Certified Robustness to Adversarial Word Substitutions

Robin Jia **Aditi Raghunathan** **Kerem Göksel** **Percy Liang**
Computer Science Department, Stanford University
{robinjia,aditir,kerem,pliang}@cs.stanford.edu

Abstract

State-of-the-art NLP models can often be fooled by adversaries that apply seemingly innocuous label-preserving transformations (e.g., paraphrasing) to input text. The number of possible transformations scales exponentially with text length, so data augmentation cannot cover all transformations of an input. This paper considers one exponentially large family of label-preserving transformations, in which every word in the input can be replaced with a similar word. We train the first models that are provably robust to *all* word substitutions in this family. Our training procedure uses Interval Bound Propagation (IBP) to minimize an upper bound on the worst-case loss that any combination of word substitutions can induce. To evaluate models' robustness to these transformations, we measure accuracy on adversarially chosen word substitutions applied to test examples. Our IBP-trained models attain 75% adversarial accuracy on both sentiment analysis on IMDB and natural language inference on SNLI. In comparison, on IMDB, models trained normally and ones trained with data augmentation achieve adversarial accuracy of only 8% and 35%, respectively.

1 Introduction

Machine learning models have achieved impressive accuracy on many NLP tasks, but they are surprisingly brittle. Adding distracting text to the input (Jia and Liang, 2017), paraphrasing the text (Iyyer et al., 2018; Ribeiro et al., 2018), replacing words with similar words (Alzantot et al., 2018), or inserting character-level "typos" (Belinkov and Bisk, 2017; Ebrahimi et al., 2017) can significantly degrade a model's performance. Such perturbed inputs are called *adversarial examples*, and have shown to break models in other domains as well, most notably in vision (Szegedy et al., 2014; Goodfellow et al., 2015). Since humans are not fooled by the same perturbations, the widespread existence of adversarial examples exposes troubling gaps in models' understanding.

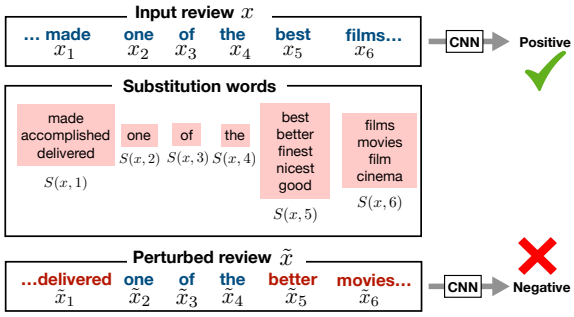

Figure 1: Word substitution-based perturbations in sentiment analysis. For an input x, we consider perturbations $\tilde{x}$, in which *every* word x_i can be replaced with any similar word from the set $S(x, i)$, without changing the original sentiment. Models can be easily fooled by adversarially chosen perturbations (e.g., changing "*best*" to "*better*", "*made*" to "*delivered*", "*films*" to "*movies*"), but the ideal model would be robust to all combinations of word substitutions.

In this paper, we focus on the word substitution perturbations of Alzantot et al. (2018). In this setting, an attacker may replace every word in the input with a similar word (that ought not to change the label), leading to an exponentially large number of possible perturbations. Figure 1 shows an example of these word substitutions. As demonstrated by a long line of work in computer vision, it is challenging to make models that are robust to very large perturbation spaces, even when the set of perturbations is known at training time (Goodfellow et al., 2015; Athalye et al., 2018; Raghunathan et al., 2018; Wong and Kolter, 2018).

Our paper addresses two key questions. First, is it possible to guarantee that a model is robust against *all* adversarial perturbations of a given in-

put? Existing methods that use heuristic search to attack models (Ebrahimi et al., 2017; Alzantot et al., 2018) are slow and cannot provide guarantees of robustness, since the space of possible perturbations is too large to search exhaustively. We obtain guarantees by leveraging Interval Bound Propagation (IBP), a technique that was previously applied to feedforward networks and CNNs in computer vision (Dvijotham et al., 2018). IBP efficiently computes a tractable *upper bound* on the loss of the worst-case perturbation. When this upper bound on the worst-case loss is small, the model is guaranteed to be robust to all perturbations, providing a *certificate* of robustness. To apply IBP to NLP settings, we derive new interval bound formulas for multiplication and softmax layers, which enable us to compute IBP bounds for LSTMs (Hochreiter and Schmidhuber, 1997) and attention layers (Bahdanau et al., 2015). We also extend IBP to handle discrete perturbation sets, rather than the continuous ones used in vision.

Second, can we train models that are robust in this way? Data augmentation can sometimes mitigate the effect of adversarial examples (Jia and Liang, 2017; Belinkov and Bisk, 2017; Ribeiro et al., 2018; Liu et al., 2019), but it is insufficient when considering very large perturbation spaces (Alzantot et al., 2018). Adversarial training strategies from computer vision (Madry et al., 2018) rely on gradient information, and therefore do not extend to the discrete perturbations seen in NLP. We instead use *certifiably robust training*, in which we train models to optimize the IBP upper bound (Dvijotham et al., 2018).

We evaluate certifiably robust training on two tasks—sentiment analysis on the IMDB dataset (Maas et al., 2011) and natural language inference on the SNLI dataset (Bowman et al., 2015). Across various model architectures (bag-of-words, CNN, LSTM, and attention-based), certifiably robust training consistently yields models which are provably robust to all perturbations on a large fraction of test examples. A normally-trained model has only 8% and 41% accuracy on IMDB and SNLI, respectively, when evaluated on adversarially perturbed test examples. With certifiably robust training, we achieve 75% adversarial accuracy for both IMDB and SNLI. Data augmentation fares much worse than certifiably robust training, with adversarial accuracies falling to 35% and 71%, respectively.

