# Secondary Publication

**Kiefer, Sebastian; Hoffmann, Mareike; Schmid, Ute**

## Semantic Interactive Learning for Text Classification : A Constructive Approach for Contextual Interactions

Date of secondary publication: 14.12.2023

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-9243

**Primary publication**

Kiefer, Sebastian; Hoffmann, Mareike; Schmid, Ute (2022): „Semantic Interactive Learning for Text Classification : A Constructive Approach for Contextual Interactions“. In: Machine learning and knowledge extraction, Vol. 4, Nr. 4, pp. 994-1010, Basel: MDPI, doi: 10.3390/make4040050.

**Legal Notice**

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

*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

check for updates

**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

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 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.

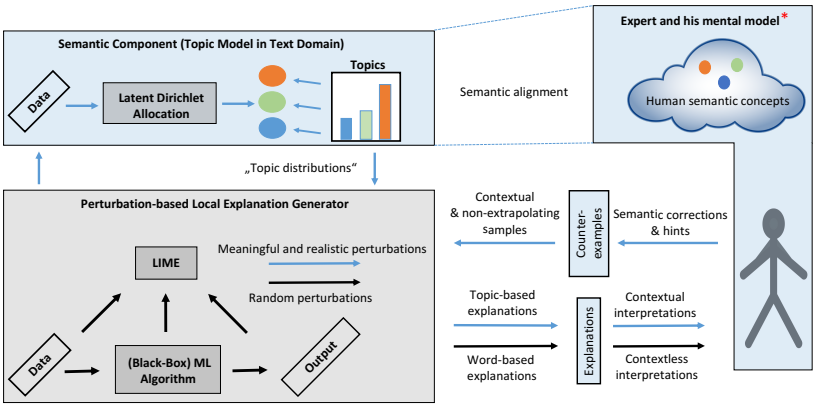



**Figure 1.** Architecture for constructive and contextual interactions. * In this work, we simulate the expert for efficient evaluation purposes using a conceptual Gold Standard as explained in Sections 3.3 and 4.3.

### *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 $\mathbf{w}$ in a corpus $\mathbf{D}$, a generative process from which the associated documents have been created is assumed as follows:

1. Choose N (the number of words) $\sim$ Poisson($\xi$).
2. Choose $\theta$ (a topic mixture) $\sim$ Dir($\alpha$).
3. For each of the N words $w_n$:
   (a) Choose a topic $z_n \sim$ Multinomial($\theta$).
   (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 $\theta$, a set of topics $\mathbf{z}$, and a set of words $\mathbf{w}$, given the hyperparameters $\alpha$ and $\beta$, 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 neighborhood 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.

**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) | |

**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. 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].

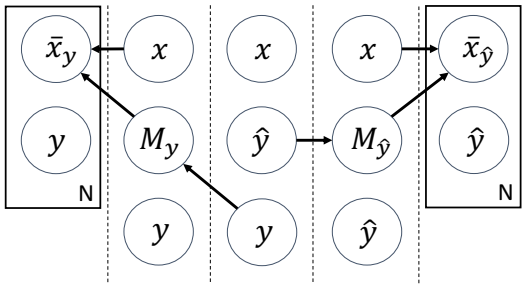



**Figure 3.** Graphical Model of SemanticPush.

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 $\lambda$

$C_{dest_x} = \{t \in \hat{z}_{xy} | t \notin GS_y\}$ ▷ 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$(\theta_{x_t})$

**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*$(\theta_{x_t})$

**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.

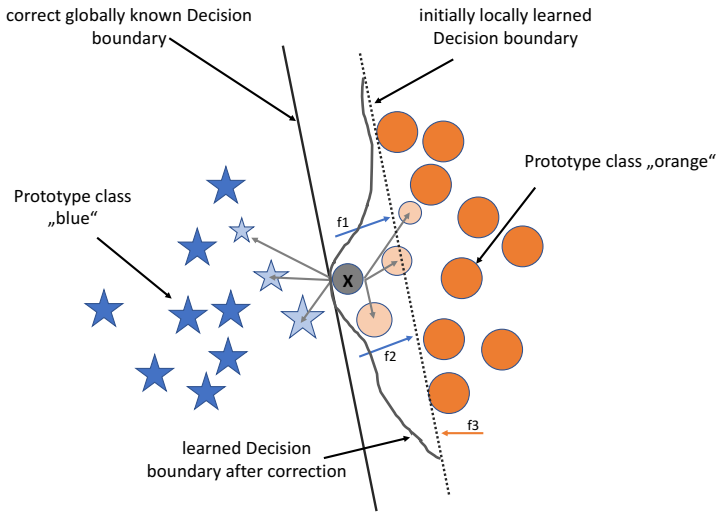
In addition, $\psi$ constitutes a normalization operator that re-normalizes the weights $t_w$ of the associated topics $t$ (either from Gold Standard or topicLIME explanation), revealing a multinomial distribution over topics $t$. SemanticPush then incorporates the concepts the classifier forgot to learn 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 $\delta_t$ 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 $\lambda$. 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' ($\text{CAIPI}_d$), 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' ($\text{CAIPI}_{d/c}$), 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 Mean$R^2$ (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.

