PERCY: A post-hoc explanation-based score for logic rule dissemination consistency assessment in sentiment classification

Shashank Gupta a, Mohamed Reda Bouadjenek a,∗, Antonio Robles-Kelly b

a School of Information Technology, Deakin University, Waurn Ponds Campus, Geelong, VIC 3216, Australia
b Defense Science and Technology Group, Australia

ARTICLE INFO

Article history:
Received 14 March 2023
Received in revised form 23 May 2023
Accepted 27 May 2023
Available online 3 June 2023

Keywords:
Logic rules dissemination
Sentiment classification
Explainable AI

ABSTRACT

Disseminating and incorporating logic rules into deep neural networks has been extensively explored for sentiment classification in recent years. In particular, most methods and algorithms proposed for this purpose rely on a specific component that aims to capture and model logic rules, followed by a sequence model to process the input sequence. While the authors of these methods claim that they effectively capture syntactic structures that affect sentiment classification, they only show improvement in accuracy to support their claims without further analysis. Focusing on various syntactic structures, particularly contrastive discourse relations such as the A-but-B structure, we introduce the PERCY score, a novel Post-hoc Explanation-based Rule Consistency Score to analyze and study the ability of several of these methods to identify these structures in a given sentence, and to make their classification decisions based on the appropriate conjunct. Specifically, we explore the use of model-agnostic post-hoc explanation frameworks to explain the predictions of any classifier in an interpretable and faithful manner. These model explainability frameworks provide feature attribution scores to estimate each word’s impact on the final classification decision. Then, they are combined to check whether the model has based its decision on the right conjunct. Our experiments show that (a) accuracy – or any other performance metric – can be misleading in assessing the ability of logic rule dissemination methods to base their decisions on the right conjunct, (b) not all analyzed methods effectively capture syntactic structures, (c) often, the underlying sequence model is what captures the structure, and (d) for the best method, less than 25% of the test examples are classified based on the appropriate conjunct, indicating that a lot of research needs to be done on this topic. Finally, we experimentally demonstrate that the PERCY scores calculated are robust and stable w.r.t. the feature-attribution frameworks used.

© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Deep Neural Networks (DNNs) provide extraordinary performance across a broad spectrum of Natural Language Processing (NLP) tasks such as Sentiment Classification [1], Machine Translation [2], Text Summarizing [3], etc. This is mainly due to their characteristic of hierarchical feature representation [4], which can be learned automatically through purely data-driven approaches (i.e., without any external supervision) using a gradient-based optimization algorithm with a task-specific objective.

However, these hierarchical representations of features, when learned through purely data-driven approaches, suffer from several drawbacks including: (i) their complexity, which often leads to the extraction of human-uninterpretable features and hinders their application in high-stakes domains where automated decision-making systems must have a human understanding of their internal process, requiring the user to trust their outputs [5], (ii) DNNs are treated as essentially black-box models, where no meaningful relationship in terms of “how?” and “why?” can be established between inputs and outputs; (iii) a huge amount of labeled training data is required to construct these models, which is both expensive and time-consuming [6]; and (iv) previous ablation studies on DNNs [7,8] have shown that purely data-driven training may also lead to the learning of spurious feature representations, which can provide unreasonable outputs and make them prone to malicious attacks based on adversarial examples [9,10].

To combat these drawbacks, several solutions aim to make these networks inherently interpretable by augmenting them with some task-specific or domain-specific expert prior knowledge. These solutions are collectively called Neural-Symbolic
methods [11] as they aim to combine symbolic knowledge represented by logical rules with Deep Neural Networks. They have been extensively explored for various NLP tasks such as sentiment classification [12], question answering [13], machine translation [14], and information extraction [15], where the ultimate goal is to model and transfer various human interpretable logical rules to a neural network in order to improve its accuracy and causal interpretability. These methods usually rely on (1) a component aimed at capturing and modeling logic rules (e.g., the teacher network in the Iterative Knowledge Distillation method [12], the ELMo component in the Contextualized Word Embeddings approach [16], or the Semantic Composition Module in SentiBERT [17]), (2) followed by a Neural Network model to process the input sequence, (e.g., 1-D CNN [18], RNN, etc.).

While authors of these methods claim that they effectively capture syntactic structures in an input sentence that affect the outcome of a particular task (e.g., sentiment classification), they have only shown improvement in terms of accuracy to support their claim with no further analysis provided. However, achieving a high classification accuracy does not necessarily indicate that a method has effectively captured and encoded such logical syntactic structures. For example, let us consider the sentence “the casting was not bad but the movie was horrible" that has an A-but-B structure – a component A being followed by but, which is followed by a component B. In this example, the conjunction is interpreted as an argument for the second conjunct, with the first functioning concessively [19,20]. While a sentiment classifier can correctly identify that this sentence has a negative sentiment, it may fail to infer its decision based only on the B part of the sentence (i.e., “the movie was horrible"), but instead, it may base its decision on individual negative words also present in Part A (i.e., “bad”). Thus, we argue in this paper that the high accuracy of a classifier does not necessarily indicate that it has effectively captured textual structures of input sentences.

Focusing on various syntactic structures, in particular contrastive discourse relations such as the A-but-B, A-yet-B, A-though-B, or A-while-B structures, we introduce the PERCY score, a novel Post-hoc Explanation-based Rule Consistency Score, which is used to evaluate both the task-specific performance and logic-rule dissemination performance of a Neural-Symbolic system. In particular, PERCY is used to analyze and study the ability of various knowledge dissemination methods to effectively identifying a syntactic structure in an input sentence, (ii) encode and model these structures in sequences, and (iii) make their classification decisions based on the appropriate conjunct. Specifically, we explore the use of post-hoc explanation frameworks that are agnostic to the underlying model such as LIME [21], SHAP [22], and Integrated-Gradients (IG) [23], which explain the predictions of any classifier by providing feature-attribution scores. Furthermore, we use these scores to evaluate the impact of each conjunct in a sentence with a syntactic structure on the decision made by a classifier. Going back to the example mentioned above (i.e., “the casting was not bad but the movie was horrible"), PERCY helps users understand if a classification model has made its decision based on the B conjunct or individual words of this sentence.

The contributions of this paper are summarized as follows:

- We conduct an exhaustive experimental evaluation on two datasets, Sentiment140 [24] and SST2 [25], on which we compare various methods for logic rule dissemination with diverse classification methods – a total of 40 sentiment classifiers are evaluated. Briefly, we demonstrate that:
  1. Accuracy or any other performance-based metric can be misleading in assessing methods for capturing logic rules.
  2. Not all methods are effectively capturing syntactic structures as they claim to do.
  3. Their sequence model is often what captures the syntactic structure.
  4. The best method makes its decision based on the appropriate conjunct in less than 25% of the test examples.

- We experimentally demonstrate that the PERCY scores calculated are robust and stable w.r.t. the Feature-attribution based Local Post-hoc Explanation frameworks used in this study.

The rest of the paper is organized as follows: Section 2 covers related work and puts our work in perspective. Section 3 describes the post-hoc explanation frameworks that we use in this paper, followed by a detailed presentation of the PERCY score. Section 4 gives a brief overview of the analyzed sentiment classifiers and a thorough presentation of the investigated logic rule dissemination methods. Sections 5 and 6 respectively provide a description of our experimental setup and an analysis of the results we obtained. Finally, in Section 7 we conclude and suggest future research directions.

