On Guaranteed Optimal Robust Explanations for NLP Models

Emanuele La Malfa[1*], **Rhiannon Michelmore**[1], **Agnieszka M. Zbrzezny**[1,2]
Nicola Paoletti[3] and **Marta Kwiatkowska**[1]
[1]University of Oxford
[2]University of Warmia and Mazury, Olsztyn
[3]Royal Holloway, University of London
{emanuele.la.malfa, rhiannon.michelmore, marta.kwiatkowska}@cs.ox.ac.uk,
agnieszka.zbrzezny@matman.uwm.edu.pl, nicola.paoletti@rhul.ac.uk

Abstract

We build on abduction-based explanations for machine learning and develop a method for computing local explanations for neural network models in natural language processing (NLP). Our explanations comprise a subset of the words of the input text that satisfies two key features: optimality w.r.t. a user-defined cost function, such as the length of explanation, and robustness, in that they ensure prediction invariance for any bounded perturbation in the embedding space of the left-out words. We present two solution algorithms, respectively based on implicit hitting sets and maximum universal subsets, introducing a number of algorithmic improvements to speed up convergence of hard instances. We show how our method can be configured with different perturbation sets in the embedded space and used to detect bias in predictions by enforcing include/exclude constraints on biased terms, as well as to enhance existing heuristic-based NLP explanation frameworks such as Anchors. We evaluate our framework on three widely used sentiment analysis tasks and texts of up to 100 words from SST, Twitter and IMDB datasets, demonstrating the effectiveness of the derived explanations[1].

1 Introduction

The increasing prevalence of deep learning models in real-world decision-making systems has made AI explainability a central problem, as we seek to complement such highly-accurate but opaque models with comprehensible explanations as to why the model produced a particular prediction [Samek *et al.*, 2017; Ribeiro *et al.*, 2016; Zhang *et al.*, 2019; Liu *et al.*, 2018; Letham *et al.*, 2015]. Amongst existing techniques, *local explanations* explain the individual prediction in terms of a subset of the input features that justify the prediction. State-of-the-art explainers such as LIME and Anchors [Ribeiro *et al.*, 2016; Ribeiro *et al.*, 2018] use heuristics to obtain short explanations, which may generalise better beyond the given input and are more easily interpretable to human experts, but lack robustness to adversarial perturbations. The abduction-based method of [Ignatiev *et al.*, 2019b], on the other hand, ensures minimality and robustness of the prediction by requiring its invariance w.r.t. any perturbation of the left-out features, meaning that the explanation is sufficient to imply the prediction. However, since perturbations are potentially unbounded, this notion of robustness may not be appropriate for certain applications.

In this paper, we focus on natural language processing (NLP) neural network models and, working in the embedding space with words as features, introduce *optimal robust explanations (OREs)*. OREs are *provably guaranteed* to be both *robust*, in the sense that the prediction is invariant for any (reasonable) replacement of the features outside the explanation, and *minimal* for a given user defined cost function, such as the length of the explanation. Our core idea shares similarities with abduction-based explanations (ABE) of [Ignatiev *et al.*, 2019b], but is better suited to NLP models, where the unbounded nature of ABE perturbations may result in trivial explanations equal to the entire input. We show that OREs can be formulated as a particular kind of ABE or, equivalently, minimal satisfying assignment (MSA). We develop two methods to compute OREs by extending existing algorithms for ABEs and MSAs [Ignatiev *et al.*, 2019b; Dillig *et al.*, 2012]. In particular, we incorporate state-of-the-art robustness verification methods [Katz *et al.*, 2019; Wang *et al.*, 2018] to solve entailment/robustness queries and improve convergence by including sparse adversarial attacks and search tree reductions. By adding suitable constraints, we show that our approach allows one to detect biased decisions [Darwiche and Hirth, 2020] and enhance heuristic explainers with robustness guarantees [Ignatiev *et al.*, 2019d].

To the best of our knowledge, this is the first method to derive local explanations for NLP models with provable robustness and optimality guarantees. We empirically demonstrate that our approach can provide useful explanations for non-trivial fully-connected and convolutional networks on three widely used sentiment analysis benchmarks (SST, Twitter and IMDB). We compare OREs with the popular Anchors method, showing that Anchors often lack prediction robustness in our benchmarks, and demonstrate the usefulness of our framework on model debugging, bias evaluation, and re-

*First Author, contact at emanuele.la.malfa@cs.ox.ac.uk

[1]Code available at https://github.com/EmanueleLM/OREs

pair of non-formal explainers like Anchors.

