Electronic Markets
https://doi.org/10.1007/s12525-022-00612-5

RESEARCH PAPER

Global reconstruction of language models with linguistic rules – Explainable AI for online consumer reviews

Markus Binder[1] · **Bernd Heinrich**[1] · **Marcus Hopf**[1] · **Alexander Schiller**[1]

Received: 30 May 2022 / Accepted: 27 October 2022

Abstract

Analyzing textual data by means of AI models has been recognized as highly relevant in information systems research and practice, since a vast amount of data on eCommerce platforms, review portals or social media is given in textual form. Here, language models such as BERT, which are deep learning AI models, constitute a breakthrough and achieve leading-edge results in many applications of text analytics such as sentiment analysis in online consumer reviews. However, these language models are "black boxes": It is unclear how they arrive at their predictions. Yet, applications of language models, for instance, in eCommerce require checks and justifications by means of global reconstruction of their predictions, since the decisions based thereon can have large impacts or are even mandatory due to regulations such as the GDPR. To this end, we propose a novel XAI approach for global reconstructions of language model predictions for token-level classifications (e.g., aspect term detection) by means of linguistic rules based on NLP building blocks (e.g., part-of-speech). The approach is analyzed on different datasets of online consumer reviews and NLP tasks. Since our approach allows for different setups, we further are the first to analyze the trade-off between comprehensibility and fidelity of global reconstructions of language model predictions. With respect to this trade-off, we find that our approach indeed allows for balanced setups for global reconstructions of BERT's predictions. Thus, our approach paves the way for a thorough understanding of language model predictions in text analytics. In practice, our approach can assist businesses in their decision-making and supports compliance with regulatory requirements.

Keywords Explainable AI · Text analytics · Language models · BERT · Linguistic rules · Online consumer reviews

JEL Classification C80

Introduction

Huge amounts of unstructured textual data are generated across various channels of information systems (IS) such as eCommerce platforms, review portals or social media every second (Potnis, 2018). Consequently, the need for techniques that automatically analyze textual data is increasing: Until 2028, the revenues from the natural language processing (NLP) market worldwide are expected to increase at a compound annual growth rate of almost 30% to over 100 billion USD, with text analytics expected to have the highest growth (Fortune Business Insights, 2021). As text analytics facilitate diverse applications such as sentiment analysis or text summarization (Young et al., 2018), various organizations in different business areas benefit from techniques of text analytics (Coheur, 2020; Zhang et al., 2020). For instance, product or service providers can use such techniques to analyze consumer sentiments in large amounts of online consumer reviews. Using this consumer feedback enables organizations

Responsible Editor: Fethi Abderrahmane Rabhi.

This article is part of the Topical Collection on Explainable and responsible artificial intelligence

✉ Bernd Heinrich
bernd.heinrich@ur.de

Markus Binder
Markus1.Binder@ur.de

Marcus Hopf
Marcus.Hopf@ur.de

Alexander Schiller
Alexander.Schiller@ur.de

[1] University of Regensburg, Germany, at the Faculty of Informatics and Data Science, Regensburg, Germany

Published online: 13 December 2022

to effectively improve their products and services (Chatterjee, 2019; Heinrich et al., 2022; Heinrich et al., 2020).

The state-of-the-art techniques of text analytics are language models, such as the popular deep learning AI model 'Bidirectional Encoder Representations from Transformers' (BERT) (Devlin et al., 2019) or its descendants (e.g., ALBERT; Lan et al., 2020), as they have achieved leading-edge results in many tasks such as aspect-based sentiment analysis (Wang et al., 2018). Language models enable a contextualized representation of textual data by assessing the conditional probability of each token (e.g., a word) given the contextual tokens surrounding it (Peters et al., 2018a). Besides coarser classification tasks for sentences, for example, these language model representations can then be used, in particular, as basis for central token-level classifications such as aspect term and sentiment term detection. Since the language model BERT is already incorporated in a plethora of business IS applications, we demonstrate our approach by means of BERT as leading exponent of language models in this paper. Amongst others, popular application scenarios of BERT in electronic markets are eCommerce, chatbots, finance or online recruiting (Coheur, 2020; Dastin, 2018; Luo et al., 2022; Repke & Krestel, 2021; Shrestha et al., 2021; S. Xu et al., 2020; Yang et al., 2020; Zhang et al., 2020). However, similar to most other state-of-the-art deep learning models, BERT is a "black box". That is, over 100 million learned parameters (Devlin et al., 2019) and various hidden layers contribute to BERT's immense complexity, making it hardly (if at all) possible to comprehend why and how BERT arrives at its predictions (Kovaleva et al., 2019). To address this black box nature of AI models, a vastly increasing focus on explainable AI (XAI) in IS research and practice has emerged (Adadi & Berrada, 2018; Förster et al., 2021; Förster et al., 2020b). Literature agrees that the need for reconstructions and justifications is urgent and a 'huge open scientific challenge' (Guidotti et al., 2018). It is even expected that "algorithmic auditing and 'data protection by design' practices will likely become the new gold standard for enterprises deploying machine learning systems" (Casey et al., 2019). Thereby, regulations such as the General Data Protection Regulation (GDPR) in the European Union impose an extensive 'right to explanation' for automated data processing systems in general and thereby lay the foundation to enforce algorithmic auditing in companies. In particular, algorithmic auditing is highly relevant for domain experts, managers and data scientists that utilize the language models' predictions for business-critical decisions or implementations and need to justify their actions. This is especially the case for application scenarios (AS) in electronic markets, as exemplarily outlined in the following and captured later on:

- eCommerce (AS1): In eCommerce, BERT is used to conduct token-level classification in the course of sentiment analyses of online consumer reviews on online platforms such as Airbnb, Yelp or TripAdvisor for product development, services offerings and forecasting future demand (Heidari & Rafatirad, 2020; Shrestha et al., 2021; S. Xu et al., 2020). Since these analyses and decisions have large impacts, they require additional validation checks and justifications, far beyond measuring only the prediction accuracy of BERT. For instance, it needs to be ensured that specific groups of consumers are not discriminated against by assigning a negative sentiment to certain countries, ethnicities or genders.
- Chatbots (AS2): In applications in consumer services (Luo et al., 2022), BERT-based chatbots conduct direct consumer interaction and embody the company's voice. Thereby, reconstructions and justifications regarding the underlying BERT model are mandatory to prevent unhelpful, rude or misleading dialogues and thus, to support consumer satisfaction.
- Financial applications (AS3): BERT descendants such as FinBERT (Yang et al., 2020) enable token-level classifications of financial entities, sentiments and their relations from texts such as social media posts (e.g., tweets from CEOs or other experts) or contract documents. The extracted information is used for key tasks in finance such as accounting, auditing, compliance and risk assessment. Furthermore, language models enable to automatically process millions of documents as contained in data leaks such as the Panama Papers (O'Donovan et al., 2019) for tax fraud detection. In particular, if legal actions are initiated based on predictions from language models (e.g., tax prosecution based on data leaks), validation checks are mandatory.
- Online recruiting (AS4): Supporting text analytics of application documents (Schiller, 2019), language models such as BERT enable pre-processing und pre-filtering of applications and candidates on online job platforms. Here, auditing and validation are required as such automated recruitment may lead to discrimination (e.g., by gender or origin; Dastin, 2018). Reconstructions of models help to avoid such discriminations.

These application scenarios show that it is crucial to reconstruct BERT's predictions to be able to justify the decisions based thereon. Here, the reconstructions and explanations in these scenarios are required on a global level as in all those application scenarios the predictions of language models are used in ongoing operations on a daily basis. This means that a vast number of decisions are made based on these predictions day-by-day for newly generated and hitherto unknown textual data (e.g., chatbots or review summarizations are applied in real-time on consumer texts). Therefore, it is not feasible to use local approaches for reconstruction, as this would require huge efforts for manual checks of each local reconstruction and could practically only be

done a-posteriori if at all. Therefore, global approaches are essential for reconstructions of language model predictions in many applications. Here, we focus on global reconstructions of BERT's predictions for token-level classifications in this work, since this constitutes popular application scenarios of BERT (e.g., AS1, AS3) and since BERT also establishes text representations based on tokens. Moreover, as Zafar et al. (2021) and Yan et al. (2022) indicate, a reconstruction approach for token-level classifications can also serve as a basis for reconstructions of coarser classification tasks, for instance, for sentence-level classifications (e.g., AS2, AS4).