2. Related work

There is a substantial body of research related to disseminating and incorporating logic rules in deep neural networks. Below, we first describe the main text syntactic structures we consider in this paper, and then, we review Neural-Symbolic models, which are DNN models augmented with symbolic domain knowledge related to a specific task. Next, we discuss Neural-Symbolic models for NLP and then we provide a description of how these augmented models are often evaluated — mainly by focusing on their performance.

2.1. Logic rules for sentence-level sentiment classification

Text sentiment classification has a long and rich history of research due to its various practical applications, e.g., e-commerce, social media analysis, etc. In particular, sentence-level sentiment classification is the task that consists of determining the sentiment of a sentence by classifying it often as Positive, Negative, or Neutral. One important challenge in this regard is to model discourse relations between phrases and clauses in a sentence and to identify which part of a sentence will determine its overall sentiment [26–28].

In linguistics, a discourse relation is a description of how two segments of a sentence are logically connected to each other through a discourse marker or connector. Prasad et al. [29] have classified discourse markers as follows:

1. **Contingency relations:** which include markers like because, therefore, if, so, since to convey cause–effect relations between segments.

2. **Contrast relations:** which include markers like but, although, though, however, whereas, while to convey contrastive sense relations between segments.

---

1 In this paper, we use the term “knowledge dissemination methods" to refer specifically to techniques for incorporating explicit logic rules into machine learning models, with the aim of improving their performance on specific tasks.
3. Temporal relations: which include markers like before, after, when, since, while to convey the order of occurrence of segments.

4. Expansion relations: which include markers like and, in addition to convey that a latter segment elaborates on the former.

Previous work has shown that Contrastive Discourse Relations (CDRs) are hard to capture by general DNN models like CNNs or RNNs for sentence-level binary sentiment classification through purely data-driven learning [16,30]. Thus, Prasad et al. [29] define such relations as sentences containing A-keyword-B syntactic structure where two clauses A and B are connected through a discourse marker (the keyword) and have contrastive polarities of sentiment. Mukherjee and Bhattacharyya [26] argue that these relations need to be learned by the model while determining the overall sentence sentiment.

Table 1 summarizes all logic rules that we consider in this paper with the PERCY score, where we show the structure, the rule conjunct, and an example sentence. We selected these structures because they are examples of contrastive discourse relations that are commonly used in natural language to express complex ideas and opinions. These structures introduce a sense of contrast or opposition between two clauses or ideas, which can have a significant impact on the overall sentiment expressed in a sentence. Our study focuses on these structures in particular because they have been shown to be particularly challenging for sentiment classification models to accurately identify and classify. As a result, we believe that exploring how different knowledge dissemination methods perform on these specific structures will help shed light on the strengths and weaknesses of these methods when dealing with complex and nuanced linguistic phenomena.

2.2. Neural symbolic models

While traditional DNN models provide state-of-the-art performance on various pattern recognition tasks, they lack reasoning capabilities and act as black-box function approximators. On the other hand, symbolic models such as Decision Trees [32] or Inductive Logic Reasoning-based approaches [33,34] are inherently interpretable as they manipulate discrete categorical variables, but they show lower performance capabilities compared to DNNs [35]. Hence, a hybrid model called Neural Symbolic Model that combines both approaches has been proposed with the aim of equipping the hierarchical feature representation learning of Neural Networks with some real-world rules to make their prediction, coherent, consistent, and easily interpretable [11,36,37]. However, even before the advent of modern Neural Networks, constructing such knowledge and rule-augmented models has been extensively explored. For example, Towell and Shavlik [38] developed Knowledge-Based Artificial Neural Networks (KBANN) to combine symbolic domain knowledge abstracted as propositional logic rules with neural networks via a three-step pipelined framework. Later on, França et al. [36] constructed a neural model called CLIP++, which learns first-order relations from structured data through Inductive Logic Programming. More recently, Evans and Grefenstette [39] proposed a differentiable Inductive Learning framework to train a neural network via back-propagation on unstructured data. Instead of integrating Logic Rules as hard constraints, Manhaeve et al. [40] and Xu et al. [41] convert them into probabilistic soft-logic and integrate them with Deep Learning frameworks as soft-constraints. More recently, Lin et al. [42] proposed to fuse domain knowledge as topology contexts and logical rules of Knowledge Graphs into Language Models as soft constraints via Knowledge Distillation [43]. In this paper, our contribution is more analytical and focused on effectively testing such methods for their ability to encode and disseminate knowledge.

2.3. Neural symbolic models for natural language processing

A lot of research has been done on developing Neural-Symbolic Models for various Natural Language Processing tasks, where performance metrics have been mainly used to report their efficacy. For example, Hu et al. [12] fused domain knowledge abstracted as First Order Logic Rules with Deep Neural Networks via EM style algorithm called Iterative Knowledge Distillation. An updated version of this algorithm called Mutual Distillation [30] introduced some learnable parameters with Logic Rules to incorporate the fuzzy nature of domain knowledge. Zhang et al. [44] introduced a critic learning framework to augment a CNN-based model with various syntactical logical rules via Knowledge Distillation [43]. Cambria et al. [45] introduces a model called SenticNet7 which builds a hierarchical knowledge graph from the input sentence using kernel methods and auto-regressive language models and uses linguistic patterns to determine the sentiment polarity. Also, Chen et al. [46] introduced a feedback masking method where redundant parts of the input sequence (i.e., A tokens of a sentence with an A-but-B structure) are masked out before being fed to a Recurrent Neural Network fusing it with logical knowledge. More recently, Wang and Fan [15,47] developed a discrepancy loss to fuse First Order Logic Rules with Neural Networks for Information Extraction tasks like Opinion Target Extraction and Relation Extraction. On the other hand, instead of imposing constraints on loss function, Li and Srikanth [48] developed constrained neural layers, where logical constraints govern the forward computation operations in each neuron. Instead of changing either the loss function or the architecture, Wang and Poon [49] and Gu et al. [50] performed manipulation on input training data to induce logical domain knowledge. Finally, while not proposed to construct Neural-Symbolic models, Krishna et al. [16] showed that contextualized word embeddings constructed from large pretrained models like ELMo [51] can inherently capture the logical relationships like A-but-B for sentiment classification, but again, they have use accuracy to prove their claims.
2.4. Evaluation of neural-symbolic models

All of the work mentioned above has reported results using performance metrics such as accuracy [12,30,44,46,48–50] or F1-score [15,47,48] to support the claim that their methods have effectively captured logical domain knowledge. In this paper, we claim that while these performance metrics reflect the ability of a model to correctly identify the true class, they may fail to assess whether the classifier has actually captured logic rules and other syntactic structures. Hence, in this paper, we follow a different approach to evaluate such logic rule-augmented models by exploring the use of model-agnostic post-hoc explanation frameworks such as LIME [21], SHAP [22], and Integrated-Gradients (IG) [23], which gives a local explanation for each output in terms of input features. This approach helps to provide a causal explanation as quantifiable feature-attribution scores for an output given an input sentence with a certain rule syntactic structure, which we use to formulate our metric. To the best of our knowledge, this is the first work to provide a quantitative evaluation of such models using post-hoc explanation methods. While there have been various proposed Neural-Symbolic models for a wide range of tasks, our paper specifically concentrates on sentence-level sentiment classification. It is important to note that the analysis of other models for different tasks is beyond the scope of our study.