2 Setup

We consider tasks where a model must predict a label $y \in \mathcal{Y}$ given textual input $x \in \mathcal{X}$. For example, for sentiment analysis, the input x is a sequence of words $x_1, x_2, \ldots, x_L$, and the goal is to assign a label $y \in \{-1, 1\}$ denoting negative or positive sentiment, respectively. We use $z = (x, y)$ to denote an example with input x and label y, and use θ to denote parameters of a model. Let $f(z, \theta) \in \mathbb{R}$ denote some loss of a model with parameters θ on example z. We evaluate models on $f^{\text{0-1}}(z, \theta)$, the zero-one loss under model θ.

2.1 Perturbations by word substitutions

Our goal is to build models that are robust to label-preserving perturbations. In this work, we focus on perturbations where words of the input are substituted with similar words. Formally, for every word x_i, we consider a set of allowed substitution words $S(x, i)$, including x_i itself. We use $\tilde{x}$ to denote a perturbed version of x, where each word $\tilde{x}_i$ is in $S(x, i)$. For an example $z = (x, y)$, let $B_{\text{perturb}}(z)$ denote the set of *all* allowed perturbations of z:

$$B_{\text{perturb}}(z) = \{(\tilde{x}, y) : \tilde{x}_i \in S(x, i) \ \ \forall i\}. \quad (1)$$

Figure 1 provides an illustration of word substitution perturbations. We choose $S(x, i)$ so that $\tilde{x}$ is likely to be grammatical and have the same label as x (see Section 5.1).

2.2 Robustness to all perturbations

Let $\mathcal{F}(z, \theta)$ denote the set of losses of the network on the set of perturbed examples defined in (1):

$$\mathcal{F}(z, \theta) = \{f(\tilde{z}, \theta) : \tilde{z} \in B_{\text{perturb}}(z)\}. \quad (2)$$

We define the *robust loss* as $\max \mathcal{F}(z, \theta)$, the loss due to worst-case perturbation. A model is robust at z if it classifies all inputs in the perturbation set correctly, i.e., the robust zero-one loss $\max \mathcal{F}^{\text{0-1}}(z, \theta) = 0$. Unfortunately, the robust loss is often intractable to compute, as each word can be perturbed independently. For example, reviews in the IMDB dataset (Maas et al., 2011) have a median of 10^{31} possible perturbations and max of 10^{271}, far too many to enumerate. We instead propose a tractable *upper bound* by constructing a set $\mathcal{O}(z, \theta) \supseteq \mathcal{F}(z, \theta)$. Note that

$$\max \mathcal{O}^{\text{0-1}}(z, \theta) = 0 \Rightarrow \max \mathcal{F}^{\text{0-1}}(z, \theta) = 0 \Leftrightarrow \text{robust at } z. \quad (3)$$

Therefore, whenever $\max \mathcal{O}^{0\text{-}1}(z, \theta) = 0$, this fact is sufficient to *certify* robustness to all perturbed examples $B_{\text{perturb}}(z)$. However, since $\mathcal{O}^{0\text{-}1}(z, \theta) \supseteq \mathcal{F}^{0\text{-}1}(z, \theta)$, the model could be robust even if $\max \mathcal{O}^{0\text{-}1}(z, \theta) \neq 0$.

3 Certification via Interval Bound Propagation

We now show how to use Interval Bound Propagation (IBP) (Dvijotham et al., 2018) to obtain a superset $\mathcal{O}(z, \theta)$ of the losses of perturbed inputs $\mathcal{F}(z, \theta)$, given z, θ, and $B_{\text{perturb}}(z)$. For notational convenience, we drop z and θ. The key idea is to compute upper and lower bounds on the activations in each layer of the network, in terms of bounds computed for previous layers. These bounds *propagate* through the network, as in a standard forward pass, until we obtain bounds on the final output, i.e., the loss f. While IBP bounds may be loose in general, Section 5.2 shows that training networks to minimize the upper bound on f makes these bounds much tighter (Gowal et al., 2018; Raghunathan et al., 2018).

Formally, let g^i denote a scalar-valued function of z and θ (e.g., a single activation in one layer of the network) computed at node i of the computation graph for a given network. Let $\text{dep}(i)$ be the set of nodes used to compute g^i in the computation graph (e.g., activations of the previous layer). Let $\mathcal{G}^i$ denote the set of possible values of g^i across all examples in $B_{\text{perturb}}(z)$. We construct an interval $\mathcal{O}^i = [\ell^i, u^i]$ that contains all these possible values of g^i, i.e., $\mathcal{O}^i \supseteq \mathcal{G}^i$. $\mathcal{O}^i$ is computed from the intervals $\mathcal{O}^{\text{dep}(i)} = \{\mathcal{O}^j : j \in \text{dep}(i)\}$ of the dependencies of g^i. Once computed, $\mathcal{O}^i$ can then be used to compute intervals on nodes that depend on i. In this way, bounds propagate through the entire computation graph in an efficient forward pass.

We now discuss how to compute interval bounds for NLP models and word substitution perturbations. We obtain interval bounds for model inputs given $B_{\text{perturb}}(z)$ (Section 3.1), then show how to compute $\mathcal{O}^i$ from $\mathcal{O}^{\text{dep(i)}}$ for elementary operations used in standard NLP models (Section 3.2). Finally, we use these bounds to certify robustness and train robust models.

