ that have a negative attribution to the predicted class) against financial concepts like ‘‘Financial Rates’’ and ‘‘Financial key figures’’ that both have positive attributions. Fig. 6 compares explanations for a wrong classification, where the class ‘‘Gold’’ is predicted although the true class is ‘‘Acquisition’’. Here the classifier to be explained obviously accumulates more confidence for words like ‘‘gold, mining, project’’ (representative for the concept ‘‘Assets’’) towards the class ‘‘Gold’’ than for the words ‘‘sold’’ and ‘‘acquisition’’ (that represent the concept ‘‘Loan and Taxes’’) towards the class ‘‘Acquisition’’.

3.5. Semantic interrogations

Another interesting property of a combination of an explanation system with a semantic approach like LDA is its *semantic interrogation ability* (refer to Fig. 1). When it comes to the point, whether to trust an ML classifier, potential questions to be answered could be: ‘‘Does a classifier behave in a manner that is expected by humans?’’, ‘‘How much does a classifier resemble human intuition?’’ or ‘‘Which hidden features has a classifier learned?’’. As described in Section 3.1, humans think and explain via semantic coherent concepts and therefore have an implicit mental model of a specific domain. In the following, another version of CaSE called *Semantic Interrogations* is proposed that helps with answering the question, how much a classifier works in semantic accordance with human thinking regarding the learned aspects in the classifier’s inductive chain. Similar to work from the authors Dong et al. [35], who analyze correspondence between neurons of a neural network and LDA-topics within a video captioning task, *Semantic Interrogations* is capable of enabling humans to identify the presence or absence of higher-level semantic features. By integrating a semantic approach like LDA as well as a model-agnostic explanation technique like LIME, CaSE is even able to provide this functionality beyond the scope of neural networks and thus in a model-agnostic way.

Algorithm 4 Semantic Interrogations

Require: Topic Model lda, Classifier f
Require: Document-Topic-Mix doc_top_mix ▷ user specified; topic probabilities must sum to one, else normalization step
Require: Topic-Word-Mix top_word_mix ▷ Distribution over words per topic; retrieved from LDA model
Require: Number of documents N to be generated
Require: Number of words W_n to be sampled per topic
  $Documents \leftarrow \{\}$
  for $i \in$ {1,2,...,N} **do**
    $Topics \leftarrow \{\}$
    for $j \in \{1,2,...,W_n\}$ **do**
      $t \leftarrow$ choose_topic_from_multinomial(doc_top_mix)
      $Topics \leftarrow Topics \cup t$
    end for
    $Words \leftarrow \{\}$
    for $topic \in Topics$ **do**
      $w \leftarrow$ choose_word_from_multinomial($top\_word\_mix_{\text{topic}}$)
      $Words \leftarrow Words \cup w$
    end for
    $Documents \leftarrow Documents \cup Words$
  end for
  $Labels \leftarrow \{\}$
  for $document \in Documents$ **do**
    $label \leftarrow f.predict(document)$ ▷ Get predicted label for each document
    $Labels \leftarrow Labels \cup label$
  end for
  return $mode(Labels)$ ▷ Return the most frequent label (mode of Labels) as the average prediction for the user-specified documents

As shown in Algorithm 4, *Semantic Interrogations* enables humans to generate documents that reveal a specific semantic content as well as semantic structure. First, a user-specified document-topic-mixture as multinomial distribution is provided by a human. Then, for each document to be generated, W_n topics are drawn from the multinomial distribution. Having retrieved the semantic topics that a single document shall contain, for each topic words are drawn from a topic-word-mix multinomial distribution. As a result, a set of documents is obtained, whose elements are similar in terms of its semantic nature. Similarity between documents can be measured using similarity kernels that are based on distance measures like *cosine distance*. Especially semantic similarity can be determined using Kullback–Leibler divergence regarding the topic and word distributions offered by LDA.

Feeding all generated documents into the classifier and getting the according classifications together with the topic-encoded explanations from topicLIME, humans can explicate, translate to computational representations and compare their inherent domain knowledge with the actual classification results similar to a ShadowBox Task, in which learners compare their knowledge with those of domain experts [36]. Classifiers that work in accordance with an expert's domain knowledge should output labels equivalent to those semantic labels that correlate with the specified semantic input structure (represented as topic distributions) within the expert's mental model. Doing so, a differentiation between classifiers that are able to distinguish between human semantic concepts (represented as topic distributions as a proxy), such that find new semantic concepts not anticipated by humans, and such that are likely to overfit the data due to a missing semantic sensitivity can be elaborated. This can be useful for model verification. In case of semantic discrepancies between human and classifier identified through such interrogations, epistemic curiosity might be stimulated and further knowledge might be gained [36]. As a side note and left for future research, the proposed algorithm to create documents that reveal a specific semantic structure might help to close the loop in the context of human-in-the-loop-learning. In case of uncertainty in a specific area of the classifier's decision boundary, one could create user-specified documents similar to those that lie in this area, interrogate the classifier with those documents, retrieve the classification results and analyze the according explanations. For all documents classified incorrectly, humans could then assign the according correct labels and integrate those newly generated samples into the training set and retrain the classifier. Doing so, classifiers could be locally forced to better consider predefined semantic aspects that are in line with existing knowledge.

4. Experimental setting and results

In the following an experimental validation of the capabilities of CaSE is performed. On the one hand, a quantitative consistency measure is proposed that is used to analyze the inner structure of CaSE and to examine the implications it has on the topic-explanations. On the other hand, standard word-based LIME, word-based LIME using *Semantic Feature Arrangements* (LIME-SFA) and topicLIME are compared in terms of Local Fidelity. LIME-SFA differs from standard LIME only in that way that it uses algorithms 1 or 2 for generating the local neighborhood instead of sampling words independently. It then proceeds just like LIME and creates word-based explanations.

For the evaluation, scikit-learn (version 0.20.2) and gensim (version 3.8.3) have been used, code is available at https://github.com/sb1990gtr/CaSE. If not explicitly noted otherwise, a Logistic Regression is used as Base Classifier, whose implementation and hyper-parameters are adopted from scikit-learn. It is applied to the Reuters R52 dataset, which consists of 9100 documents (6532 train documents and 2568 test documents) and 52 classes. The according LDA model (implemented via the python-package *gensim*) that CaSE used comprised $k = 19$ topics, which is on the one hand the global optimum for C_v coherence and on the other hand also the outcome of applying Hierarchical Dirichlet Process upstream of LDA for inferring an adequate "number of topics k" automatically. This choice has been made to obtain topic-encoded explanations, whose elements (topics consisting of single words) form a maximal semantically interpretable set with regard to the global topic structure. Another variant could be to use a higher number of topics, for example $k = 80$ topics (which constitutes a local optimum with regard to C_v coherence), to further account for coherent words in the explanations on a finer and more local level of detail. This choice in general depends on the domain characteristics of the according classification task as well as on the requirements to the explanations.

4.1. Consistency property

For providing a common foundation on which word-based and topic-based explanations can be compared taking into account the different mechanisms of perturbation as mentioned in Fig. 2, a *consistency property* is proposed (see Eq. (7)). It is useful for analyzing the effects that different types of perturbation (i.e. word-based and topic-based) can have on the attributions of the according explanation features depending on the base classifier at hand.

$$p(y = 1) = \frac{1}{1 + e^{-(a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b)}}. \tag{6}$$

One difference can result from using classifiers with non-linear activations. When a Logistic Regression, where the predictors appear in the exponent and therefore are not combined linearly (additively) but non-linearly (multiplicatively, see Eq. (6) for its formalization, where a_1 to a_n represent the weights of the according features x_1 to x_n, b represents the bias, and y the target categorical dependent variable) is used as base classifier, the attributions of predictors to a classifier's confidence towards a certain label are not the same when certain predictors are removed sequentially from a model and the according attributions are summed up compared to a simultaneous removal. Another difference can be caused by interaction effects, where features interact with each other with regard to the prediction. As the prediction confidences generated by the base classifier together with the instances

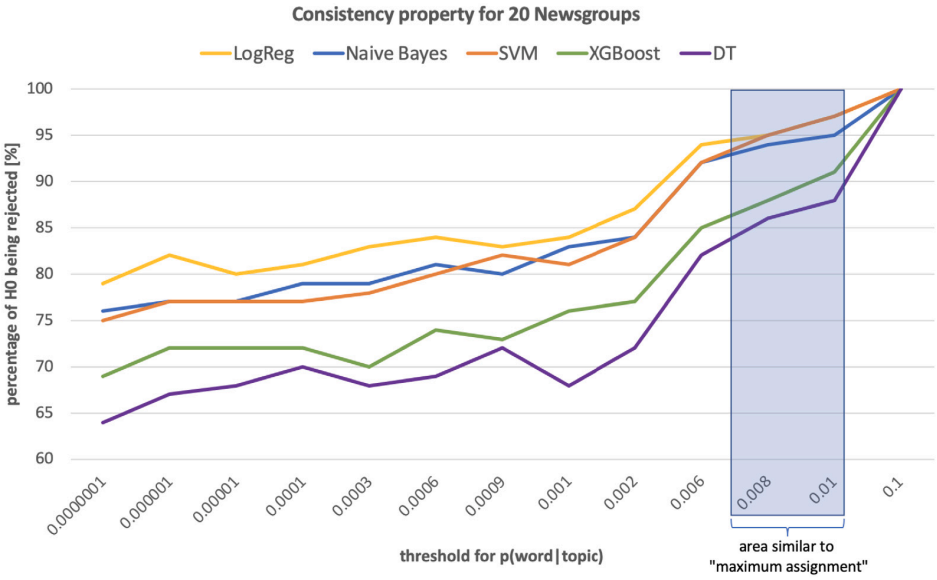


Fig. 7. Analysis of consistency property for Reuters R52 dataset.

perturbed by LIME or topicLIME are used as input for the interpretable LIME or topicLIME models, the learned local interpretable models from both might differ not solely in terms of granularity and cohesiveness of the used explanation features. Also differences in the according feature attributions might result from the characteristics of the base classifier that shall be explained, for example non-linearity or consideration of feature interactions during the learning process.

CaSE is intended to generate explanations that are reliable, meaningful, provide contextual information and reveal a coherent structure by sampling from a realistic local and semantically meaningful perturbation distribution considering feature dependence. Thus, it might be insightful to evaluate topicLIME by comparing to LIME regarding the impact that the different sampling strategies have on the way the interpretable models reflect the original classifier's decision characteristics. Of special interest is the question, how topic-encoded explanations differ from word-based explanations depending on the base classifier. Do they only differ in the arrangement of explanation features and the degree of included higher-level context? Under which circumstances do topic-attributions and word-attributions reveal a consistent order within the explanation chain despite stemming from different sampling strategies? For answering these questions, the *consistency property* from Eq. (7) analyzes the attributions of individual explanation features (the "LIME-scores" denoted by L) and especially the order of the explanation chain and compares the results for both for different types of base classifiers.

$$\Big(L_{\text{topic}}(T_{\text{one}}) \geq L_{\text{topic}}(T_{\text{two}})\Big) \Rightarrow \Big(\sum_{w \in T_{\text{one}}} L_{\text{word}}(w) \geq \sum_{w \in T_{\text{two}}} L_{\text{word}}(w)\Big). \quad (7)$$

Eq. (7) constitutes the hypothesis that there exists a monotone, that is order-preserving or isotone, relationship between topic-based explanation features and the according aggregated word-based explanation features regarding their LIME score. This score approximates the importance of features with regard to the classification confidence for a specific class label and can thus be referred to as attributed impact of the according feature. In the further course of this paper, let $L_{\text{word}}(x)$ be the original LIME attribution score per word. Furthermore let $L_{\text{topics}}(x)$ be the individual score of a single topic-attribution generated by topicLIME. T represents a single topic, w a word respectively.

Topic-explanations in this evaluation are based on CaSE's "assignment by threshold" (refer to algorithm 1), where words from documents of the Reuters R52 dataset are assigned to 19 LDA topics for subsequent perturbation. In the next step, LIME and topicLIME explanations are compared as described in Eq. (7) according to their LIME attribution score. Summing up the results of the individual comparisons for a predefined number of samples, hypothesis tests are performed testing whether Eq. (7) holds true or not. Therefore, the H_0 hypothesis is stated as H_a from Eq. (7) not holding true. Finally, the number of times the H_0 hypothesis can be rejected is computed.

Fig. 7 summarizes the results as the number of times H_0 has been rejected in percent, depending on the choice of the threshold p_{min} (see algorithm 1) which determines the probability $p(word|topic)$ (as estimated by LDA) of a word being assigned to one or more topics during algorithm 1. For five different base classifiers each, 100 documents from the Reuters Test-Dataset have been evaluated and for each document, 50 comparisons have been performed between each two topic attributions and its associated aggregate word attributions. All classifiers have been fine-tuned w.r.t. parameterization using 10-fold-cross-validation and F1-score.

Fig. 7 indicates a trend towards a monotone relationship between topic-based explanations and the according aggregated word-based explanations. Especially when the probability $p(word|topic)$ of a word being assigned to a topic, which is negatively correlated with the threshold p_{min}, is low, the order of the explanation chain is preserved with high chance. With increasing $p(word|topic)$, the order of the explanation chain provided by LIME and topicLIME diverges. This can be explained by the fact, that, with decreasing threshold p_{min}, words will be assigned to more than one topic, which leads to more heterogeneous and huger topics used for explanations that potentially consider polysemy of words. On the other side, as the threshold reaches the area between 0.008 and 0.01 (in that case, a word is only assigned to few topics or even a single topic which is the same as the outcome from "maximum assignment" (refer to algorithm 2)), for all types of base classifiers, LIME and topicLIME explanations reveal a monotone relationship according to Eq. (7) most of the time. Increasing the threshold further, topics contain only single words or no words at all, which leads to a "perfect" monotonicity. Remarkable is the steadily increasing divergence between tree-based classifiers like CART or XGBoost, which is a boosted ensemble of trees, with regard to monotonicity as p_{min} decreases. For the other types of classifiers, topicLIME explanations differ almost exclusively from the word explanations in their inner coherent structure, which provides contextual information for humans,

Table 1
Comparison of LIME, topicLIME and LIME-SFA w.r.t. local fidelity via local approximation error and R^2 for Reuters R52 dataset.

	Approx. Error			R^2		
	Lime	topicLime	Lime-SFA	Lime	topicLime	Lime-SFA
XGBoost	0.0195	0.0076 (−61%)	0.0053	0.864	0.951	0.952
Log. Regr.	0.0545	0.0409 (−25%)	0.0233	0.697	0.847	0.821
SVM	0.0371	0.0277 (−25%)	0.0158	0.733	0.862	0.856

Table 2
Comparison of LIME, topicLIME and LIME-SFA w.r.t. local fidelity via local approximation error and R^2 for 20 Newsgroups dataset.

	Approx. Error			R^2		
	Lime	topicLime	Lime-SFA	Lime	topicLime	Lime-SFA
XGBoost	0.0071	0.0032 (−55%)	0.0021	0.755	0.798	0.791
Log. Regr.	0.0321	0.0169 (−47%)	0.0111	0.692	0.738	0.713
SVM	0.0240	0.0156 (−35%)	0.0112	0.733	0.765	0.0752

Table 3
Comparison of LIME and topicLIME w.r.t. area over the perturbation curve.

	AOPC		
	Lime	topicLime	Difference in %
Reuters R 52	0.271	0.302	+11%
20 Newsgroups	0.097	0.112	+15%

but not with regard to the order of the explanation chain. In the according topics, words that are polysemous or reveal semantic interrelationships with each other, what specific types of classifiers are able to learn as interaction effects, are included. Decision trees are biased towards interaction even more than ensembles of trees that naturally can include both variables that do interact and variables whose effects do not interact [37]. This behavior constitutes one reason, why topic-explanations for tree-based classifiers diverge more from word-based explanations in general and especially as the size of topics increases.

Summing up, topics can be used as basic explanation features and especially resemble word-representations with regard to order if the average topic size is small. Non-linear activations do not distort topicLIME explanations with regard to order in the explanation chain. If tree-based algorithms for classification are used, semantic interaction effects between coherent features with regard to the prediction potentially are included in the topic-based explanations leading to a different order of the explanation chain stemming from the "the whole is greater than the sum of its parts" phenomenon. Results from this evaluation enable drawing direct conclusions with respect to designing the word-topic-assignment function t_{map} and its hyper-parameter p_{min} depending on the explanation context as well the base classifiers to be explained. Using "maximum assignment", maximally coherent, meaningful topic-explanations are generated by topicLIME that are likely to reveal order-preserving characteristics, which makes comparing LIME and topicLIME explanations regarding human comprehensibility easy.

4.2. Local fidelity and approximation accuracy

Another important question to be answered is whether explanation models that use realistic and meaningful local distributions for perturbation reveal higher Local Fidelity to the base classifier. Local Fidelity is said to be achieved if an explanation model $g \in G$ is found such that $f(z) \approx g(z')$ for $z, z' \in Z$ where Z constitutes the vicinity of x. In the course of this comparison between standard word-based LIME, LIME-SFA and topicLIME, Mean Local Approximation Error (MLAE, see Eq. (8)) and MeanR^2 (see Eq. (9)) are used as a proxy for the accuracy of the local approximation with regard to the original decision boundary and therefore for Local Fidelity. During this evaluation, words from a document are assigned to LDA topics with the help of "maximum assignment" (see algorithm 2) during LIME-SFA and topicLIME.

$$MLAE(f,g) = \frac{\sum_i^N |f(x_i) - g_i(x_i)|}{N}, \tag{8}$$

$$MeanR^2(g) = \frac{\sum_i^N R^2(g_i)}{N}, R^2(f,g) = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (f(z_i) - g(z_i'))^2}{\frac{1}{n}\sum_{i=1}^n (f(z_i) - f_{mean})^2} \tag{9}$$

In both cases, N is the number of samples in the according test dataset.

Tables 1 and 2 summarize the results of the quantitative analysis for the two datasets *Reuters R52* (the according LDA model comprises 19 topics) and *20 Newsgroups* (the according LDA model comprises 20 topics). Per dataset, three different base classifiers are compared with regard to Local Approximation Error and R^2. In each comparison run between LIME, LIME-SFA and topicLIME, for each instance from the according test dataset, topicLIME generates five topics with coherent words as topic-based explanation and the total number of words n_w that are part of this explanation is calculated. Then, an explanation consisting of n_w words is generated via the LIME and LIME-SFA explainers. Finally, R^2 and Local Approximation Error are calculated for each explanation model and the resulting values are normalized by dividing by the number of instances from the test dataset. It turns out, that topic-based explanations and the ones created using *Semantic Feature Arrangements* consistently perform better in terms of Local Fidelity and R^2. On the one hand, coherent sampling enables the explanation models to consider feature dependencies and doing so approximate the base classifier locally using more realistic input data. As a consequence resulting explanations are being less likely misinterpreted as they are generated based on a realistic local perturbation distribution that leads to semantically meaningful inputs for local explanation generation. Furthermore, higher R^2-values reached by topicLIME and LIME-SFA indicate that the topic-encoded features used within the locally approximated interpretable model are more predictive with regard to the dependent label.