A promising way to obtain such a reconstruction and thus justify BERT's predictions is to conduct a rule-based XAI approach. On the one hand, rules are highly concrete, which also has been emphasized by Förster et al. (2020a) as decisive XAI characteristic. Indeed, studies have shown that users "prefer, trust and understand rules better than alternatives" (Ribeiro et al., 2018; cf. also Arrieta et al., 2020). On the other hand, rule-based approaches preserve the AI model itself and thus, its high performance, while offering post-hoc reconstructions for explanations (Adadi and Berrada, 2018). Here, local rule-based approaches focus on explaining each prediction for a specific input separately, for instance, by using specific words to predict the sentiment term in a single sentence of an online consumer review. In contrast, global approaches aim at reconstructing the model's predictions as a whole (Danilevsky et al., 2020). A global approach ideally requires a smaller rule set for reconstructing multiple predictions of a language model compared to local approaches that establish a separate and highly specific rule for each individual prediction and therefore are not really generalizable (Danilevsky et al., 2020).

To enable such a global approach, our idea is to build rules based on linguistic information (so-called linguistic rules) which generalize specific words and sentences and can be modeled by NLP building blocks such as part-of-speech tags or dependency relations (Qi et al., 2020). Using NLP building blocks instead of single words as rule arguments is promising for global reconstruction, as they allow for rule arguments and rules analyzing (much) more than, for instance, one single sentence in an online consumer review. Moreover, NLP relation building blocks allow to account for the contextual information in a sentence (i.e., relations between words), which is crucial for the reconstruction of language model predictions for token-level classifications, since language models also use contextual information. Thus, we focus on the following main research question:

> **RQ1:** How can language model predictions for token-level classifications be globally reconstructed by means of an XAI approach based on linguistic rules?

Analogous to local reconstructions, a global reconstruction has to be analyzed regarding its fidelity (Danilevsky et al., 2020; Gilpin et al., 2018) and comprehensibility (Guidotti et al., 2018). In case of rule-based approaches, the comprehensibility of the rule set depends on the complexity (with respect to the length of the rules; cf. Guidotti et al., 2018) and the generalizability (words vs. NLP building blocks as discussed above) of the rules. Thereby, our approach allows for different setups regarding the comprehensibility of the rule set (e.g., by varying rule length), which is in general outlined as an important requirement of an XAI approach (Gilpin et al., 2018). This enables to analyze the trade-off between these two objectives in a reconstruction, which further supports adoption in IS. Thus, the second research question is as follows:

> **RQ2:** How can the trade-off between fidelity and comprehensibility of global reconstructions of language model predictions by linguistic rules be analyzed?

Hence, our contribution is twofold: (1) We are the first to propose a global XAI approach for reconstructing predictions of language models by linguistic rules. In particular, (2) this paper is thus the first to analyze the trade-off between fidelity and comprehensibility (i.e., complexity and generalizability) in this setting.

For our analysis, we focus on the highly relevant tasks of aspect term detection and sentiment term detection in online consumer reviews. To that end, we use two recognized online consumer review datasets from the domains of laptops and restaurants to account for different types of goods (i.e., laptops as *search* goods and restaurants as *experience* goods). We find that our linguistic rules are indeed suited for a global reconstruction of BERT's predictions in online consumer reviews and in particular allow for balanced setups with respect to the trade-off between comprehensibility and fidelity of the reconstruction.

The remainder of this paper is structured as follows. The next section presents the background of our research. Subsequently, we discuss how to globally reconstruct language models such as BERT with linguistic rules. Thereafter, we analyze different global reconstructions of BERT, discuss their results and outline implications for research and practice. Finally, we summarize the paper and provide an outlook on future research directions.

Background

In this section, we first outline which different types of XAI approaches exist in the context of language models. Second, several NLP building blocks recognized by literature are introduced forming the basis for our approach. The section concludes with a discussion of related work yielding the addressed research gap.

Types of XAI approaches in the context of language models

To clarify the notion of XAI (i.e., what explainable AI really means), a characterization in opaque systems, interpretable systems and comprehensible systems has been proposed (Doran et al., 2017). Here, *opaque systems* offer no insights into the system's reasoning on how inputs are mapped to the corresponding outputs. In that line, modern language models such as BERT are opaque systems, as it is not possible to comprehend its mappings, for instance, comprising over 100 million learned parameter values in the case of BERT. Based on that, there are two separate notions of addressing this problem. First, *interpretable systems* allow to understand how inputs are mapped to outputs by subdividing the mapping. This is not feasible for language models such as BERT due to its large amount of parameters and layers, which results in highly complex concatenated functions (Devlin et al., 2019). Second, *comprehensible systems* allow to relate *properties* of the inputs, for instance, single terms of an input sentence, to their output such as a classification of sentiment terms (Doran et al., 2017). While research in both areas is important, it has to be pointed out that the resulting XAI approaches are not "actually" explanation systems (Doran et al., 2017). For instance, rule-based approaches mostly give insights on how, but not why specific predictions are made (Doran et al., 2017). That is, causality cannot be directly established. To account for these different notions, we deliberately refer to "reconstructing" BERT rather than "explaining" in this paper.

Related to the two notions of interpretable and comprehensible systems, there are, in general, two main approaches in XAI (Adadi and Berrada 2018): On the one hand, *intrinsic XAI approaches* 'force' the AI model (during training) to produce interpretable mappings from input to output (Adadi and Berrada 2018). The drawback of these intrinsic approaches is that they are limited in the type of interpretations they can provide, as they need to restrict the model to obtain interpretable mappings, thus usually worsening the model's performance (Adadi and Berrada 2018). Due to its complexity, BERT would have to be extremely simplified to enable interpretable mappings. On the other hand, *post-hoc XAI approaches* aim to comprehensibly reconstruct the mappings from input to output of an AI model. These approaches do not require to restrict the model during training (Adadi and Berrada 2018). Here, a popular method is rule extraction, since rules can potentially exhibit a high degree of comprehensibility (Ribeiro et al., 2018). In general, there are two categories of rule extraction techniques (Adadi and Berrada 2018): 1) *Decompositional rule extraction* aims at extracting rules at selected, often single nodes within a neural network. To comprehend the predictions of a language model, it is then necessary to concatenate multiple extracted rules for various hidden layers. Thus, the drawback of this technique is that concatenations of rules are highly complex for deep neural networks such as BERT (Augasta & Kathirvalavakumar, 2012). Since the resulting rules would again be difficult to comprehend, decompositional rule extraction is not feasible for comprehensibly reconstructing language models. 2) In contrast, *pedagogical rule extraction* aims at extracting rules considering only the inputs and outputs. In particular, rules are extracted based on properties of the inputs and the corresponding outputs to reconstruct the mappings of the AI model. Thus, this approach can contribute to a comprehensible reconstruction even for language models such as BERT, since the extracted rules do not have to be concatenated through the various hidden layers.

Additionally, a further important differentiation within post-hoc XAI research is between *global* and *local approaches* (Danilevsky et al., 2020). Here, global approaches aim at reconstructing the predictions of an AI model by means of one single global model (Danilevsky et al., 2020). In contrast, local approaches create separate, highly specific reconstruction models for each prediction (e.g., in a single sentence of an online consumer review). To enable local reconstructions for IS text analytics applications, rules solely based on specific words are used by extant literature (e.g., Ribeiro et al., 2018). However, such rules lack the ability to generalize. In contrast, linguistic rules based on NLP building blocks are more promising for the global reconstruction of language models. Indeed, rule arguments with NLP building blocks generalize much better than rule arguments with specific words, and NLP relation building blocks enable to incorporate contextual information, which is a main component of language models.

Both objectives *fidelity* and *comprehensibility* are crucial for global post-hoc XAI approaches (Arrieta et al., 2020; Guidotti et al., 2018; Szczepański et al., 2021). Indeed, on the one hand, a global reconstruction needs to match the predictions of an AI model to avoid false conclusions, which is measured by fidelity (Gilpin et al., 2018). On the other hand, comprehensibility (i.e., complexity and generalizability; commonly measured in terms of model size) enables the use of the reconstruction (Guidotti et al., 2018). Thus, we analyze the reconstruction of BERT regarding its fidelity and its comprehensibility and strive to enable different setups between the two objectives.

NLP building blocks

To enable a reconstruction using linguistic rules, our idea is to use different semantical and syntactical NLP building blocks (cf. Introduction). Thus, we briefly outline NLP building blocks that are widely recognized in the literature (Fellbaum, 2013; Kamps et al., 2004; Tenney et al., 2019b) and that constitute a basis for our reconstruction. Table 1

summarizes these different building blocks. Thereby, the column 'type' characterizes a building block as *tag* or *relation* (as described in the following). In addition, the column 'linguistic information' shows whether a building block provides semantic or syntactic information. For each building block, an example is given in the last column.