3.1 Bounds for the input layer

Previous work (Gowal et al., 2018) applied IBP to continuous image perturbations, which are naturally represented with interval bounds (Dvijotham et al., 2018). We instead work with discrete word substitutions, which we must convert into interval bounds $\mathcal{O}^{\text{input}}$ in order to use IBP. Given input words $x = x_1, \ldots, x_L$, we assume that the model embeds each word as $g^{\text{input}} = [\phi(x_1), \ldots, \phi(x_L)] \in \mathbb{R}^{L \times d}$, where $\phi(x_i) \in \mathbb{R}^d$ is the word vector for word x_i. To compute $\mathcal{O}^{\text{input}} \supseteq \mathcal{G}^{\text{input}}$, recall that each input word x_i can be replaced with any $\tilde{x}_i \in S(x, i)$. So, for each coordinate $j \in \{1, \ldots, d\}$, we can obtain an interval bound $\mathcal{O}^{\text{input}}_{ij} = [\ell^{\text{input}}_{ij}, u^{\text{input}}_{ij}]$ for g^{input}_{ij} by computing the smallest axis-aligned box that contains all the word vectors:

$$\ell^{\text{input}}_{ij} = \min_{w \in S(x,i)} \phi(w)_j, \quad u^{\text{input}}_{ij} = \max_{w \in S(x,i)} \phi(w)_j. \tag{4}$$

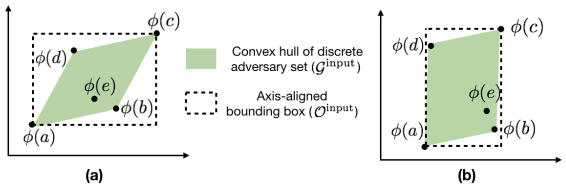

Figure 2: Bounds on the word vector inputs to the neural network. Consider a word (sentence of length one) $x = a$ with the set of substitution words $S(x, 1) = \{a, b, c, d, e\}$. (a) IBP constructs axis-aligned bounds around a set of word vectors. These bounds may be loose, especially if the word vectors are pre-trained and fixed. (b) A different word vector space can give tighter IBP bounds, if the convex hull of the word vectors is better approximated by an axis-aligned box.

Figure 2 illustrates these bounds. We can view this as relaxing a set of discrete points to a convex set that contains all of the points. Section 4.2 discusses modeling choices to make this box tighter.

3.2 Interval bounds for elementary functions

Next, we describe how to compute the interval of a node i from intervals of its dependencies. Gowal et al. (2018) show how to efficiently compute interval bounds for affine transformations (i.e., linear layers) and monotonic elementwise nonlinearities (see Appendix 3). This suffices to compute interval bounds for feedforward networks and CNNs. However, common NLP model components like LSTMs and attention also rely on softmax (for attention), element-wise multiplication (for LSTM gates), and dot product (for computing attention scores). We show how to compute interval bounds for these new operations. These building blocks can be used to compute interval bounds

not only for LSTMs and attention, but also for any model that uses these elementary functions.

For ease of notation, we drop the superscript i on g^i and write that a node computes a result $z^{\text{res}} = g(z^{\text{dep}})$ where $z^{\text{res}} \in \mathbb{R}$ and $z^{\text{dep}} \in \mathbb{R}^m$ for $m = |\text{dep}(i)|$. We are given intervals $\mathcal{O}^{\text{dep}}$ such that $z_j^{\text{dep}} \in \mathcal{O}_j^{\text{dep}} = [\ell_j^{\text{dep}}, u_j^{\text{dep}}]$ for each coordinate j and want to compute $\mathcal{O}^{\text{res}} = [\ell^{\text{res}}, u^{\text{res}}]$.

Softmax layer. The softmax function is often used to convert activations into a probability distribution, e.g., for attention. Gowal et al. (2018) uses unnormalized logits and does not handle softmax operations. Formally, let z^{res} represent the normalized score of the word at position c. We have $z^{\text{res}} = \frac{\exp(z_c^{\text{dep}})}{\sum_{j=1}^m \exp(z_j^{\text{dep}})}$. The value of z^{res} is largest when z_c^{dep} takes its largest value and all other words take the smallest value:

$$u^{\text{res}} = \frac{\exp(u_c^{\text{dep}})}{\exp(u_c^{\text{dep}}) + \sum_{j \neq c} \exp(\ell_j^{\text{dep}})}. \quad (5)$$

We obtain a similar expression for ℓ^{res}. Note that ℓ^{res} and u^{res} can each be computed in a forward pass, with some care taken to avoid numerical instability (see Appendix A.2).

Element-wise multiplication and dot product. Models like LSTMs incorporate gates which perform element-wise multiplication of two activations. Let $z^{\text{res}} = z_1^{\text{dep}} z_2^{\text{dep}}$ where $z^{\text{res}}, z_1^{\text{dep}}, z_2^{\text{dep}} \in \mathbb{R}$. The extreme values of the product occur at one of the four points corresponding to the products of the extreme values of the inputs. In other words,

$$\begin{aligned} \mathcal{C} = \{&\ell_1^{\text{dep}} \ell_2^{\text{dep}}, \quad \ell_1^{\text{dep}} u_2^{\text{dep}} \\ &u_1^{\text{dep}} \ell_2^{\text{dep}}, \quad u_1^{\text{dep}} u_2^{\text{dep}}\} \\ \ell^{\text{res}} = \min(\mathcal{C}) \quad & u^{\text{res}} = \max(\mathcal{C}). \end{aligned} \quad (6)$$

Propagating intervals through multiplication nodes therefore requires four multiplications.