2 Related Work

Interpretability of machine learning models is receiving increasing attention [Chakraborty *et al.*, 2017]. Existing methods broadly fall in two categories: explanations via globally interpretable models (e.g. [Wang and Rudin, 2015; Zhang *et al.*, 2018]), and local explanations for a given input and prediction (to which our work belongs). Two prominent examples of the latter category are LIME [Ribeiro *et al.*, 2016], which learns a linear model around the neighbourhood of an input using random local perturbations, and Anchors [Ribeiro *et al.*, 2018] (introduced in Section 3). These methods, however, do not consider robustness, making them fragile to adversarial attacks and thus insufficient to imply the prediction. Repair of non-formal explainers has been studied in [Ignatiev *et al.*, 2019d] but only for boosted trees predictors. [Narodytska *et al.*, 2019] assesses the quality of Anchors' explanations by encoding the model and explanation as a propositional formula. The explanation quality is then determined using model counting, but for binarised neural networks only. Other works that focus on binarised neural networks, Boolean classifiers or similar representations include [Shi *et al.*, 2020; Darwiche and Hirth, 2020; Darwiche, 2020]. Methods tailored to (locally) explaining NLP model decisions for a given input include [Li *et al.*, 2015; Singh *et al.*, 2018]. These identify input features, or clusters of input features, that most contribute to the prediction, using saliency and agglomerative contextual decomposition respectively. Layer-wise relevance propagation [Bach *et al.*, 2015] is also popular for NLP explanations, and is used in [Arras *et al.*, 2016; Arras *et al.*, 2017; Ding *et al.*, 2017]. Similarly to the above, these methods do not consider robustness. Robustness of neural network NLP models to adversarial examples has been studied in [Huang *et al.*, 2019; Jia *et al.*, 2019; La Malfa *et al.*, 2020]. We note that robustness verification is a different (and arguably simpler) problem from deriving a robust explanation, as the latter requires performing multiple robustness verification queries (see Section 4). Existing neural network verification approaches include symbolic (SMT) [Katz *et al.*, 2019], relaxation [Ko *et al.*, 2019; Wang *et al.*, 2018], and global optimisation [Ruan *et al.*, 2018]. Research utilising hitting sets can be seen in [Ignatiev *et al.*, 2019c], which relates explanations and adversarial examples through a generalised form of hitting set duality, and [Ignatiev *et al.*, 2019a], which works on improving model-based diagnoses by using an algorithm based on hitting sets to filter out non-subset-minimal sets of diagnoses.

3 Optimal Robust Explanations for NLP

Preliminaries. We consider a standard NLP classification task where we classify some given input text t into a plausible class y from a finite set $\mathcal{Y}$. We assume that t is a fixed length sequence of words (i.e., *features*) l, $t = (w_1, \ldots, w_l)$, where $w_i \in W$ with W being a finite vocabulary (possibly including padding). Text inputs are encoded using a continuous *word embedding* $\mathcal{E} : W \to \mathbb{R}^d$, where d is the size of the embedding [Mikolov *et al.*, 2013]. Thus, given a text $t = (w_1, \ldots, w_l)$, we define the embedding $\mathcal{E}(t)$ of t as the sequence $x = (x_{w_1}, \ldots, x_{w_l}) \in \mathbb{R}^{l \cdot d}$, where $x_{w_i} = \mathcal{E}(w_i)$. We denote with $W_{\mathcal{E}} \subseteq W$ the vocabulary used to train $\mathcal{E}$. We consider embedding vectors trained from scratch on the sentiment task, a technique that enforces words that are positively correlated to each of the output classes to be gathered closer in the embedding space [Baroni *et al.*, 2014], which is considered a good proxy for semantic similarity with respect to the target task compared to count-based embeddings [Alzantot *et al.*, 2018]. For classification we consider a *neural network* $M : \mathbb{R}^{l \cdot d} \to \mathcal{Y}$ that operates on the text embedding.

Robust Explanations. In this paper, we seek to provide *local explanations* for the predictions of a neural network NLP model. For a text embedding $x = \mathcal{E}(t)$ and a prediction $M(x)$, a local explanation E is a subset of the features of t, i.e., $E \subseteq F$ where $F = \{w_1, \ldots, w_l\}$, that is sufficient to imply the prediction. We focus on deriving *robust explanations*, i.e., on extracting a subset E of the text features F which ensure that the neural network prediction remains invariant for any perturbation of the other features $F \setminus E$. Thus, the features in a robust explanation are *sufficient to imply the prediction* that we aim to explain, a clearly desirable feature for a local explanation. In particular, we focus on explanations that are *robust w.r.t. bounded perturbations in the embedding space of the input text*. We extract word-level explanations by means of word embeddings: we note that OREs work, without further extensions, with diverse representations (e.g., sentence-level, characters-level, etc.). For a word $w \in W$, with embedding $x_w = \mathcal{E}(w)$ we denote with $\mathcal{B}(w) \subseteq \mathbb{R}^d$ a generic set of word-level perturbations. We consider the following kinds of perturbation sets, depicted also in Fig. 1.