3. Methodology

As mentioned earlier, our main goal in this paper is to assess a sentiment classifier for its ability to correctly classify a test example with a logical syntactic structure on the basis of the appropriate conjunct. There are many methods proposed for generating explanations and incorporating interpretability and transparency into machine learning models [52]. Some methods provide explanations prior to their training, which typically involves designing models that are inherently more interpretable, such as rule-based systems [53,54] or decision trees [55–57], or incorporating specific features or constraints into the learning process to ensure that the resulting model is more transparent and easier to understand [58]. On the other hand, model-specific explainable AI involve developing techniques for analyzing and interpreting the internal workings of these models to better understand how they arrive at their predictions. Some common approaches include model simplification [59,60], visual attribution methods [61–63], feature relevance estimation [64,65], and other attention-based methods [66]. In contrast to these explainable methods, our approach requires generating explanations post-modeling. This is because we aim to answer the question of why a generated model produces a certain output for a given input, regardless of the specific model architecture used. Therefore, we use local post-hoc explanation frameworks, whose output is a causal mapping from the input datapoint to the model prediction. We distinguish diverse frameworks that are different depending upon the nature of the explanation provided, for instance: feature attribution scores [21–23], natural language explanations like Counterfactuals [67,68], or if–then–else type logical rules such as Scoped-rules in [69].

In this paper, we rely on Feature Attribution (also called Feature Importance) to calculate our PERCY score. Feature Attribute scores are obtained using model-agnostic Local Post-hoc explanation frameworks, which operate at the level of an individual input/prediction pair, producing an explanation for why a model predicted an output for a particular input. Fig. 1 provides a visual representation of the method we propose in this paper. It offers a general overview of the steps involved in our approach and how they relate to each other. By referring to this figure, readers can get a better understanding of the PERCY score calculation process.

3.1. Feature attribution based local post-hoc explanation frameworks

Local Post-hoc Explanation Frameworks have been used to explain outputs of various machine learning models ranging from a simple logistic regression to complex deep neural networks like Inception network [70–73]. The output of these frameworks is usually a list of weights, where each reflects the contribution of a particular feature to the prediction of a test datapoint. This provides local interpretability, and it also allows to determine which feature changes will have most impact on the prediction. Such approaches can be built on different types of features, such as manual features obtained from feature engineering, lexical features including words/tokens and n-gram, or latent features learned by NNs. In the next sections, we provide details about three local post-hoc explanation frameworks used in this paper and how feature-attribute scores are calculated using these frameworks.

3.1.1. LIME

Local Interpretable Model-agnostic Explanations (LIME) is a framework developed by Ribeiro et al. [21] that can explain the output prediction of any classifier or a regressor in a faithful way, by approximating it locally with a simpler and interpretable model on an input instance. LIME learns surrogate models using an operation called input perturbation and can be used to achieve either local [21] or global explanations [74].

Let’s consider a model $f$ and a sentence $s \in S$ represented with an n-dimensional token sequence vector $s = \{t_1,t_2,\ldots,t_n\}$. LIME proceeds by assigning to each token $t_i$ a weight $w_i$ that reveals its importance in influencing the output prediction of $f$. LIME assigns these weights as “Sparse Linear Models”, which are surrogate linear models learned in the vicinity of the input $s$. These surrogate models are computed by solving the following optimization:

$$\arg\min_{g \in G} \sum_{z \in \mathbb{Z}} \exp(-\cos(s, z)^2/\sigma^2)(f(z) - g(z'))^2 + \Omega(g)$$

where $\mathbb{Z}$ is a set of all perturbations computed for $s$, $cos$ is the cosine distance, $G$ is the set of interpretable linear surrogate models, and $\Omega(g)$ denotes the measure of complexity of $g$ (for explanations as linear models, it is the number of weights in every model). The optimal solution of Eq. (1) denotes the preciseness of surrogate model $g$ in approximating model $f$ around the locality of $s$ defined by $\Omega(g)$.

3.1.2. SHAP

Shapley Additive Explanations (SHAP) is a framework developed by Lundberg and Lee [22] to provide model-agnostic local explanations based on feature-attribution. SHAP is based on the game theoretically optimal Shapley values. Specifically, given a model $f$ and an input sentence $s$, to produce an interpretable model, SHAP defines an output model $g(s')$ with simplified input $s'$ as a linear addition of all input tokens as follows:

$$f(s) = g(s') = \phi_0 + \sum_{t_i \in s'} \phi_i t_i$$

where $s$ is the original sentence, $s'$ is a simplified input sentence with a mapping function $s = h_i(s')$ between $s$ and $s'$, and $\phi_0 = f(h_i(0))$ is the model output without all of the simplified inputs. A detailed description of SHAP and a possible solution for Eq. (2) can be found in the literature [22].
3.1.3. IG

Integrated Gradients (IG) is a simple, yet powerful axiomatic attribution method developed by [23], which provides feature importance scores using product of their gradients and values. While the previous two frameworks are based on local perturbations, IG is based on a Gradient perturbation method to calculate Feature Attribution scores. Specifically, let’s suppose we aim to explain the prediction of a model $f$ for an input sentence $s$. The integrated gradient for the token $t_i$ of the input sentence is defined as follows:

$$IG(s) = (t_i - t_i') \int_{a=0}^{1} \frac{\partial f(t'_i + a(t_i - t_i'))}{\partial t_i} \, da$$  \hspace{1cm} (3)$$

where the gradient of $f$ for the token $t_i$ is denoted by $\frac{\partial f}{\partial t_i}$, and $t_i'$ is the $i$th token in a selected sentence baseline $s'$. For most models, it is recommended to choose a baseline such that the prediction at the baseline is near zero ($f(s') \approx 0$). A more detailed explanation of IG can be found in [23].

3.2. PERCY: Post-hoc explanation-based rule consistency score

Let’s consider a model $f$ and a sentence $s \in S$ that is represented with an $n$-dimensional token sequence vector $s = [t_1 \cdots t_n]$. All feature attribution frameworks described above assign a weight $w_i$ for each term $t_i \in s$ to estimate its contribution to the prediction of $f$. Often, a positive weight $w_i > 0$ indicates that $t_i$ contributes and supports the positive class, whereas a negative weight $w_i < 0$ indicates a contribution of $t_i$ towards a negative prediction. Hence, given a sentence $s$ that contains an $A$-but-$B$ syntactic structure, we first define the sub-sequences $a = \{t_1 \cdots t_{k-1}\}$ and $b = \{t_{k+1} \cdots t_n\}$ as respectively the left and right sub-sequences w.r.t. the word ”but” indexed by $k$.