A *tag building block* provides tag labels for selected tokens (e.g., words or punctuation marks) of a sentence. Tag labels describe a certain syntactic or semantic information of tokens in consideration of the whole sentence. Part-of-speech (POS) tags provide information on the *syntactic* structure of a sentence. Thereby, the POS tag, such as noun (NN), adjective (JJ) or verb (VB), is assigned to a single token. The building block synsets (SYN) considers the *semantic* information of tokens. In particular, SYN labels (e.g., derived from the lexical database WordNet) indicate words which share the same or a similar meaning (Fellbaum, 2013) taking into account its word context in a sentence.

A *relation building block* provides a label for a pair of tokens in a sentence describing a certain syntactic or semantic relation between these tokens. These relation building blocks enable to account for the contextual information in a sentence (i.e., the relation between tokens in a sentence), which is crucial for a reconstruction of BERT as BERT also considers contextual information. A basic *syntactic* information is the distance between two tokens, which is covered by the proximity (PROX) building block. For instance, if two tokens are next to each other in a sentence, their distance is 1. Dependencies (DEP) also link two tokens based on their *syntactical* relationship, such as the adjectival modifier (amod) or nominal subject (nsubj) dependencies (Manning et al., 2014). *Semantic* information is provided by the building blocks semantic role labeling (SRL) and coreference (COREF). SRL relations identify combinations of predicates and semantic arguments in a sentence (Tenney et al. 2019b). COREF links two tokens referring to the same entity (Tenney et al. 2019a; b). Consequently, information referring to one part of the relation can be traced back to the other part.

Related work

Our goal is to reconstruct the language model BERT by means of linguistic (pedagogical) rules composed of NLP building blocks. Hence, XAI approaches analyzing language models regarding NLP building blocks (category A), XAI approaches analyzing pedagogical rules for reconstructing language models (category B) and XAI approaches for language models based on other techniques (category C) constitute the related work. In contrast, general rule-based XAI approaches (cf. Adadi and Berrada 2018) and XAI approaches (Ramon et al., 2020; Sushil et al., 2018) relying on a simple 'bag-of-words' analysis – both without any focus on language models – are not in the scope for our research.

Ad category A): Several existing works analyze language models by using their (contextualized) word embeddings or internal states as input to *predict* NLP building blocks (Coenen et al., 2019; Hewitt & Manning, 2019; Jumelet & Hupkes, 2018; Kim et al., 2019; Peters et al., 2018b; Tenney et al., 2019a; Tenney et al. 2019b; Van Aken et al., 2019). Then, the quality of these predictions is used as an indication whether a certain NLP building block is encoded in particular word embeddings (i.e., vector representations) or specific layers of the language models. That is, instead of reconstructing predictions of language models for NLP tasks in IS (e.g., sentiment term detection), an analysis of the general word embeddings themselves is aimed for in these works. For instance, different NLP building blocks have been predicted by word embeddings of the language models ELMo (Peters et al. 2018b) and BERT (Tenney et al. 2019a; b). However, the aim of our research is a different one. As discussed in the Introduction, our focus is to better comprehend BERT's predictions on NLP tasks in IS, for instance, to be able to justify decisions made based on its results. To enable that, it is necessary to reconstruct the predictions of BERT for relevant NLP tasks (such as the extracted sentiment terms in online consumer reviews), since these predictions and not particular word embeddings in form of vector representations are the foundation for further decisions. In

Table 1 Overview of NLP building blocks

Building block	Type	Linguistic information	Example labels for the sentence "The waiter of The Burger House was nice, he smiled at us."
Part-of-speech tags (POS)	Tags	Syntactic	POS-label ("*waiter*") = NN (Noun)
Synsets (SYN)	Tags	Semantic	SYN-label ("*nice*") = nice.a.01 (Synset description: "pleasant or pleasing or agreeable in nature or appearance")
Dependencies (DEP)	Relations	Syntactic	DEP-label ("*waiter*", "*nice*") = amod (adjectival modifier)
Semantic role labeling (SRL)	Relations	Semantic	SRL-label ("*he*", "*smiled*") = agent-predicate-relation
Coreferences (COREF)	Relations	Semantic	COREF-label ("*waiter*", "*he*") = True (referring to the same entity)
Proximity (PROX)	Relations	Syntactic	PROX-label ("*waiter*", "*nice*") = 6

that line, none of the approaches in this category considers pedagogical rules to enable a reconstruction of predictions of a language model for NLP tasks in IS.

Ad category B): There also exist recent, interesting works that analyze language models by means of pedagogical rules in a local manner (i.e., for single predictions). In Ribeiro et al. (2018), individual predictions of simple recurrent neural network-based language models are reconstructed by separate if–then rules. Building on this work, BERT's predictions in an application of fake news detection on social media are analyzed in Szczepański et al. (2021). Both works hardly incorporate contextual information for reconstructions. That is, only information of the previous token is considered to obtain local reconstruction rules. Thus, both works consider only short rules of low complexity. In addition, rules based on individual tokens (e.g., specific words) are used. Hence, both works do not discuss the composition of tag and relation building blocks when extracting rules for reconstruction and as a result, the proposed rules exhibit only low generalizability. In particular, relation building blocks such as DEP or COREF, which enable rules to comprise vital contextual information, are not considered.

Ad category C): Moreover, local non-rule-based XAI approaches have been proposed to reason language model predictions. In Malkiel et al. (2022), saliency maps are used to reason similarity predictions of online consumer reviews by a BERT-based model, aiming to highlight important word-pairs for specific similarity predictions. Moreover, different visualizations with respect to neuron activations in the hidden layers have been applied to reason specific language model predictions (Brasoveanu & Andonie, 2022). In Kokalj et al. (2021), the known feature importance XAI approach 'shapley additive explanations' (Lundberg & Lee, 2017) has been adapted to account for the contextualized (token-based) text representation in language models. Further, approaches based on the attention weights in language models have been recently proposed (Ali et al., 2022; S. Liu et al., 2021), similarly establishing feature importance scores for language model predictions. As an application case, these approaches aim to determine important words for sentence sentiment classifications of a BERT-based model. However, all of these works focus on local reconstructions, for instance, for individual sentences, for which they do not consider NLP building blocks. That is, global (token-level) reconstructions by linguistic rules are out of their scope.

Overall, while the approaches in category A) give interesting indications on how NLP building blocks may be encoded in contextualized word embeddings, they do not enable to reconstruct the predictions of language models in NLP tasks in IS. In contrast, the approaches in category B) indeed analyze rules for reconstructing specific predictions, but only enable *local* reconstructions and do not incorporate different NLP building blocks comprising contextual linguistic information. Thus, they exhibit only low generalizability. Similarly, the approaches in category C) focus on reasoning specific language model predictions locally by non-rule-based approaches and do not incorporate different NLP building blocks either.

Summing up, there are very interesting contributions in the field of XAI regarding language models. However, literature lacks an approach for global reconstructions of language model predictions for NLP tasks in IS (e.g., sentiment term detection in online consumer reviews) based on pedagogical rules. To address this research gap, this paper proposes, to the best of our knowledge, the first global XAI approach for reconstructing token-level language model predictions by linguistic (pedagogical) rules. In particular, this paper is thus the first to enable an analysis of the trade-off between fidelity and comprehensibility (i.e., complexity and generalizability) in this setting.

Global reconstruction of BERT with linguistic rules

In this section, we introduce our approach by postulating the formal structure of linguistic rules for the global reconstruction of BERT's predictions and then outline appropriate measures to analyze this reconstruction.

Formal structure of linguistic rules for reconstructing BERT's predictions

We begin by deriving the formal structure of linguistic rules. Thereby, for illustration purpose, the language model BERT is applied for the token classification tasks aspect term detection and sentiment term detection that are frequently used in online consumer reviews (Dai & Song, 2019; Sun et al., 2019; H. Xu et al. 2019). More precisely, each sentence in a document comprises a string value and can be split up by tokenization into disjunct substrings (so-called tokens), which have a linguistic meaning, such as (sub)words or punctuation marks. The precise tokenization of sentences depends on specific tokenization policies. For this work, we used w. l. o. g. the widely applied tokenization of the python package NLTK (cf. https://www.nltk.org). The goal of the token classification tasks performed by BERT is to assign class labels to such tokens. For example, the second token 'fish' in the tokenized sentence ('The', 'fish', 'was', 'good', '!') is assigned with the class label *ASP* indicating an aspect term. The following postulates P1)-P3) provide the foundation for linguistic rules based on NLP building blocks, which enable a global reconstruction of BERT's predictions (i.e., the predicted class labels for the tokens of a sentence).