The consistency analysis from Section 4.1 indicated a greater divergence w.r.t. monotonicity between explanations from LIME and topicLIME especially for tree-based classifiers. To test, whether this divergence in the order of the explanation chain is likely to be caused by their ability to learn feature interactions, which are captured by semantically interrelated words in the topicLIME explanations, the *Area Over The Perturbation Curve* (AOPC) [38] is analyzed. It measures Local Fidelity of individual explanations and is defined as:

$$AOPC(k) = \frac{1}{N}\sum_{i=1}^{N} \{p(\hat{y}|x_i) - p(\hat{y}|\tilde{x}_i^{(k)})\}, \tag{10}$$

where the top k% explanation features are removed from x_i yielding $\tilde{x}_i^{(k)}$, $\hat{y}$ denotes the predicted label for x_i and N is the number of samples in the according test dataset. Since XGBoost is a high-performing ensemble and thus tree-based classification algorithm that can intrinsically learn feature interactions, AOPC was evaluated for LIME and topicLIME (see Table 3).

As a higher AOPC value indicates that the removed explanation features are considered more important by the base classifier, topicLIME explanations reveal higher Local Fidelity and are also likely to capture semantic feature interactions that lead to the change in the explanation chain. This hypothesis is also supported by the fact, that for both test datasets used in Section 4.2, XGBoost significantly benefits the most with regard to Approximation Error of the explanation model when topic-based sampling is used compared to the other classification algorithms (topicLIME: −61% for Reuters R52 Dataset and −55% for 20 Newsgroups Dataset; LIME-SFA: −73% for Reuters R52 Dataset and −70% for 20 Newsgroups Dataset). LIME-SFA that also uses topic-based sampling is even slightly better than topicLIME because the according linear approximation models are characterized by higher degrees of

freedom as they use a higher number of features (words) in the linear models. Consequently, topic-based sampling in general leads to a lower Local Approximation Error and topicLIME is capable of addressing the "the whole is greater than the sum of its parts" phenomenon especially when being applied to classification algorithms that learned interaction effects.

5. Summary and conclusion

A novel proposal for obtaining coherent and semantically meaningful explanations has been worked out. One main merit of CaSE is that it builds explanations in a way, in which its explanation features are made of coherent words offering the explainee contextual information leading to increased human interpretability. Explanations generated by topicLIME are also adjustable w.r.t. granularity by incorporating prior knowledge into the LDA process of topic extraction and allow for presentation of explanations at an individual semantic level of detail. Due to the resulting Semantic Alignment between ML model and human users, human mental models are likely to be updated more efficiently which makes topic explanations more informative. Additionally, as topicLIME explanations consist of a lower number of explanation units compared to LIME explanations while pertaining the same amount of information, the complexity $\Omega(g)$ of the explanation models and therefore the length of the explanation chain is reduced. Thus, the number of mental calculation steps that humans have to perform in order to understand the explanations is reduced. Worth mentioning is also the fact, that the proposed architecture is not limited to specific classifiers nor explanation systems nor application domains, but can be used for a variety of explanation and interpretation techniques. CaSE can offer a realistic and meaningful local perturbation distribution by avoiding extrapolation, which is a line of research that, according to the authors from Ribeiro et al. "would benefit multiple explanation methods" [11]. As a consequence, explanation models like topicLIME built from such a local neighborhood reveal higher Local Fidelity compared to LIME, which results in reliable and contextual explanations. Furthermore, a new consistency measure is proposed, which enables comparisons of word- and topic-based explanation systems, in this case of LIME and topicLIME, depending on the type of the underlying base classifier and its individual characteristics like non-linearity or consideration of interaction effects.

Another contribution of CaSE enables semantic interactivity. Such an approach can complement the rising repertoire of Interactive Machine Learning that aims at building systems that improve their learning outcome with the help of a human expert who interacts with them [39]. By techniques like *Semantic Interrogations*, a human expert has the opportunity to interact with a classifier and its data and thus can help with improving prediction outcome in order to "enable what neither a human nor a computer could do on their own" [39].

Beneath all the benefits listed so far, there are also mainly two prerequisites of CaSE that should be mentioned. First, as the entire semantic functionality is based on an LDA approach, some expertise in Topic modeling is required, for instance in order to implement a suitable data preprocessing or to find an adequate "number of topics k". The latter can be bypassed be using Hierarchical Dirichlet Process upstream of LDA, as it is capable of inferring a suitable "number of topics k" automatically. Additionally, even if it makes the use of CaSE even more flexible, specific choices of the function t_{map} that performs Semantic Feature Arrangements have to be made and implemented depending on the input domain.

This work leaves some perspectives for further studies. First, a user experiment can be conducted that evaluates explanations generated by topicLIME with regard to efficiency and human interpretability and compares topicLIME to LIME for different text documents of different length. Second, a supervised version of LDA named *sLDA* [40] or another version called *GuidedLDA* [41] could be integrated into the CaSE architecture in order to either allow for Topic Models capable of inferring latent topics predictive of a certain response type or the injection of *seed words* for human-guided topic identification. The former might allow for further inclusion of semantic feature interactions with regard to the response type into the explanations. Third, further and improved sampling strategies can be developed that might lead to even better Local Fidelity. As an example, the sampling mechanism used within the algorithm *Semantic Interrogations* can be adapted and used for generation of a meaningful and realistic local neighborhood. Fourth, it could be interesting to train a bimodal LDA model that estimates the joint probability of words and parts of images and to use the obtained insights to generate bimodal explanations for a classification task. Last but not least, a semantic component like LDA is not only able to help with generating meaningful explanations, but can also be harnessed as evaluation basis. In such a scenario, the output of different explanation systems can be compared regarding its cohesiveness and its semantic expressiveness and based on that, new measures for analyzing semantic explanation quality can be developed. Also, new evaluation measures for interactively analyzing semantic accordance between humans and classification algorithms (called *human intuition resembling* in the course of this paper) might be enabled by CaSE that potentially lead to better evaluation of trust that humans develop into ML systems.

CRediT authorship contribution statement

Sebastian Kiefer: Conceptualization, Design, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I would like to say many thanks to Jonas Amling, who helped with developing the code for CaSE. In addition, I am grateful to Ute Schmid and Alisa Münsterberg, who contributed with fruitful discussions. Special thanks goes to DATEV eG, who provided computing resources and sample data for experiments.

References

[1] A. Holzinger, C. Biemann, C. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, 2017, arXiv preprint arXiv:1712.09923.
[2] J. Sliwinski, M. Strobel, Y. Zick, A characterization of monotone influence measures for data classification, in: IJCAI-17 Workshop on Explainable AI (XAI) Proceedings, 2017.
[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), in: IEEE Access, Vol. 6, 2018, pp. 52138–52160.
[4] S. Bruckert, B. Finzel, U. Schmid, The next generation of medical decision support: A roadmap toward transparent expert companions, in: Frontiers in Artificial Intelligence, Vol. 3, 2020.
[5] M.T. Ribeiro, S. Singh, C. Guestrin, Why Should I Trust You?, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[6] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, in: Artificial Intelligence, Vol. 267, 2019, pp. 1–38.
[7] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 399–408.
[8] S. Syed, M. Spruit, Full-text or abstract? examining topic coherence scores using latent dirichlet allocation, in: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017, pp. 165–174.
[9] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing, 2017, pp. 4768–4777.
[10] J. Rabold, G. Schwalbe, U. Schmid, Expressive explanations of dnns by combining concept analysis with ilp, in: KI 2020: Advances in Artificial Intelligence ; 43rd German Conference on AI, 2020.

[11] M.T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: AAAI Conference on Artificial Intelligence, 2018, pp. 1527–1535.
[12] C. Molnar, Interpretable machine learning: a guide for making black box models explainable, 2019.
[13] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, 2018, arXiv preprint arXiv:1805.10820.
[14] D. Alvarez-Melis, T.S. Jaakkola, On the robustness of interpretability methods, 2018, arXiv preprint arXiv:1806.08049.
[15] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Earth, Vol. 1, 2004.
[16] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, in: Information Fusion, Vol. 71, 2021.
[17] F.C. Keil, Explanation and understanding, in: Annual Review of Psychology, Vol. 57, 2006, pp. 227–254.
[18] D. Dennett, The Intentional Stance, MIT Press, Cambridge MA, 1987.
[19] R.A. Wilson, F. Keil, The shadows and shallows of explanation, in: Minds and Machines, Vol. 8, 1998, pp. 137–159.
[20] W.-K. Ahn, N.S. Kim, M.E. Lassaline, M.J. Dennis, Causal status as a determinant of feature centrality, in: Cognitive Psychology, Vol. 41, 2000, pp. 361–416.
[21] S.A. Sloman, B.C. Love, W.-K. Ahn, Feature centrality and conceptual coherence, in: Cognitive Science, Vol. 22, 1998, pp. 189–228.
[22] J.H. Danovitch, C.M. Mills, Understanding when and how explanation promotes exploration, in: Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms, 2018.
[23] P. Thagard, Coherence in Thought and Action, MIT Press, Cambridge MA, 2000.
[24] D. Gentner, C. Toupin, Systematicity and surface similarity in the development of analogy, in: Cognitive Sciences, Vol. 10, 1986, pp. 277–300.
[25] D.M. Blei, A.Y. NG, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[26] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 952–961.
[27] Z.S. Harris, Distributional structure, in: WORD, Vol. 10, 1954, pp. 146–162.
[28] C. Molnar, G. König, B. Bischl, G. Casalicchio, Model-agnostic feature importance and effects with dependent features – a conditional subgroup approach, 2020, arXiv preprint arXiv:2006.04628v1.
[29] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical dirichlet processes, J. Amer. Statist. Assoc. 101 (2006) 1566–1581.
[30] F. Morstatter, H. Liu, In search of coherence and consensus: Measuring the interpretability of statistical topics, J. Mach. Learn. Res. 18 (2018) 1–32.
[31] N. Aletras, M. Stevenson, Labelling topics using unsupervised graph-based methods, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 2, 2014, pp. 631–636.
[32] J.H. Lau, K. Grieser, D. Newman, T. Baldwin, Automatic labelling of topic models, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 1536–1545.
[33] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, X. Li, Automatic labeling hierarchical topics, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 2383–2386.
[34] D. Stutz, A. Hermans, B. Leibr, Superpixels: An evaluation of the state-of-the-art, in: Computer Vision and Image Understanding, Vol. 166, 2017, pp. 1–27.
[35] Y. Dong, H. Su, J. Zhu, B. Zhang, Improving interpretability of deep neural networks with semantic information, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4306–4314.
[36] R. Hoffman, S. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, 2019, arXiv preprint arXiv:1812.04608.
[37] K. Goyal, S. Dumancic, H. Blockeel, Feature interactions in xgboost, 2020, arXiv preprint arXiv:2007.05758v1.
[38] D. Nguyen, Comparing automatic and human evaluation of local explanations for text classification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2018, pp. 1069–1078.
[39] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? in: Brain Informatics, Vol. 3, (ISSN: 2198-4026) 2016, pp. 119–131.
[40] D.M. Blei, J.D. McAuliffe, Supervised topic models, in: Advances in Neural Information Processing Systems, 21, 2007, pp. 121–128.
[41] J. Jagarlamudi, H.D. III, R. Udupa, Incorporating lexical priors into topic models, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 204–213.

A.2.2 Kiefer et al. "Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations." In: KI 2021: Advances in Artificial Intelligence

Full Reference of Paper

Sebastian Kiefer and Günter Pesch. "Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations." In *Edelkamp, S., Möller, R., Rueckert, E. (eds) KI 2021: Advances in Artificial Intelligence. KI 2021. Lecture Notes in Computer Science()*, vol 12873. Springer, Cham. DOI: 10.1007/978-3-030-87626-5_22.

My Scientific Contributions

Related research question(s): **RQ3**.
Type of contribution: **technical, applied**.

During the research underlying this dissertation, I made the following scientific contributions published in this paper:

- Identification of the research gap of missing solutions to the problem of selective and receiver-dependent model-agnostic explainability of unsupervised ML approaches.
- Collection and structuring of literature on current types of anomaly detection technologies and especially on according intrinsic and global explanation strategies for such algorithms.
- Identification of major drawbacks related to the field of unsupervised learning, like missing applicability of model-agnostic local additive feature attribution methods.
- Proposition of an integrated architecture that allows model-agnostic selective explainability for different types of unsupervised ML tasks. It works by integrating an intermediate supervised classification task as global approximation of the unsupervised model at hand.
- Implementation of an ensemble-based architecture for unsupervised anomaly detection for financial auditing with receiver-dependent LIME-explanations.
- Outlining future research, like investigating in model-agnostic explainers that are directly applicable to unsupervised ML and thus avoid extra complexity and additional inductive bias.

My Written Contents

I wrote all of the contents except parts of section 3 (details on an annual audit and details on the used datasets). My written contribution to the paper is 85%.

Unsupervised Anomaly Detection for Financial Auditing with Model-Agnostic Explanations

Sebastian Kiefer[1,2(✉)] and Günter Pesch[1]

[1] DATEV eG, Paumgartnerstr. 6-14, 90429 Nürnberg, Germany
{sebastian.kiefer,guenter.pesch}@datev.de
[2] Cognitive Systems, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany
sebastian.kiefer@uni-bamberg.de

Abstract. Explainable Artificial Intelligence (AI) has emerged to be a key component for Black-Box Machine Learning (ML) approaches in domains with a high demand for transparency. Besides medical expert systems, which inherently need to be interpretable, transparent, and comprehensible as they deal with life-changing decision tasks, other application domains like financial auditing require trust in ML as well. The European General Data Protection Regulation (GDPR) also applies to such highly regulated areas where an auditor evaluates financial transactions and statements of a business. In this paper we propose an ML architecture that shall help financial auditors by transparently detecting anomalous datapoints in the absence of ground truth. While most of the time Anomaly Detection (AD) is performed in a supervised manner, where model-agnostic explainers can be easily applied, unsupervised AD is hardly comprehensible especially across different algorithms. In this work we investigate how to dissolve this: We describe an integrated architecture for unsupervised AD that identifies outliers at different levels of granularity using an ensemble of independent algorithms. Furthermore, we show how model-agnostic explanations can be generated for such an ensemble using supervised approximation and Local Interpretable Model-Agnostic Explanations (LIME). Additionally, we propose techniques for explanation-post-processing that allow explanations to be selective, receiver-dependent, and easily understandable. In a nutshell, our architecture paves the way for model-agnostic explainability for the task of unsupervised AD. It can further be transferred smoothly to other unsupervised ML problems like clustering problems.

Keywords: Anomaly Detection · Outlier Detection · Unsupervised Learning · Explainable Artificial Intelligence · Human-like explanations

Supported by organization DATEV eG.

S. Edelkamp et al. (Eds.): KI 2021, LNAI 12873, pp. 291–308, 2021.
https://doi.org/10.1007/978-3-030-87626-5_22

1 Introduction

Besides ML research, especially supervised ML is broadly applied to solve data-driven problems in business environments. Many of those "applied" ML domains, whether they are concerned with strongly regulated scenarios like in financial auditing, with decision-critical situations like in self-driving cars, or even with life-changing situations like in tasks of medical diagnosis, require more than purely a high prediction accuracy. Instead, concepts such as transparency and comprehensibility are needed so that humans affected by ML decisions can develop trust into the systems. Especially Black-Box ML classifiers lack the ability to provide an explicit declarative knowledge representation and hide the underlying explanatory structure [15]. To address these shortcomings, two scientific fields have emerged, namely Interpretable Machine Learning (IML) and Explainable Artificial Intelligence (XAI). While IML focuses on creating global and model-intrinsic interpretability by providing intrinsic, ex ante- understanding of the whole ML model's logic, XAI strives for enabling local and model-agnostic interpretability to achieve an ex post- understanding of the models' specific behavior [1]. Combining both results in Comprehensible Artificial Intelligence (cAI), which strives for generating results that are transparent and comprehensible [7]. Those two properties form the basis of trust in AI. Comparatively little attention, in general and in particular from interpretability point of view, has been paid to unsupervised learning approaches. Nevertheless, many data-driven problems in business environments like the one described in the following are hard to tackle with supervised ML, mostly because of missing label annotations that represent the ground truth. The research results, that are presented in the further course of this paper, have been achieved during work for the company DATEV eG. DATEV eG is a software house and IT service provider for tax consultants, auditors and lawyers as well as for their clients. Since the company provides software solutions for automated processes in financial accounting, among other things, quality assurance is of great importance. One possibility to continuously monitor and improve such solutions is the use of AD mechanisms. Therefore, we develop an integrated and explainable AD architecture and demonstrate its functionality prototypical for the domain of financial auditing. As a wide range of customers and companies from different sectors is involved, which all reveal tax and financial singularities, an exhaustive labeling of instances as normal or anomalous is not feasible. Therefore, our approach strives for several objectives. On the one hand, the proposed architecture shall perform efficient AD using unsupervised ML that suggests specific presumably anomalous instances for further inspection. As financial accounting and auditing represent highly regulated processes, the system on the other hand shall provide intuitive and human-like explanations for the ML system's decisions allowing humans to comprehend the system and understand its individual decisions. In such a critical area the ability to explain an AD algorithm might almost be as important as the model's prediction accuracy [3]. By the absence of ground truth and by the usage of different AD algorithms, purposeful feature engineering is complicated (what are good features w.r.t. detection performance and are there differences in

feature quality across different algorithms?). Thus, the system shall also be capable of generating more detailed explanations and visualizations for Data Scientists to create predictive features by introspection of the explanations (the feature attributions). Also a restriction to specific AD algorithms like Deep Learning (DL) algorithms (such as Autoencoders) is not desired when lacking class labels (what are outliers and how many outliers are there?). Instead, an ensemble-based architecture shall be harnessed that combines different algorithms and performs the varying AD tasks from different perspectives and in a robust manner while keeping the false positive rate of the system low. As a consequence, the explanation module for such an ensemble cannot be based on global and intrinsic explanations, but rather must comprise a model-agnostic way of generating helpful explanations that describe the AD ensemble's behavior locally. Such model-agnostic explanation generators like LIME [31] or SHapley Additive exPlanations (SHAP) [21] typically are designed for supervised ML approaches, which have access to ground-truth-information. Therefore, we propose to integrate a supervised ML model into our AD architecture that approximates the unsupervised AD ensemble globally. It can then be used as a basis for LIME to generate local and model-agnostic explanations at different levels of granularity.

2 Related Work

Anomaly Analysis aims at identifying datapoints or regions from the data whose characteristics differ from expected values [9]. Those subsets that differ significantly from the remainder of the data are called anomalies and in the following also referred to as outliers interchangeably [23]. As AD is applied to a wide variety of application domains, which all might differ in terms of the nature of the data (be it continuous or discrete data), the type of occurring anomalies (Point Anomalies, Collective Anomalies, or Contextual Anomalies), the availability of data (labeled or unlabeled data), and the evaluation criteria (like Overall Accuracy, Precision, Recall), different AD technologies have been proposed [8,11]. These range from kernel-based over distance-based, clustering-based, density-based, to ensemble-based algorithms [26]. Since a single AD technique usually cannot discover all anomalies in a low dimensional subspace due to data complexity, ensemble-based algorithms have its justification in many application domains [35]. Often, AD approaches are also divided into ML approaches (where DL represents a special case) and statistical approaches, which use stochastic models and the according assumptions of certain data distributions [8]. Approaches from both types can be combined or integrated with visualization techniques (like scatter plot matrix or parallel coordinate plots) and with dimension reduction techniques (like Principal Component Analysis (PCA), Multidimensional Scaling (MDS), or t-stochastic neighbor embedding (t-SNE)) [34].