ϵ-ball: $\mathcal{B}(w) = \{x \in \mathbb{R}^d \mid \|x - x_w\|_p \leq \epsilon\}$, for some $\epsilon > 0$ and $p > 0$. This is a standard measure of local robustness in computer vision, where ϵ-variations are interpreted as manipulations of the pixel intensity of an image. It has also been adopted in early NLP robustness works [Miyato *et al.*, 2016], but then replaced with better representations based on actual word replacements and their embeddings, see below.

k-NN box closure: $\mathcal{B}(w) = BB(\mathcal{E}(NN_k(w)))$, where $BB(X)$ is the minimum bounding box for set X; for a set $W' \subseteq W$, $\mathcal{E}(W') = \bigcup_{w' \in W'} \{\mathcal{E}(w')\}$; and $NN_k(w)$ is the set of the k closest words to w in the embedding space, i.e., words w' with smallest $d(x_w, \mathcal{E}(w'))$, where d is a valid notion of distance between embedded vectors, such as p-norms or cosine similarity, even though the box closure can be calculated for any set of embedded words. This provides an over-approximation of the k-NN convex closure, for which constraint propagation (and thus robustness checking) is more efficient [Jia *et al.*, 2019; Huang *et al.*, 2019].

For some word-level perturbation $\mathcal{B}$, set of features $E \subseteq F$, and input text t with embedding $(x_1, \ldots, x_l)$, we denote with $\mathcal{B}_E(t)$ the set of *text-level* perturbations obtained from t by keeping constant the features in E and perturbing the

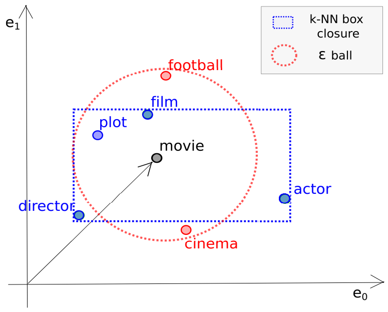

Figure 1: A graphical representation of the perturbation sets we define in the embedding space.

others according to $\mathcal{B}$:

$$\mathcal{B}_E(t) = \{(x'_1, \ldots, x'_l) \in \mathbb{R}^{l \cdot d} \mid x'_w = x_w \text{ if } w \in E; \quad x'_w \in \mathcal{B}(w) \text{ otherwise}\}. \quad (1)$$

A robust explanation $E \subseteq F$ ensures prediction invariance for any point in $\mathcal{B}_E(t)$, i.e., any perturbation (within $\mathcal{B}$) of the features in $F \setminus E$.

Def. 1 (Robust Explanation)**.** *For a text $t = (w_1, \ldots, w_l)$ with embedding $x = \mathcal{E}(t)$, word-level perturbation $\mathcal{B}$, and classifier M, a subset $E \subseteq F$ of the features of t is a* robust explanation *iff*

$$\forall x' \in \mathcal{B}_E(t).\ M(x') = M(x). \quad (2)$$

We denote (2) *with predicate* $\mathsf{Rob}_{M,x}(E)$.

Optimal Robust Explanations (OREs). While robustness is a desirable property, it is not enough alone to produce useful explanations. Indeed, we can see that an explanation E including all the features, i.e., $E = F$, trivially satisfies Definition 1. Typically, one seeks short explanations, because these can generalise to several instances beyond the input x and are easier for human decision makers to interpret. We thus introduce *optimal robust explanations (OREs)*, that is, explanations that are both robust and optimal w.r.t. an arbitrary cost function that assign a penalty to each word.

Def. 2 (Optimal Robust Explanation)**.** *Given a cost function $\mathcal{C} : W \to \mathbb{R}^+$, and for $t = (w_1, \ldots, w_l)$, x, $\mathcal{B}$, and M as in Def. 1, a subset $E^* \subseteq F$ of the features of t is an* ORE *iff*

$$E^* \in \arg\min_{E \subseteq F} \sum_{w \in E} \mathcal{C}(w) \text{ s.t. } \mathsf{Rob}_{M,x}(E). \quad (3)$$

Note that (3) is always feasible, because its feasible set always includes at least the trivial explanation $E = F$. A special case of our OREs is when $\mathcal{C}$ is *uniform* (it assigns the same cost to all words in t), in which case E^* is (one of) the *robust explanations of smallest size*, i.e., with the least number of words.

Relation with Abductive Explanations. Our OREs have similarities with the *abduction-based explanations (ABEs)* of [Ignatiev *et al.*, 2019b] in that they also derive minimal-cost explanations with robustness guarantees. For an input text $t = (w_1, \ldots, w_l)$, let $C = \bigwedge_{i=1}^{l} \chi_i = x_{w_i}$ be the *cube* representing the embedding of t, where χ_i is a variable denoting the i-th feature of x. Let $\mathcal{N}$ represent the logical encoding of the classifier M, and $\hat{y}$ be the formula representing the output of $\mathcal{N}$ given $\chi_1, \ldots, \chi_l$.