Dot products between activations are often used to compute attention scores.[1] The dot product $(z_1^{\text{dep}})^\top z_2^{\text{dep}}$ is just the sum of the element-wise product $z_1^{\text{dep}} \odot z_2^{\text{dep}}$. Therefore, we can bound the dot product by summing the bounds on each element of $z_1^{\text{dep}} \odot z_2^{\text{dep}}$, using the formula for element-wise multiplication.

[1] This is distinct from an affine transformation, because both vectors have associated bounds; in an affine layer, the input has bounds, but the weight matrix is fixed.

3.3 Final layer

Classification models typically output a single logit for binary classification, or k logits for k-way classification. The final loss $f(z, \theta)$ is a function of the logits $s(x)$. For standard loss functions, we can represent this function in terms of element-wise monotonic functions (Appendix 3) and the elementary functions described in Section 3.2.

1. Zero-one loss: $f(z, \theta) = \mathbb{I}[\max(s(x)) = y]$ involves a max operation followed by a step function, which is monotonic.

2. Cross entropy: For multi-class, $f(z, \theta) = \text{softmax}(s(x))$. In the binary case, $f(z, \theta) = \sigma(s(x))$, where the sigmoid function σ is monotonic.

Thus, we can compute bounds on the loss $\mathcal{O}(z, \theta) = [\ell^{\text{final}}, u^{\text{final}}]$ from bounds on the logits.

3.4 Certifiably Robust Training with IBP

Finally, we describe certifiably robust training, in which we encourage robustness by minimizing the upper bound on the worst-case loss (Dvijotham et al., 2018; Gowal et al., 2018). Recall that for an example z and parameters θ, $u^{\text{final}}(z, \theta)$ is the upper bound on the loss $f(z, \theta)$. Given a dataset D, we optimize a weighted combination of the normal loss and the upper bound u^{final},

$$\min_\theta \sum_{z \in D} (1 - \kappa) f(z, \theta) + \kappa\, u^{\text{final}}(z, \theta), \quad (7)$$

where $0 \leq \kappa \leq 1$ is a scalar hyperparameter.

As described above, we compute u^{final} in a modular fashion: each layer has an accompanying function that computes bounds on its outputs given bounds on its inputs. Therefore, we can easily apply IBP to new architectures. Bounds propagate through layers via forward passes, so the entire objective (7) can be optimized via backpropagation.

Gowal et al. (2018) found that this objective was easier to optimize by starting with a smaller space of allowed perturbations, and make it larger during training. We accomplish this by artificially shrinking the input layer intervals $\mathcal{O}_{ij}^{\text{input}} = [\ell_{ij}^{\text{input}}, u_{ij}^{\text{input}}]$ towards the original value $\phi(x_i)_j$ by a factor of ϵ:

$$\begin{aligned} \ell_{ij}^{\text{input}} &\leftarrow \phi(x_i)_j - \epsilon(\phi(x_i)_j - \ell_{ij}^{\text{input}}) \\ u_{ij}^{\text{input}} &\leftarrow \phi(x_i)_j + \epsilon(u_{ij}^{\text{input}} - \phi(x_i)_j). \end{aligned}$$

Standard training corresponds to $\epsilon = 0$. We train for T^{init} epochs while linearly increasing ϵ from 0

to 1, and also increasing κ from 0 up to a maximum value of $\kappa^\star$, We then train for an additional T^{final} epochs at $\kappa = \kappa^\star$ and $\epsilon = 1$.

To summarize, we use IBP to compute an upper bound on the model's loss when given an adversarially perturbed input. This bound is computed in a modular fashion. We efficiently train models to minimize this bound via backpropagation.

4 Tasks and models

Now we describe the tasks and model architectures on which we run experiments. These models are all built from the primitives in Section 3.

4.1 Tasks

Following Alzantot et al. (2018), we evaluate on two standard NLP datasets: the IMDB sentiment analysis dataset (Maas et al., 2011) and the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). For IMDB, the model is given a movie review and must classify it as positive or negative. For SNLI, the model is given two sentences, a premise and a hypothesis, and is asked whether the premise entails, contradicts, or is neutral with respect to the hypothesis. For SNLI, the adversary is only allowed to change the hypothesis, as in Alzantot et al. (2018), though it is possible to also allow changing the premise.

4.2 Models

IMDB. We implemented three models for IMDB. The bag-of-words model (BOW) averages the word vectors for each word in the input, then passes this through a two-layer feedforward network with 100-dimensional hidden state to obtain a final logit. The other models are similar, except they run either a CNN or bidirectional LSTM on the word vectors, then average their hidden states. All models are trained on cross entropy loss.

SNLI We implemented two models for SNLI. The bag-of-words model (BOW) encodes the premise and hypothesis separately by summing their word vectors, then feeds the concatenation of these encodings to a 3-layer feedforward network. We also reimplement the Decomposable Attention model (Parikh et al., 2016), which uses attention between the premise and hypothesis to compute richer representations of each word in both sentences. These context-aware vectors are used in the same way BOW uses the original word vectors to generate the final prediction. Both models are trained on cross entropy loss. Implementation details are provided in Appendix A.4.