Recently, the question how AD algorithms can be explained arouses certain interest. On the one hand, there are special explainers for Isolation Forests like an adaptation of the SHAP Tree Explainer [21] and extensions of Kernel SHAP for explaining Autoencoders for AD [3]. On the other hand, there is

work using centroids as representation for a cluster of points for cluster-based AD techniques [30], which works well if clusters are compact or isotropic, but malfunctions otherwise. In addition, harnessing PCA or t-SNE enables visualization of identified clusters in two-dimensional space [16,22]. Furthermore, one can use a decision tree per cluster after the clustering process for explaining certain clusters or use approaches like Interpretable Clustering via Optimal Trees (ICOT), where the individual clusters represent the leaves of a decision tree [5]. Another possibility is to apply Deep Taylor Decomposition of a One-Class Support Vector Machine (OCSVM) and to explain the outcomes using support vectors or input features [17]. Summing up, a lot of research either focussed on intrinsic and global explanations for AD algorithms or adapted local and model-agnostic explainers like SHAP to specific AD algorithms like Isolation Forests or Autoencoders. By doing so, the full power of real model-agnostic explainers is often lost. Many times, research on explainability even entirely concentrated on supervised AD approaches. As a consequence, we propose to conduct further research especially on real model-agnostic explanation strategies for unsupervised AD models. Although developed independently from each other, we found work from Morichetta et al. [25] that partially aligns well with ours. In their work the authors propose an architecture for generating model-agnostic explanations for the unsupervised problem of Network Traffic Analysis. Our main idea is to build a supervised ML model that approximates the unsupervised AD ensemble globally and acts as a basis for subsequent model-agnostic local explanation systems like LIME or SHAP. LIME is a method that explains an individual model's prediction by locally approximating the model's decision boundary in the neighborhood of the given instance [31]. It uses a local linear explanation model and can thus be characterized as an additive feature attribution method [21]. For even more expressive explanations it can be combined with Inductive Logic Programming (ILP) in order to generate first-order rules as explanations [29].

3 Explainable Anomaly Detection in the Context of Auditing

An annual audit covers all transactions of a client with all business partners in one year. Due to the huge amount of transactions, the auditor is caught in a dilemma: On the one hand, he is required to conduct the audit very thoroughly. On the other hand, the contract's monetary structure doesn't allow him to run a full examination. Therefore, he restricts the examination on a sample of transactions carefully selected based on his experience. This procedure, however, still creates many blind spots with potential irregularities staying unrevealed. The aim of this paper is to dissolve this dilemma. It provides an integrated approach to detect anomalies in the audit process among the whole amount of transactions. Additionally, it restricts the quantity of anomalies to a feasible size and gives receiver-dependent model-agnostic explanations for the reasons why datapoints were identified as anomalies. Figure 1 gives an overview of the integrated

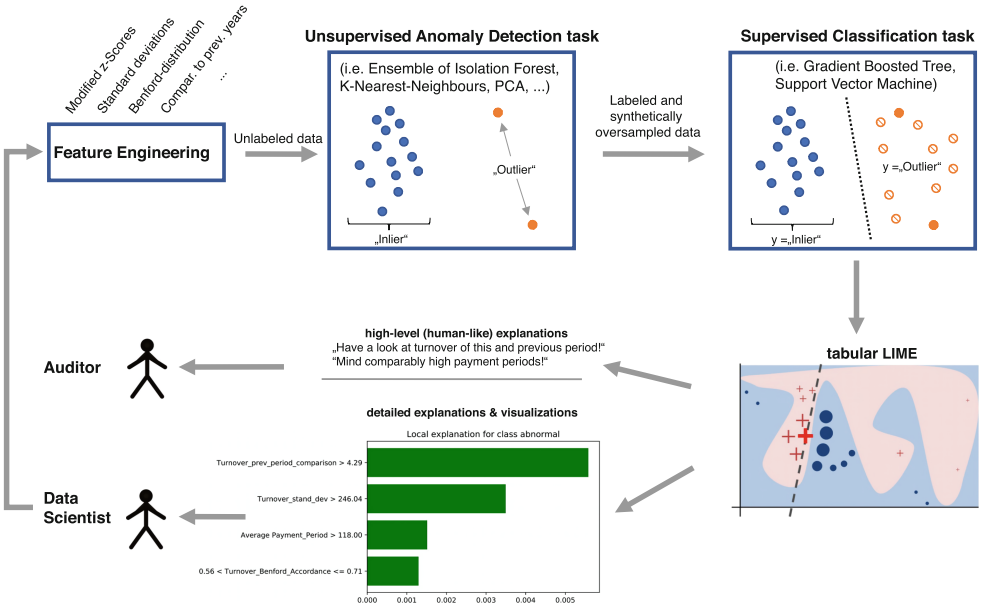


Fig. 1. Architecture of the Anomaly Detection and Explanation process

architecture of our proposed approach. Each step is described in detail in the following subsections.

3.1 Data

This work is based on two different DATEV datasets. The first one is generally used to train aspiring tax advisors during their education at university. Although it consists of artificial data, it provides a good impression exemplary for one client of how the data and the technical issues are made of. The second dataset consists of real client data covering the full bandwidth from small to big clients with many transactions. Table 1 depicts both datasets' characteristics.

Table 1. Characteristics of the used datasets

	University	Real-World
Clients	1	220
Accounts	258	14687
Transactions	6346	998085
Maximum (Transact. per Account)	717	96353
Median (Transact. per Account)	19	22

In general, the used datasets comprise data from one or more clients. A client's bookkeeping is structured in different accounts, which have different

intentions. While there are accounts collecting internal transactions only, there are other accounts containing transactions between the client and its business partners. Due to the reason that transactions with business partners are one of the main fields of fraud, our work focuses on accounts with business partners exclusively. In particular, we evaluate accounts that cover sales revenue and incoming goods. Furthermore, we distinguish between ingoing and outgoing accounting entries. As an example, in an incoming goods account an invoice from the business partner represents the ingoing accounting entry, whereas the accounting entry for paying the invoice represents the outgoing one. Both accounting entries build a logical entry couple comprising the whole transaction. Overall, the described datasets are multivariate, unlabeled, and supposedly highly imbalanced with regard to the distribution of normal and anomalous datapoints.

3.2 Feature Engineering

As already mentioned, the smallest unit of consideration in our context is a transaction. It consists of various data fields such as invoice number, date of transaction, discount, amount, currency, tax rate, and a descriptive text. In order to find crucial features for AD, we interviewed several domain experts, who gave us valuable insights. It turned out that it is useful to take two different views when dealing with such data in the context of financial auditing:

Transaction
Examine one transaction among all other transactions with a business partner. In the following, this view is called *TA*.

Account
Examine one account among all other accounts. In the following, this view is called *ACC*. As there are no meaningful singular facts for an account, we gain the features from the set of transactions belonging to the account.

Furthermore, AD should be conducted stationary and thus should leave changes over time aside, with a certain feature that compares turnover values to the previous year's values as the only exception. Since the complete feature description is intellectual property of DATEV eG, we are unable to describe each feature in detail. Nevertheless, among others the following features have been engineered during our work for *TA*:

Turnover - Modified Z-Score: Deviation to Median Absolute Deviation (MAD), which is an alternative and more robust measure of dispersion compared to sample variance or standard deviation [18].

Turnover - Smoothness: Smooth turnover values can be a hint for obscure transactions. Therefore, we developed a measure to express the smoothness of turnover.

Tax Rate: In Germany, the general tax rate is 19%. However, for some goods the rate is reduced to 10.7%, 7%, 5.5% or 0%. In the wake of the corona crisis some rates were changed from 19% to 16% and from 7% to 5%.

Payment Period: The period between invoicing and payment.

For *ACC*, the following features are included:

Turnover Standard Deviation: The Standard Deviation of the turnover per account is calculated.

Turnover Maximum Deviation: The maximum of the modified z-scores per account, which internally use the Median Absolute Deviation of the turnover, is calculated.

Turnover - Degree of Smoothness: The share of smooth turnover values is taken as feature for AD.

Turnover - Degree of Accordance with Benford Distribution: Benford's law states that in many sets of numbers the leading digit *1* tends to occur with probability approximately 30%, although the expected probability is 11.1%. It is a common analytical tool when probing financial fraud [4,14].

Turnover - Accordance with Period of Previous Year: This feature describes a value that compares current turnover to previous year's turnover.

Tax Rate: This feature analyzes homogeneity of used tax rates in the account.

Payment Period: A measure that specifies the average payment period per account.

3.3 Ensemble-Based Architecture

Due to the fact that there is no labeled data (i.e., existing anomalies in historical data), we are challenged with an unsupervised task. Moreover, as we suppose that there is a very small share of anomalies, we deal with a very imbalanced dataset. These observations lead us to the question how many anomalies should be detected. A simple answer would be one datapoint per investigation, i.e., one transaction per account. Unfortunately, this approach has several limitations. First, there is not always an anomaly in the account's data. Second, due to the huge number of accounts, there would be detected far more anomalies as an auditor would be able to check. Third, there could be more than one anomaly within an account. As a consequence, a flexible and customizable solution is needed.

There are lots of different AD algorithms available and each of the algorithms has its own strengths and weaknesses. The idea is to use this variety of approaches by connecting different AD algorithms to an ensemble. Ensemble techniques are a proven technology in ML, its advantage is to create a better prediction by linking multiple weak predictions [28]. To achieve this, we constructed an ensemble of nine AD algorithms (see Fig. 2) with the intention that a datapoint is flagged as anomaly only if most of the algorithms (connected by a meta detector) agree with. The hyper-parameter *min_count* usually is set to a value near to the number of different algorithms in the ensemble (in our case, it is set to 8 or 9). Thus, an anomaly is detected only on few accounts, which helps

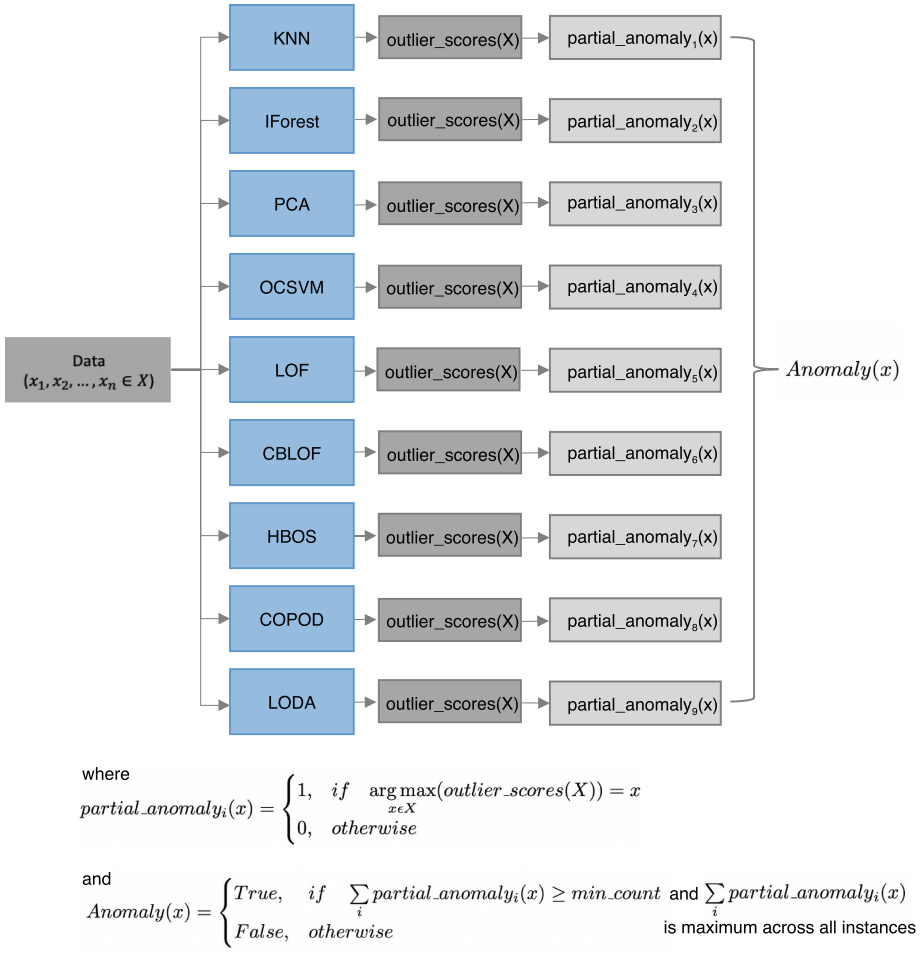


Fig. 2. Architecture of the Anomaly Detection Ensemble

to restrict the number of identified anomalies in total to a human-manageable size. The remaining partial anomalies give the auditor a good hint where he potentially has to further investigate.

From technical point of view, we use the Python library PyOD. PyOD includes more than 30 cutting-edge AD algorithms that have been used in various academic and commercial projects [36]. Once a detector has been fitted on a dataset, the corresponding outlier scores can be accessed. We construct our ensemble approach by simply iterating through nine preselected AD algorithms. The functionality of the anomaly detector is shown in Algorithm 1 as pseudo-code.

The applied nine algorithms are listed in Table 2 in Appendix A. Those algorithms have been carefully selected according to various criteria as they should be heterogeneous, fast, and stable. Particular attention has been taken on the variety of approaches. While some are proximity-based, others base on a linear

Algorithm 1. Ensemble-based Anomaly Detection

Require: m methods ▷ user specified; an array of selected AD algorithms to be applied
Require: data ▷ unlabeled data; one object of consideration, either data describing individual accounting transactions or individual accounts
Require: anomalies ▷ an array of length $len(data)$ initialized with 0
Require: min_count ▷ user specified threshold for anomalies
  for $i \in \{1,2,...,m\}$ **do**
    methods[i].fit(data)
    partial_anomaly $= \arg\max_{x \epsilon data} \; methods[i].outlier\_scores(data)$
    anomalies [partial_anomaly] $+= 1$
  end for
  anomaly $\leftarrow \{\}$
  if $\max(anomalies) \geq min\_count$ **then**
    $anomaly = \arg\max \; anomalies$
  end if
  return anomaly

or probabilistic model. Other promising algorithms haven't been considered as they were either too slow (i.e., Autoencoders) or required specific characteristics of the data like normally distributed data.

3.4 Model-Agnostic and Receiver-Dependent Explanations

AD for financial auditing requires the ability to explain identified outliers. In the following we present a new process of generating receiver-dependent and truly model-agnostic explanations for unsupervised learning. It comprises four steps: Oversampling to account for class imbalances, supervised approximation of the AD ensemble, tabular LIME for explaining the supervised model, and explanation-post-processing to tailor the explanations to the needs of the individual receivers. Each step described in the following is applied to each object of consideration (for views TA and ACC).

Synthetic Oversampling. Dependent on the choice of the threshold min_count for the AD ensemble (refer to Subsect. 3.3), a certain number of outliers is detected. As usual, the class comprising the anomalies clearly represents the minority class. On average, the AD ensemble identified between 0.007% and 0.17% of all datapoints as anomalous for the view ACC and between 0.007% and 0.15% for the view TA. For a subsequent approximation via a supervised ML model, which takes the features used by the AD ensemble as well as the decisions of the ensemble (normal or abnormal class) as input, the classes should be balanced at least to some degree. Therefore, we perform a synthetic oversampling harnessing Synthetic Minority Over-sampling Technique (SMOTE) [10]. SMOTE represents an over-sampling approach in which the minority class is oversampled by generating synthetic datapoints rather than oversampling with

replacement as proposed earlier. Samples for the minority class are extended by creating new examples along a line that joins the k nearest neighbors of the minority class. Combined with undersampling the majority class, SMOTE is known to significantly improve classification performance in Receiver Operating Characteristic (ROC) space [10]. Therefore, we adopt and extend this procedure as follows: First, we oversample the minority class using Random Oversampling (with a sampling strategy of 0.3) in order to reach a sufficient number of k neighbors for SMOTE. Then, the minority class is oversampled synthetically using SMOTE (with a sampling strategy of 0.5). In the end, we undersample the majority class with a sampling strategy of 0.5. As a result, we receive a class distribution which is approximately evenly distributed.

Supervised Approximation of Anomaly Detector. In the next step, the AD ensemble is approximated globally using supervised ML in order to learn the dependencies between the original input features and the classes *normal* and *anomalous* (as provided by the AD ensemble). Therefore, we experimented with two discriminating classifiers for our supervised approximation task. On the one hand, we tried a Support Vector Machine (SVM) and on the other hand an XGBoost model. The reason for this selection lies in the fact that we initially wanted to avoid extensive oversampling and undersampling and see how effective the supervised algorithms are in approximating the AD ensemble. As a consequence, we looked for classification algorithms that can intrinsically deal with imbalanced data w.r.t. class distribution through hyper-parametrization. Using class-weighted algorithms, the resulting models give classification errors made on the minority class more impact. As the cross-validated macro-averaged F1-score was only around 0.65 due to highly imbalanced data often comprising only one anomaly, we decided to perform over- and undersampling as described above and reached a satisfying F1-score of around 0.9. In the end, we decided to use XGBoost with SMOTE and no train-test-split or cross-validation at all, since our goal is not to make inference on unseen examples, but to globally imitate the AD ensemble using approximation.

Receiver-Dependent Explanations. According to Fig. 1, we provide two different kinds of explanations for all identified anomalies, depending on the characteristics of the receivers (refer to Appendix B for both explanation types). For Data Scientists, detailed and built-in-LIME visualizations are provided as explanations, which show the individual feature attributions (w.r.t. to the supervised approximation of the AD ensemble) to the anomaly-class. We include both categorical as well as continuous features. For the latter, feature importances as well as discretized intervals for the according values are shown. Alternatively, we offer higher-level and more human-like explanations comprising natural-language-elements in order to give auditors hints for further investigation. Such human-friendly explanations shall not only be truthful, general, probable, and consistent with prior believes, but also selective [24]. As a consequence, they do not need to cover the complete list of causes. Instead, one to three causes

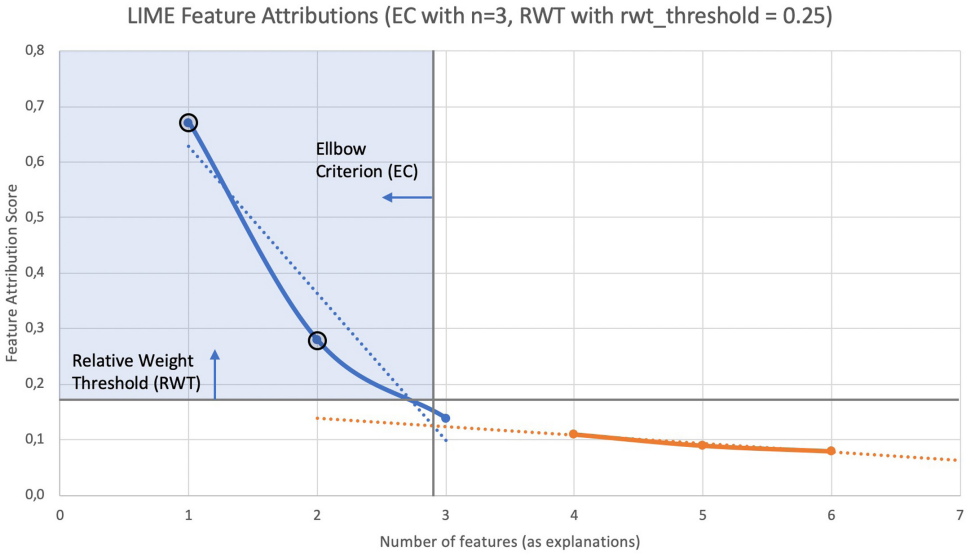


Fig. 3. Selective explanations through explanation-post-processing

should be selected and presented as the explanation. To achieve this, we propose a technique for explanation-post-processing, that comprises two concepts: *Ellbow Criterion* (EC) and *Relative Weight Threshold* (RWT), both depicted in Fig. 3. The EC is found by sorting the absolute feature attributions in descending order and then fitting regression lines through n adjacent points each (depicted in blue and orange in Fig. 3). In a next step, the intersections between all regression lines are calculated and the left-most intersection is taken as EC. Finally, all features lying left of EC are included in the explanation. RWT constitutes a predefined threshold. All features that have a higher relative attribution (compared to the first one, which is the most important one) than the *rwt_threshold* are considered relevant. Therefore, all features lying above the RWT are included in the explanation. Combining both EC and RWT results in approximately one to four features as explanation depending on the hyper-parameters (n for EC and *min_threshold* for RWT).