Def. 3 ([Ignatiev *et al.*, 2019b])**.** *An* abduction-based explanation (ABE) *is a minimal cost subset C^* of C such that $C^* \wedge \mathcal{N} \models \hat{y}$.*

Note that the above entailment is equivalently expressed as $C^* \models (\mathcal{N} \to \hat{y})$. Let $B = \bigwedge_{i=1}^{l} \chi_i \in \mathcal{B}(w_i)$ be the constraints encoding our perturbation space. Then, the following proposition shows that OREs can be defined in a similar abductive fashion and also in terms of *minimum satisfying assignments (MSAs)* [Dillig *et al.*, 2012]. In this way, we can derive OREs via analogous algorithms to those used for ABEs [2019b] and MSAs [Dillig *et al.*, 2012], as explained in Section 4. Moreover, we find that every ORE can be formulated as a prime implicant [Ignatiev *et al.*, 2019b], a property that connects our OREs with the notion of sufficient reason introduced in [Darwiche and Hirth, 2020].

Prop. 1. *Let E^* be an ORE and C^* its constraint encoding. Define $\phi \equiv (B \wedge \mathcal{N}) \to \hat{y}$. Then, all the following definitions apply to C^*:*

1. *C^* is a minimal cost subset of C such that $C^* \models \phi$.*
2. *C^* is a minimum satisfying assignment for ϕ.*
3. *C^* is a prime implicant of ϕ.*

Proof. See supplement[2] □

The key difference with ABEs is that our OREs are robust to *bounded* perturbations of the excluded features, while ABEs must be robust to *any* possible perturbation. This is an important difference because it is hard (often impossible) to guarantee prediction invariance w.r.t. the entire input space when this space is continuous and high-dimensional, like in our NLP embeddings. In other words, if for our NLP tasks we allowed any word-level perturbation as in ABEs, in most cases the resulting OREs will be of the trivial kind, $E^* = F$ (or $C^* = C$), and thus of little use. For example, if we consider ϵ-ball perturbations and the review *"the gorgeously elaborate continuation of the lord of the rings"*, the resulting smallest-size explanation is of the trivial kind (it contains the whole review) already at $\epsilon = 0.1$.

Exclude and include constraints. We further consider OREs E^* derived under constraints that enforce specific features F' to be included/excluded from the explanation:

$$E^* \in \arg\min_{E \subseteq F} \sum_{w \in E} \mathcal{C}(w) \text{ s.t. } \mathsf{Rob}_{M,x}(E) \wedge \phi(E), \quad (4)$$

where $\phi(E)$ is one of $F' \cap E = \emptyset$ (*exclude*) and $F' \subseteq E$ (*include*). Note that adding *include* constraints doesn't affect the feasibility of our problem, because the feasible region of (4) always contains at least the explanation $E^* \cup F'$, where E^* is a solution of (3) and F' are the features to include. See Def. 1. Conversely, *exclude* constraints might make the problem infeasible when the features in F' don't admit perturbations,

[2]For this and any subsequent reference to the supplement/appendix, please refer to the full version of the paper on the arXiv.

i.e., they are necessary for the prediction, and thus cannot be excluded. Such constraints can be easily accommodated by any solution algorithm for non-constrained OREs: for *include* ones, it is sufficient to restrict the feasible set of explanations to the supersets of F'; for *exclude* constraints, we can manipulate the cost function so as to make any explanation with features in F' strictly sub-optimal w.r.t. explanations without, that is, we use cost $\mathcal{C}'$ such that $\forall_{w \in F \setminus F'} \mathcal{C}'(w) = \mathcal{C}(w)$ and $\forall_{w' \in F'} \mathcal{C}'(w') > \sum_{w \in F \setminus F'} \mathcal{C}(w)$. The ORE obtained under cost $\mathcal{C}'$ might still include features from F', which implies that (4) is infeasible (i.e., no robust explanation without elements of F' exists).

Constrained OREs enable two crucial use cases: *detecting biased decisions*, and *enhancing non-formal explainability frameworks*.

Detecting bias. Following [Darwiche and Hirth, 2020], we deem a classifier decision *biased* if it depends on protected features, i.e., a set of input words that should not affect the decision (e.g., a movie review affected by the director's name). In particular, a decision $M(x)$ is biased if we can find, within a given set of text-level perturbations, an input x' that agrees with x on all but protected features and such that $M(x) \neq M(x')$.

Def. 4. *For classifier M, text t with features F, protected features F' and embedding $x = \mathcal{E}(t)$, decision $M(x)$ is* biased *w.r.t. some word-level perturbation $\mathcal{B}$, if*

$$\exists x' \in \mathcal{B}_{F \setminus F'}(t).M(x) \neq M(x').$$

The proposition below allows us to use exclude constraints to detect bias.

Prop. 2. *For M, t, F, F', x and $\mathcal{B}$ as per Def. 4,* decision $M(x)$ is biased iff (4) is infeasible under $F' \cap E = \emptyset$.

Proof. See supplement □

Enhancing non-formal explainers. The local explanations produced by heuristic approaches like LIME or Anchors do not enjoy the same robustness/invariance guarantees of our OREs. We can use our approach to *minimally extend* (w.r.t. the chosen cost function) any non-robust local explanation F' in order to make it robust, by solving (4) under the *include* constraint $F' \subseteq E$. In particular, with a uniform $\mathcal{C}$, our approach would identify the smallest set of extra words that make F' robust. Being minimal/smallest, such an extension retains to a large extent the original explainability properties.