3.2.1. Calculating the conjunct contribution

Next, we estimate the contribution of each sub-sequence $a$ and $b$ to the prediction of $f$ as an expectation over $a$ and $b$ using a weighted average of all tokens in each part as follows:

$$E[a] = \sum_{i=1}^{k-1} w_i \times p(y = 1 | s) + \sum_{i=1}^{k-1} |w_i| \times p(y = 0 | s)$$  \hspace{1cm} (4)$$

$$E[b] = \sum_{i=k+1}^{n} w_i \times p(y = 1 | s) + \sum_{i=k+1}^{n} |w_i| \times p(y = 0 | s)$$

where $k$ is the index of the “keyword”, $p(y = 0 | s)$ and $p(y = 1 | s)$ are the probabilities to predict respectively the class 0 and 1 given a sentence $s$, $w_i$ is the feature attribution weight given to a term $t_i$, and $\mathbb{E}[^+ \cdot | \mathbb{E}[^+]$ is the expected value over terms contributing to the positive class (resp. negative class). Following these estimations, the PERCY score of a sentence $s$ is calculated depending on the rule as detailed in Table 2.

In Table 2, $f(s) = y$ aims to check that the prediction is correct — accuracy, “$\mathbb{E}[a] < \mathbb{E}[b]$” ensures that the sub-sequence $b$ has contributed more to the prediction of $f$, and the $p$-value aims to make sure that the difference between $\mathbb{E}[a]$ and $\mathbb{E}[b]$ is statistically significant. When a sentence contains multiple syntactic structures, such as a combination of “A-but-B” and “not only”, the PERCY score can be calculated for each syntactic structure separately.

Finally, the PERCY score of a collection of sentences $S$ is calculated by averaging as follows:

$$\text{PERCY}(S) = \frac{1}{|S|} \sum_{s \in S} \text{PERCY}(s)$$  \hspace{1cm} (7)$$

We note that in the experimental evaluation we present in Section 6, we mainly report the PERCY at the collection level (Eq. 7) to compare and contrast the different sentiment classification methods we describe in the next section. In Section 6.3, we provide justification behind each step involved in the calculation of PERCY score in 2 using qualitative analysis of A-but-B type sentences.

4. Sentiment classification methods

In this section, we provide a succinct description of the sentiment classification methods used in our experimental analysis.

4.1. Logic rules dissemination methods

In this section, we describe the main methods we analyze to disseminate logic rule knowledge into the Neural Network models described in Section 4.2.

4.1.1. Iterative knowledge distillation

The Iterative rule knowledge distillation method proposed by Hu et al. [12] aims to transfer the domain knowledge encoded in first order logic rules into a neural network defined by a conditional probability $p(y | x)$ where $\theta$ is a parameter to
learn. To integrate the information encoded in the rules, Hu et al. [12] have proposed to train the network via knowledge distillation as proposed in Hinton et al. [43] where hard targets are provided through labeled training data and soft targets are constructed through rule constrained projection of posterior $p_{y|x}$. Specifically, during training, a posterior $q_{y|x}$ is constructed by projecting $p_{y|x}$ into a subspace constrained by the rules to encode the desirable properties as follows:

$$\min_{q_{y|x}} KL(q_{y|x} \| p_{y|x}) + C \sum_{x} \xi_{x}$$

s.t. $$(1 - E_{y} \leftarrow q_{y|x}(r_{0}(x,y)) \leq \xi_{x}$$

where $q_{y|x}$ denotes the distribution of $(x, y)$ when $x$ is drawn uniformly from the train set $X$ and $y$ is drawn according to $q_{y|x}$, $r_{0}(x, y) \in [0, 1]$ is a variable that indicates how well labeling $x$ with $y$ satisfies the rule, $\xi_{x} \leq 0$ is the slack variable for respective logic constraint, and $C$ is the regularization parameter. The closed form solution for $q_{y|x}$ is used as soft targets to imitate the outputs of a rule-regularized projection of $p_{y|x}$, which explicitly includes rule knowledge as regularization terms.

Next, the rule knowledge is transferred to the posterior $p_{y|x}$ through knowledge distillation optimization objective:

$$(1 - \pi) \times L(p_{y}, P_{true}) + \pi \times L(p_{y}, q)$$

where $P_{true}$ denotes the distribution implied by the ground truth, $L(\bullet, \bullet)$ denotes the cross-entropy function, and $\pi$ is a hyper-parameter that needs to be tuned to calibrate the relative importance of the two objectives. Following the terminologies used by authors in Hinton et al. [43], $p_{y}$ is called a “student” network and $q$ is called a “teacher” network, which is intuitively analogous to human education where a teacher is aware of systematic general rules and instructs students. Overall, the Iterative rule knowledge distillation method is agnostic to the network architecture, and thus is applicable to general types of neural models such as those depicted in Fig. 2.

### 4.1.2. Word embeddings

Traditional word embedding methods like Word2vec [76] and Glove [77] provide a unique and fixed vector for each word in the vocabulary. However, language is complex and context can completely change the meaning of a word in a sentence. Hence, contextual word embeddings methods have emerged as a way to capture the different nuances of the meaning of words given the surrounding text. Krishna et al. [16] have advocated that word embeddings when fine-tuned with downstream sentiment analysis task might capture logic rules and thus disseminate that latent information, for example in the 1D CNN sequence models of the neural network in Fig. 2(a). In this paper, we experiment with the following word embedding methods:

1. **Word2vec**: which is one of the most popular methods to efficiently create word embeddings developed by Mikolov et al. [76]. Briefly, word2vec embeddings are computed from a two-layer neural network. Word2vec maps each token to a vector space, typically of several hundred dimensions, where word vectors are positioned in the vector space such that words that share common contexts (semantically similar) are located close to each other in the space.

2. **Glove**: is an unsupervised learning algorithm for obtaining vector representations for words developed by Pennington et al. [77]. Training is performed on the non-zero entries of a global word-to-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. A matrix factorization algorithm is applied to efficiently extract the embeddings.

3. **ELMo**: stands for Embeddings from Language Models is a pre-trained model developed by Peters et al. [51]. Instead of using a fixed embedding for each word, ELMo looks at the entire sentence before assigning each word in it an embedding. It uses a bi-directional LSTM trained on a specific task to be able to create those embeddings. Krishna et al. [16] proposed to use ELMo in their method.

4. **BERT**: stands for Bidirectional Encoder Representations from transformers. This is also a pre-trained model developed by Devlin et al. [78]. Briefly, the BERT is a model based on Encoder Transformer blocks [79], which processes each element of the input sequence by incorporating and estimating the influence of other elements in the sequence to create embeddings.

### 4.1.3. Modeling semantic composition using self-attention

Large pre-trained models like BERT [78] and GPT [80] have achieved state-of-the-art performance on various NLP tasks. Usually, these models follow a pre-training step on a large language corpus and then fine-tuning on the downstream NLP task coupled with a smaller Neural Network model. Recent work [81,82] has shown that inducing domain or task specific knowledge during their pre-training phase improves performance on the downstream task. Following this line of research, other work [50,83–85] has sought to develop methods and frameworks to induce domain-specific or task knowledge into pre-training of large language models.

In particular, SentiBERT developed by Yin et al. [17] focuses on sentence-level sentiment analysis task and develops a self-attention based mechanism on top of BERT to capture rulesyntactic structures like *A-but-B* in input sentences. The authors argue that combining contextual information generated from a language model like BERT [78] with constituency parse-trees like that generated by Socher et al. [25] can better capture composition semantic relations in an input sentence. In this paper, we use the pre-trained weights of SentiBERT provided by the authors instead of training the Language Model from scratch.