Evaluation. As we describe an integrated architecture for unsupervised AD, we are not able to evaluate our approach against any kind of ground truth. Instead, we decided to perform an evaluation by interviewing domain experts and showing them the detected anomalies. We discussed the results with both financial auditors and business experts for the financial domain. All domain experts were supportive of the practical use and especially highlighted three major benefits: First, they considered the ensemble architecture as very effective, as it enables a high variability in the kind of AD algorithms. This statement is also supported by Fig. 6 and 7 in Appendix A. Both figures visualize the correlation between the different algorithms w.r.t. to the identified anomalies. Second, the experts appreciated the possibility to individually and easily specify the granularity of

the meta detector and therefore achieve different levels of granularity in the detected anomalies. Figure 4 and 5 in Appendix A show the effects of varying the *min_count* threshold for the meta detector. In our special use case they decided to choose $min\_count = 8$ such that an anomaly is detected if at least 8 AD algorithms of the ensemble agree. Third, all experts agreed that the provided explanations help a lot in understanding the ensemble's functionality and can provide the auditor with reasonable hints for further inspection. Especially the last point is strengthened by the possibility to choose between different kinds and granularity levels of explanations (see Fig. 8, 9, 10 and 11 in Appendix B).

4 Conclusion and Future Work

A novel proposal for performing efficient and comprehensible unsupervised AD for financial auditing has been worked out. One main merit of the described architecture is the possibility to detect anomalies at different levels of granularity. Worth mentioning is also the fact that all decisions made by the AD ensemble can be comprehended and thus, auditors can develop trust into the system. This is achieved by generating model-agnostic and receiver-dependent explanations despite the lack of class labels. Furthermore, offering different kinds of explanations contributes in two ways. First, Data Scientists might use insights gained from the explanations for constructing purposeful features in an unsupervised setting. This might pave the way for better human involvement in the feature engineering process in the absence of ground truth. Second, persons active in regulatory matters need concise explanations to develop trust in the AD system. Therefore, we developed a technique for explanation-post-processing that consists of an Ellbow Criterion and a Relative Weight Threshold. The combination of both enables the generation of selective explanations that can be easily understood by auditors. Besides all the benefits listed so far, there are also a few prerequisites and limitations of the approach. Although the ensemble-architecture yields many advantages, finding a suitable meta detector often is not straightforward. Furthermore, explaining the ensemble via supervised approximation as intermediate step adds some extra complexity, requires synthetic oversampling, and comes with an additional inductive bias. Lastly, a quantitative evaluation of the whole system is hard to conduct in the absence of ground truth and requires domain experts' input. Summing up, this work leaves some perspectives for further studies. First, adding more types of AD algorithms to the ensemble could improve the AD system. In case of sufficient amount of data, especially DL approaches for AD might add some benefits. Second, optimization of the meta detector, maybe using meta-learning, could be helpful for further adjusting the granularity of the detected anomalies. Finally, building model-agnostic explainers that are directly suitable for unsupervised ML without intermediate supervised approximation might be an interesting research direction.

Acknowledgments. We say many thanks to DATEV eG (Markus Decker, Jörg Schaller, Dr. Thilo Edinger, Gregor Fischer) and the University of Bamberg (Prof. Dr. Ute Schmid, head of Cognitive Systems Group) for professional and organizational support.

A Anomaly Detection Ensemble

Table 2. Characteristics of the different AD algorithms included in the ensemble

Algorithm	Abbreviation	Principle	References
K Nearest Neighbors	KNN	Proximity-based	[2]
Isolation Forest	IForest	Ensemble-based	[20]
Principal Component Analysis	PCA	Dimension reduction	[33]
One-class Support Vector Machine	OCSVM	Linear model	[32]
Local Outlier Factor	LOF	Proximity-based	[6]
Clustering-Based Local Outlier Factor	CBLOF	Proximity-based	[13]
Histogram-Based Outlier Score	HBOS	Proximity-based	[12]
Copula-Based Outlier Detection	COPOD	Probabilistic model	[19]
Lightweight On-line Detector of Anomalies	LODA	Ensemble-based	[27]

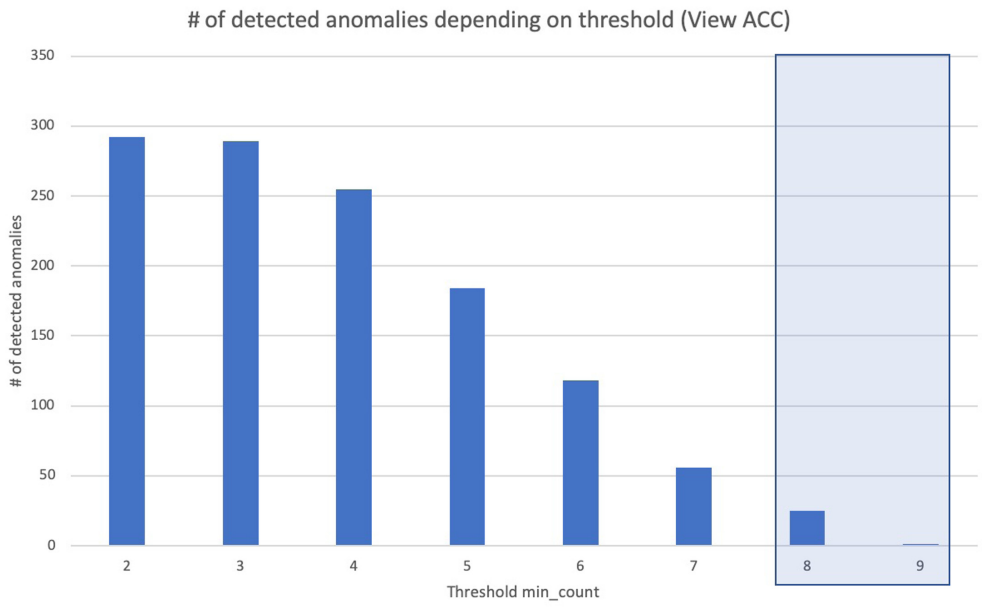


Fig. 4. Number of detected anomalies depending on threshold *min_count* for view *ACC*

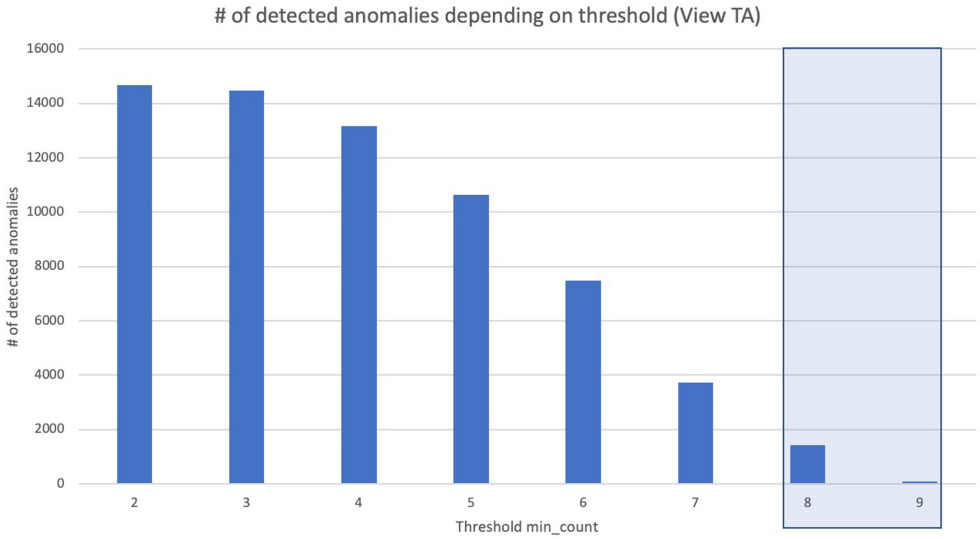


Fig. 5. Number of detected anomalies depending on threshold *min_count* for view *TA*

	KNN	IForest	PCA	OCSVM	LOF	CBLOF	HBOS	COPOD	LODA
KNN	1,00	0,63	0,73	0,85	0,35	0,57	0,34	0,44	0,03
IForest	0,63	1,00	0,58	0,59	0,19	0,39	0,42	0,58	0,07
PCA	0,73	0,58	1,00	0,71	0,31	0,53	0,31	0,42	0,02
OCSVM	0,85	0,59	0,71	1,00	0,32	0,54	0,31	0,40	0,01
LOF	0,35	0,19	0,31	0,32	1,00	0,35	0,14	0,12	0,10
CBLOF	0,57	0,39	0,53	0,54	0,35	1,00	0,19	0,24	0,03
HBOS	0,34	0,42	0,31	0,31	0,14	0,19	1,00	0,47	0,16
COPOD	0,44	0,58	0,42	0,40	0,12	0,24	0,47	1,00	0,12
LODA	0,03	0,07	0,02	0,01	0,10	0,03	0,16	0,12	1,00

Fig. 6. Correlations between different AD methods for view *ACC*

	KNN	IForest	PCA	OCSVM	LOF	CBLOF	HBOS	COPOD	LODA
KNN	1,00	0,73	0,64	0,80	0,35	0,53	0,42	0,60	0,03
IForest	0,73	1,00	0,60	0,74	0,29	0,48	0,47	0,62	0,04
PCA	0,64	0,60	1,00	0,67	0,28	0,49	0,34	0,50	0,01
OCSVM	0,80	0,74	0,67	1,00	0,34	0,55	0,39	0,59	0,01
LOF	0,35	0,29	0,28	0,34	1,00	0,33	0,21	0,22	0,16
CBLOF	0,53	0,48	0,49	0,55	0,33	1,00	0,26	0,36	0,04
HBOS	0,42	0,47	0,34	0,39	0,21	0,26	1,00	0,48	0,14
COPOD	0,60	0,62	0,50	0,59	0,22	0,36	0,48	1,00	0,05
LODA	0,03	0,04	0,01	0,01	0,16	0,04	0,14	0,05	1,00

Fig. 7. Correlations between different AD methods for view *TA*

B Explanations

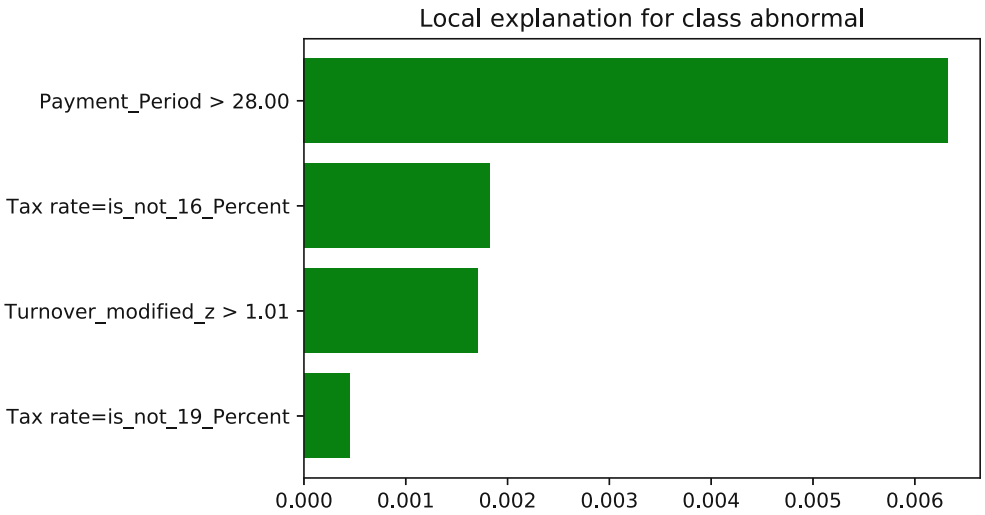


Fig. 8. Detailed explanation for view *TA*

High-level & human-like explanations for view *TA*:
„Mind a comarably long payment period!“
“Have a look at the tax rate, which is neither 16 nor 19 percent!“
“Mind a comparably high turnover!“

Fig. 9. Human-like explanation for view *TA*

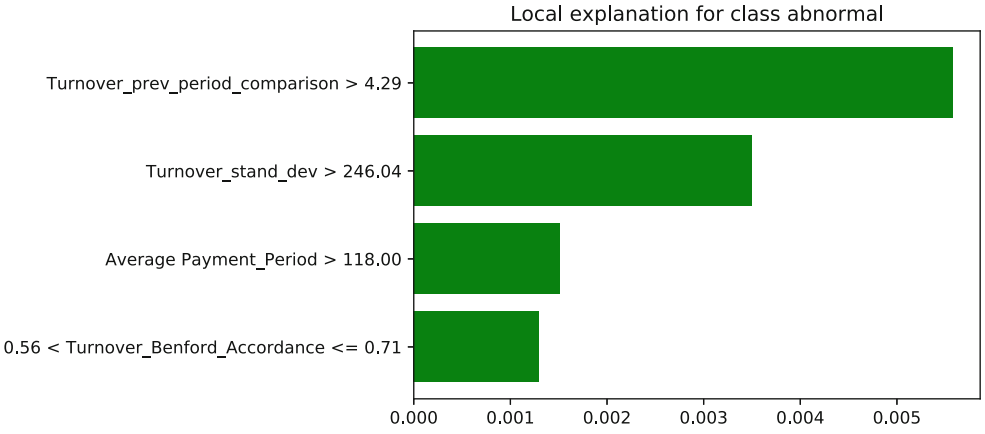


Fig. 10. Detailed explanation for view *ACC*

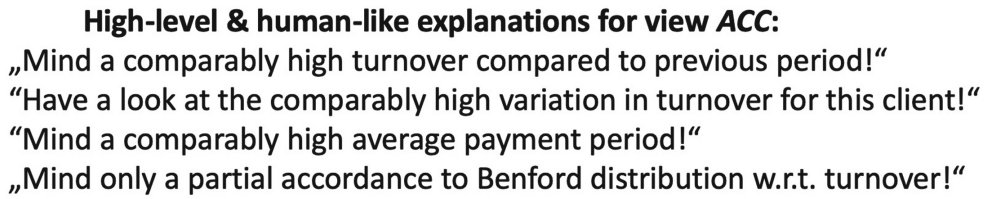


Fig. 11. Human-like explanation for view *ACC*

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6** (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS, vol. 2431, pp. 15–27. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45681-3_2
3. Antwarg, L., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv (2019)
4. Benford, F.: The law of anomalous numbers. Proc. Am. Philos. Soc. **78**, 551–572 (1938)
5. Bertsimas, D., Dunn, J.: Optimal classification trees. Mach. Learn. **106**(7), 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9
6. Breuniq, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. SIGMOD Rec. (ACM Special Interest Group on Management of Data) **29** (2000). https://doi.org/10.1145/335191.335388
7. Bruckert, S., Finzel, B., Schmid, U.: The next generation of medical decision support: a roadmap toward transparent expert companions. Front. Artif. Intell. **3** (2020). https://doi.org/10.3389/frai.2020.507973
8. Böhmer, K., Rinderle-Ma, S.: Anomaly detection in business process runtime behavior – challenges and limitations. arXiv (2017)
9. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41** (2009). https://doi.org/10.1145/1541880.1541882
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16** (2002). https://doi.org/10.1613/jair.953
11. Fahim, M., Sillitti, A.: Anomaly detection, analysis and prediction techniques in IoT environment: a systematic literature review. IEEE Access **7** (2019). https://doi.org/10.1109/ACCESS.2019.2921912
12. Goldstein, M., Dengel, A.: Histogram-based outlier score (hbos): a fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track (2012)
13. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. **24** (2003). https://doi.org/10.1016/S0167-8655(03)00003-5
14. Henselmann, K., Scherr, E., Ditter, D.: Applying Benford's law to individual financial reports: an empirical investigation on the basis of SEC XBRL filings. Working papers in accounting valuation auditing (2012)
15. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv (2017)

16. Jolliffe, I.T.: Principal component analysis, second edition. Encyclopedia of Statistics in Behavioral Science **30** (2002). https://doi.org/10.2307/1270093
17. Kauffmann, J., Müller, K.R., Montavon, G.: Towards explaining anomalies: a deep Taylor decomposition of one-class models. Pattern Recogn. **101** (2020). https://doi.org/10.1016/j.patcog.2020.107198
18. Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L.: Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. **49** (2013). https://doi.org/10.1016/j.jesp.2013.03.013
19. Li, Z., Zhao, Y., Botta, N., Ionescu, C., Hu, X.: COPOD: copula-based outlier detection. In: Proceedings - IEEE International Conference on Data Mining, ICDM (2020). https://doi.org/10.1109/ICDM50108.2020.00135
20. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data **6** (2012). https://doi.org/10.1145/2133360.2133363
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (2017)
22. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
23. Mehrotra, K.G., Mohan, C.K., Huang, H.: Anomaly Detection Principles and Algorithms. Book (2017)
24. Molnar, C.: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Book (2019)
25. Morichetta, A., Casas, P., Mellia, M.: Explain-it: towards explainable AI for unsupervised network traffic analysis. In: Big-DAMA 2019 - Proceedings of the 3rd ACM CoNEXT Workshop on Big Data, Machine Learning and Artificial Intelligence for Data Communication Networks, Part of CoNEXT 2019 (2019). https://doi.org/10.1145/3359992.3366639
26. Munir, M., Chattha, M.A., Dengel, A., Ahmed, S.: A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In: Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019 (2019). https://doi.org/10.1109/ICMLA.2019.00105
27. Pevný, T.: Loda: lightweight on-line detector of anomalies. Mach. Learn. **102**(2), 275–304 (2015). https://doi.org/10.1007/s10994-015-5521-0
28. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag. **6** (2006). https://doi.org/10.1109/MCAS.2006.1688199
29. Rabold, J., Schwalbe, G., Schmid, U.: Expressive explanations of DNNs by combining concept analysis with ILP. In: Schmid, U., Klügl, F., Wolter, D. (eds.) KI 2020. LNCS (LNAI), vol. 12325, pp. 148–162. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58285-2_11
30. Radev, D.R., Jing, H., Styś, M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manag. **40** (2004). https://doi.org/10.1016/j.ipm.2003.10.006
31. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939778
32. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13** (2001). https://doi.org/10.1162/089976601750264965
33. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. In: 3rd IEEE International Conference on Data Mining (2003)

34. Thudumu, S., Branch, P., Jin, J., Singh, J.J.: A comprehensive survey of anomaly detection techniques for high dimensional big data. J. Big Data **7**(1), 1–30 (2020). https://doi.org/10.1186/s40537-020-00320-x
35. Xu, X., Liu, H., Yao, M.: Recent progress of anomaly detection. Hindawi Complex. **2019** (2019). https://doi.org/10.1155/2019/2686378
36. Zhao, Y., Nasrullah, Z., Li, Z.: Pyod: a Python toolbox for scalable outlier detection. J. Mach. Learn. Res. **20**, 1–7 (2019)

A.3 Concept-based Explanatory Interactive Machine Learning for Text Classification

A.3.1 Kiefer et al. "Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions." In: Machine Learning and Knowledge Extraction 2022

Full Reference of Paper

Sebastian Kiefer, Mareike Hoffmann and Ute Schmid. "Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions." In *Machine Learning and Knowledge Extraction, 4* (2022), pp. 994-1010. DOI: 10.3390/make4040050.