Relation with Anchors. Anchors [Ribeiro *et al.*, 2018] are a state-of-the-art method for ML explanations. Given a perturbation distribution $\mathcal{D}$, classifier M and input x, an anchor A is a predicate over the input features such that $A(x)$ holds and A has high *precision* and *coverage*, defined next.

$$\mathsf{prec}(A) = \Pr_{\mathcal{D}(x' | A(x'))}(M(x) = M(x'));\ \mathsf{cov}(A) = \Pr_{\mathcal{D}(x')}(A(x')) \tag{5}$$

In other words, $\mathsf{prec}(A)$ is the probability that the prediction is invariant for any perturbation x' to which explanation A applies. In this sense, precision can be intended as a robustness probability. $\mathsf{cov}(A)$ is the probability that explanation A applies to a perturbation. To discuss the relation between Anchors and OREs, for an input text t, consider an arbitrary distribution $\mathcal{D}$ with support in $\mathcal{B}_{\emptyset}(t)$ (the set of all possible text-level perturbations), see (1); and consider anchors A defined as subsets E of the input features F, i.e., $A_E(x) = \bigwedge_{w \in E} x_w = \mathcal{E}(w)$. Then, our OREs enjoy the following properties.

Prop. 3. *If E is a robust explanation, then* $\mathsf{prec}(A_E) = 1$.

Proof. See supplement □

Note that when $\mathcal{D}$ is continuous, $\mathsf{cov}(A_E)$ is always zero unless $E = \emptyset$, in which case $\mathsf{cov}(A_\emptyset) = 1$ (as $A_\emptyset = \mathsf{true}$). Indeed, for $E \neq \emptyset$, the set $\{x' \mid A_E(x')\}$ has $|E|$ fewer degrees of freedom than the support of $\mathcal{D}$, and thus has both measure and coverage equal to zero. We thus illustrate the next property assuming that $\mathcal{D}$ is discrete (when $\mathcal{D}$ is continuous, the following still applies to any empirical approximation of $\mathcal{D}$).

Prop. 4. *If $E \subseteq E'$, then* $\mathsf{cov}(A_E) \geq \mathsf{cov}(A_{E'})$.

Proof. See supplement □

The above proposition suggests that using a uniform $\mathcal{C}$, i.e., minimizing the explanation's length, is a sensible strategy to obtain high-coverage OREs.

4 Solution Algorithms

We present two solution algorithms to derive OREs, respectively based on the hitting-set (HS) paradigm of [Ignatiev *et al.*, 2019b] and the MSA algorithm of [Dillig *et al.*, 2012]. Albeit different, both algorithms rely on repeated entailment/robustness checks $B \wedge E \wedge \mathcal{N} \models \hat{y}$ for a candidate explanation $E \subset C$. For this check, we employ two state-of-the-art neural network verification tools, Marabou [Katz *et al.*, 2019] and Neurify [Wang *et al.*, 2018]: they both give provably correct answers and, when the entailment is not satisfied, produce a counter-example $x' \in \mathcal{B}_E(t)$, i.e., a perturbation that agrees with E and such that $B \wedge C' \wedge \mathcal{N} \not\models \hat{y}$, where C' is the cube representing x'. We now briefly outline the two algorithms. A more detailed discussion (including the pseudo-code) is available in the supplement.

Minimum Hitting Set. For a counterexample C', let I' be the set of feature variables where C' does not agree with C (the cube representing the input). Then, every explanation E that satisfies the entailment must hit all such sets I' built for any counter-examples C' [Ignatiev *et al.*, 2016]. Thus, the HS paradigm iteratively checks candidates E built by selecting the subset of C whose variables form a minimum HS (w.r.t. cost $\mathcal{C}$) of said I's. However, we found that this method often struggles to converge for our NLP models, especially with large perturbations spaces (i.e., large ϵ or k). We solved this problem by extending the HS approach with a sub-routine that generates batches of *sparse adversarial attacks* for the input C. This has a two-fold benefit: 1) we reduce the number of entailment queries required to produce counter-examples, and 2) sparsity results in small I' sets, which further improves convergence.

'# this movie is really stupid and very boring most of the time there are almost no ghoulies in it at all there is nothing good about this movie on any level just more bad actors pathetically attempting to make a movie so they can get enough money to eat avoid at all costs.' (**IMDB**)	'# well I am the target market I loved it furthermore my husband also a boomer with strong memories of the 60s liked it a lot too i haven't read the book so i went into it neutral i was very pleasantly surprised its now on our highly recommended video list br br.' (**IMDB**)
'The main story ... is compelling enough but it is difficult to shrug off the annoyance of that chatty fish.' (**SST**)	'Still this flick is fun and host to some truly excellent sequences.' (**SST**)
'i couldn't bear to watch it and I thought the UA loss was embarrassing ...' (***Twitter***)	'Is delighted by the beautiful weather.' (**Twitter**)

Figure 2: OREs for IMDB, SST and Twitter datasets (all the texts are correctly classified). Models employed are FC with 50 input words each with accuracies respectively 0.89, 0.77 and 0.75. OREs are highlighted in blue. Technique used is kNN boxes with k=15.