P1) "Label assignments": In our approach, we assign labels only to single tokens or token pairs. Hence, we do not

consider label assignments for whole sentences, documents nor for single character values. This focus is promising for reconstructing BERT, as BERT internally also establishes text representations on a token level.

P1.1) "Tag label assignments": A tag building block $tbb \in TBB$ (where TBB is the set of tag building blocks) assigns at most one *tag label* $tbb(t_i) \in L_{tbb}$ to a token t_i (L_{tbb} is the set of all labels from tbb). For instance, the tag building block POS with $L_{POS} = \{NN, VB, JJ, \dots\}$ assigns the label $POS(t_2) == NN$ (= 'noun') to the token t_2 ='fish' in the exemplary sentence above.

P1.2) "Relation label assignments": A relation building block $rbb \in RBB$ (where RBB is the set of relation building blocks) assigns at most one *relation label* $rbb(t_i, t_j) \in L_{rbb}$ to a token pair (t_i, t_j) (L_{rbb} is the set of all labels from rbb). For example, the relation building block DEP with $L_{DEP} = \{amod, nsubj, \dots\}$ assigns the label $DEP(t_2, t_4) == nsubj$ (= 'nominal subject') to the token pair (t_2, t_4) = ('fish', 'good'). In particular, relation building blocks enable to capture contextual information in a sentence, which is a main component of language models such as BERT.

P1.3) "Class label assignments": BERT assigns a *class label* $l_\tau(t_i) \in L_\tau$ to each token t_i (L_τ is the set of all class labels in a token classification task τ). For instance, in the aspect term detection task with class labels $L_{ASP} = \left\{ASP, \overline{ASP}\right\}$, the token t_2 ='fish' is assigned with the class label ASP by BERT indicating that 'fish' is an aspect term.

P2) "Feasible arguments for rules": In our approach, *feasible arguments* in the antecedent and consequents of a rule only reference to labels for tokens or token pairs as postulated in P1).

P2.1) "Feasible arguments in rule antecedents": A feasible argument in the rule antecedent only contains conditions regarding tag labels of tokens (cf. P1.1)) and relation labels of token pairs (cf. P1.2)).

P2.2) "Feasible arguments in rule consequents": A feasible argument in the rule consequent only contains class label assignments of tokens (cf. P1.3). Considering the classification task of sentiment term detection, the argument $l_{SENT}(t_4) \rightarrow SENT$ assigns the class label $SENT$ to the token t_4 ='good', indicating that 'good' is labelled as a sentiment term by BERT in the sentence 'The fish was good!'.

P3) "Conflicting classification results of multiple rules": Multiple rules $R_1, \dots, R_{n_R}$ ($n_R \in \mathbb{N}$) may result in conflicting classification results $l_\tau^1(t_i), \dots, l_\tau^{n_R}(t_i) \in L_\tau$ for the same token t_i. To resolve such conflicting classification results for a token t_i, it is sensible to assign the class of the rule with the highest precision (cf. next section).

Given the postulates P1)-P3), the structure of linguistic rules can be defined. A linguistic rule R is an "if–then-else" rule in the form of **IF** antecedent **THEN** "then"-consequent (**ELSE** "else"-consequent). Here, the antecedent is an arbitrary combination of feasible arguments as postulated in P2.1) by means of logical operators such as AND (i.e., "$\wedge$"), OR (i.e., "$\vee$") and NOT (i.e., "$\neg$"). Further, each "then"-consequent and each "else"-consequent consists of one feasible argument as postulated in P2.2). Thus, a rule R outputs the class assignments of the "then"-consequent in case that the antecedent is TRUE (otherwise and if an "else"-consequent is contained in the rule, it outputs the class assignments of the "else"-consequent). Moreover, rules can be characterized by their length, which is given by the number of tokens that are connected by a relation building block in the antecedent of a rule. A brief example of a rule of length two is given by:

IF $([POS(t_i) == NN] \vee \neg[POS(t_j) == VB]) \wedge [DEP(t_i, t_j) == nsubj]$
THEN $l_{ASP}(t_i) \rightarrow ASP$

This rule can be applied to the tokenized sentence ('The', 'fish', 'was', 'good', '!') from above. For this sentence, the antecedent of the rule is only TRUE if $t_i = t_2$ ='fish' and $t_j = t_4$ ='good'. For any other selection of t_i and t_j, the antecedent is FALSE since only the token pair ('fish', 'good') has the relation "nsubj" in this sentence. Hence, this linguistic rule correctly detects the aspect term 'fish'. Rules of the outlined formal structure based on the postulates P1)-P3) constitute the foundation for our approach for reconstructing BERT.

Assessing fidelity and comprehensibility of global reconstructions

To globally reconstruct BERT, all predictions of BERT for a token classification task have to be considered. Here, fidelity and comprehensibility are the most relevant measures (cf. Section "Types of XAI approaches in the context of language models") and assessing both measures is required to analyze the trade-off between fidelity and comprehensibility. Since we focus on global reconstructions of language models, we outline in detail how both measures can be assessed for global reconstructions in the following.

To measure fidelity, we consider the predictions of BERT for each class label. More precisely, the set of token ids (i.e., the positions of tokens in the text corpus) predicted by BERT as class $C \in L_\tau$ is given by $I_{C,BERT} = \left\{i \in I | l_{BERT}(t_i) = C\right\}$, where I is the set of all token ids. These token ids $I_{C,BERT}$ are used as the basis for extracting the linguistic rules on training data $I_{train,C,BERT}$ and validation data $I_{validation,C,BERT}$ as well as for assessing their fidelity of globally reconstructing BERT on test data $I_{test,C,BERT}$. Once a set Σ of linguistic rules is extracted, the F1 score is appropriate to assess the fidelity of the rule set (Sushil et al., 2018) as - in contrast to the accuracy measure - it accounts for imbalanced class distributions. The F1 score (i.e., based on precision and recall) of

the rule set Σ for reconstructing BERT's predictions $I_{C,BERT}$ is given by:

$$Pr_C(\Sigma) = \frac{\left|I_{test,C,BERT} \cap I_{test,C,\Sigma}\right|}{\left|I_{test,C,\Sigma}\right|} \tag{1}$$

$$Rec_C(\Sigma) = \frac{\left|I_{test,C,BERT} \cap I_{test,C,\Sigma}\right|}{\left|I_{test,C,BERT}\right|} \tag{2}$$

$$F1_C(\Sigma) = \frac{2 * Pr_C(\Sigma) * Rec_C(\Sigma)}{Pr_C(\Sigma) + Rec_C(\Sigma)} \tag{3}$$

Here, $I_{test,C,\Sigma} = \left\{i \in I_{test} | l_\Sigma(t_i) == C\right\}$ is the set of token ids from the test data that are assigned with class C by the rule set Σ. In case of multiclass classification the fidelity is then assessed by the average F1 score per class label C, denoted as $\overline{F_1}(\Sigma)$ (i.e., by the macro-averaged F1 score (Sushil et al., 2018)). In contrast to the regular formulas for classifier evaluation, which aim to evaluate the predictions of a classifier regarding the true class labels, the formulas (1)-(3) enable to evaluate the linguistic rules regarding the predicted class labels by BERT and hence, to assess the fidelity of reconstructing BERT by certain sets of linguistic rules Σ.

In contrast to the comprehensibility of local reconstructions (e.g., complexity of single rules), literature suggests to assess the comprehensibility of a global reconstruction by its model size (Guidotti et al., 2018). Since our model is a set of rules Σ, both the number of rules $NR(\Sigma)$ in the rule set and the number of unique argument values $NUAV(\Sigma)$ in the antecedents in the rule set (Vilone & Longo, 2021) determine its comprehensibility. These measures are given by:

$$NR(\Sigma) = |\Sigma| \tag{4}$$

$$NUAV(\Sigma) = |\{v \in AAV | \exists R \in \Sigma : v \in R\}| \tag{5}$$

Here, $AAV = L_{POS} \cup L_{SYN} \cup L_{DEP} \cup L_{SRL} \cup L_{COREF} \cup L_{PROX}$ is the set of all argument values of all NLP building blocks. For both measures, a lower value indicates higher comprehensibility. That is, we leverage two different measures which capture two important perspectives on comprehensibility. Overall, based on the measures (1) – (5) the fidelity and comprehensibility of global reconstructions can be assessed.