My Scientific Contributions

Related research question(s): **RQ4a - RQ4c**.
Type of contribution: **technical**.

During the research underlying this dissertation, I made the following scientific contributions published in this paper:

- Identifying the research gap of missing constructive and conceptual feedback during human corrections in the scope of interactive machine learning.
- Structuring the literature on active and interactive learning approaches and motivating the new semantic interactive ML approach by referencing insights from human-computer interaction research and psychology.
- Identification of major drawbacks of CAIPI, which does not allow for constructive and continuous feedback and does not include context in its corrections. In addition, CAIPI only works in the "right for the wrong reasons" case.
- Proposition of a novel interaction framework called *semantic interactive learning* that allows humans to incorporate conceptual corrections also for false predictions.
- Specification of a new interaction strategy called *Semantic Push* as an instantiation of the *semantic interactive learning* framework.
- Definition of the evaluation methodology comprising several integrated measures that analyze predictive performance and local explanation quality from both directions of interaction.

- Outlining future research, like including language models to generate linguistically, especially syntactically correct counterexamples.

My Written Contents

I wrote all of the contents except parts of section 2 on Related Work. My written contribution to the paper is 95%.

machine learning & knowledge extraction

MDPI

Article

Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions

Sebastian Kiefer *, Mareike Hoffmann and Ute Schmid

Cognitive Systems, University of Bamberg, 96047 Bamberg, Germany
* Correspondence: sebastian.kiefer@uni-bamberg.de

Abstract: Interactive Machine Learning (IML) can enable intelligent systems to interactively learn from their end-users, and is quickly becoming more and more relevant to many application domains. Although it places the human in the loop, interactions are mostly performed via mutual explanations that miss contextual information. Furthermore, current model-agnostic IML strategies such as CAIPI are limited to 'destructive' feedback, meaning that they solely allow an expert to prevent a learner from using irrelevant features. In this work, we propose a novel interaction framework called *Semantic Interactive Learning* for the domain of document classification, located at the intersection between Natural Language Processing (NLP) and Machine Learning (ML). We frame the problem of incorporating constructive and contextual feedback into the learner as a task involving finding an architecture that enables more semantic alignment between humans and machines while at the same time helping to maintain the statistical characteristics of the input domain when generating user-defined counterexamples based on meaningful corrections. Therefore, we introduce a technique called SemanticPush that is effective for translating conceptual corrections of humans to non-extrapolating training examples such that the learner's reasoning is pushed towards the desired behavior. Through several experiments we show how our method compares to CAIPI, a state of the art IML strategy, in terms of Predictive Performance and Local Explanation Quality in downstream multi-class classification tasks. Especially in the early stages of interactions, our proposed method clearly outperforms CAIPI while allowing for contextual interpretation and intervention. Overall, SemanticPush stands out with regard to data efficiency, as it requires fewer queries from the pool dataset to achieve high accuracy.

Keywords: human-centric machine learning; interactive machine learning; CAIPI; explainable artificial intelligence; local surrogate explanation models; contextual and semantic explanations; locally faithful explanations; topic modeling

Citation: Kiefer, S.; Hoffmann, M.; Schmid, U. Semantic Interactive Learning for Text Classification: A Constructive Approach for Contextual Interactions. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 994–1010. https://doi.org/10.3390/make4040050

Academic Editor: Andreas Holzinger

Received: 20 October 2022
Accepted: 4 November 2022
Published: 13 November 2022

1. Introduction

Although modern ML approaches have improved tremendously with regard to prediction accuracy, and even exceed human performance in many tasks, they often lack the ability to allow humans to develop an understanding of the whole logic or of the model's specific behavior [1–3]. Additionally, most systems do not allow the integration of corrective feedback for use in model adaptation.

Consequently, different research disciplines have emerged that provide first solutions. Both *Interpretable Machine Learning* and *Explainable Artificial Intelligence*, which can be summarized as *Comprehensible Artificial Intelligence* [4] when combined, allow for global or local interpretability as well as transparent and comprehensible ML results [5]. In general, global interpretability refers to providing intrinsic ex ante understanding of the whole logic of the corresponding models. The explanandum is therefore the ML model itself, with the *rules of reasoning* as the explanans providing information about how all of the different possible outcomes are connected to the inputs. In contrast, local interpretability provides ex post understanding of the model's specific behavior [1]. The accompanying explanations

for individual decisions strive to make the input–output correlations clear to the users without the need for them to know the internal structure of the model [1].

Nevertheless, the explanations used for better transparency and human comprehensibility during human–machine interactions are mostly considered unidirectional from the AI system to the human, and often lack contextual information [1]. Therefore, any correction of erroneous behavior or any inclusion of domain-specific knowledge through human experts is not possible in a model-agnostic way [4]. *Explanatory Interactive Machine Learning* addresses this shortcoming, with the intention of 'closing the loop' by allowing humans to correct the prediction and explanations of a query and thus to provide feedback [6]. The authors of [6] demonstrated that both the predictive and explanatory performance of the learner and the process of building trust in the learner can benefit from interacting through explanations. Except in systems such as EluciDebug [7] or Crayon [8], which use feedback to adapt a learner (albeit model-specifically), there are few possibilities at present for holistic, meaningful, and model-agnostic interventions to correct learner mistakes by incorporating expert knowledge.

Based on this research gap, we phrase the following research questions (RQ): (1) How can we develop a model-agnostic Interactive ML approach that offers semantic (constructive, meaningful, contextual, and realistic) means for performing corrections and providing hints? Concretely, how can conceptual human corrections be integrated into ML classifiers while avoiding counterexamples that are considered 'Out-of-Distribution'? (2) Is the elaborated interactive system with contextual interpretation and intervention support comparable to the state-of-the-art methods in terms of predictive performance of downstream multi-class classification tasks? (3) Can our method generate explanations that are comparable to the state-of-the-art methods with regard to the conclusiveness of the explanations?

Based on our research, in this paper we propose an architecture called Semantic Interactive Learning and instantiate it with a technique named SemanticPush. Technically, this approach contributes to the field of Interactive Machine Learning by allowing humans to correct all possible types of a text classifier's reasoning and prediction errors. We showcase how the proposed method harnesses the generative process of the input domain's data learned by a Latent Dirichlet Allocation Model in order to transfer human conceptual corrections to non-extrapolating counterexamples. These counterexamples can then be used to incorporate the corrections into the learner's inductive process in a model-agnostic way. Finally, we propose new context-based evaluation metrics for explanations and evaluate our approach with regard to the research questions mentioned above.

2. Related Work

Human-Centered Machine Learning can be summarized as methods for aligning machine learning systems with human goals, contexts, concerns, and ways of working [9]. It is strongly connected with Interactive Machine Learning as an interaction paradigm in which a user or user group iteratively trains a model by selecting, labeling, and/or generating training examples to deliver a desired function [10]. It can be assumed that a learner is better aligned with human goals when the end user knows more about its behavior (Explanatory Interactive Machine Learning). Kulesza et al. (2015) [7] proved this intuition with their Explanatory Debugging approach. They additionally showed that not only does the machine benefit from corrections based on transparent explanations, but also the user is able to build a more accurate mental model about its behavior. Furthermore, in Koh et al. (2020) [11] the authors found that concept bottleneck models applied to image classification tasks support intervention and interpretation while competing on predictive performance of downstream tasks such as x-ray grading and bird identification. Thus, they can enable effective human–model collaboration by allowing practitioners to reason about the underlying models in terms of higher-level concepts that humans are typically familiar with.

Hence, interactions between humans and machines via mutual explanations [4,12] have the potential to adequately bring humans into the loop in a model-agnostic way. The overall process should work as a Training–Feedback–Correction cycle that enables a Machine Learning model to quickly focus on a desired behavior [8]. Users should be able to iteratively integrate corrective feedback into a Machine Learning model after having analyzed its decisions [13].

Consequently, Teso and Kersting (2019) [6] included a local explainer called *Local Interpretable Model-Agnostic Explanations* (LIME) into an active learning (AL) setting. Their framework proposes a method called CAIPI which enables users to correct a learner when its predictions are right for the wrong reasons by adding counterexamples in a 'destructive' manner. The correction approach is based on Zaidan et al. (2007) [14]. As an example from the text domain, words which are falsely identified as relevant are masked from the original document, then the resulting counterexamples recur as additional training documents.

Although CAIPI has paved the way for model-agnostic and explanatory IML, its use has revealed a number of significant drawbacks. First, it only operates by deleting irrelevant explanatory features, i.e., those that have been incorrectly learned. Thus, it is limited to 'destructive' feedback about incorrectly-learned correlations; an active learning setting might rarely contain correct predictions made for the wrong reasons. Second, CAIPI uses contextless explanations as a basis, and in turn applies contextless feedback by independently removing irrelevant explanatory features. In this manner, human conceptual knowledge may hardly be considered during interactions, even it is known that harnessing conceptual knowledge "as a guiding model of reality" might help to develop more explainable and robust ML models which are less biased [15]. A first step towards this was suggested by Kiefer (2022) [16], who proposed topicLIME as an extension of LIME that offers contextual and locally faithful explanations by considering higher-level semantic characteristics of the input domain within the local surrogate explanation models. A third drawback of CAIPI is that it enables only 'discrete' feedback. In the textual domain, this is based on mutual explanations in a bag-of-words representation, in which words are either present as explanatory features or are not. Therefore, continuous feedback is not possible.

When explaining and correcting a classifier in the way described above, neighborhood extrapolation to feature areas with low data density, especially in cases of dependent features [17], causes a classifier to train on contextless counterexamples sampled from unrealistic local perturbation distributions. This circumstance might lead to generalization errors.

Therefore, the overall goal of this work is to enable more realistic and constructive interactions via semantic alignment between humans and ML models across all possible types of a learner's reasoning and prediction errors.

3. Method

Figure 1 depicts our proposal for answering RQ1 from the architectural point of view. This approach extends previous research called *Contextual and Semantic Explanations* (CaSE) [16]. CaSE suggests a framework that allows contextual interpretations of ML decisions by humans in a model-agnostic way via topic-based explanations. While CaSE solely refers to the process of explanation generation, our research aims at closing the loop and enabling humans to integrate domain knowledge via semantic corrections and hints. The following subsections briefly describe the components contained in our framework, and especially introduce our new IML strategy called SemanticPush.

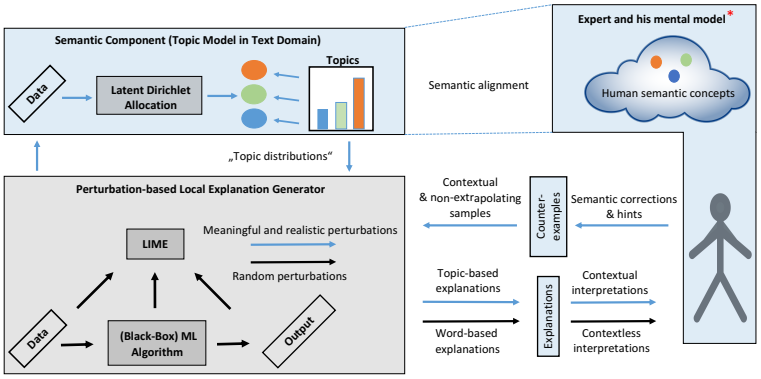


Figure 1. Architecture for constructive and contextual interactions. * In this work, we simulate the expert for efficient evaluation purposes using a conceptual Gold Standard as explained in Sections 3.3 and 4.3.

3.1. Latent Dirichlet Allocation

We instantiated the semantic component of our framework using a method called Latent Dirichlet Allocation (LDA), which can be described as a hierarchical Bayesian model for collections of discrete data [18]. Used in text modeling, it finds short representations of the documents in a corpus and preserves essential statistical relationships necessary for making sense of the input data. After training, each document can be characterized as a multinomial distribution over so-called topics. For each document **w** in a corpus **D**, a generative process from which the associated documents have been created is assumed as follows:

1. Choose N (the number of words) $\sim$ Poisson(ξ).
2. Choose θ (a topic mixture) $\sim$ Dir(α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ).
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

The joint distribution of a topic mixture θ, a set of topics **z**, and a set of words **w**, given the hyperparameters α and β, is characterized by

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta). \quad (1)$$

The latent multinomial variables are referred to as topics, and enable LDA to capture text-oriented intuitions and global statistics in a corpus. Thus, it is able to make sense of the input data due to its generative probabilistic semantic properties.

We combined LDA with a coherence measure called C_v coherence, which is used for finding an appropriate hyperparameter *number of topics k* that LDA then infers. Röder et al. found the coherence measure to be the best in terms of its correlation with respect to human topic interpretability [19,20].

Within our Semantic Interactive Learning framework (refer to Figure 1), we use LDA as a measure to address Research Question 1. LDA provides the basis to enable interactions that are deemed constructive (by harnessing its generative process to create user-specified documents), semantically meaningful (by making sense of the input data and identifying coherent words as topics), and contextual and reliable (by capturing statistical characteristics of the input domain such as word dependencies). In this way, counterexamples generated by LDA can be considered 'In-Distribution' of the input domain.

3.2. LIME and topicLIME

Ribeiro et al. [21] developed LIME, a method that explains a prediction by locally approximating the classifier's decision boundary in the neighborhbood of the given instance.

LIME uses a local linear explanation model, and can thus be characterized as an additive feature attribution method [22]. Given the original representation $x \epsilon \mathbb{R}^d$ of an instance to be explained, $x' \epsilon \{0,1\}^{d'}$ denotes a binary vector for its interpretable input representation. Furthermore, let an explanation be represented as a model $g \epsilon G$, where G is a class of potentially interpretable models such as linear models or decision trees. Additionally, let $\Omega(g)$ be a measure of complexity of the explanation $g \epsilon G$, for example, the number of non-zero weights of a linear model. The original model for which explanations are searched is denoted as $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A measure $\pi_x(z)$ defining the locality around x is used to capture the proximity between an instance z and x. The final objective is to minimize a measure $\mathcal{L}(f, g, \pi_x(z))$ that evaluates how unfaithful g (the local explanation model) is at approximating f (the model to be explained) in the locality defined by $\pi_x(z)$. Striving for both interpretability and local fidelity, an explanation is obtained by minimizing $\mathcal{L}(f, g, \pi_x(z))$ as well as by keeping $\Omega(g)$ low enough to ensure an interpretable model:

$$\xi(x) = \underset{g \epsilon G}{\arg\min}\ \mathcal{L}(f, g, \pi_x(z)) + \Omega(g). \tag{2}$$

To be a model-agnostic explainer, the local behavior of f must be learned without making any assumptions about f. Therefore, $\mathcal{L}(f, g, \pi_x(z))$ needs to be approximated by drawing random samples weighted by $\pi_x(z)$; instances around x' (a binary vector for the interpretable input representation of x) are sampled by drawing nonzero elements of x' uniformly at random, then a perturbed sample z' is obtained.

Recovering z from z' and applying $f(z)$ then yields a label, which is used as label for the explanation model. The last step consists of optimizing Equation (2) by making use of dataset $\mathcal{Z}$, which includes all perturbed samples with the associated labels. For a sample word-based explanation generated by LIME, please refer to Figure 2.

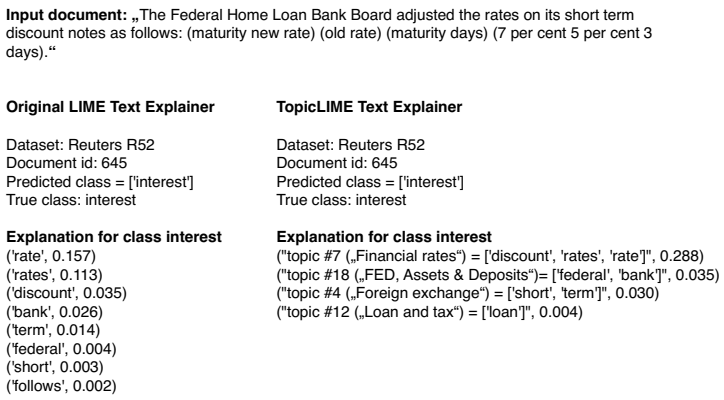


Figure 2. Textual comparison of original LIME text explainer (**left**) and topicLIME text explainer (**right**). A contextual interpretation of the word-explanations generated by LIME is complicated as the semantic "links" of a word are not reflected in the explanations. For topicLIME explanations, coherent and most likely, at least semantically, related words are considered at once including the semantic "links" that in turn provide the context in the explanations.

In contrast to LIME, topicLIME, developed by Kiefer (2022) [16], generates a local neighborhood of a document to be explained by removing coherent words. It is therefore capable of including the distributional, contextual, and semantic information of the input

domain in the resulting topic-based explanations. As such, it offers realistic and meaningful local perturbation distributions by avoiding extrapolation when generating the local neighborhood, leading to higher local fidelity of the local surrogate models. For a sample topic-based explanation generated by topicLIME, please refer to Figure 2.

3.3. Our Method: SemanticPush

Our proposed method, called SemanticPush, enables model-agnostic Interactive Machine Learning at a higher level of semantic detail. Therefore, it extends the idea of CAIPI (refer to Algorithm 1), which offers model-agnostic, albeit contextless, interactions for humans in the form of word-based explanations and 'destructive' corrections.

Algorithm 1 CAIPI [6]

Require: a set of labelled examples L, a set of unlabelled instances U, and an iteration budget T.
$f \leftarrow FIT(L)$
repeat
  $x \leftarrow$ **S***elect* **Q***uery* (f, U)
  $\hat{y} \leftarrow f(x)$
  $\hat{z} \leftarrow$ **E***xplain* $(f, x, \hat{y})$
  Present $x, \hat{y}$, and $\hat{z}$ to the user
  Obtain y and explanation correction C
  $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{c} \leftarrow$ **T***o* **C***ounterexamples*(C)
  $L \leftarrow L \cup \{(x, y)\} \cup \{\bar{x}_i, \bar{y}_i)\}_{i=1}^{c}$
  $U \leftarrow U \setminus (\{x\} \cup \{\bar{x}_i\}_{i=1}^{c})$
  $f \leftarrow FIT(L)$
until budget T is exhausted or f is good enough
return f

From IML research, it is known that humans want to demonstrate how learners *should* behave. According to Amershi et al. (2014) [13] and Odom and Natarajan (2018) [23], people do not want to simply teach 'by feedback'; we want to teach 'by demonstration', that is, by providing examples of a concept. Therefore, interaction techniques should move away from limited learner-centered ways of interacting and instead proceed to more natural modes of feedback, such as suggesting alternative or new features [24].

SemanticPush provides this knowledge in practice, as depicted by the graphical model in Figure 3.

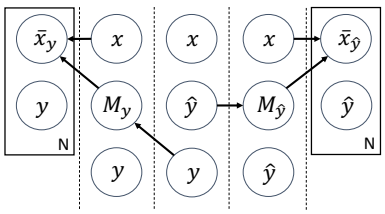


Figure 3. Graphical Model of SemanticPush.