---

2 The implementation and weights can be found here: https://github.com/WadeYin0712/SentiBERT.
4.2. Backbone models

We use in this paper the following two backbone neural network models:

**CNN model**: which is depicted in Fig. 2(a) and used in [18,86] for sentence-level sentiment classification. It takes as input a sequence of tokens, which are first processed by an embedding layer and converted into dense vectors of fixed size. Next, three 1D CNN sequence models (kernel size of 3, 4, and 5) process the embeddings in parallel in order to extract diverse features from the input sequence. These 1D CNN sequence models may learn various internal properties of the sequence that are useful for sentiment classification. Finally, the outputs of the three 1D CNNs are concatenated before being fed into a feed-forward binary classification layer with a sigmoid activation to extract the sentiment of the input sentence – 0 for a negative sentiment and 1 for a positive sentiment.

**LSTM model**: which is illustrated in Fig. 2(b) and is based on recurrent neural networks [87]. Similar to the CNN model, it also takes as input a sequence of tokens, which are converted into dense vectors by an embedding layer. Next, the token embeddings are passed to a many-to-one sequence model layer consisting of 128 LSTM units, which learn hidden features in the sequence relevant to the understanding of the sentiment of the sentence. Finally, the output corresponding to the last token of the sequence model layer is fed to a dense layer consisting of a single sigmoid activation unit which classifies the entire sequence as – 0 for a negative sentiment and 1 for a positive sentiment.

4.3. Sentiment classification methods

To conduct a thorough evaluation, we consider all possible configuration options that we discussed above as follows: (CNN, LSTM) × (Distillation, No Distillation) × (Fine-tuning, No Fine-tuning) × (word2vec, glove, elmo, bert, sentibert), which gives a total of 40 sentiment classifiers that are summarized in Table 3. For example, the classifier CDB in Table 3 indicates that the base neural network used is the CNN model, word embeddings are created using BERT, which is fine-tuned on the downstream sentiment classification task, and the training was done using Iterative Knowledge Distillation method.

5. Experimental setup

In this section, we describe the experimental setup we use in our evaluations, including a description of the datasets, the metrics used, and the details of our implementation.

5.1. Datasets

We train the sentiment classification models discussed in the previous section on two popular sentence-level sentiment classification datasets.

**Stanford Sentiment Treebank (SST2)**: This dataset proposed in [25] is a binary sentiment classification dataset and consists of 9,613 single sentences extracted from movie reviews, where sentences are labeled as either positive or negative each accounting for about 51.6% and 48.3%. A total of 1,078 sentences contain a syntactic structure, which accounts for about 11.2% of the dataset. We report our results only on test examples that contain a syntactic structure to demonstrate the ability of a classifier to capture the pattern.

**Sentiment140**: Since SST2 dataset contains low amount of sentences containing a syntactic structure, we complement it with Sentiment140 dataset, which contains a significant proportion of such sentences. Constructed from twitter corpus, Go et al. [24] released this dataset to perform sentence-level sentiment analysis on public domain tweets. It consists of 1.6M tweets scrapped from twitter using their API divided into 3 categories—positive, negative and neutral. For our evaluation, we rejected the neutral tweets and randomly selected 50,000 tweets containing equiproportion distribution of positive and negative sentiment tweets. Out of these 50,000 tweets, approximately 51% tweets contain a syntactic structure. Again, we report our results only on test examples that contain these syntactic structures.

Fig. 3 shows the complete distribution of these two datasets.
Table 3
Summary of the sentiment classification methods used in our experimental evaluation.

<table>
<thead>
<tr>
<th>Model no.</th>
<th>Classifier</th>
<th>Base model</th>
<th>Distillation</th>
<th>Fine-tuning</th>
<th>WE</th>
<th>LRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CW</td>
<td>CNN (C)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>2</td>
<td>CG</td>
<td>CNN (C)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>glove (G)</td>
</tr>
<tr>
<td>3</td>
<td>CE</td>
<td>CNN (C)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>elmo (E)</td>
</tr>
<tr>
<td>4</td>
<td>CB</td>
<td>CNN (C)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>bert (B)</td>
</tr>
<tr>
<td>5</td>
<td>CsB</td>
<td>CNN (C)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>6</td>
<td>CFW</td>
<td>CNN (C)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>7</td>
<td>CFG</td>
<td>CNN (C)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>8</td>
<td>CFE</td>
<td>CNN (C)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>9</td>
<td>CFB</td>
<td>CNN (C)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>10</td>
<td>CFSB</td>
<td>CNN (C)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>11</td>
<td>CDW</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>12</td>
<td>CDFG</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>13</td>
<td>CDFE</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>14</td>
<td>CDB</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>15</td>
<td>CDsB</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>16</td>
<td>CDFW</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>17</td>
<td>CDFG</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>18</td>
<td>CDFE</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>19</td>
<td>CDFB</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>20</td>
<td>CDFsB</td>
<td>CNN (C)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>21</td>
<td>LW</td>
<td>LSTM (L)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>22</td>
<td>LG</td>
<td>LSTM (L)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>23</td>
<td>LE</td>
<td>LSTM (L)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>24</td>
<td>LB</td>
<td>LSTM (L)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>25</td>
<td>LSB</td>
<td>LSTM (L)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>26</td>
<td>LFW</td>
<td>LSTM (L)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>27</td>
<td>LFG</td>
<td>LSTM (L)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>28</td>
<td>LFE</td>
<td>LSTM (L)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>29</td>
<td>LFB</td>
<td>LSTM (L)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>30</td>
<td>LFSB</td>
<td>LSTM (L)</td>
<td>x</td>
<td>✓ (F)</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>31</td>
<td>LDW</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>32</td>
<td>LDG</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>33</td>
<td>LDE</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>34</td>
<td>LDB</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>35</td>
<td>LDSB</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>x</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
<tr>
<td>36</td>
<td>LDFW</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>word2vec (W)</td>
</tr>
<tr>
<td>37</td>
<td>LDFG</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>glove (G)</td>
</tr>
<tr>
<td>38</td>
<td>LDFE</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>elmo (E)</td>
</tr>
<tr>
<td>39</td>
<td>LDFB</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>bert (B)</td>
</tr>
<tr>
<td>40</td>
<td>LDFSB</td>
<td>LSTM (L)</td>
<td>✓ (D)</td>
<td>✓ (F)</td>
<td></td>
<td>sentibert (sB)</td>
</tr>
</tbody>
</table>

WE = Word Embeddings used by the model.
LRD = Whether model is proposed for Logic Rule Dissemination or not.
F = Word embeddings were refined tuned on the downstream task.
D = Trained via Iterative Knowledge Distillation [12].

Fig. 3. Distributions of datasets used in our experimental evaluation. The inner-most layer gives the total number of instances in each dataset, 2nd layer indicates the total number of instances with positive and negative sentiment labels, and the outer-most layer gives the number of instances containing syntactic structures for each sentiment label.
5.2. Metrics

The performance is measured using the following conventional classification evaluation metrics: (i) Accuracy, (ii) Precision, (iii) Recall, and (iv) F1-score. In addition, we report the PERCY scores as described in Section 3 using the three explanation frameworks: LIME [69], SHAP [22] and IG [23], which we refer to as L-PERCY, S-PERCY, and I-PERCY scores respectively. We aim to assess the robustness of PERCY w.r.t. the explanation framework used and to compare the classification performance metrics with the PERCY score to ultimately assess how correlated these metrics are.