Minimum Satisfying Assignment. This algorithm exploits the duality between MSAs and maximum universal subsets (MUSs): for cost $\mathcal{C}$ and formula $\phi \equiv (B \wedge \mathcal{N}) \rightarrow \hat{y}$, an MUS X is a set of variables with maximum $\mathcal{C}$ such that $\forall X.\phi$, which implies that $C \setminus X$ is an MSA for ϕ [Dillig *et al.*, 2012] and, in turn, an ORE. Thus, the algorithm of [Dillig *et al.*, 2012] focuses on deriving an MUS, and it does so in a recursive branch-and-bound manner, where each branch adds a feature to the candidate MUS. Such an algorithm is exponential in the worst-case, but we mitigated this by selecting a good ordering for feature exploration and performing entailment checks to rule out features that cannot be in the MUS (thus reducing the search tree).

'# I've seen Foxy Brown, Coffy Friday Foster Bucktown, and Black Mama White Mama of these this is Pam Griers worst movie poor acting bad script boring action scenes theres just nothing there avoid this and rent Friday Foster Coffy or Foxy Brown instead' (**IMDB, predicted as *negative***)
'# I gave this a 2 and it only avoided a 1 because of the occasional unintentional laugh the film is excruciatingly. Boring and incredibly cheap its even worse if you know anything at all about the Fantastic Four.', (***IMDB, predicted as negative***)
'# a few words for the people here in cine club the worst crap ever seen on this honorable cinema a very poor script a very bad actors and a very bad movie dont waste your time looking this movie see the very good or any movie have been good commented by me say no more' (**IMDB, predicted as *negative***)

Figure 3: Examples of Optimal Robust Explanations - highlighted in blue -. OREs were extracted using kNN boxes with 25 neighbors per-word: fixing words in an ORE guarantees the model to be locally robust. The examples come from the IMDB dataset, model employed is a FC network with 100 input words (accuracy 0.81).

5 Experimental Results

Settings. We have trained fully connected (FC) and convolutional neural networks (CNN) models on sentiment analysis datasets that differ in the input length and difficulty of the learning task[3]. We considered 3 well-established benchmarks for sentiment analysis, namely SST [Socher *et al.*, 2013], Twitter [Go *et al.*, 2009] and IMDB [Maas *et al.*, 2011] datasets. From these, we have chosen 40 representative input texts, balancing *positive* and *negative* examples. Embeddings are pre-trained on the same datasets used for classification [Chollet and others, 2018]. Both the HS and

[3]Experiments were parallelized on a server with two 24-core Intel Xenon 6252 processors and 256GB of RAM, but each instance is single-threaded and can be executed on a low-end laptop.

'Star/producer Salma Hayek and director Julie Taymor have infused Frida with a visual style unique and inherent to the titular character paintings and in the process created a masterful work of art of their own.' (**SST**)
'The film just might turn on many people to opera in general, an art form at once visceral and spiritual wonderfully vulgar and sublimely lofty and as emotionally grand as life.' (**SST**)
'Nah I haven't received my stimulus yet.' (**Twitter**)
ORE, FC — ORE, FC ∩ CNN — ORE, CNN

Figure 4: Comparison of OREs for SST and Twitter texts on FC (red) vs CNN (blue) models (common words in magenta). The first two are *positive* reviews, the third is *negative* (all correctly classified). Accuracies of FC and CNN models are, respectively, 0.88 and 0.89 on SST, 0.77 on Twitter. Models have input length of 25 words, OREs are extracted with kNN boxes (k=25).

MSA algorithms have been implemented in Python and use Marabou [Katz *et al.*, 2019] and Neurify [Wang *et al.*, 2018] to answer robustness queries. Marabou is fast at verifying ReLU FC networks, but it becomes memory intensive with CNNs. On the other hand, the symbolic interval analysis of Neurify is more efficient for CNNs. A downside of Neurify is that it is less flexible in the constraint definition (inputs have to be represented as squared bi-dimensional grids, thus posing problems for NLP inputs which are usually specified as 3-d tensors). In the experiments below, we opted for the kNN-box perturbation space, as we found that the k parameter was easier to interpret and tune than the ϵ parameter for the ϵ-ball space, and improved verification time. Further details on the experimental settings, including a selection of ϵ-ball results, are given in the supplement.