Let X and Y be the input and output space for a binary classification, where $x \in X$ represents a query instance, $y \in Y$ is the accompanying true label, and $\hat{y} \in Y$ is the predicted label. The overall goal is to find a matrix M depending on labels y or $\hat{y}$ that adequately incorporates human feedback into the classifier's reasoning in a model-agnostic way by generating counterexamples $\bar{x}$ based on x. Thus, we seek a set of L input manipulations $M = \{m_1, ..., m_L\}$ as well as a manipulation function $q : M \times X \rightarrow \bar{X}$. Here, $q(m, x)$ is a local function such that it only affects a part of the input x. This is the case because user

input in IML shall be focused (i.e., it shall only affect a certain part/aspect of the model) as well as incremental (i.e., each user input shall only result in a small change to the model) [13]. Algorithms 2 and 3 describe in detail the different semantic manipulations of X (corrections and completions) performed by SemanticPush.

Algorithm 2 SemanticPush

Require: a destructive correction set C_{dest_x}, a topicLIME explanation $\hat{z}_{xy}$ for query instance x with true class y, expert knowledge (here simulated via Gold Standard GS), and a balancing parameter λ

$C_{dest_x} = \{t \in \hat{z}_{xy} | t \notin GS_y\}$ ▷ Set of falsely explained topics
if $\hat{y} = y \wedge C_{dest_x} \neq \emptyset$ **then** ▷ Right for partially wrong reasons
  $\bar{x}_i \leftarrow x \setminus C_{dest_x} \cup$ **S***emantic* **C***ompletion*($x, GS_y, \hat{z}_{xy}, \lambda$) ▷ Add a concept the classifier forgot to learn
  $\bar{y} \leftarrow y$
else if $\hat{y} \neq y$ **then** ▷ False prediction
  $\bar{x}_{iy} \leftarrow$ **S***emantic* **C***orrection*$_y(x, GS_y, \hat{z}_{xy})$ ▷ Provide feedback/hints for the true class
  $\bar{y}_{iy} \leftarrow y$
  $\bar{x}_{i\hat{y}} \leftarrow$ **S***emantic* **C***orrection*$_{\hat{y}}(x, GS_{\hat{y}}, \hat{z}_{x\hat{y}})$ ▷ Provide feedback/hints for the predicted class
  $\bar{y}_{i\hat{y}} \leftarrow \hat{y}$
end if

Algorithm 3 S*emantic* C*orrection*

Require: a Topic Model *lda*

$\theta_x \leftarrow lda$.**G***et* **T***opic* **M***ixture*(x)
for $t \in \theta$ **do** ▷ t represents a topic as explanation unit
  if $t \in \hat{z}^+ \cap GS^+ \vee t \in \hat{z}^- \cap GS^- \vee t \in \hat{z}^+ \cap GS^- \vee (t \notin \hat{z} \wedge t \notin GS)$ **then** ▷ Topics either correctly used or incorrectly used (but hard to reverse polarity and still important) or correctly ignored
    $\hat{\theta}_{x_t} \leftarrow$ KeepProbability(θ_{x_t})
  else if $t \in \hat{z}^- \cap GS^+ \vee (t \notin \hat{z} \wedge t \in GS^+)$ **then** ▷ Topics either incorrectly learned (but easy to reverse polarity) or forgotten to learn
    $\hat{\theta}_{x_t} \leftarrow$ **I***ncrease* **P***robability*($\theta_{x_t}, GS, \lambda$)
  else if $(t \in \hat{z} \wedge t \notin GS)$ **then** ▷ Irrelevant topics were used
    $\hat{\theta}_{x_t} \leftarrow$ **D***ecrease* **P***robability*(θ_{x_t})
  end if
end for
return *lda*.**S***ample* **I***nstance*($\psi(\hat{\theta}_x)$) ▷ sampling from the multinomial distribution harnessing the generative process of LDA

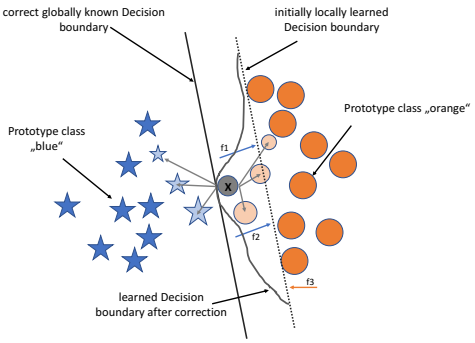
S*emantic* **C***ompletion*$(x, GS_y, \hat{z}_{xy}, \lambda)$ from Algorithm 2 is defined as $\sim [\lambda * \psi(C_{add_x}) + (1 - \lambda) * \psi(x_{add})]$, where $C_{add_x} = \{(t, t_w) \in GS_y^+ | t \notin \hat{z}_{xy}^+\}$ and $x_{add} = \{(t, t_w) \in x | t \in C_{add_x}\}$. Here, C_{add_x} contains relevant and positively attributed topics for the predicted label y weighted according to Gold Standard GS_y^+ that are (thus far) missing in the classifier's explanation.

In addition, ψ constitutes a normalization operator that re-normalizes the weights t_w of the associated topics t (either from Gold Standard or topicLIME explanation), revealing a multinomial distribution over topics t. SemanticPush then incorporates the concepts the classifier forgot to learn by adding text parts via sampling ($\sim$) from the multinomial distribution and harnessing the generative process of LDA (see Section 3.1).

I*ncrease* **P***robability*() from Algorithm 3 carries out probability change of a topic δ_t in the following way: $\delta_t = \theta_{x_t} + \lambda * GS_{y_t} + (1 - \lambda) * \theta_{x_t}$.

D*ecrease* **P***robability*() from Algorithm 3 in our scenario sets the probability of a topic to zero, as the topic is assumed to be irrelevant for the class decision.

In order to more efficiently evaluate and optimize SemanticPush, we consciously decided to use a simulated oracle that can be replaced by a human expert in a practical real-life scenario. Therefore, SemanticPush is based on a newly developed *conceptual Gold Standard GS* that works as a proxy for an expert's knowledge. Specifically, GS_y contains concepts in the form of LDA-retrieved topics that should be informative for a specific class y. We obtain this kind of Gold Standard using intrinsic feature selection, especially by extracting the weights of a Logistic Regression Model trained on all available topic-represented data from the datasets described in Section 4.2. The details of how GS is implemented can be found in Section 4.3. The superscripts $+$ and $-$ (of Gold Standard GS or explanations z, respectively) indicate positive and negative attributions for a specific class. In addition to the algorithmic descriptions, Figure 4a,b illustrates SemanticPush conceptually and with an example application.



(**a**)

1) Input document: „ CRA sold Forrest Gold for 8 mln dlrs. Whim Creek Consolidated NL said, the consortium it is leading will pay 50 mln dlrs for the acquisition of CRA LTDs Forrest Gold. PTY LTD Unit reported yesterday CRA and Whim Creek did not disclose the price yesterday. Whim Creek will hold 10 pct of the consortium while Austwhim resources NL will hold 5 pct and Croesus Mining NL 5 pct it said in a statement. As reported Forrest Gold owns two mines in western australia producing a combined 50 ounces of gold a year. It also owns an undeveloped gold project.**"**

2) Predicted class: **gold** <> **True** class: **acquisition**

3) TopicLIME explanation for class **gold**:
("topic #18 („Assets & Deposits"), 0.383)
("topic #12 („Merger & Acquisition"), -0.086)

3) TopicLIME explanation for class **acquisition**:
("topic #12 („Merger & Acquisition"), 0.302)
("topic #18 („Assets & Deposits"), -0.099)

4) Corrections/Hints for class **gold**:
- topic #18 („Assets & Deposits") → increase ↑
- topic #12 („Merger & Acquisition") → decrease / remove ↓
- topic #4 („Foreign exchange") → add +

4) Corrections/Hints for class **acquistion**:
- topic #12 („Merger & Acquisition") → increase ↑
- topic #18 („Assets & Deposits") → decrease / remove ↓
- topic #2 („Key figures") → add +

5) Generate counterexamples for classes **gold** and **acquisition**
-based on document-topic-attributions and incorporating corrections/hints
-using relative importance scores
-balancing, how global feedback shall be applied locally

(**b**)

Figure 4. (**a**) Conceptualization of SemanticPush: The grey query instance in the middle is predicted as class "blue", but should be "orange" instead according to ground truth. Local explanation features f1 and f2 are features used by the classifier locally to assign the query instance to class "blue". According to expert knowledge, those features push the learned local decision boundary too far towards the class "orange". Feature f3 also constitutes expert knowledge as it is, among others, significantly used globally by the classifier to assign instances to class "orange". SemanticPush incorporates this information by generating new instances (shown in light color) for both classes and eventually weighs them by their distance to the query instance. The degree of locality of applying the expert knowledge to the query instance is controlled by the hyperparameter λ. Sampling new instances only based on global expert knowledge might result in prototypical instances (located in dense regions) which might not lead to great benefit for the classifier. (**b**) An exemplary application of SemanticPush to document ID 9 of the Reuters R 52 Dataset.

4. Experimental Setup

4.1. Baseline: Active Learning and CAIPI

In this section, we compare our SemanticPush approach against three baseline approaches. First, we use a standard ActiveLearner that internally harnesses Maximum Classification Uncertainty with regard to a pool dataset as a sampling strategy. Classification uncertainty is defined as $U(x) = 1 - P_\theta(\hat{y}|x)$, where x is the instance to be predicted and $\hat{y}$ is the most likely prediction. Second, we apply the original CAIPI method, as described in [6] (refer to Algorithm 1), which provides explanation corrections for the 'right for the wrong reasons' ($\hat{y} = y$) case. We call this setup 'CAIPI destructive' (CAIPI$_d$), as it is only capable of removing those components that have been identified by a local LIME explanation $\epsilon(x)$ as relevant even though an oracle believes those components to be irrelevant. Third, we extend CAIPI such that it is additionally able to deal with false predictions ($\hat{y} \neq y$). We call this setting 'CAIPI destructive + constructive' (CAIPI$_{d/c}$), as we additionally generate new documents comprising words that could have been used to predict the associated true class. We therefore sample words from a set $GS^{+}_{local}(x)$ (where $GS^{+}_{local}(x) = GS^{(k^+)}_{global}(y) \cap x$) that contains the top k positive words from a global Gold Standard of the true class y (see Section 3.3) that are part of the document x.

4.2. Datasets

We evaluated SemanticPush on two multiclass classification tasks harnessing the following datasets: the *AG News Classification* Dataset [25] and *Reuters R52* Dataset [26]. The *AG News* Dataset (127,600 documents) is constructed by selecting the four largest classes from the original AG Dataset, which is a collection of more than one million news articles. The average document length is 25 words, and the classes to be distinguished are 'Business News', 'Science-Technology News', 'Sports News', and 'World News'. The *Reuters R52* Dataset (9100 documents) originally comprises 52 classes. Due to strong imbalance between the classes, we selected the ten most represented classes ('Earn', 'Acquisition', 'Coffee', 'Sugar', 'Trade', 'Ship', 'Crude', 'Interest', and 'Money-Foreign-Exchange'), leading to a corpus comprising 7857 documents. The average document length is 60 words. From now on, we refer to this dataset as the *Reuters R10* Dataset.

For both datasets, we performed standard NLP preprocessing steps such as Tokenization, Lemmatization, Stemming, Lower-Casing, and Removal of Stopwords.

4.3. Models

Our architecture comprises a semantic component that provides contextual information about the input domain. Here, we showcase how we instantiated the **Latent Dirichlet Allocation Models** for the two datasets. For this research, we used scikit-learn (version 0.20.2) and gensim (version 3.8.3). For the *AG News* Dataset, several LDA models were trained on the preprocessed corpus with different values for the *number of topics* hyperparameter k. A final selection was made by determining the optimal number K^* of topics $t = 1, \ldots, K$ by solving $\arg\max_K \frac{1}{K} \sum_{t=1}^{K} C_v(t)$, where C_v is the C_v coherence as introduced in Section 3.1. We set K to 30 and determined $K^* = 13$, meaning an optimal number of thirteen topics. These topics, together with their most representative words, are described in Table 1.

We proceeded analogously with the *Reuters R10* Dataset; however, in contrast to the *AG News* Dataset, we could not solely rely on C_v coherence to find a suitable number of topics. As the LDA model in our framework serves as both the semantic component and is used to build a topic-based Gold Standard model (see next paragraph), we had to trade off C_v coherence against learning performance. In order to achieve sufficient predictive performance for *Reuters R10* while preserving high coherence, the optimal number of topics K^* was set to 100.

Table 1. Learned LDA topics and most representative words for the *AG News* Dataset.

Topic	Representative Words
0	Iraq, Baghdad, Nuclear, Iran, Force, Military
1	Microsoft, Company, Software, IBM, System
2	European, United, Bank, Million, Trade, Deal
3	Bush, President, Press, Washington, John, Kerry
4	Internet, Search, Service, Phone, Online, Google
5	Oil, Price, Percent, Sale, Profit, Rate
6	Court, Company, Charge, Million, Trial, Drug
7	World, Cup, Win, Gold, Final, Champion
8	Game, Season, Team, League, Coach, Sport
9	New, York, Stock, Dollar, Share, Investor
10	Game, India, Australia, Fan, Video, Cricket
11	Police, People, Killed, Attack, Palestinian, Bomb
12	Minister, Election, Leader, President, Vote, Party

As described in Section 3.3, a Logistic Regression model is harnessed as an approximation for the oracle's expert knowledge required in any Active Learning setting. To obtain that kind of *Gold Standard GS* for CAIPI, we trained the regression model on the bag-of-words-represented documents and obtained the following results.

For the *AG News* Dataset, a macro-averaged F1 score of 0.85 was achieved, while for *Reuters R10* the regression model reached a score of 0.8.

In order to include contextual and higher-level semantic information (simulating the conceptual knowledge of a human expert) in the *GS* used for SemanticPush, we represented the documents as multinomial distributions over topics (features of the regression model) using the LDA model described above. The associated model achieved a macro-averaged F1 score of 0.74 for *AG News* and of 0.71 for *Reuters R10*. Due to the reduced number of features when representing documents via topics, the topic-based *GS* obviously performs slightly worse than the word-based *GS* due to reduced degrees of freedom of the regression model.

During our experiments, we primarily used an XGBoost model as the **Base Learner**, as it constitutes a high-performing ensemble and tree-based classification algorithm. We consciously made that decision because a tree-based learner is biased towards feature interaction and is able to naturally and intrinsically include both variables that interact and variables with effects that do not interact [27]. This choice allows us to compare interactions based on both context-less mutual explanations and contextual mutual explanations. The latter are based on topics that contain words that can be polysemous or can exhibit semantic interrelationships with each other.

In addition, we experimented with a Support Vector Machine (SVM) with a linear kernel. SVMs can be described as max-margin classifiers that try to maximize a *margin*. When learning a linear decision boundary, maximizing the margin intuitively means searching for a decision boundary that maximizes the distance to those datapoints that are closest to the boundary. Adding counterexamples in a separable case can be compared to enforcing an orthogonality constraint during learning. In that case, counterexamples amount to additional max-margin constraints [28] that can help to obtain a better model (please refer to Figure 4a). For this reason, we chose to additionally include an SVM as the base learner for cases in which model-agnostic and local corrections via counterexamples develop their potential in an inherently interpretable way.

For instantiating the Active Learner, we chose the modAL python framework [29]. As the query strategy, we used Maximum Classification Uncertainty. For both datasets, a stratified split into training, pool, and test sets was performed (training 1%, pool 79%, and test 20% of the data). We therefore accounted for a standard Active Learning setting where a small number of labeled data and a huge amount of unlabeled data were available. All experiments were performed over 200 iterations each.

4.4. Evaluation Metrics

To evaluate the quality of our framework and answer Research Questions 2 and 3 (see Section 1), we performed two kinds of experiments. First, we measured the **Predictive Performance** of the different IML strategies with regard to a downstream classification task on the testset during 200 iterations. As performance metrics for evaluation of Research Question 2, we chose the macro-averaged F1 score (after each AL iteration) and the Average Classification Margin between the predicted and true class (after every tenth AL iteration). The Average Classification Margin between predicted and true class is defined as $M(x) = \frac{1}{N}\sum_{i=1}^{N} P(\hat{y}|x_i) - P(y|x_i)$, where $\hat{y}$ is the predicted class, y is the true class, $\hat{x_i}$ is a certain instance of the testset to be predicted, and N is the total number of instances in the testset. Accordingly, this measure analyzes the classifier's confidence towards false predictions for all test instances and then finds the average over them.

To answer Research Question 3, **Local Explanation Quality** was analyzed in two ways: (a) with regard to local fidelity and approximation accuracy (the quality of the local explanation generators itself before any interactions), and (b) with regard to the 'Explanation Ground Truth' of the downstream classification tasks (the quality of local explanations for all test instances compared to the bag-of-words represented Gold Standard described in Section 4.3).

Local fidelity is said to be achieved if an explanation model $g \, \epsilon \, G$ is found such that $f(z) \approx g(z')$ for $z, z' \, \epsilon \, Z$, where Z constitutes the vicinity of x and f is the model to be explained. Here, we use the Mean Local Approximation Error (MLAE, Equation (3)) and MeanR^2 (Equation (4)) as a proxy to measure the local fidelity of the whole explanation models to be compared.

$$MLAE = \frac{\sum_{i=1}^{N} |f(x_i) - g_i(x_i)|}{N}. \tag{3}$$

$$MeanR^2 = \frac{\sum_{i=1}^{N} R^2(g_i)}{N}, R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} (f(z_i) - g(z_i'))^2}{\frac{1}{n}\sum_{i=1}^{n} (f(z_i) - f_{mean})^2}. \tag{4}$$

In both cases, N is the number of instances in the associated test dataset.

Furthermore, we analyzed a modified variant of the *Area Over The Perturbation Curve* (AOPC), which measures the local fidelity of individual explanations. We call this the Combined Removal Impact (CRI), defined as follows:

$$CRI = \frac{1}{N}\sum_{i=1}^{N} p(\hat{y}|x_i) - p(\hat{y}|\tilde{x}_i^{(k)}), \tag{5}$$

where the top k% explanation features are removed from x_i to yield $\tilde{x}_i^{(k)}$, $\hat{y}$ denotes the predicted label for x_i, and N is the number of instances in the associated test dataset. For both evaluation metrics, please refer to [16] for details on how these metrics have been applied to compare word-based and topic-based contextual explanations.

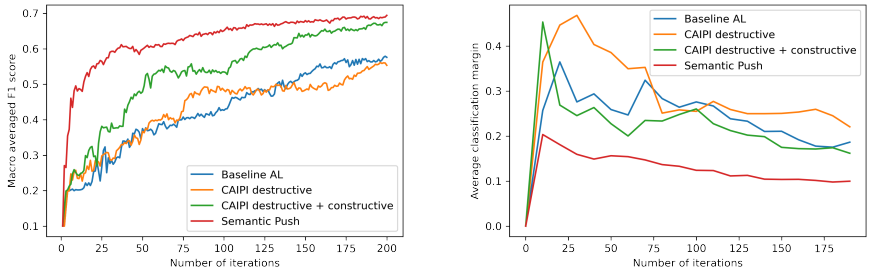
In order to analyze the development of Local Explanation Quality after applying the different IML strategies, we calculated a measure called 'Explanatory Accuracy'. First, we took $k = 10\%$ of the most relevant words from the global Gold Standard $GS_{global}^{(k)}(y)$ and intersected them with words ($GS_{global}^{(k)}(y) \cap x$) from a document x, resulting in a local Gold Standard ($GS_{local}(x)$) per document x. For each test document x, a local explanation $\epsilon(x)$ was subsequently generated using LIME. The Average Explanatory Accuracy was then defined as

$$ExplanatoryAccuracy_{AVG} = \frac{1}{N}\sum_{i=1}^{N} \frac{|GS_{local}(x_i) \cap \epsilon(x_i)|}{|GS_{local}(x_i)|}, \tag{6}$$

with N being the number of documents in the test dataset. We restricted the complexity of the local surrogate models (number of explanatory words) to $\Omega(g) = |GS_{local}(x)|$, such that the LIME explanations were theoretically capable of finding all relevant explanations according to local GS. We measured the Average Explanatory Accuracy of the test instances after every 20th iteration.