Word vector layer. The choice of word vectors affects the tightness of our interval bounds. We choose to define the word vector $\phi(w)$ for word w as the output of a feedforward layer applied to a fixed pre-trained word vector $\phi^{\text{pre}}(w)$:

$$\phi(w) = \text{ReLU}(g^{\text{word}}(\phi^{\text{pre}}(w))), \qquad (8)$$

where g^{word} is a learned linear transformation. Learning g^{word} with certifiably robust training encourages it to orient the word vectors so that the convex hull of the word vectors is close to an axis-aligned box. Note that g^{word} is applied *before* bounds are computed via (4).[2] Applying g^{word} after the bound calculation would result in looser interval bounds, since the original word vectors $\phi^{\text{pre}}(w)$ might be poorly approximated by interval bounds (e.g., Figure 2a), compared to $\phi(w)$ (e.g., Figure 2b). Section 5.7 confirms the importance of adding g^{word}. We use 300-dimensional GloVe vectors (Pennington et al., 2014) as our $\phi^{\text{pre}}(w)$.

5 Experiments

5.1 Setup

Word substitution perturbations. We base our sets of allowed word substitutions $S(x, i)$ on the substitutions allowed by Alzantot et al. (2018). They demonstrated that their substitutions lead to adversarial examples that are qualitatively similar to the original input and retain the original label, as judged by humans. Alzantot et al. (2018) define the neighbors $N(w)$ of a word w as the $n = 8$ nearest neighbors of w in a "counter-fitted" word vector space where antonyms are far apart (Mrkšić et al., 2016).[3] The neighbors must also lie within some Euclidean distance threshold. They also use a language model constraint to avoid nonsensical perturbations: they allow substituting x_i with $\tilde{x}_i \in N(x_i)$ if and only if it does not decrease the log-likelihood of the text under a pre-trained language model by more than some threshold.

We make three modifications to this approach. First, in Alzantot et al. (2018), the adversary applies substitutions one at a time, and the neighborhoods and language model scores are computed

[2] Equation (4) must be applied before the model can combine information from multiple words, but it can be delayed until after processing each word independently.

[3] Note that the model itself classifies using a different set of pre-trained word vectors; the counter-fitted vectors are only used to define the set of allowed substitution words.

relative to the current altered version of the input. This results in a hard-to-define attack surface, as changing one word can allow or disallow changes to other words. It also requires recomputing language model scores at each iteration of the genetic attack, which is inefficient. Moreover, the same word can be substituted multiple times, leading to semantic drift. We define allowed substitutions relative to the original sentence x, and disallow repeated substitutions. Second, we use a faster language model that allows us to query longer contexts; Alzantot et al. (2018) use a slower language model and could only query it with short contexts. Finally, we use the language model constraint only at test time; the model is trained against all perturbations in $N(w)$. This encourages the model to be robust to a larger space of perturbations, instead of specializing for the particular choice of language model. See Appendix A.3 for further details.

Analysis of word neighbors. One natural question is whether we could guarantee robustness by having the model treat all neighboring words the same. We could construct equivalence classes of words from the transitive closure of $N(w)$, and represent each equivalence class with one embedding. We found that this would lose a significant amount of information. Out of the 50,000 word vocabulary, 19,122 words would be in the same equivalence class, including the words "good", "bad", "excellent", and "terrible." Of the remaining words, 24,389 (79%) have no neighbors.

Baseline training methods. We compare certifiably robust training (Section 3) with both standard training and data augmentation, which has been used in NLP to encourage robustness to various types of perturbations (Jia and Liang, 2017; Belinkov and Bisk, 2017; Iyyer et al., 2018; Ribeiro et al., 2018). In data augmentation, for each training example z, we augment the dataset with K new examples $\tilde{z}$ by sampling $\tilde{z}$ uniformly from $B_{\text{perturb}}(z)$, then train on the normal cross entropy loss. For our main experiments, we use $K = 4$. We do not use adversarial training (Goodfellow et al., 2015) because it would require running an adversarial search procedure at each training step, which would be prohibitively slow.

Evaluation of robustness. We wish to evaluate robustness of models to all word substitution perturbations. Ideally, we would directly measure *robust accuracy*, the fraction of test examples z for which the model is correct on all $\tilde{z} \in B_{\text{perturb}}(z)$. However, evaluating this exactly involves enumerating the exponentially large set of perturbations, which is intractable. Instead, we compute tractable upper and lower bounds:

1. Genetic attack accuracy: Alzantot et al. (2018) demonstrate the effectiveness of a genetic algorithm that searches for perturbations $\tilde{z}$ that cause model misclassification. The algorithm maintains a "population" of candidate $\tilde{z}$'s and repeatedly perturbs and combines them. We used a population size of 60 and ran 40 search iterations on each example. Since the algorithm does not exhaustively search over $B_{\text{perturb}}(z)$, accuracy on the perturbations it finds is an *upper bound* on the true robust accuracy.

2. Certified accuracy: To complement this upper bound, we use IBP to obtain a tractable lower bound on the robust accuracy. Recall from Section 3.3 that we can use IBP to get an upper bound on the zero-one loss. From this, we obtain a *lower bound* on the robust accuracy by measuring the fraction of test examples for which the zero-one loss is guaranteed to be 0.

Experimental details. For IMDB, we split the official train set into train and development subsets, putting reviews for different movies into different splits (matching the original train/test split). For SNLI, we use the official train/development/test split. We tune hyperparameters on the development set for each dataset. Hyperparameters are reported in Appendix A.4.