Effect of classifier's accuracy and robustness. We find that our approach generally results in meaningful and compact explanations for NLP. In Figure 2, we show a few OREs extracted for *negative* and *positive* texts, where the returned OREs are both concise and semantically consistent with the predicted sentiment. However, the quality of our OREs depends on that of the underlying classifier. Indeed, enhanced models with better accuracy and/or trained on longer inputs tend to produce higher quality OREs. We show this in Figures 4, 3 and 5, where we observe that enhanced models tend to result in more semantically consistent explanations. For lower-quality models, some OREs include seemingly

'# what a waste of talent a very poor semi coherent script cripples this film rather unimaginative direction too some very faint echoes of Fargo here but it just doesnt come off.' (**IMDB**)
'# a few words for the people here in cine club the worst crap ever seen on this honorable cinema a very poor script a very bad actors and a very bad movie [...]' (**IMDB**)
'I couldn't bear to watch it and I thought the UA loss was embarrassing ...' (**Twitter**)
ORE, FC 25 Inp. Words; ORE, FC 100 Inp. Words; ORE, FC 25 ∩ FC 50 ∩ FC 100; ORE, FC 50 Inp. Words; ORE, FC 25 ∩ FC 50

Figure 5: Comparison of OREs on *negative* IMDB and Twitter inputs for FC models. The first and third examples are trained with 25 (red) VS 50 (blue) input words (words in common to both OREs are in magenta). The second example further uses an FC model trained with 100 input words (words in common to all three OREs are in orange). Accuracy is respectively 0.7 and 0.77 and 0.81 for IMDB, and 0.77 for both Twitter models. All the examples are classified correctly. OREs are extracted with kNN boxes (k=25).

irrelevant terms (e.g., *"film"*, *"and"*), thus exhibiting shortcomings of the classifier.

Detecting biases. As per Prop. 2, we applied exclude constraints to detect biased decisions. In Figure 6, we provide a few example instances exhibiting such a bias, i.e., where *any* robust explanation contains at least one protected feature. These OREs include proper names that shouldn't constitute a sufficient reason for the model's classification. When we try to exclude proper names, no robust explanation exists, indicating that a decision bias exists.

Austin Powers in Goldmember has the right stuff for silly [...]' (**SST, FC 10 Input Words, k-NN (k=27)**)
'Star/producer Salma Hayek and director Julie Taymor have infused Frida [...]' (**SST, FC 10 Input Words, k-NN (k=375)**)
ORE; Words to exclude

Figure 6: Two examples of decision bias from an FC model with an accuracy of 0.80.

'This one is not nearly as dreadful as expected.' (**SST, predicted as negative**)
'Morning!! Beautiful isn't it! What you got planned for today?' (**Twitter, predicted as negative**)
ORE; ORE's *polarized words*

Figure 7: Two examples of over-sensitivity to polarized terms (in red). Other words in the OREs are highlighted in green. Models used are FC with 25 input words (accuracy 0.82 and 0.74). Method used is kNN with k respectively equal to 8 and 10.

Debugging prediction errors. An important use-case for OREs is when a model commits a *misclassification*. Misclassifications in sentiment analysis tasks usually depend on over-sensitivity of the model to polarized terms. In this sense, knowing a minimal, sufficient reason behind the model's prediction can be useful to debug it. As shown in the first example in Figure 7, the model cannot recognize the *double negation* constituted by the terms *not* and *dreadful* as a syntax construct, hence it exploits the negation term *not* to classify the review as *negative*.

Comparison to Anchors. We evaluate the robustness of Anchors for FC and CNN models on the SST and Twitter datasets: accuracies are 0.89 for FC+SST, 0.82 for FC+Twitter, 0.89 for CNN+SST, and 0.77 for CNN+Twitter. To compute robustness, we assume a kNN-box perturbation space $\mathcal{B}$ with $k = 15$ for FC and $k = 25$ for CNN models. To extract Anchors, we set $\mathcal{D}$ to the standard perturbation distribution of [Ribeiro *et al.*, 2018], defined by a set of context-wise perturbations generated by a powerful language model. Thus defined $\mathcal{B}$s are small compared to the support of $\mathcal{D}$, and so one would expect high-precision Anchors to be relatively robust w.r.t. said $\mathcal{B}$s. On the contrary, the Anchors extracted for the FC models attain an average precision of 0.996 on SST and 0.975 on Twitter, but only 12.5% of them are robust for the SST case and 7.5% for Twitter. With CNN models, high-quality Anchors are even more brittle: 0% of Anchors are robust on SST reviews and 5.4% on Twitter, despite an average precision of 0.995 and 0.971, respectively.

We remark, however, that Anchors are not designed to provide such robustness guarantees. Our approach becomes useful in this context, because it can *minimally extend* any local explanation to make it robust, by using *include constraints* as explained in Section 3. In Figure 8 we show a few examples of how, starting from non-robust Anchors explanations, our algorithm can find the minimum number of words to make them provably robust.

`The film just might turn on many people to opera in general, an art form at once visceral and spiritual wonderfully vulgar and sublimely lofty.` (***SST***)
`There are far worse messages to teach a young audience which will probably be perfectly happy with the sloppy slapstick comedy.` (***SST***)
`This one is not nearly as dreadful as expected.` (**SST**)
Anchors; Minimal Robust Extension

Figure 8: Examples of Anchors explanations (in blue) along with the minimal extension required to make them robust (in red). Examples are classified (without errors) with a 25-input-word CNN (accuracy 0.89). OREs are extracted for kNN boxes and k=25.