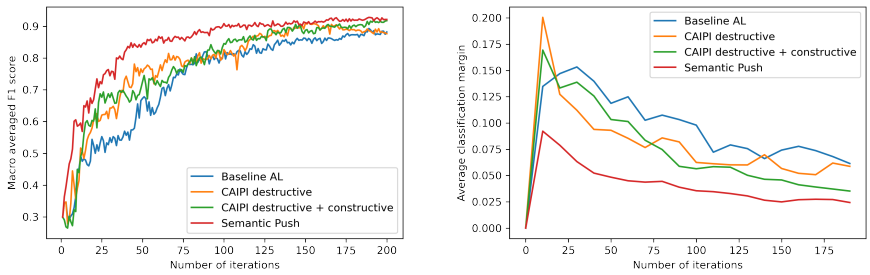
5. Experiment 1: Predictive Performance

We conducted the first experiment by measuring the **Predictive Performance** of the different IML strategies. Figures 5a–7a show the convergence of the macro-averaged F1 score on the two testsets over 200 iterations for our SemanticPush approach along with its baselines. For both datasets, SemanticPush clearly outperforms the standard ActiveLearner and the two versions of CAIPI when using XGBoost as the base learner, despite a Gold Standard that is around ten percent worse than that used for CAIPI. In the early stages of interaction, this holds true for the SVM base learner as well.



(**a**) (**b**)

Figure 5. (**a**) Learning performance of different IML strategies for *AG News* Dataset (XGBoost as base learner). (**b**) Average Classification Margin of different IML strategies for *AG News* Dataset (XGBoost as base learner).

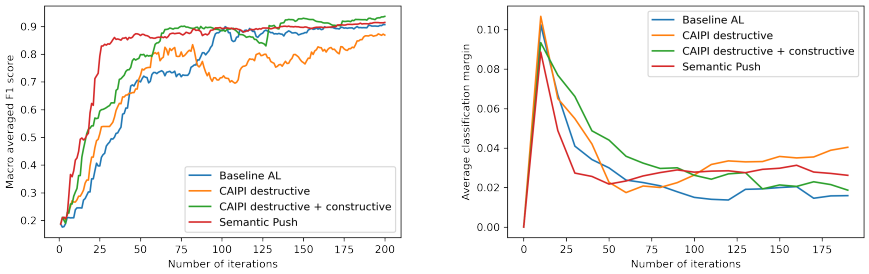


(**a**) (**b**)

Figure 6. (**a**) Learning performance of different IML strategies for Reuters R10 Dataset (XGBoost as base learner). (**b**) Average Classification Margin of different IML strategies for Reuters R10 Dataset (XGBoost as base learner).

More generally, SemanticPush shows high data efficiency with respect to queries from the pool dataset, as it incorporates the oracle's expert knowledge efficiently at a much earlier stage (around 90 percent of final F1 score reached already after only 50 iterations). In the middle range of the iterations, SemanticPush has already applied much of the correct knowledge; therefore, its performance starts to increase more slowly. For classifiers such as the Support Vector Machine, which reach high classification accuracy earlier (in the realm of the conceptual Gold Standard's performance), the performance of

SemanticPush begins to stagnate during later iterations, as it partially has applied 'incorrect corrections'. *CAIPI destructive* is not able to consistently beat the ActiveLearner's baseline, while our constructive extension performs better. Figures 5b–7b confirm the above observations from the point of view of the Average Classification Margin between predicted and true class, where SemanticPush on average provides false predictions less frequently and/or with less confidence than its baselines.



(**a**) (**b**)

Figure 7. (**a**): Learning performance of different IML strategies for Reuters R10 Dataset (SVM as base learner). (**b**): Average Classification Margin of different IML strategies for Reuters R10 Dataset (SVM as base learner).

Across all experiments, we kept the hyperparameters constant. At each iteration, we allowed the different methods to generate $N = 10$ counterexamples incorporating the corrective knowledge. Furthermore, we set the length (number of words) of each counterexample to the average document length of the respective corpora (25 for the *AG News* Dataset and 60 for *Reuters R10*). We allowed LIME to generate explanations containing 7 words (for *AG News*) and 15 words (for *Reuters R10).* The topicLIME explanations included three and five topics, respectively. This limitation was enforced due to the fact that in the real world humans are only able to perceive, process, and remember a limited number of pieces of information. According to Miller's law [30], this capacity is somewhere between seven plus or minus two. Additionally, we set λ (from Algorithm 2) to 0.95 to simulate the effect of global expert knowledge.

Experiment 2: Local Explanation Quality

We performed experiments by analyzing the **Local Explanation Quality** of the different IML strategies in both directions of interaction with the oracle. Table 2 compares the quality of the local surrogate models and the resulting explanations generated by LIME and topicLIME. The related measures are the Approximation Error, $MeanR^2$, and Combined Removal Impact (CRI) of the two different test datasets. It is noticeable that both the surrogate explanation models and the local explanations itself are more faithful towards the model to be explained when using contextual explanations generated from realistic local perturbation distributions. Therefore, the resulting explanations are regarded as more reliable.

Tables 3 and 4 take up the topic of Local Explanation Quality from the other direction (after the interactions with the oracle).

It is striking that only SemanticPush is capable of clearly transferring the expert knowledge in a way that it is adequately adopted by the base learner. The two versions of CAIPI do not reveal better results than the standard ActiveLearner.

To sum up, our proposed approach improves Learning Performance, especially in the early stages of interactions, pushing the reasoning of the learner towards the desired behavior.

Table 2. Comparison of LIME and topicLIME with respect to local fidelity (with XGBoost as the base learner).

	AG News		
	Lime	**TopicLIME**	**Difference**
Approx. Error	0.0394	0.0342	**−13%**
R^2	0.863	0.884	**+2.5%**
CRI	0.229	0.277	**+21%**
	Reuters R52		
Approx. Error	0.0195	0.0076	**−61%**
R^2	0.864	0.951	**+10%**
CRI	0.271	0.302	**+11%**

Table 3. Local explanation quality with respect to 'Ground Truth' of downstream classification tasks (with XGBoost as the base learner).

	Explanatory Accuracy$_{AVG}$			
	AL	**CAIPI**$_d$	**CAIPI**$_{d/c}$	**Sem.Push**
AG News	0.690	0.683	0.685	**0.711**
Reuters R10	0.741	0.739	0.742	**0.768**

Table 4. Local explanation quality with respect to 'Ground Truth' of downstream classification task (with SVM as the base learner).

	Explanatory Accuracy$_{AVG}$			
	AL	**CAIPI**$_d$	**CAIPI**$_{d/c}$	**Sem.Push**
Reuters R10	0.786	0.785	0.788	**0.796**

6. Subsumption and Discussion

As social beings, humans engage in interactions, often attempting to communicate an understanding between individuals. Therefore, humans are naturally driven to acquire and provide explanations as well as to receive explanations in order to expand their understanding [31]. As human explanations are often framed by stances or modes of construal, and are therefore interpretative and diverse in nature, humans need to perform mental calculations in order to understand such explanations [32]. Often, the human capability to flexibly use contextual and background information as well as intuition and feeling are consulted in order to distinguish 'brilliant' and 'real' intelligence [33] from Artificial Intelligence, as computers generally are deemed 'stupid' with regard to such tasks.

Therefore, we developed SemanticPush in order to account for the inclusion of contextual and background knowledge during interactions between humans and ML systems. We illuminate the topic of semantic interactivity from both directions: from machines to humans by enabling ML explanations to be coherent, semantically meaningful, and locally faithful, and from the other direction by enabling humans to include expert knowledge in a conceptual manner.

As a result, SemanticPush differs from state of the art approaches such as CAIPI in (a) using contextual topicLIME explanations instead of LIME explanations, (b) internally using a conceptually meaningful Gold Standard that allows corrections on higher semantic detail, (c) additionally enabling constructive feedback, and (d) being able to locally correct the reasoning used to arrive at false predictions. Transferred to a real-world interaction setting, human annotators are capable of indicating and correcting (a) components that a learner wrongly identified as relevant (as CAIPI does), (b) components that the learner has forgotten to learn, and (c) relevant components that have been incorrectly used.

In a practical text classification scenario, humans could teach a learner by generating documents that exhibit a specific semantic content and structure together with a target class. As an example, a human domain expert could analyze the reasoning of a learner by locally harnessing the contextual topicLIME explanations for a document of interest. In the next step, the expert could gradually manipulate the document's concept composition by analogy to his or her conceptual knowledge. Hence, the expert would be able to underweight, overweight, remove, or add higher-level concepts of the according input domain via manipulation (decreasing, increasing, removing, or adding) of individual topic attributions (see Figure 4b). As a result of this interaction it is possible to maintain statistical characteristics of the input domain, leading to non-extrapolation of training examples comprising the annotator's corrections, and thereby forcing the classifier's reasoning to converge to the desired behavior.

7. Summary and Conclusions

In this paper, we introduced a novel IML architecture called Semantic Interactive Learning that helps to bring humans into the loop and allows for richer interactions. We instantiated it with SemanticPush, the first IML strategy enabling semantic and constructive corrections of a learner, also for completely false predictions. Our approach offers locally faithful and contextual explanations; on this basis, it qualifies humans to provide conceptual corrections that can be considered as continuous. The corrections are in turn integrated into the learner's reasoning via non-extrapolating and contextual additional training instances. As a consequence of combining richer explanations with more extensive semantic corrections, our proposed interaction paradigm outperforms its baselines with regard to learning performance as well as local explanation quality of downstream classification tasks in the majority of our experiments. Please note that our constructive extension for CAIPI outperforms original CAIPI as well in most experiments w.r.t. Learning Performance.

In addition to all the listed benefits, there are two main prerequisites for our approach that should be mentioned. First, as its entire semantic functionality is based on an LDA approach, a certain level of expertise in topic modeling is required; for instance, in order to implement suitable data preprocessing or to find an adequate number of topics k. Therefore, additional analysis is necessary upfront. However, if helpful semantic concepts have been identified, then fewer interactions with the oracle might be required. This allows model developers to trade off more potentially costly interactions with an oracle against the cost of extra data preprocessing and topic modeling. Second, as for all interactive scenarios, efficient access to an oracle is needed, be it simulated or based on human annotators.

Therefore, this work can provide new perspectives for further studies. For our experiments, where the simulation of expert knowledge via a global Gold Standard is a crucial aspect, we plan to improve the simulation accuracy as well as to evaluate its quality using inter-rater reliability. Furthermore, we intend to conduct experiments with human experts. Additionally, we intend to include a language model such as BERT into our architecture to ensure that generated counterexamples are meaningful both semantically and linguistically, and especially that they are syntactically correct. Masked Language Modeling could be harnessed to check for linguistically sensible counterexamples, while Autoencoders could be used to identify 'Out-of-Distribution' counterexamples by analyzing the reconstruction error.

In summary, this work takes a step towards Human-Centered Machine Learning by allowing contextual interpretation and intervention in an interactive setting. Effective and efficient co-work between users and an ML learner is enabled, allowing the learner to take advantage of the richness of human expertise.

Author Contributions: Conceptualization, S.K., M.H. and U.S.; methodology, S.K., M.H. and U.S.; software, S.K. and M.H.; validation, S.K., M.H. and U.S.; formal analysis, S.K. and M.H.; investigation, S.K., M.H. and U.S.; resources, S.K., M.H. and U.S.; data curation, S.K. and M.H.; writing—original draft preparation, S.K. and M.H.; writing—review and editing, S.K., M.H. and U.S.; visualization,

S.K.; supervision, U.S.; project administration, S.K. and U.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Please refer to Section 4.2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.
2. Holzinger, A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inform.* **2016**, *3*, 119–131.
3. Holzinger, A.; Biemann, C.; Pattichis, C.; Kell, D.B. What do we need to build explainable AI systems for the medical domain? *arXiv* **2017**, arXiv:1712.09923.
4. Bruckert, S.; Finzel, B.; Schmid, U. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Front. Artif. Intell.* **2020**, *3*, 507973.
5. Akata, Z.; Balliet, D.; de Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* **2020**, *53*, 18–28. https://doi.org/10.1109/MC.2020.2996587.
6. Teso, S.; Kersting, K. Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27 January–1 February 2019.
7. Kulesza, T.; Burnett, M.; Wong, W.K.; Stumpf, S. Principles of explanatory debugging to personalize interactive machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces, Atlanta, GA, USA, 29 March–1 April 2015. ACM Press: New York, NY, USA, 2015. https://doi.org/10.1145/2678025.2701399.
8. Fails, J.A.; Olsen, D.R., Jr. Interactive machine learning. In Proceedings of the 8th International Conference on Intelligent User Interfaces, Miami, FL, USA, 12–15 January 2003; ACM: New York, NY, USA, 2003; Volume 3, pp. 39–45.
9. Gillies, M.; Fiebrink, R.; Tanaka, A.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; Amershi, S.; Lee, B.; et al. Human-centered machine learning. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016.
10. Dudley, J.J.; Kristensson, P.O. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* **2018**, *8*, 1–37 https://doi.org/10.1145/3185517.
11. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept bottleneck models. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5338–5348.
12. Schmid, U.; Finzel, B. Mutual Explanations for Cooperative Decision Making in Medicine. *KI KüNstliche Intell.* **2020**, *34*, 227–233. https://doi.org/10.1007/s13218-020-00633-2.
13. Amershi, S.; Cakmak, M.; Knox, W.B.; Kulesza, T. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* **2016**, *35*, 105–120
14. Zaidan, O.; Eisner, J.; Piatko, C. Using "annotator rationales" to improve machine learning for text categorization. In Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, USA, 22–27 April 2007; pp. 260–267.
15. Holzinger, A.; Malle, B.; Saranti, A.; Pfeifer, B. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf. Fusion* **2021**, *71*, 28–37. https://doi.org/10.1016/j.inffus.2021.01.008.
16. Kiefer, S. CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge. *Inf. Fusion* **2022**, *77*, 184–195. https://doi.org/https://doi.org/10.1016/j.inffus.2021.07.014.
17. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Independently Published. 2019. ISBN13: 978-0244768522.
18. Blei, D.M.; NG, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
19. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 31 January–3 February 2015; pp. 399–408.
20. Syed, S.; Spruit, M. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 19–21 October 2017; pp. 165–174.
21. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you? In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
22. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; pp. 4768–4777.

23. Odom, P.; Natarajan, S. Human-Guided Learning for Probabilistic Logic Models. *Front. Robot. AI* **2018**, *5*, 56. https://doi.org/10.3389/frobt.2018.00056.
24. Stumpf, S.; Rajaram, V.; Li, L.; Burnett, M.; Dietterich, T.; Sullivan, E.; Drummond, R.; Herlocker, J. Toward harnessing user feedback for machine learning. In Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI, Honolulu, HI, USA, 28–31 January 2007; pp. 82–91. https://doi.org/10.1145/1216295.1216316.
25. Zhang, X.; Zhao, J.; Le Cun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, MIT Press: Cambridge, MA, USA; 2015, Volume 28.
26. Lewis, D. REUTERS-21578. 1993. Available online: https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection (accessed on 5 September 2022).
27. Goyal, K.; Dumancic, S.; Blockeel, H. Feature Interactions in XGBoost. *arXiv* **2020**, arXiv:2007.05758v1.
28. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
29. Danka, T. modAL: A modular active learning framework for Python. *arXiv* **2018**, arXiv:1805.00979v2.
30. Miller, G.A. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63 2*, 81–97.
31. Keil, F.C. Explanation and Understanding. *Annu. Rev. Psychol.* **2006**, 57, 227–254.
32. Dennett, D. *The Intentional Stance*; MIT Press: Cambridge, MA, USA, 1987.
33. Bergstein, B. *AI Isn't Very Smart Yet. But We Need to Get Moving to Make Sure Automation Works for More People*; MIT Technology: Cambridge, MA, USA, 2017.

B. Additional Results

B.1 Empirical Evaluation of Topic-based Explanations

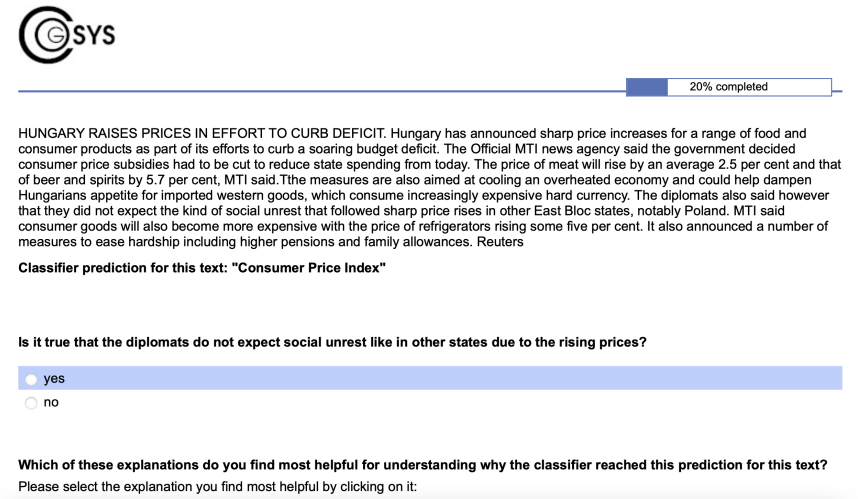


Figure B.1: Preference selection task - example 1: this figure shows a text document, the according predicted class, an attention check and a why-question for preference.

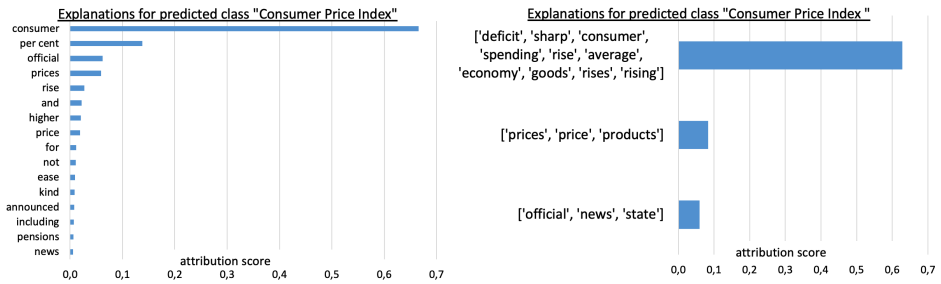


Figure B.2: Preference selection task - example 1: participants can indicate their preference - either standard word-based LIME explanation (left) or concept-based topicLIME explanation (right). The positioning of the two modalities (left versus right) has been randomized during the experiment.