Analysis

In this section we analyze the reconstruction of BERT's predictions by our approach. First, we outline the selected tasks, datasets and the conducted automated extraction of linguistic rules for global reconstruction. Then, we demonstrate how our approach based on linguistic rules can reconstruct predictions of BERT. After that, we present and discuss the results as well as implications for research and practice.

Task selection, data preparation and rule extraction

For a meaningful analysis of the reconstruction of BERT's predictions, we selected the NLP tasks aspect term detection and sentiment term detection as these tasks are frequently analyzed in the IS field and constitute common applications for BERT and text analytics (Dai and Song, 2019; Sun et al., 2019; H. Xu et al., 2019), in particular in electronic markets (Chatterjee et al., 2021; Steur et al., 2022). Also, we chose two publicly available datasets that exhibit different characteristics – with restaurants reviews from the platform *Yelp* (Yelp Dataset Challenge; cf. https://www.yelp.com/dataset) as experience goods vs. laptop reviews from the platform *Amazon* (Ni et al., 2019) as search goods – to enable broader insights independent of specific item domains. To extract linguistic rules based on the formal structure postulated in the previous section, we used state-of-the-art toolkits for annotating both datasets with the NLP building blocks discussed in Section "NLP building blocks" and leveraged and extended rule generation and rule selection techniques from the literature. The following paragraphs provide more details.

The goal of aspect term detection and sentiment term detection is to classify tokens in online consumer reviews that express aspects or sentiments. An aspect term (e.g., 'laptop screen') represents an item aspect for which an opinion polarity is expressed by a sentiment term (e.g., 'very good') (Sun et al., 2019). The task of token classification is to assign a class label $C \in L_\tau$ (i.e., $L_{ASP} = \left\{ASP, \overline{ASP}\right\}$ and $L_{SENT} = \left\{SENT, \overline{SENT}\right\}$) to tokens of a sentence. To conduct aspect term detection and sentiment term detection, we used the publicly available state-of-the-art language model BERT. In particular, we used pre-trained BERT models, which were specifically adapted to the domains of restaurant reviews and laptop reviews, respectively (H. Xu et al., 2019). We fine-tuned these BERT models for the tasks aspect term and sentiment term detection on both domains using the publicly available, labeled dataset SemEval2014 provided by Fan et al. (2019). After that, the fine-tuned BERT models were used in this work to predict aspect terms and sentiment terms in the two review datasets. That is, the tokens of both review datasets were assigned with the class labels of BERT's predictions. An overview of the (randomly sampled) dataset excerpts used for analysis, including the predictions of BERT regarding both tasks, is given in Table 2.

For annotation of NLP building blocks on these datasets, we used the state-of-the-art toolkits Stanza (Qi et al., 2020) and AllenNLP (Gardner et al., 2018) as well as the lexical database WordNet (Fellbaum, 2013). More precisely, POS tags and DEP relations were annotated based

on Stanza, SRL relations and COREF relations based on AllenNLP; PROX relations were directly tangible and SYN tags could be extracted from WordNet.

To prepare the datasets for the analysis, we randomly split the sentences of the datasets into 65% training data, 15% validation data and 20% test data. Then, the extraction of linguistic rules comprises two steps. Firstly, automated rule generation determines linguistic rules that appear at minimum ten times in the training data to avoid rules that are only applicable for very few and highly specific sentences. Secondly, the rule selection assembles a subset of these linguistic rules by iteratively adding rules to a (initially empty) rule set if the F1 score of the rule set is thereby enhanced on the validation data (Q. Liu et al., 2015).

To conduct the extraction of rules, we used and extended existing techniques for automated rule generation (Dai and Song 2019) and automated rule selection (Q. Liu et al., 2015) to enable an integration and combination of different NLP building blocks. That is, Dai and Song (2019) proposed a rule generation algorithm for aspect and sentiment term extraction based on POS tags and DEP relations, which we extended to allow for further NLP building blocks – including the combination of different NLP building blocks – in a single rule. Based on further extensions to allow for an evaluation of these rules and the generated rule sets by means of precision and recall on validation data, the automated rule selection approach of Q. Liu et al. (2015) could be applied.

Then, we assessed the F1 score of the extracted set of linguistic rules on the test data. To assess comprehensibility of the extracted rules, we focused on rules with antecedents containing *at most* two arguments regarding tag building blocks and *at most* one argument regarding a relation building block. Hence, the rules are of at most length two. In that line, we only used the logical operator "AND" to preserve comprehensibility (Askira-Gelman, 1998).

We made the annotated datasets and our source code available at https://github.com/BertRules/Global_reconstruction_of_language_models_with_linguistic_rules.

Demonstration of reconstructing BERT's predictions with linguistic rules

Before discussing the results based on the introduced datasets and tasks, we give a brief preliminary demonstration of how our approach based on linguistic rules can be utilized to reconstruct predictions of BERT. Thereby, we consider the following three exemplary sentences of real restaurant reviews and highlight the extracted sentiment terms of BERT by bold font: "*The Homeburger was* ***huge***.", "*Moreover, John is* ***friendly*** *and* ***welcoming***.", "*Overall, the BurgerBarn is* ***amazing***.". A linguistic rule proposed by our approach that reconstructs these predicted sentiment terms is given by:

IF $[POS(t_i) == NNP] \wedge [POS(t_k) == JJ] \wedge [DEP(t_i, t_k) == nsubj]$
THEN $l_{SENT}(t_k) \rightarrow SENT$

This single rule detects the adjectives ("JJ"), which are in a nominal subject relation ("nsubj") with a proper noun ("NNP"), as sentiment terms. The application of this rule for the three sentences is given in Table 3.

As illustrated in Table 3, the rule reconstructs the sentiment terms detected by BERT in these example sentences and constitutes a generalizing, plausible rule, which is important for online consumer reviews, as special product/service names or attributes (e.g., special dishes or waiters in restaurant reviews) are often referenced by proper nouns. Overall, this rule alone already reconstructs around 350 sentiment terms in the restaurant dataset with a precision of 89% with respect to BERT's predictions. In contrast, reconstructing these sentiment terms by means of rules with specific tokens instead of NLP tag building blocks, a separate rule for each instantiation in Table 3 would be required for each of the sentiment terms. For instance, the rule.

IF
$[TOKEN(t_i) == John] \wedge [TOKEN(t_k) == friendly] \wedge [DEP(t_i, t_k) == nsubj]$
THEN $l_{SENT}(t_k) \rightarrow SENT$

is obviously highly specific and cannot reconstruct the sentiment terms 'huge, 'welcoming' or 'amazing'. Therefore, this example emphasizes that linguistic rules with NLP building blocks enable to achieve higher generalizability for

Table 2 Datasets for analysis

Dataset characteristic	Restaurants (Yelp reviews)	Laptops (Amazon reviews)
# of sentences	150,000	150,000
# of tokens	2,320,726	2,575,492
# of predicted aspect tokens by BERT	230,505	236,692
# of predicted sentiment tokens by BERT	186,204	166,109
Relative frequency of predicted aspect tokens by BERT (relative to # of tokens or # of sentences)	0.099 (rel. to tokens); 1.537 (rel. to sentences)	0.092 (rel. to tokens); 1.578 (rel. to sentences)
Relative frequency of predicted sentiment tokens by BERT (relative to # of tokens or # of sentences)	0.080 (rel. to tokens); 1.241 (rel. to sentences)	0.064 (rel. to tokens); 1.107 (rel. to sentences)

Table 3 Application of a linguistic rule to reconstruct BERT's predictions in exemplary sentences

Example sentence	Application of the above linguistic rule
*"The Homeburger was **huge**."*	**IF** $[POS(Homeburger) == NNP] \wedge [POS(\mathbf{huge}) == JJ] \wedge [DEP(Homeburger, \mathbf{huge}) == nsubj]$ **THEN** $l_{SENT}(\mathbf{huge}) \rightarrow SENT$
*"Moreover, John is **friendly** and welcoming."*	**IF** $[POS(John) == NNP] \wedge [POS(\mathbf{friendly}) == JJ] \wedge [DEP(John, \mathbf{friendly}) == nsubj]$ **THEN** $l_{SENT}(\mathbf{friendly}) \rightarrow SENT$
*"Moreover, John is friendly and **welcoming**."*	**IF** $[POS(John) == NNP] \wedge [POS(\mathbf{welcoming}) == JJ] \wedge [DEP(John, \mathbf{welcoming}) == nsubj]$ **THEN** $l_{SENT}(\mathbf{welcoming}) \rightarrow SENT$
*"Overall, the BurgerBarn is **amazing**."*	**IF** $[POS(BurgerBarn) == NNP] \wedge [POS(\mathbf{amazing}) == JJ] \wedge [DEP(BurgerBarn, \mathbf{amazing}) == nsubj]$ **THEN** $l_{SENT}(\mathbf{amazing}) \rightarrow SENT$

a reconstruction of the predictions of language models (e.g., in online consumer reviews).