5.3. Implementation details

We divide the Sentiment140 dataset into train, val, and test splits using 60%, 20%, and 20% proportion of sentences respectively. Each split contains similar distributions for various subsets – no structure-positive, no structure-negative, syntactic structure-positive, syntactic structure-negative – as present in the complete dataset in Fig. 3(b). For SST2 dataset, since the sample size of test instances is very small (1,078 sentences), all classifier are trained, tuned, and tested using stratified nested k-fold cross-validation and evaluated primarily according to accuracy. This is done to increase the size of test instances and to reduce high variance. For Sentiment140, we use standard training procedure with Early-Stopping.

We optimize all models using mini-batch gradient descent with batch size = 50 using an Adam optimizer [88] with learning rate $\eta = 3e^{-5}$. We also use early stopping and a dropout = 0.5 regularization techniques to get the best weights for the models. For nested k-fold cross-validation in SST2, we set the value of $k = 5$ and the value of inner fold $l = 3$. The code of our implementation can be found at: https://github.com/shashgpt/PERCY.

6. Experimental evaluation

In this section, we discuss and analyze the results obtained for the sentiment classification models described in Section 4.3. Briefly, we first discuss their performance using conventional classification performance metrics as detailed in Section 5.2 and PERCY scores. Next, we analyze the correlation of PERCY with respect to the classification performance metrics. Finally, we discuss and evaluate the consistency and robustness of the PERCY score across the different explainability frameworks we use — LIME, SHAP, and IG.

6.1. Performance evaluation

Tables 4 and 5 show the performance of all classifiers described in Table 3 on SST2 and Sentiment140 respectively. We show the performance using all sentiment classification metrics (Classification Accuracy, Weighted Precision, Weighted Recall and Weighted F1-scores) and the PERCY score with three explainability frameworks (LIME, SHAP and IG), which are called L-PERCY, S-PERCY and I-PERCY respectively. Briefly, we make the following key observations:

1. For all classifiers, PERCY score values are less than 25%, which indicates that less than 25% of the test examples are correctly classified based on the correct conjunct. This suggests that the performance claimed by the logic rule dissemination methods analyzed in Sections 4.1.1, 4.1.2, and 4.1.3 is far from being achieved and that there is a lot of research that needs to be done on this topic.

2. There is major discrepancy between the classification performance metrics and the PERCY score values — the values of the classification performance metrics are much higher. The reason for this difference is discussed with anecdotal examples further in Section 6.3. In short, we observe that often, the two conjuncts contain the same number of sentiment-sensitive words. Hence, we argue that the classifiers are using those individual tokens to base their sentiment decision.

3. The BERT word-embeddings dissemination method (Section 4.1.2) provides the best classification performance values, whereas ELMo word-embeddings provides the best PERCY score values. This indicates that contextualized word-embeddings are better at capturing and disseminating logic rules.

4. We note that the classifiers that use the Iterative Knowledge-Edge Distillation method show almost no improvement on all metrics, e.g., in Table 4, CW and CDW classifiers provide similar values for all metrics. This simply suggests that [12] is not efficient and that it is the underlying sequence model that is capturing to some extent the syntactic structure.

5. Finally, we also observe that the SentiBERT method (Section 4.1.3) is not efficient at capturing syntactic structures as the performance using the PERCY score is always very low.

6.2. Correlation between PERCY scores and performance metrics

To analyze the correlation between performance metric values and PERCY score values, we show in Figs. 4(a) and 4(b) scatterplots with best fit linear regression of the rankings obtained using PERCY scores vs. Accuracy scores of all classifiers described above. The correlation is quantified using the Kendall’s correlation coefficient ($\tau$) [89].

Briefly, at first glance we observe that in all plots there is a little to no correlation between the rankings obtained using PERCY scores and Accuracy scores — the highest Kendall $\tau$ is 0.1487. This indicates clearly that higher accuracy cannot be used to claim that a classifier is performing well on rule dissemination since there is no-correlation between them. Probable reason for no correlation can be attributed to the classifiers using individual sentiment sensitive tokens in A conjunct. We note that we do not show the rank correlation plots for precision, recall and F1-scores vs. PERCY scores as we found them to be identical to those we show in Figs. 4(a) and 4(b).

6.3. Qualitative analysis

In order to provide insight into the factors behind specific PERCY scores, explore the lack of correlation between PERCY scores and Accuracy scores, and strengthen our analysis, we present in Table 6 examples of sentences exhibiting an A-but-B syntactic structure divided into three categories:

1. Examples where the sentiment prediction of the classifier is correct and the decision was based on the “B” conjunct to provide intuitive sense on why the sentiment of sentences containing A-but-B syntactic structures should be based on the “B” conjunct. These examples are shown in Table 6a.

2. Examples where the sentiment prediction of the classifier is correct but the decision was based on the “A” conjunct according to the PERCY score to show that accuracy can be misleading to assess rule-dissemination performance. These examples are shown in Table 6b.
3. Examples where the sentiment prediction is correct and conjunct contribution of "A" is greater than "B" but the "B" conjunct contains a single token having a higher score than all "A" conjunct tokens, i.e., $E[a] > E[b]$ but $\max[a] < \max[b]$. These examples show that using an additive operation like "expectation" is better suited from a robustness point of view than using any other operation like "max" to calculate the conjunct contribution in PERCY scores. These examples are shown in Table 6c.

For each category, we provide two examples in which one has a positive ground-truth sentiment and the other has a negative ground-truth sentiment. Briefly, we observe that:

1. In Table 6a, A and B conjuncts of both examples contain a comparable amount of sentiment-sensitive tokens to each other and the feature attribution scores assigned to B conjunct tokens are higher than the scores assigned to tokens in the conjunct A. Observing the nature of the sentiment switch from A to B, we can see that it makes sense to base the decision on the B conjunct to determine the sentence-level sentiment. This observation is consistent with the general Linguistics study of contrastive discourse relations like A-but-B done in [19,20]. Thus, there are neural-symbolic methods (Section 4.1) proposed to dissemminate this A-but-B rule knowledge into a general DNN model (Section 4.2) to force the model to make sentiment-prediction as per the B conjunct.

2. In Table 6b, we observe that A conjunct contain more sentiment-sensitive tokens than B conjunct and have a similar sense of sentiment i.e. they do not have any contrastive sentiment polarities. Thus, the classifier uses the individual tokens in A conjunct to base its decision, which is consistent with the ground-truth sentiment. This observation proves that A-but-B rule-dissemination performance and sentiment classification performance cannot be interlinked as the former checks whether the methods of rule-dissemination actually enable the classifier to learn and recognize A-but-B syntactic structures and forces the model to base its decision on the B conjunct.