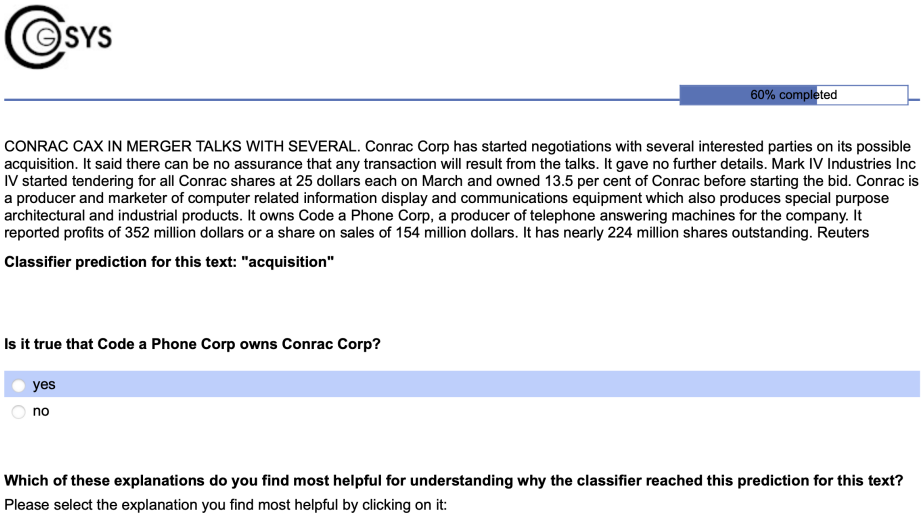


Figure B.3: Preference selection task - example 2: this figure shows a text document, the according predicted class, an attention check and a why-question for preference.

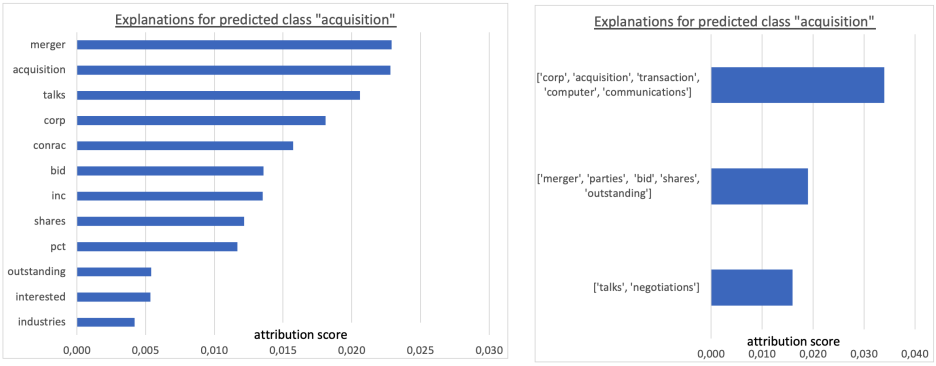


Figure B.4: Preference selection task - example 2: participants can indicate their preference - either standard word-based LIME explanation (left) or concept-based topicLIME explanation (right). The positioning of the two modalities (left versus right) has been randomized during the experiment.

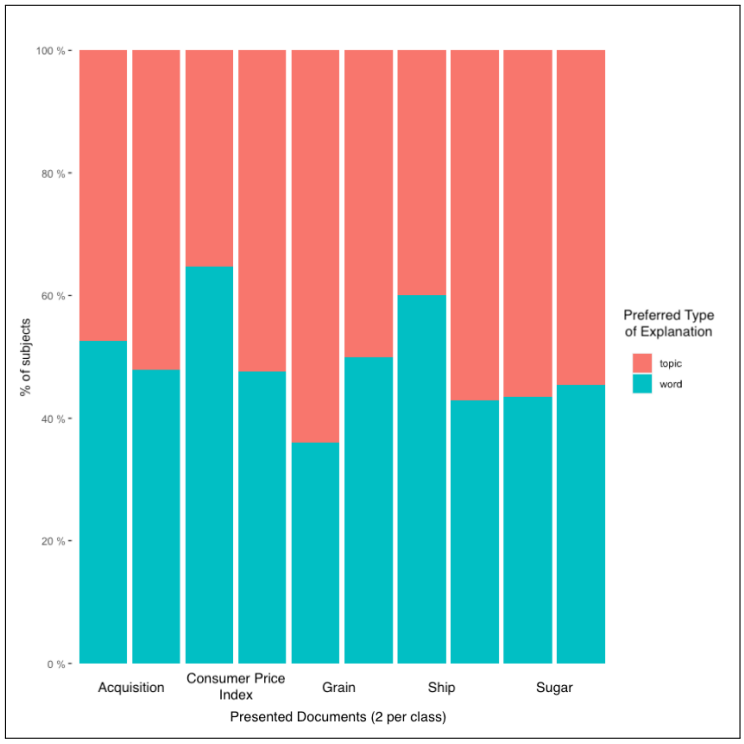


Figure B.5: Descriptive analysis of the answers provided during the preference selection task: per presented document (two documents per class depicted on the x-axis), it is indicated on the y-axis, in percent, how many participants preferred topic-based and word-based explanations. A general preference cannot be identified.

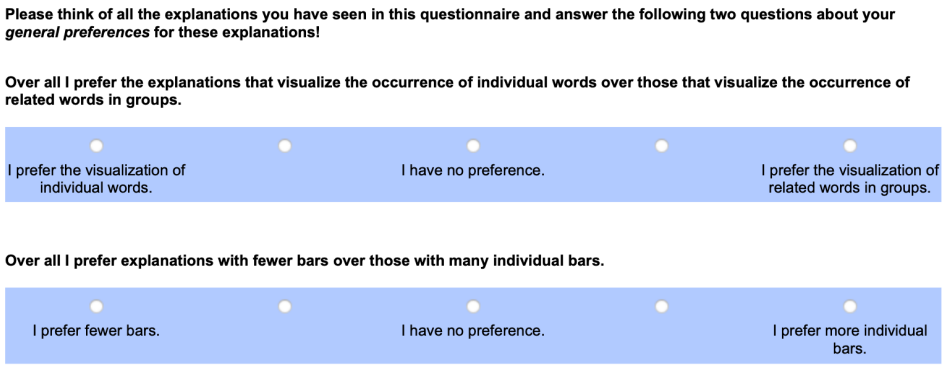


Figure B.6: General preference selection task: participants can indicate their general preference - whether they prefer independent words or groups of related words as explanations and whether they prefer fewer or more bars in the visualization of the explanations.

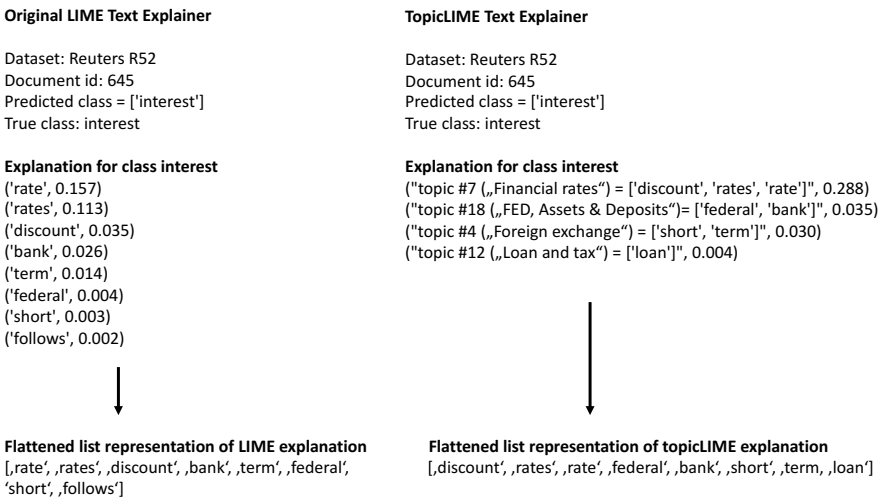


Figure B.7: Forward prediction task: to compare LIME and topicLIME explanations with regard to the included explanatory features - independently of the way of presenting the explanations - the explanatory features are flattened into a list representation.

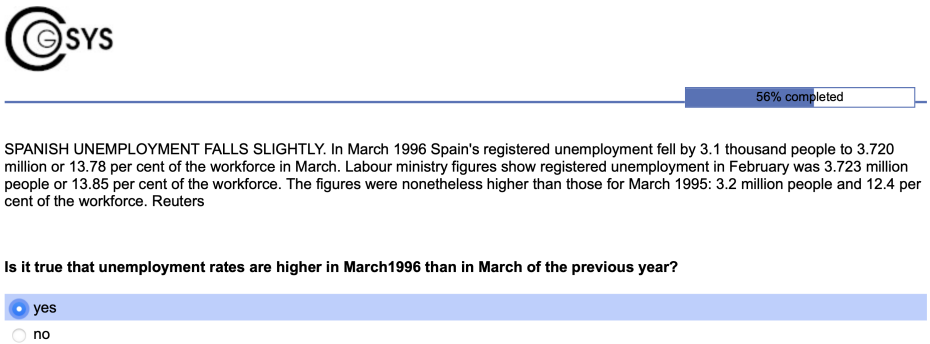


Figure B.8: Forward prediction task - example 1: this figure shows a text document and an attention check.

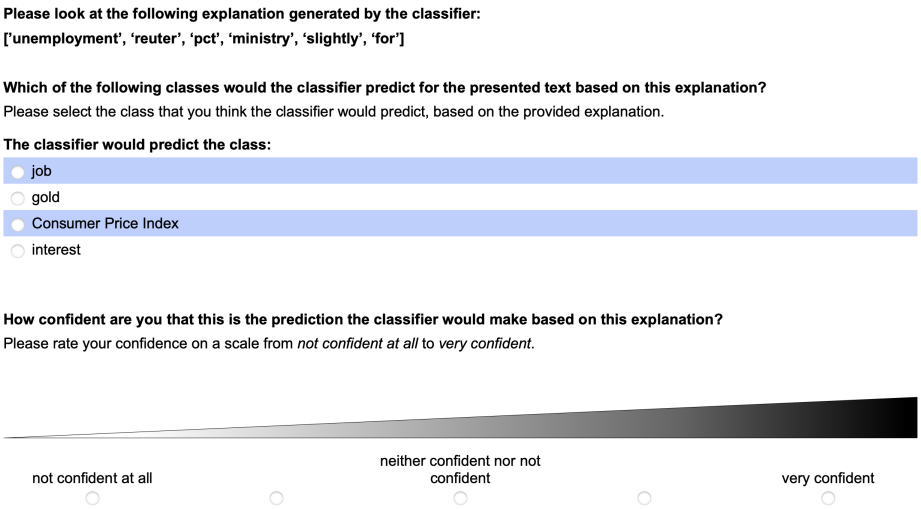


Figure B.9: Forward prediction task - example 1: participants can select the class they think the classifier will predict and indicate their confidence towards their selection - based on a flattened explanation generated by standard LIME.

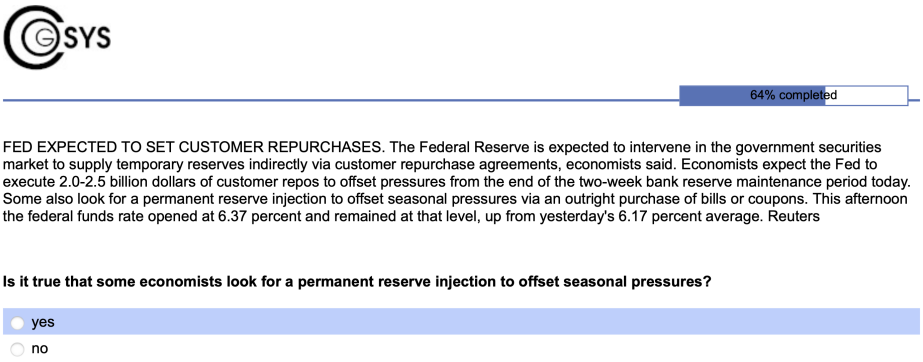


Figure B.10: Forward prediction task - example 2: this figure shows a text document and an attention check.

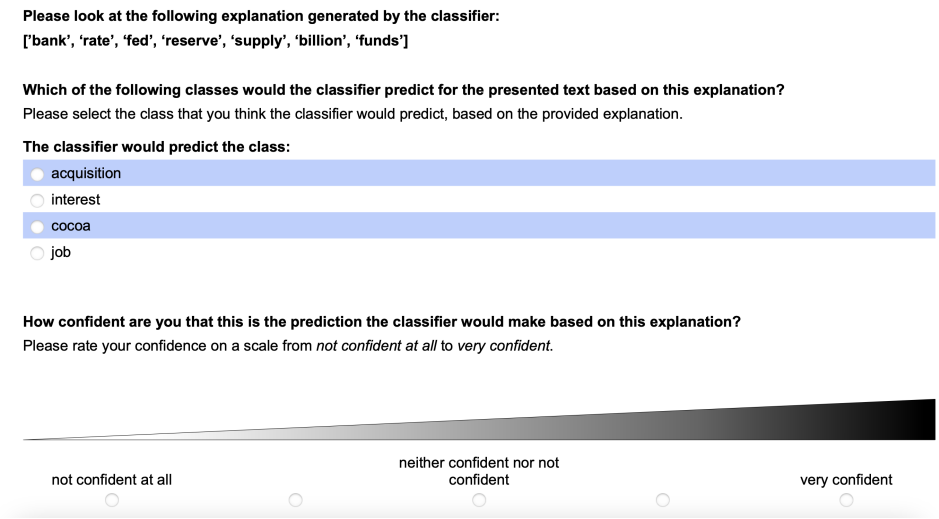


Figure B.11: Forward prediction task - example 2: participants can select the class they think the classifier will predict and indicate their confidence towards their selection - based on a flattened explanation generated by topicLIME.

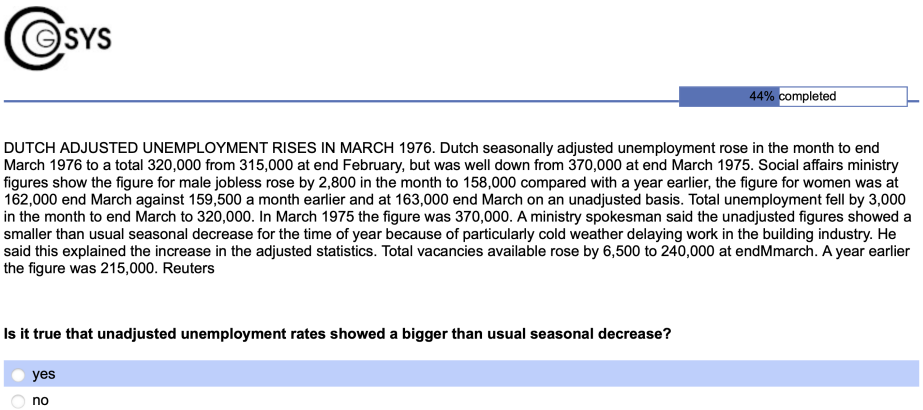


Figure B.12: Forward prediction task - example 3: this figure shows a text document and an attention check.

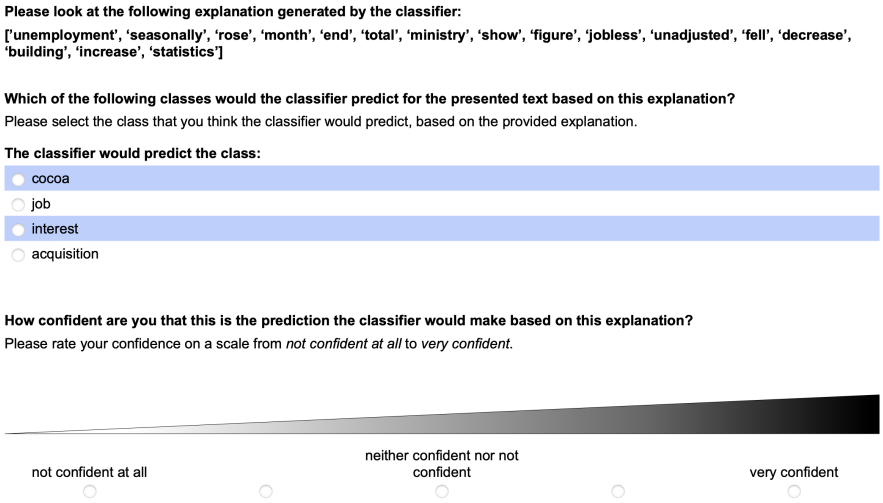


Figure B.13: Forward prediction task - example 3: participants can select the class they think the classifier will predict and indicate their confidence towards their selection - based on a flattened explanation generated by topicLIME.

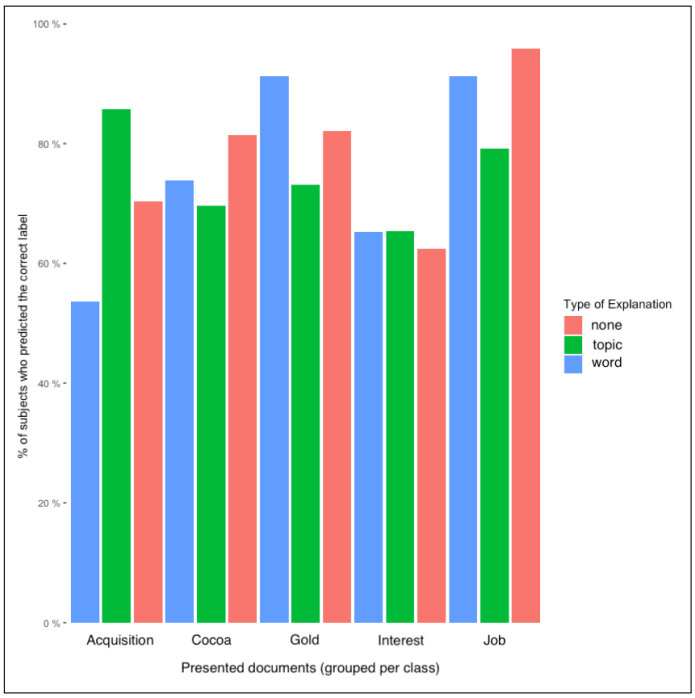


Figure B.14: Descriptive analysis of the answers provided during the forward prediction task with regard to task performance: per presented documents grouped per class on the x-axis, it is indicated on the y-axis, in percent, how many participants predicted the correct label based on the different types of explanations.

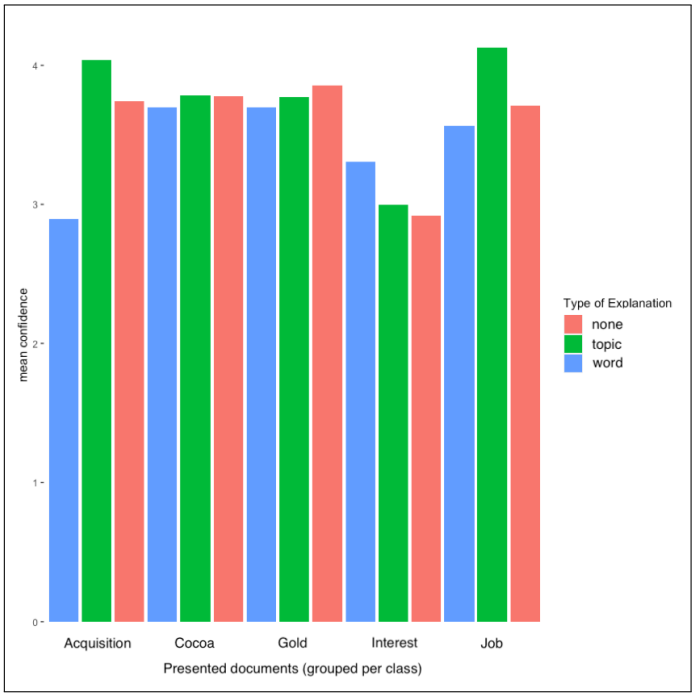


Figure B.15: Descriptive analysis of the answers provided during the forward prediction task with regard to confidence: per presented documents grouped per class on the x-axis, it is indicated on the y-axis, using a 5-point Likert scale, how confident the participants were on average based on the different types of explanations.