Results

In this section, we present the results of the proposed approach for the reconstruction of BERT's predictions. In particular, we analyze the fidelity and the comprehensibility to which an extracted set of rules is able to globally reconstruct BERT. As outlined in the Section "Assessing fidelity and comprehensibility of global reconstructions" in detail, we determined the fidelity by the F1 score between the token classification of the linguistic rules and BERT's predictions and assessed comprehensibility by *NR* and *NUAV*. To account for the objectives of high fidelity and high comprehensibility, we consider four different setups of (low vs. high) rule complexity and (low vs. high) rule generalizability: To analyze rule complexity, we distinguish between "L1-rules" containing rules of length one and "L2-rules" comprising rules of length at most two (i.e., every L1-rule is also a L2-rule, but not vice versa). We point out that L2-rules contain relation labels and thus consider contextual information, while this is not possible for L1-rules. To analyze rule generalizability, we compare "rules with specific tokens" as arguments (low generalizability) against "rules with (only) NLP building blocks" (high generalizability). Given this, the comprehensibility of the four setups is shown in the Tables 4, 5, 6 and 7 regarding both tasks on the respective datasets.

Discussion of the results

We elaborate on the major findings of applying our approach for global reconstruction of language model predictions by discussing the results related to the research questions RQ1 and RQ2:

> **Ad RQ1:** *Our approach based on linguistic rules allows for the global reconstruction of language model predictions (e.g., in online consumer reviews).*

Our analysis shows that the predictions of BERT can be globally reconstructed by our approach with a fidelity of 78%-82% based on L2-rules with tokens (i.e., rules with high complexity and low generalizability) on the considered tasks for online consumer reviews (cf. Tables 5 and 7). In more detail, the recall of these global reconstructions (i.e., how many classified tokens of BERT could be reconstructed) ranges between 79%-83%, while the precision ranges between 76%-83%. This shows that incorporating relation building blocks by means of rules of length two, which enables capturing contextual information, is indeed helpful to globally reconstruct BERT's predictions with higher fidelity. In comparison, the rule sets with NLP building blocks (i.e., rules with high generalizability) yield higher comprehensibility, which is indicated by low numbers of unique argument values (298 at most) compared to over 165,000 classified tokens by BERT. At the same time,

Table 4 Comprehensibility of the global reconstruction of BERT's predictions for aspect term detection measured by *NR*; *NUAV*

Aspect term detection	Restaurants (Yelp reviews)		Laptops (Amazon reviews)	
	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)
Low complexity (i.e., L1-rules)	1,169; 1,169	27; 27	944; 944	35; 35
High complexity (i.e., L2-rules)	9,770; 2,565	2,791; 237	9,004; 2,201	2,718; 298

Table 5 Fidelity of the global reconstruction of BERT's predictions for aspect term detection

Aspect term detection	Restaurants (Yelp reviews)		Laptops (Amazon reviews)	
	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)
Low complexity (i.e., L1-rules)	75.0% (75.5%,74.5%)	58.4% (44.1%,86.1%)	78.0% (80.1%,76.0%)	53.4% (38.9%,85.0%)
High complexity (i.e., L2-rules)	78.2% (75.6%,81.1%)	66.0% (56.9%,78.6%)	82.4% (82.1%,82.8%)	62.8% (55.3%,72.5%)

Fidelity is measured by F1 Score (numbers in parentheses indicate Precision, Recall)

Table 6 Comprehensibility of the global reconstruction of BERT's predictions for sentiment term detection measured by *NR*; *NUAV*

Sentiment term detection	Restaurants (Yelp reviews)		Laptops (Amazon reviews)	
	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)
Low complexity (i.e., L1-rules)	757; 757	15; 15	700; 700	21; 21
High complexity (i.e., L2-rules)	5,627; 1,615	1,787; 258	5,491; 1,434	1,973; 288

Table 7 Fidelity of the global reconstruction of BERT's predictions for sentiment term detection

Sentiment term detection	Restaurants (Yelp reviews)		Laptops (Amazon reviews)	
	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)	Low generalizability (i.e., rules with specific tokens)	High generalizability (i.e., rules with NLP building blocks)
Low complexity (i.e., L1-rules)	76.8% (83.8%,70.9%)	65.9% (60.1%,72.8%)	75.0% (77.8%,72.3%)	57.8% (50.2%,68.1%)
High complexity (i.e., L2-rules)	81.7% (83.1%,80.4%)	69.9% (66.8%,73.2%)	79.1% (79.5%,78.6%)	65.7% (65.8%,65.6%)

Fidelity is measured by F1 Score (numbers in parentheses indicate Precision, Recall)

these rule sets also maintain fidelities of up to 70%. Here, it could be substantiated that the reconstruction of BERT's predictions is constituted by transparent and well-generalizing rules. For instance, the rule "**IF** a term is a synset of 'good' and an adjectival modifier (DEP-relation 'amod') of a noun (POS-tag 'NN'), **THEN** that token is labelled as a sentiment term by BERT." achieved 99% precision and enables to reconstruct over 1,000 sentiment terms in the restaurant dataset. Furthermore, in an additional analysis, no discriminating factors such as specific synsets regarding gender, origin or neglected negative sentiments for specific products/services were detected. In total, this yields that our global reconstruction approach by means of linguistic rules is suitable to support algorithmic auditing including validation checks in application scenarios such as discussed in the introduction (AS1-4).

Ad RQ2: *Our approach enables to establish a balanced trade-off between fidelity and comprehensibility.*

As the proposed linguistic rules allow to vary their rule complexity (e.g., L2-rules vs. L1-rules) and their rule generalizability (e.g., rules with NLP tag building blocks vs. specific tokens), it is possible to create setups for global reconstructions with different comprehensibility (cf. Tables 4 and 6). Our analysis of these setups shows that higher fidelity is achieved by reducing comprehensibility and vice versa. This yields that fidelity and comprehensibility are two conflicting objectives, which has also been a topic of discussion in general XAI literature (Arrieta et al., 2020; Gilpin et al., 2018). Indeed, the reconstruction by means of linguistic rules can either have a higher fidelity or a higher comprehensibility, while both objectives cannot be achieved

simultaneously. In particular, our results show that L1-rules with NLP building blocks, which have low complexity and high generalizability, yield the global reconstruction with the highest comprehensibility (i.e., NR and NUAV are at most 35; cf. Tables 4 and 6) in comparison. These rule sets achieve fidelities between 53 to 66% (cf. Tables 5 and 7). This means that BERT's predictions on the tasks of aspect term detection and sentiment term detection can already be partly reconstructed in a very comprehensible manner with a small set of rules of only one tag building block as argument. Conversely, when utilizing specific tokens instead of NLP building blocks as arguments in L1-rules, a higher fidelity of 75% to 78% is achieved (cf. rules with low generalizability in Tables 5 and 7). However, such rules (e.g., the rule "flavorful is a sentiment term") are highly specific and have low generalizability, which results in rule sets with at least 700 rules and unique argument values in the antecedents (cf. Tables 4 and 6). Furthermore, Tables 5 and Table 7 indicate that the fidelity increases when the rules become more complex, but this is accompanied by a decreasing comprehensibility as indicated in Tables 4 and 6. Here, L2-rules with NLP building blocks achieve fidelities between 63 to 70% (cf. Tables 5 and 7) with at most 298 unique argument values. Contrarily, L2-rules with tokens achieve the highest fidelities with values from 78% up to 82% (cf. Tables 5 and 7), but they exhibit the lowest generalizability and thus, global reconstructions with low comprehensibility which is indicated by multiple thousands of rules and between about 1,400 and 2,600 unique argument values (cf. Tables 4 and 6) in the rule sets. These different setups show that either higher fidelity or higher comprehensibility can be achieved by reconstructing BERT's predictions with linguistic rules. However, if both objectives are crucial and focused equally, the best setup may be L2-rules with NLP building blocks, which exhibit decent fidelity and comprehensibility at the same time. The advantage of these L2-rules compared to L1-rules with tokens is the much lower number of unique argument values, which is based on the higher generalizability of NLP building blocks compared to specific tokens, and in particular, the use of contextual information in form of relation building blocks. Overall, our linguistic rules enable to establish different relevant setups with respect to fidelity and comprehensibility depending on the requirements for an XAI approach in practice.