5.2 Main results

Table 1 and Table 2 show our main results for IMDB and SNLI, respectively. We measure accuracy on perturbations found by the genetic attack (upper bound on robust accuracy) and IBP-certified accuracy (lower bound on robust accuracy) on 1000 random test examples from IMDB,[4] and all 9824 test examples from SNLI. Across many architectures, our models are more robust to perturbations than ones trained with data augmentation. This effect is especially pronounced on IMDB, where inputs can be hundreds of words long, so many words can be perturbed. On IMDB, the best IBP-trained model gets 75.0% accuracy on perturbations found by the genetic at-

[4] We downsample the test set because the genetic attack is slow on IMDB, as inputs can be hundreds of words long.

System	Genetic attack (Upper bound)	IBP-certified (Lower bound)
Standard training		
BOW	9.6	0.8
CNN	7.9	0.1
LSTM	6.9	0.0
Robust training		
BOW	70.5	68.9
CNN	**75.0**	**74.2**
LSTM	64.7	63.0
Data augmentation		
BOW	34.6	3.5
CNN	35.2	0.3
LSTM	33.0	0.0

Table 1: Robustness of models on IMDB. We report accuracy on perturbations obtained via the genetic attack (upper bound on robust accuracy), and certified accuracy obtained using IBP (lower bound on robust accuracy) on 1000 random IMDB test set examples. For all models, robust training vastly outperforms data augmentation ($p < 10^{-63}$, Wilcoxon signed-rank test).

System	Genetic attack (Upper bound)	IBP-certified (Lower bound)
Normal training		
BOW	40.5	2.3
DECOMPATTN	40.3	1.4
Robust training		
BOW	**75.0**	**72.7**
DECOMPATTN	73.7	72.4
Data augmentation		
BOW	68.5	7.7
DECOMPATTN	70.8	1.4

Table 2: Robustness of models on the SNLI test set. For both models, robust training outperforms data augmentation ($p < 10^{-10}$, Wilcoxon signed-rank test).

tack, whereas the best data augmentation model gets 35.2%. Normally trained models are even worse, with adversarial accuracies below 10%.

Certified accuracy. Certifiably robust training yields models with tight guarantees on robustness—the upper and lower bounds on robust accuracy are close. On IMDB, the best model is *guaranteed* to be correct on all perturbations of 74.2% of test examples, very close to the 75.0% accuracy against the genetic attack. In contrast, for data augmentation models, the IBP bound cannot guarantee robustness on almost all examples. It is possible that a stronger attack (e.g., exhaustive search) could further lower the accuracy of these models, or that the IBP bounds are loose.

LSTM models can be certified with IBP, though they fare worse than other models. IBP bounds may be loose for RNNs because of their long computation paths, along which looseness of bounds can get amplified. Nonetheless, in Appendix A.7, we show on synthetic data that robustly trained LSTMs can learn long-range dependencies.

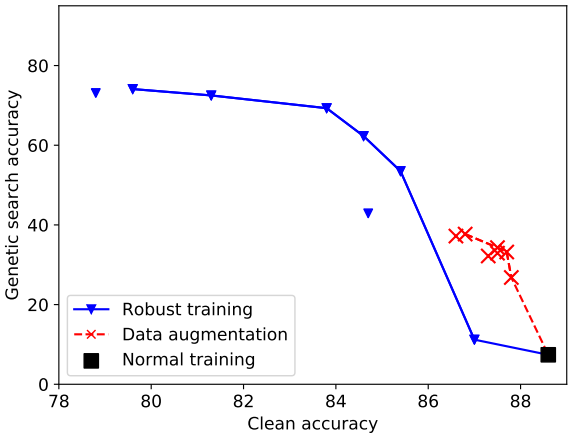

Figure 3: Trade-off between clean accuracy and genetic attack accuracy for CNN models on IMDB. Data augmentation cannot achieve high robustness. Certifiably robust training yields much more robust models, though at the cost of some clean accuracy. Lines connect Pareto optimal points for each training strategy.

5.3 Clean versus robust accuracy

Robust training does cause a moderate drop in clean accuracy (accuracy on unperturbed test examples) compared with normal training. On IMDB, our normally trained CNN model gets 89% clean accuracy, compared to 81% for the robustly trained model. We also see a drop on SNLI: the normally trained BOW model gets 83% clean accuracy, compared to 79% for the robustly trained model. Similar drops in clean accuracy are also seen for robust models in vision (Madry et al., 2017). For example, the state-of-the-art robust model on CIFAR10 (Zhang et al., 2019) only has 85% clean accuracy, but comparable normally-trained models get $> 96\%$ accuracy.

We found that the robustly trained models tend to underfit the training data—on IMDB, the CNN model gets only 86% clean training accuracy, lower than the *test* accuracy of the normally trained model. The model continued to underfit when we increased either the depth or width of the network. One possible explanation is that the attack surface adds a lot of noise, though a large enough model should still be able to overfit the training set. Better optimization or a tighter way to compute bounds could also improve training accuracy. We leave further exploration to future work.