6 Conclusions

We have introduced optimal robust explanations (OREs) and applied them to enhance interpretability of NLP models. OREs provide concise and sufficient reasons for a particular prediction, as they are guaranteed to be both minimal w.r.t. a given cost function and robust, in that the prediction is invariant for any bounded replacement of the left-out features. We have presented two solution algorithms that build on the relation between our OREs, abduction-based explanations and minimum satisfying assignments. We have demonstrated the usefulness of our approach on widely-adopted sentiment analysis tasks, providing explanations for neural network models beyond reach for existing formal explainers. Detecting biased decisions, debugging misclassifications, and repairing non-robust explanations are some of key use cases that our OREs enable. Future research plans include exploring more general classes of perturbations beyond the embedding space.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115) and the EPSRC Programme Grant on Mobile Autonomy (EP/M019918/1).

References

[Alzantot *et al.*, 2018] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

[Arras *et al.*, 2016] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining predictions of non-linear classifiers in nlp. *arXiv preprint arXiv:1606.07298*, 2016.

[Arras *et al.*, 2017] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206*, 2017.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[Baroni *et al.*, 2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

[Chakraborty *et al.*, 2017] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[Chollet and others, 2018] François Chollet et al. Keras: The python deep learning library. *Astrophysics Source Code Library*, pages ascl–1806, 2018.

[Darwiche and Hirth, 2020] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. *arXiv preprint arXiv:2002.09284*, 2020.

[Darwiche, 2020] Adnan Darwiche. Three modern roles for logic in ai. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 229–243, 2020.

[Dillig *et al.*, 2012] Isil Dillig, Thomas Dillig, Kenneth L McMillan, and Alex Aiken. Minimum satisfying assignments for SMT. In *CAV*, volume 7358 of *LNCS*, pages 394–409. Springer, 2012.

[Ding *et al.*, 2017] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, 2017.

[Go *et al.*, 2009] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[Huang *et al.*, 2019] Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *arXiv preprint arXiv:1909.01492*, 2019.

[Ignatiev *et al.*, 2016] Alexey Ignatiev, Alessandro Previti, and Joao Marques-Silva. On finding minimum satisfying assignments. In *International Conference on Principles and Practice of Constraint Programming*, pages 287–297. Springer, 2016.

[Ignatiev *et al.*, 2019a] Alexey Ignatiev, António Morgado, Georg Weissenbacher, Joao Marques-Silva, and ISDCT SB RAS. Model-based diagnosis with multiple observations. In *IJCAI*, pages 1108–1115, 2019.

[Ignatiev *et al.*, 2019b] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1511–1519, 2019.

[Ignatiev *et al.*, 2019c] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. 2019.

[Ignatiev *et al.*, 2019d] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On validating, repairing and refining heuristic ml explanations. *arXiv preprint arXiv:1907.02509*, 2019.

[Jia *et al.*, 2019] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*, 2019.

[Katz *et al.*, 2019] Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *International Conference* *on Computer Aided Verification*, pages 443–452. Springer, 2019.

[Ko *et al.*, 2019] Ching-Yun Ko, Zhaoyang Lyu, Lily Weng, Luca Daniel, Ngai Wong, and Dahua Lin. Popqorn: Quantifying robustness of recurrent neural networks. In *International Conference on Machine Learning*, pages 3468–3477. PMLR, 2019.

[La Malfa *et al.*, 2020] Emanuele La Malfa, Min Wu, Luca Laurenti, Benjie Wang, Anthony Hartshorn, and Marta Kwiatkowska. Assessing robustness of text classification through maximal safe radius computation. *arXiv preprint arXiv:2010.02004*, 2020.

[Letham *et al.*, 2015] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[Li *et al.*, 2015] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

[Liu *et al.*, 2018] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Improving the interpretability of deep neural networks with knowledge distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 905–912. IEEE, 2018.

[Maas *et al.*, 2011] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Miyato *et al.*, 2016] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

[Narodytska *et al.*, 2019] Nina Narodytska, Aditya Shrotri, Kuldeep S Meel, Alexey Ignatiev, and Joao Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 267–278. Springer, 2019.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Ruan *et al.*, 2018] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. *arXiv preprint arXiv:1805.02242*, 2018.

[Samek *et al.*, 2017] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[Shi *et al.*, 2020] Weijia Shi, Andy Shih, Adnan Darwiche, and Arthur Choi. On tractable representations of binary neural networks. *arXiv preprint arXiv:2004.02082*, 2020.

[Singh *et al.*, 2018] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[Wang and Rudin, 2015] Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022. PMLR, 2015.

[Wang *et al.*, 2018] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. *arXiv preprint arXiv:1809.08098*, 2018.

[Zhang *et al.*, 2018] Quanshi Zhang, Yu Yang, Yuchen Liu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised learning of neural networks to explain neural networks. *arXiv preprint arXiv:1805.07468*, 2018.

[Zhang *et al.*, 2019] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6261–6270, 2019.