Implications for research and practice

Our work contributes to the comprehensibility of opaque AI models in text analytics, as it allows for comprehensible global reconstructions of language models. Therefore, our work is not only valuable for multiple different research strands, but it is also highly relevant for applications and supports the adoption of language models, as outlined in the following.

Implications for research

1) *Linguistic rules enable global reconstructions of high fidelity for language model predictions in text analytics.*

Existing literature on XAI (e.g., Arrieta et al., 2020) discusses that rule-based XAI models can exhibit high comprehensibility but tend to lack high fidelity for reconstructions of complex AI models. Our findings extend this existing body of knowledge, as our analysis shows that our approach based on linguistic rules enables reconstructions with higher fidelity as well as reconstructions with higher comprehensibility for language model predictions. In particular, linguistic rules can achieve high fidelity by means of the contained relation building blocks capturing contextual information which is relevant for many text analytics tasks (Devlin et al., 2019; Geng et al., 2021; Peters et al. 2018a). As both, a global reconstruction approach by linguistic rules and an analysis of the trade-off between fidelity and comprehensibility thereby, do not exist in the field of text analytics so far, we extend the existing body of knowledge for rule-based XAI approaches.

2) *Global reconstruction by means of linguistic rules paves the way for a thorough understanding of language models.*

In contrast to the existing body of knowledge from local reconstruction approaches, the proposed approach based on linguistic rules enables a global reconstruction of language models (cf. Section "Discussion of the results"). Hence, linguistic rules constitute a first step for global and thorough understanding of these black boxes, which cannot be achieved by local reconstruction approaches (cf. Section "Introduction"). With linguistic rules as vital instrument, researchers in the field of XAI can now focus on how to thoroughly justify predictions of language models for text analytics tasks (e.g., by leveraging tests of statistical significance for linguistic rules in a global reconstruction for language model predictions). Moreover, researchers can aim to improve language models in text analytics tasks based on our approach. That is, our approach could be used to additionally reconstruct and analyze false predictions of language models, to detect its flaws and by means of that, to enhance these language models. Furthermore, an analysis of linguistic rules reconstructing a language model's predictions could enable to derive deeper insights regarding effects of different types of review texts (e.g., reviews for search goods vs. experience goods or reviews of different consumer segments). That is, such analyses could support

to analyze whether language model predictions for reviews of different review types vary in the NLP building blocks contained in the rules for global reconstruction. In particular, our approach allows for assessing the contribution of specific NLP building blocks to global reconstructions of language model predictions, which supports enhancing the understanding and use of language model predictions in text analytics.

3) *Global reconstructions help to understand language model-detected features*.

Our work also has implications for other research strands such as text analytics of online consumer reviews regarding star ratings (e.g., Binder et al., 2019; Goeken et al., 2020; Heinrich et al. 2021) or review helpfulness (e.g., Yin et al., 2014). Here, many IS researchers aim at analyzing and explaining the relations between (aspect-based) sentiments and a target variable (e.g., star ratings or review helpfulness). To enable such analyses, it is necessary to extract high-quality features from large review datasets by means of state-of-the-art language models in a first step (e.g., to extract aspect-based sentiments from review texts). Similar as in the practical application scenarios (AS1-4), it is also vital for researchers to base their analyses and insights on reliable and comprehensible features. Hence, a comprehensible global reconstruction of language model predictions detecting such features may further enable a better understanding of the target variable based on the review texts as it reduces the opacity of the feature detection in the first step of such analyses of online consumer reviews. That is, our approach can help to shed light on black-box language models used for feature extraction in IS text analytics research.

Implications for practice

1) *Global reconstructions with high comprehensibility can improve acceptance of language models, and support their adoption in practice*.

The language model BERT is already used in various applications (cf. AS1-AS4 in Section "Introduction"). Here, reconstructions with higher comprehensibility by means of our approach can help to shed light on language model predictions in these applications, and thereby, to improve acceptance of such models. In particular, a reconstruction by our approach allows to verify that a language model applied in an electronic marketplace does not discriminate against specific groups. For instance, when online consumer reviews are analyzed (cf. AS1), our approach can help to prevent that specific groups of consumers are discriminated against (e.g., by assigning a negative sentiment to certain countries, ethnicities or genders). In text analytics-assisted recruitment processes (cf. AS4), the rules provided by the presented approach can be examined whether they contain arguments regarding gender or other discriminating attributes (detected by particular synsets) indicating undesired biases or discriminations. Similarly, our approach helps to reconstruct and justify BERT's predictions in chatbots (AS2) and finance applications (AS3). Further, the rules provided by our approach support an algorithmic auditing based on the GDPR and thus to comply with regulatory requirements. For instance, a data scientist has to be able to show that the data processing is fair according to the GDPR, which can be supported by analyses with respect to discriminations as outlined above. This is especially relevant since algorithmic auditing will likely become the gold standard for companies deploying AI models (Casey et al., 2019).

2) *Linguistic rules enable different relevant setups with respect to the trade-off between fidelity and comprehensibility depending on the requirements of different stakeholders for XAI approaches in practice.*

Our approach based on linguistic rules is particularly promising, as it enables to establish different setups with respect to the trade-off between fidelity and comprehensibility, allowing for more profound analyses (Gilpin et al., 2018). That is, reconstructions with higher fidelity might be leveraged by data scientists to analyze language model predictions in detail. In addition, domain experts might leverage reconstructions with higher comprehensibility to assess, disclose and communicate the justifications (e.g., of BERT's aspect term detection) in a given domain. In particular, AI text analytics models in practice can thus be analyzed with different setups by means of our approach, which can be combined to gain more robust insights and to comply with regulatory requirements.

Conclusion

Global reconstruction of language model predictions such as for the state-of-the-art model BERT is an important issue in both research and practice, since it can enable to justify decisions based thereon in many application scenarios (e.g., in eCommerce or finance) and thereby allow to comply with necessary algorithmic auditing. In this paper, we thus proposed a global XAI approach in text analytics for reconstructing predictions of language models for token-level classifications by linguistic rules. Further, we discussed the trade-off between fidelity and comprehensibility for the global reconstructions. For the analysis of our approach and the trade-off, we considered aspect term and sentiment term detection in two datasets of different domains. That

is, we considered laptops as search goods and restaurants as experience goods. The results for both domains showed that linguistic rules enable global reconstructions of higher fidelity for language models, which paves the way for a thorough understanding of language models in text analytics in the future. Further, our approach helps to understand language model-detected features used for further analytics in research. For practical application scenarios such as eCommerce, finance or online recruitment, our approach can improve acceptance of language models and thus support their adoption in text analytics. Here, our approach also supports compliance with regulatory requirements.

Nevertheless, our research has some limitations, which could be starting points for future works. In this paper, we focused on the predictions of BERT without further considering the correctness of these predictions. Thus, our research could also be transferred to an analysis of BERT's prediction errors aiming towards a further enhancement of language models (i.e., by using linguistic rules to specifically reconstruct false predictions). Moreover, as we focused on the tasks of aspect and sentiment detection for search and experience goods in eCommerce, other NLP tasks in different domains would be possible for examination and could further substantiate our findings. Here, our work provides the necessary first step toward such insights.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

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

Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., & Wolf, L. (2022). XAI for transformers: Better explanations through conservative propagation. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2202.07304

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Askira-Gelman, I. (1998). Knowledge discovery: Comprehensibility of the results. *Proceedings of the thirty-first Hawaii international conference on system sciences* (Vol. 5, pp. 247–255). IEEE.

Augasta, M. G., & Kathirvalavakumar, T. (2012). Rule extraction from neural networks – A comparative study. *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)* (pp. 404–408). IEEE.

Binder, M., Heinrich, B., Klier, M., Obermeier, A. A., & Schiller, A. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Brasoveanu, A. M. P., & Andonie, R. (2022). Visualizing and explaining language models. *Integrating Artificial Intelligence and Visualization for Visual Knowledge Discovery* (pp. 213–237). Springer, Cham.

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's "right to explanation" debate and the rise of algorithmic audits in enterprise. *Berkeley Tech. LJ, 34*, 143.

Chatterjee, S. (2019). Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents. *Decision Support Systems, 119*, 14–22. https://doi.org/10.1016/j.dss.2019.02.008

Chatterjee, S., Goyal, D., Prakash, A., & Sharma, J. (2021). Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. *Journal of Business Research, 131*, 815–825. https://doi.org/10.1016/j.jbusres.2020.10.043

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Coheur, L. (2020). From Eliza to Siri and beyond. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 29–41). Springer, Cham.