Next, we analyzed the trade-off between clean and robust accuracy by varying the importance

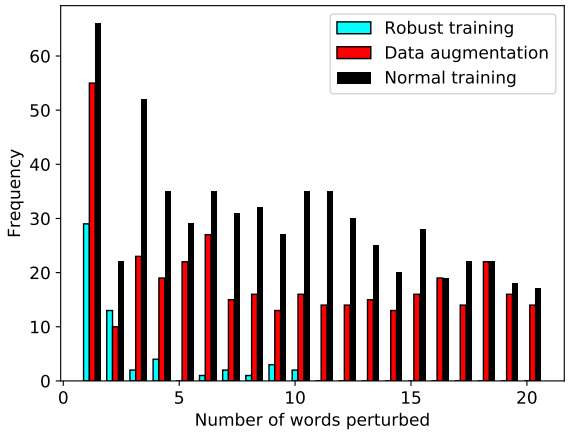

Figure 4: Number of words perturbed by the genetic attack to cause errors by CNN models on 1000 IMDB development set examples. Certifiably robust training reduces the effect of many simultaneous perturbations.

placed on perturbed examples during training. We use accuracy against the genetic attack as our proxy for robust accuracy, rather than IBP-certified accuracy, as IBP bounds may be loose for models that were not trained with IBP. For data augmentation, we vary K, the number of augmented examples per real example, from 1 to 64. For certifiably robust training, we vary $\kappa^\star$, the weight of the certified robustness training objective, between 0.01 and 1.0. Figure 3 shows trade-off curves for the CNN model on 1000 random IMDB development set examples. Data augmentation can increase robustness somewhat, but cannot reach very high adversarial accuracy. With certifiably robust training, we can trade off some clean accuracy for much higher robust accuracy.

5.4 Runtime considerations

IBP enables efficient computation of $u^{\text{final}}(z, \theta)$, but it still incurs some overhead. Across model architectures, we found that one epoch of certifiably robust training takes between $2\times$ and $4\times$ longer than one epoch of standard training. On the other hand, IBP certificates are much faster to compute at test time than genetic attack accuracy. For the robustly trained CNN IMDB model, computing certificates on 1000 test examples took 5 seconds, while running the genetic attack on those same examples took over 3 hours.

5.5 Error analysis

We examined development set examples on which models were correct on the original input but incorrect on the perturbation found by the genetic attack. We refer to such cases as *robustness errors*. We focused on the CNN IMDB models trained normally, robustly, and with data augmentation. We found that robustness errors of the robustly trained model mostly occurred when it was not confident in its original prediction. The model had $> 70\%$ confidence in the correct class for the original input in only 14% of robustness errors. In contrast, the normally trained and data augmentation models were more confident on their robustness errors; they had $> 70\%$ confidence on the original example in 92% and 87% of cases, respectively.

We next investigated how many words the genetic attack needed to change to cause misclassification, as shown in Figure 4. For the normally trained model, some robustness errors involved only a couple changed words (e.g., "*I've finally found a movie worse than . . .*" was classified negative, but the same review with "*I've finally* ***discovered*** *a movie worse than. . .*" was classified positive), but more changes were also common (e.g., part of a review was changed from "*The creature looked very cheesy*" to "*The creature* ***seemed supremely dorky***", with 15 words changed in total). Surprisingly, certifiably robust training nearly eliminated robustness errors in which the genetic attack had to change many words: the genetic attack either caused an error by changing a couple words, or was unable to trigger an error at all. In contrast, data augmentation is unable to cover the exponentially large space of perturbations that involve many words, so it does not prevent errors caused by changing many words.

5.6 Training schedule

We investigated the importance of slowly increasing ϵ during training, as suggested by Gowal et al. (2018). Fixing $\epsilon = 1$ during training led to a 5 point reduction in certified accuracy for the CNN. On the other hand, we found that holding κ fixed did not hurt accuracy, and in fact may be preferable. More details are shown in Appendix A.5.

5.7 Word vector analysis

We determined the importance of the extra feed-forward layer g^{word} that we apply to pre-trained word vectors, as described in Section 4.2. We compared with directly using pre-trained word vectors, i.e. $\phi(w) = \phi^{\text{pre}}(w)$. We also tried using g^{word} but applying interval bounds on $\phi^{\text{pre}}(w)$, then computing bounds on $\phi(w)$ with the IBP for-

mula for affine layers. In both cases, we could not train a CNN to achieve more than 52.2% certified accuracy on the development set. Thus, transforming pre-trained word vectors and applying interval bounds *after* is crucial for robust training. In Appendix A.6, we show that robust training makes the intervals around transformed word vectors smaller, compared to the pre-trained vectors.

6 Related Work and Discussion

Recent work on adversarial examples in NLP has proposed various classes of perturbations, such as insertion of extraneous text (Jia and Liang, 2017), word substitutions (Alzantot et al., 2018), paraphrasing (Iyyer et al., 2018; Ribeiro et al., 2018), and character-level noise (Belinkov and Bisk, 2017; Ebrahimi et al., 2017). These works focus mainly on demonstrating models' lack of robustness, and mostly do not explore ways to increase robustness beyond data augmentation. Data augmentation is effective for narrow perturbation spaces (Jia and Liang, 2017; Ribeiro et al., 2018), but only confers partial robustness in other cases (Iyyer et al., 2018; Alzantot et al., 2018). Ebrahimi et al. (2017) tried adversarial training (Goodfellow et al., 2015) for character-level perturbations, but could only use a fast heuristic attack at training time, due to runtime considerations. As a result, their models were still be fooled by running a more expensive search procedure at test time.

Provable defenses have been studied for simpler NLP models and attacks, particularly for tasks like spam detection where real-life adversaries try to evade detection. Globerson and Roweis (2006) train linear classifiers that are robust to adversarial feature deletion. Dalvi et al. (2004) analyzed optimal strategies for a Naive Bayes classifier and attacker, but their classifier only defends against a fixed attacker that does not adapt to the model.