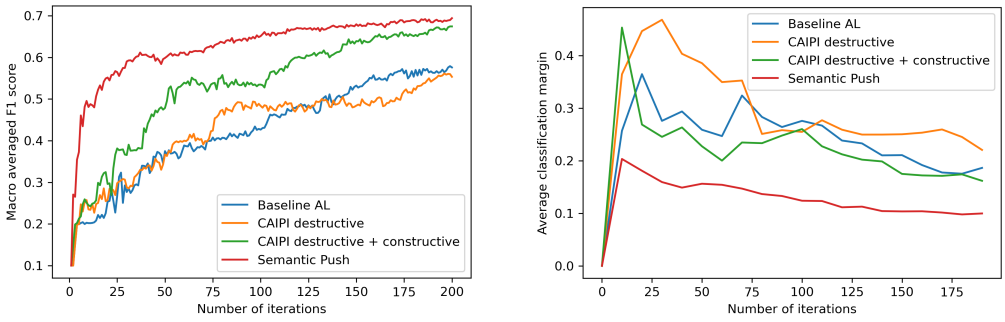
In order to analyze the development of Local Explanation Quality after applying the different IML strategies, we calculated a measure called 'Explanatory Accuracy'. First, we took $k = 10\%$ of the most relevant words from the global Gold Standard $GS_{global}^{(k)}(y)$ and intersected them with words ($GS_{global}^{(k)}(y) \cap x$) 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.

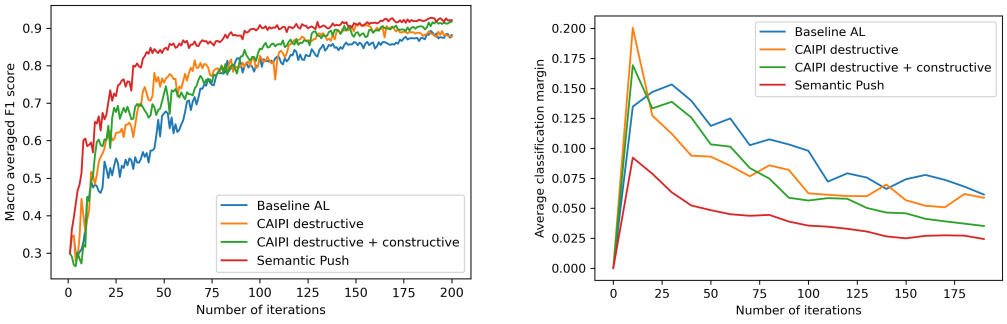
## 5. Experiment 1: Predictive Performance

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).

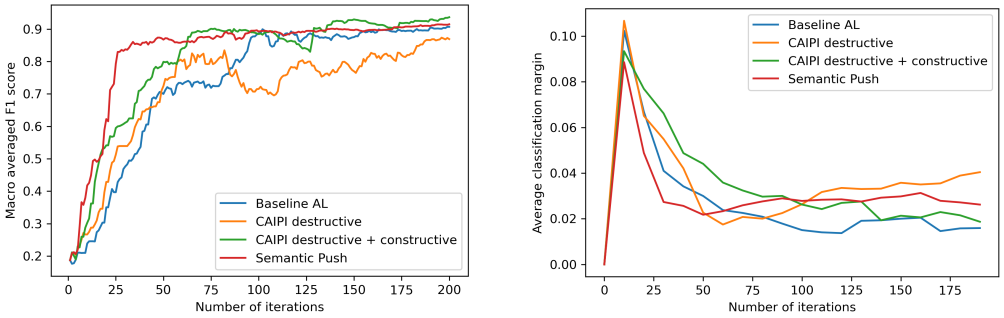



(**a**) (**b**)

**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 $\lambda$ (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. [PubMed]
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. [CrossRef] [PubMed]
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. [CrossRef]
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. [CrossRef]
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 [CrossRef]
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. [CrossRef]
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 [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef] [PubMed]
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. [CrossRef]
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. [CrossRef]
31. Keil, F.C. Explanation and Understanding. *Annu. Rev. Psychol.* **2006**, 57, 227–254. [CrossRef] [PubMed]
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.