Dai, H., & Song, Y. (2019). Neural aspect and opinion term extraction with mined rules as weak supervision. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5268–5277). ACL.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2010.00711

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 30 Aug 2022.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 NAACL* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *ArXiv preprint*. https://doi.org/10.48550/arXiv.1710.00794

Fan, Z., Wu, Z., Dai, X., Huang, S., & Chen, J. (2019). Target-oriented opinion words extraction with target-fused neural sequence labeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2509–2518).

Fellbaum, C. (2013). *Wordnet in the encyclopedia of applied linguistics*. Boston: Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1285

Förster, M., Hühn, P., Klier, M., & Kluge, K. (2021). Capturing users' reality: A novel approach to generate coherent counterfactual

explanations. *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 1274).

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020a). Evaluating explainable artifical intelligence-What users really appreciate. *Proceedings of the 28th European Conference on Information Systems (ECIS)*.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020b). Fostering human agency: A process for the design of user-centric XAI systems. *ICIS 2020 Proceedings*.

Fortune Business Insights (2021). *Natural Language Processing (NLP) Market size, share and Covid-19 impact analysis*. Retrieved from https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933. Accessed 30 Aug 2022.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.1803.07640

Geng, Z., Zhang, Y. [Yanhui], & Han, Y. (2021). Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, *429*, 132–140. https://doi.org/10.1016/j.neucom.2020.12.037

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80–89). IEEE.

Goeken, T., Tsekouras, D., Heimbach, I., & Gutt, D. (2020). The rise of robo-reviews-The effects of chatbot-mediated review elicitation on review valence. *ECIS 2020 Proceedings*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42. https://doi.org/10.1145/3236009

Heidari, M., & Rafatirad, S. (2020). Semantic convolutional neural network model for safe business investment by using BERT. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1–6). IEEE. https://doi.org/10.1109/SNAMS52053.2020.9336575

Heinrich, B., Hollnberger, T., Hopf, M., & Schiller, A. (2022). Long-term sequential and temporal dynamics in online consumer ratings. *ECIS 2022 Proceedings*.

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2020). Something's missing? A procedure for extending item content data sets in the context of recommender systems. *Information Systems Frontiers*, *24*, 267–286. https://doi.org/10.1007/s10796-020-10071-y

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2021). Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets*, *31*(2), 389–409. https://doi.org/10.1007/s12525-019-00366-7

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in Word representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) (pp. 4129–4138).

Jumelet, J., & Hupkes, D. (2018). Do language models understand anything? On the ability of LSTMs to understand negative polarity items. *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP@EMNLP 2018)* (pp. 222–231). ACL.

Kamps, J., Marx, M., Mokken, R. J., & de Rijke, M. (2004). Using WordNet to measure semantic orientations of adjectives. In *LREC* (Vol. 4, pp. 1115–1118). ACL.

Kim, N., Patel, R., Poliak, A., Wang, A., Xia, P., McCoy, R. T., Tenney, I., Ross, A., Linzen, T., Van Durme, B., Bowman, S. R., & Pavlick, E. (2019). Probing what different NLP tasks teach machines about function word comprehension. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. ACL.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021). BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation* (pp. 16–21).

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of BERT. In *EMNLP-IJCNLP* (pp. 4365–4374). ACL. https://doi.org/10.18653/v1/D19-1445

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for self-supervised learning of language representations. *Proceedings of the International Conference on Learning Representations 2020 (ICLR)*.

Liu, Q., Gao, Z., Liu, B., & Zhang, Y. [Yuanlin] (2015). Automated rule selection for aspect extraction in opinion mining. *Twenty-Fourth international joint conference on artificial intelligence*. AAAI.

Liu, S., Le, F., Chakraborty, S., & Abdelzaher, T. (2021). On exploring attention-based explanation for transformer models in text classification. *2021 IEEE International Conference on Big Data (Big Data)* (pp. 1193–1203). *IEEE*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems,* 30.

Luo, B., Lau, R. Y. K., Li, C., & Si, Y.-W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1). https://doi.org/10.1002/widm.1434

Malkiel, I., Ginzburg, D., Barkan, O., Caciularu, A., Weill, J., & Koenigstein, N. (2022). Interpreting BERT-based text similarity via activation and saliency maps. *Proceedings of the ACM Web Conference 2022* (pp. 3259–3268).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations* (pp. 55–60). ACL. Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010. Accessed 30 Aug 2022.

Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 188–197).

O'Donovan, J., Wagner, H. F., & Zeume, S. (2019). The value of offshore secrets: Evidence from the Panama Papers. *The Review of Financial Studies, 32*(11), 4117–4155. https://doi.org/10.1093/rfs/hhz017

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018a). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers) (pp. 2227–2237).

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018b). Dissecting contextual word embeddings: Architecture and representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL.

Potnis, A. (2018). *Illuminating insight for unstructured data at scale*. Retrieved from https://www.ibm.com/downloads/cas/Z2ZBAY6R. Accessed 30 Aug 2022.

Qi, P., Zhang, Y. [Yuhao], Zhang, Y. [Yuhui], Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *ACL System Demonstrations*

(pp. 101–108). ACL. Retrieved from https://arxiv.org/pdf/2003.07082. Accessed 30 Aug 2022.

Ramon, Y., Martens, D., Evgeniou, T., & Praet, S. (2020). Metafeatures-based rule-extraction for classifiers on behavioral and textual data. *ArXiv Preprint*. Accessed 30 Aug 2022.https://doi.org/10.48550/arXiv.2003.04792

Repke, T., & Krestel, R. (2021). Extraction and representation of financial entities from text. In S. Consoli, D. Reforgiato Recupero, & M. Saisana (Eds.), *Springer eBook Collection. Data science for economics and finance: Methodologies and applications* (pp. 241–263). Cham, Switzerland: Springer k. https://doi.org/10.1007/978-3-030-66891-4_11

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Schiller, A. (2019). Knowledge discovery from CVs: A topic modeling procedure. *Proceedings of the 14th International Conference on business informatics (Wirtschaftsinformatik).*

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research, 123*, 588–603. https://doi.org/10.1016/j.jbusres.2020.09.068

Steur, A. J., Fritzsche, F., & Seiter, M. (2022). It's all about the text: An experimental investigation of inconsistent reviews on restaurant booking platforms. *Electronic Markets*, *32*(3), 1187–1220. https://doi.org/10.1007/s12525-022-00525-3

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *Conference of the North American Chapter of the ACL* (pp. 380–385). ACL. https://doi.org/10.18653/v1/N19-1035

Sushil, M., Šuster, S., & Daelemans, W. (2018). Rule induction for global explanation of trained models. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 82–97). ACL.

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Nature Scientific Reports, 11*(1), 1–13. https://doi.org/10.1038/s41598-021-03100-6

Tenney, I., Das, D., & Pavlick, E. (2019a). Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* ACL.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., & Pavlick, E. (2019b). What do you learn from context? Probing for sentence structure in contextualized word representations. *International Conference on Learning Representations 2019 (ICLR)*.

Van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How does BERT answer questions? A layer-wise analysis of transformer representations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 1823–1832).

Vilone, G., & Longo, L. (2021). A Quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.717899

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP Workshop BlackboxNLP* (pp. 353–355). ACL. https://doi.org/10.18653/v1/W18-5446

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *Conference of the North American Chapter of the ACL* (pp. 2324–2335). ACL. https://doi.org/10.18653/v1/N19-1242

Xu, S., Barbosa, S. E., & Hong, D. (2020). BERT feature based model for predicting the helpfulness scores of online customers reviews. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Intelligent Systems and Computing. Advances in Information and Communication* (Vol. 1130, pp. 270–281). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-39442-4_21

Yan, H., Gui, L., & He, Y. (2022). Hierarchical interpretation of neural text classification. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2202.09792

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2006.08097

Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539–560. https://doi.org/10.25300/MISQ/2014/38.2.10

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligence magazine, 13(3),* 55–75. https://doi.org/10.1109/MCI.2018.2840738

Zafar, M. B., Schmidt, P., Donini, M., Archambeau, C., Biessmann, F., Das, S. R., & Kenthapadi, K. (2021). More than words: Towards better quality interpretations of text classifiers. *ArXiv Preprint.* https://doi.org/10.48550/arXiv.2112.12444

Zhang, R., Yang, W., Lin, L., Tu, Z., Xie, Y., Fu, Z., Xie, Y., Tan, L., Xiong, K., Lin, J. (2020). Rapid adaptation of BERT for information extraction on domain-specific business documents. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2002.01861