3. Finally, in Table 6c, we observe that in both examples, A conjuncts contain more sentiment-sensitive tokens than B conjuncts and the conjuncts do not have contrastive sentiment polarities. Moreover, in both examples, the tokens that get the highest feature-attribution score in B conjunct are non-sentiment sensitive tokens (e.g., "value" and "too"). We note that the sentiment of the sentence is consistent with the sentiment of A conjunct and it makes...
6.4. Robustness of explanation frameworks for PERCY

Feature attribution based local post-hoc explanation \( E \) on a sentence \( s \in S \) for the model \( f \) can be viewed on a higher level as a function of both \( s \) and \( f \) [80] as follows:

\[
E_s = g(s, f)
\]

This is true for all frameworks that we use in our analysis — LIME [21], SHAP [22] and IG [23]. As we can see in Eq. (8), the explanation \( E \) depends on both the sentence \( s \) and the model to be explained \( f \).

Previous studies like [90,91] have shown that these explanation frameworks suffer from non-robustness issues, i.e., they provide substantially different explanations on a locally perturbed sample \( z \) of the input sentence \( s \) (the sample sentence \( z \) is locally perturbed to \( s \) if \( \|s - z\| \approx 0 \)). Alvarez-Melis and S. Jaakkola [90] argues that the explanation can only be considered meaningful or valid if they fulfill the criteria of being robust to the local perturbations of the input sentence \( s \). Intuitively, similar inputs should provide similar explanations. Hence, in the following, we analyze the robustness of explanation frameworks for PERCY using two methods.

6.4.1. Local lipschitz estimates

Mathematically, a post-hoc explanation \( E \) in Eq. (8) is robust if \( \|E_s - E_z\| \approx 0 \) for \( \|s - z\| \approx 0 \) given \( |P(y|s; w) - P(y|z; w)| \approx 0 \). To quantify this robustness, Alvarez-Melis and S. Jaakkola [90] propose to calculate the Local Lipschitz Estimate of \( E \). Inspired by Lipschitz continuity in calculus, which measures relative changes in function output with respect to function input in the entire domain, Alvarez-Melis and S. Jaakkola [90] propose to calculate the point-wise, neighborhood-based Local Lipschitz Estimate of \( E \) on an input sentence \( s \) of interest. Specifically, given a set of \( N \) sentences \( S_s = \{s(0), s(1), \ldots, s(K)\} \), they propose to define for every sentence \( s(i) \in S_s \) a set of all local perturbations to \( Z(i) \) as follows:

\[
Z_s(i) = \{z(i) \mid \|s(i) - z(i)\| \leq \epsilon \}
\]

where they set \( \epsilon = 0.1 \) to obtain local perturbations. Then, they propose to calculate the Lipschitz Estimate for each sentence \( s(i) \in S_s \) as follows:

\[
L_E(s(i)) = \arg \max_{z(i)} \frac{\|E_s - E_z\|}{\|s(i) - z\|^2}
\]
S. Gupta, M.R. Bouadjenek and A. Robles-Kelly

Knowledge-Based Systems 275 (2023) 110685

Fig. 4. Rank correlation scatter plots between sentiment classification accuracy values vs. PERCY score values on SST2 and Sentiment140 datasets. Green points represent classifiers proposed for Logic Rule Dissemination (LRD) in Table 3 whereas red points represent classifiers constructed for general sentiment classification (non-LRD).

Intuitively, the fraction $\frac{\|E_s - E_z\|_2}{\|s - z\|_2}$ in Eq. (10) should be bounded by a constant value $L$ where $L \approx 0$ for all $s \in S$. Hence, the higher the value of $L$, the lower the stability of explanations $E$, which in turn means that the explanatory framework that generated $E$ is less robust. Although the Lipschitz Estimate is a unit-less quantity, Alvarez-Melis and S. Jaakkola [90] states that it has no “ideal” universally desirable value and its acceptable value will depend on the end use of the generated explanations. In our case, we use the generated explanations to compute PERCY scores as shown in Section 3. Moreover, as it can be seen in Eq. (8), the generated explanations are dependent upon the dataset and model, which means lower values of Lipschitz scores on one set of (dataset, model) does not mean it will be lower on another set of (dataset, model).

In Fig. 5, we show the obtained local Lipschitz scores for all the classifiers in Table 3 using the three explanation frameworks used in our analysis – LIME, SHAP, and IG – on our two datasets – Sentiment140 and SST2. Overall, we make the following key observations:

1. We observe that SHAP has the lowest overall scores as the median values are lowest and thus, has the highest stability among all frameworks.
2. Moreover, the two other frameworks provide comparable scores as well which are low as compared to the ones reported in Alvarez-Melis and S. Jaakkola [90]. We note that these results are not contradictory but as stated by Alvarez-Melis and S. Jaakkola [90] and can be seen in Eq. (8), Local Lipschitz Estimates depend on the dataset and model to be explained, which are different in our paper compared to [90]. Thus, in our case, we can reach the conclusion that the explanation frameworks seem to provide robust-enough explanations on datasets and classifiers used.

6.4.2. Correlation between PERCY scores

As mentioned earlier, the end use of the generated explanations from all frameworks is to calculate PERCY scores as detailed in Section 3. While Lipschitz scores in Fig. 5 are quite low, they are still unable to tell whether the generated explanations are robust-enough so as not to influence the final PERCY score calculation. To measure the impact of explanations instability on final PERCY scores calculations, we compute correlations between PERCY scores calculated from all three explanation frameworks – LIME, SHAP, and IG – as shown in Figs. 6 and 7.

In particular, we show in Fig. 6 scatterplots with best fit linear regression of the rankings obtained using PERCY scores with the different explanation frameworks of all classifiers described above. Also, the correlation is quantified using Kendall’s correlation coefficient ($\tau$) [89]. On top of these plots, we also calculate the Pearson’s correlation between PERCY scores from different frameworks as distributions, i.e., PERCY scores calculated from each explanation framework – LIME, SHAP, and IG – are represented respectively as:

$$\text{PERCY}_{f_{\text{dist}}} = \begin{cases} 1, & \text{if } \text{PERCY}_f(x_i) = 1 \\ 0, & \text{if } \text{PERCY}_f(x_i) = 0 \end{cases} \quad \forall x_i \in X_{\text{test}} = \{X_{\text{test}1} + X_{\text{test}2} + \cdots + X_{\text{test}n}\}$$

(a) Classification Accuracy vs PERCY scores rank correlation plots on SST2 dataset

(b) Classification Accuracy vs PERCY scores rank correlation plots on Sentiment140 dataset

Fig. 4. Rank correlation scatter plots between sentiment classification accuracy values vs. PERCY score values on SST2 and Sentiment140 datasets. Green points represent classifiers proposed for Logic Rule Dissemination (LRD) in Table 3 whereas red points represent classifiers constructed for general sentiment classification (non-LRD).
Table 6
Anecdotal examples containing A-but-B syntactic structures. In each conjunct, we highlight tokens based on their feature-attribution weights assigned from an explanation framework. Darker color indicates a higher token score while lighter color indicates a lower token score. We show the scores obtained for the CDE classifier in Table 4 since it is an LRD classifier and it has the highest PERCY score values on all three explanation frameworks.

(a) Examples where the predicted sentiment was correct and the decision was based on the B conjunct.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Ground truth sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>lots of effort and intelligence are on display but in execution it is all awkward rumblings.</td>
<td>Negative</td>
</tr>
<tr>
<td>often messy and frustrating, but very pleasant at its best moments, it’s much like life itself.</td>
<td>Positive</td>
</tr>
</tbody>
</table>

(b) Examples where the predicted sentiment was correct but the decision was based on the A conjunct.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Ground truth sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>“analyze that” is one of those that not only fails on its own, but makes you second guess your affection for the original.</td>
<td>Negative</td>
</tr>
<tr>
<td>a gorgeously strange movie, heaven is deeply concerned with morality, but it refuses to spell things out for viewers.</td>
<td>Positive</td>
</tr>
</tbody>
</table>

(c) Examples to support using the “Expectation” operation instead of the “Max” for calculating conjunct contribution. The sentiments of these examples were correctly predicted.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Ground truth sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>a fine, rousing, g rated family film, aimed mainly at little kids, but with plenty of entertainment value to keep grown ups from squirming in their seats.</td>
<td>Positive</td>
</tr>
<tr>
<td>tries to add some spice to its dull sentiments but the taste is all too familiar.</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Fig. 5. Lipschitz scores of the explanation frameworks we use in our analysis. Each box plot denotes the Lipschitz values for all the classifiers in Table 3 of all the test data points on a particular dataset from a particular explanation framework.

where \( \text{PERCY}_{fr} \) means PERCY score calculated from a particular explanation framework (LIME, SHAP, or IG) and \( X_{test,fr} \) means set of test-datapoints for a \( r \)th classifier in Table 3 on which PERCY scores were calculated from \( \text{PERCY}_{fr} \). The results are shown in Fig. 7.

Overall, we observe that there is a significant correlation between the PERCY scores estimated using the three explanation methods as shown in Fig. 6 – all Kendall’s \( \tau \) values are above 0.5 indicating the classifiers share similar ranks of PERCY scores and the explanations instability do not influence the final PERCY scores calculation. These findings can be further supplemented with more granular level correlation results between PERCY scores as shown in Fig. 7 where LIME & SHAP, LIME & IG and SHAP & IG frameworks have significant positive correlation denoted by high positive Pearson’s values. These values further denote that even on the data-point level, the frameworks provide similar PERCY score values.

6.5. Limitations and discussion

While we believe that PERCY can provide valuable insights into the ability of knowledge dissemination methods to identify and classify contrastive discourse relations, we acknowledge that there are several limitations to our method that should be noted.

One limitation is that PERCY relies on post-hoc explanation frameworks to analyze the predictions of a given classifier. While these frameworks provide valuable insights into how a model arrived at its decision, they are not perfect and may not capture all the relevant syntactic structures in a given sentence. Additionally, PERCY assumes that the correct conjunct can be identified and extracted from the sentence, which may not always be the case in practice. Another limitation is that our study focuses specifically on contrastive discourse relations, and may not generalize to other syntactic structures or linguistic phenomena. For example, PERCY may not be effective at identifying and classifying more complex syntactic structures such as:
Fig. 6. Ranked correlation scatter plots between L-PERCY, S-PERCY, and I-PERCY scores on SST2 and Sentiment140 datasets — each dot point represents a method in Table 3 with its ranking on each axis. The line $y = x$ (black) is also included to convey whether the rank of a particular method is consistent with respect to the different explainability frameworks used. Also, we include the regression line (green) to show the general trend of the data points, making it easier to observe the positive relationship between the rankings.

Fig. 7. Pearson Correlation Heatmaps between L-PERCY, S-PERCY, and I-PERCY scores distributions on SST2 and Sentiment140 datasets respectively.
• **Complex subordination structures:** PERCY relies on identifying and extracting the correct conjunct in a sentence, which may be difficult or impossible in cases where the sentence contains complex subordination structures such as relative clauses or nested clauses.

• **Ambiguous discourse relations:** Some sentences may contain ambiguous discourse relations where it is not clear which conjunct should be considered as the main clause. For example, consider the sentence “Although he was tired, he went for a run and felt better.” It is not clear whether the main clause is “he went for a run” or “he felt better”, which may make it difficult to identify the appropriate conjunct to use for sentiment classification.

• **Negation and polarity:** Our method assume that the sentiment expressed in a sentence can be determined based on the sentiment words and discourse relations present in the sentence. However, in cases where the sentence contains negation or conflicting polarity, the overall sentiment may not be easily inferred from these features alone.

• **Irony and sarcasm:** Our method focuses on identifying and classifying sentiment expressed in a straightforward manner. However, some sentences may contain irony, sarcasm, or other forms of figurative language that may require a more nuanced approach to sentiment analysis.

It is important to note that there may be many other types of syntactic structures that PERCY may not be able to process effectively. We believe that further research is needed to address these limitations and to develop more robust and effective methods for incorporating logic rules into machine learning models.

7. Conclusion

This paper provides an analysis and a study of neural-symbolic methods focused on their ability to effectively disseminate logic rule knowledge in a DNN model for sentence-level binary sentiment classification task. This includes enabling a DNN model to effectively identify syntactic structures in a sentence and force the DNN model to base its decision on the appropriate conjunct. We show that accuracy or task-specific performance metric can be misleading in effectively assessing this ability. Hence, we proposed an alternative metric called PERCY, which stands for Post-hoc Explanation-based Rule Consistency Score to effectively assess the ability of a method to encode syntactic structures. We conducted an exhaustive set of experiments to support our hypothesis and concluded that the high performance of sentiment classification metrics does not necessarily indicate high rule-dissemination performance. Specific findings of our paper include that (a) accuracy – or any other performance metric – can be misleading in assessing the ability of logic rule dissemination methods to base their decisions on the right conjunct; (b) not all analyzed methods effectively capture syntactic structures; (c) often, the underlying sequence model is what captures the syntactic structure; and (d) for the best method less than 25% of test examples are classified based on the right conjunct indicating a lot of research needs to be done on this topic. Last but not least, we experimentally demonstrate that the PERCY scores calculated are robust and stable w.r.t. the feature-attribution frameworks used.

Our experiments demonstrated that in cases where a weaker sentiment is expressed in the rule conjunct of a discourse relation (e.g., after the “but” in a sentence), a naive model (e.g., CW in Table 3) that is solely based on the number or intensity of sentiment words may incorrectly classify the sentiment of the sentence based on the stronger sentiment that comes in the other conjunct (e.g., before the “but”). To address this issue, a model could incorporate a mechanism that takes into account the discourse relations in the sentence. One interesting approach to explore is to use a Rule-Mask Mechanism with, which given an input sequence predicts a vector that captures there exists an applicable logic rule on the input sequence [92]. Another approach is to use attention mechanisms that selectively focus on the important parts of the sentence, taking into account the discourse relations. For example, a model could use self-attention to weigh the importance of different words in the sentence based on their relation to other words, allowing the model to give more weight to the sentiment expression that is in the rule conjunct (e.g., after the “but”). Incorporating discourse relations and attention mechanisms can help improve the PERCY score of sentiment classification. Future work includes exploring the use of light-wise explanation frameworks to ease the calculation of the PERCY score.

CRediT authorship contribution statement

Shashank Gupta: Methodology, Software, Formal analysis, Validation, Data curation, Writing – original draft, Visualization. Mohamed Reda Bouadjenek: Methodology, Formal analysis, Validation, Writing – review & editing, Supervision. Antonio Robles-Kelly: Methodology, Formal analysis, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

The data used is already publicly available.

References