Recent work in computer vision (Szegedy et al., 2014; Goodfellow et al., 2015) has sparked renewed interest in adversarial examples. Most work in this area focuses on L_∞-bounded perturbations, in which each input pixel can be changed by a small amount. The word substitution attack model we consider is similar to L_∞ perturbations, as the adversary can change each input word by a small amount. Our work is inspired by work based on convex optimization (Raghunathan et al., 2018; Wong and Kolter, 2018) and builds directly on interval bound propagation (Dvijotham et al., 2018; Gowal et al., 2018), which has certified robustness of computer vision models to L_∞ attacks. Adversarial training via projected gradient descent (Madry et al., 2018) has also been shown to improve robustness, but assumes that inputs are continuous. It could be applied in NLP by relaxing sets of word vectors to continuous regions.

This work provides certificates against word substitution perturbations for particular models. Since IBP is modular, it can be extended to other model architectures on other tasks. It is an open question whether IBP can give non-trivial bounds for sequence-to-sequence tasks like machine translation (Belinkov and Bisk, 2017; Michel et al., 2019). In principle, IBP can handle character-level typos (Ebrahimi et al., 2017; Pruthi et al., 2019), though typos yield more perturbations per word than we consider in this work. We are also interested in handling word insertions and deletions, rather than just substitutions. Finally, we would like to train models that get state-of-the-art clean accuracy while also being provably robust; achieving this remains an open problem.

In conclusion, state-of-the-art NLP models are accurate on average, but they still have significant blind spots. Certifiably robust training provides a general, principled mechanism to avoid such blind spots by encouraging models to make correct predictions on all inputs within some known perturbation neighborhood. This type of robustness is a necessary (but not sufficient) property of models that truly understand language. We hope that our work is a stepping stone towards models that are robust against an even wider, harder-to-characterize space of possible attacks.

Acknowledgments

This work was supported by NSF Award Grant no. 1805310 and the DARPA ASED program under FA8650-18-2-7882. R.J. is supported by an NSF Graduate Research Fellowship under Grant No. DGE-114747. A.R. is supported by a Google PhD Fellowship and the Open Philanthropy Project AI Fellowship. We thank Allen Nie for providing the pre-trained language model, and thank Peng Qi, Urvashi Khandelwal, Shiori Sagawa, and the anonymous reviewers for their helpful comments.

Reproducibility

All code, data, and experiments are available on Codalab at `https://bit.ly/2KVxIFN`.

References

M. Alzantot, Y. Sharma, A. Elgohary, B. Ho, M. Srivastava, and K. Chang. 2018. Generating natural language adversarial examples. In *Empirical Methods in Natural Language Processing (EMNLP)*.

A. Athalye, N. Carlini, and D. Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*.

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Y. Belinkov and Y. Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

S. Bowman, G. Angeli, C. Potts, and C. D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. 2004. Adversarial classification. In *International Conference on Knowledge Discovery and Data Mining (KDD)*.

K. Dvijotham, S. Gowal, R. Stanforth, R. Arandjelovic, B. O'Donoghue, J. Uesato, and P. Kohli. 2018. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*.

J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

A. Globerson and S. Roweis. 2006. Nightmare at test time: robust learning by feature deletion. In *International Conference on Machine Learning (ICML)*, pages 353–360.

I. J. Goodfellow, J. Shlens, and C. Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, T. Mann, and P. Kohli. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *North American Association for Computational Linguistics (NAACL)*.

R. Jia and P. Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

N. F. Liu, R. Schwartz, and N. A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *North American Association for Computational Linguistics (NAACL)*.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 2011. Learning word vectors for sentiment analysis. In *Association for Computational Linguistics (ACL)*.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2017. Towards deep learning models resistant to adversarial attacks (published at ICLR 2018). *arXiv*.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

P. Michel, X. Li, G. Neubig, and J. M. Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *North American Association for Computational Linguistics (NAACL)*.

N. Mrkšić, D. Ó Séaghdha, B. Thomson, M. Gašić, L. Rojas-Barahona, P. Su, D. Vandyke, T. Wen, and S. Young. 2016. Counter-fitting word vectors to linguistic constraints. In *North American Association for Computational Linguistics (NAACL)*.

A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*.

J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

D. Pruthi, B. Dhingra, and Z. C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Association for Computational Linguistics (ACL)*.

A. Raghunathan, J. Steinhardt, and P. Liang. 2018. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Association for Computational Linguistics (ACL)*.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

E. Wong and J. Z. Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*.

H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*.

A Supplemental material

A.1 Additional interval bound formulas

Gowal et al. (2018) showed how to compute interval bounds for affine transformations and monotonic element-wise functions. Here, we review their derivations, for completeness.

Affine transformations. Affine transformations are the building blocks of neural networks. Suppose $z^{\text{res}} = a^\top z^{\text{dep}} + b$ for weight $a \in \mathbb{R}^m$ and bias $b \in \mathbb{R}$. z^{res} is largest when positive entries of a are multiplied with u^{dep} and negative with ℓ^{dep}:

$$\begin{aligned} u^{\text{res}} &= \underbrace{0.5(a+|a|)^\top}_{\text{positive}} u^{\text{dep}} + \underbrace{0.5(a-|a|)^\top}_{\text{negative}} \ell^{\text{dep}} + b \\ &= \mu + r, \end{aligned} \tag{9}$$

where $\mu = 0.5a^\top(\ell^{\text{dep}} + u^{\text{dep}}) + b$ and $r = 0.5|a|^\top(u-l)$. A similar computation yields that $\ell^{\text{res}} = \mu - r$. Therefore, the interval $\mathcal{O}^{\text{res}}$ can be computed using two inner product evaluations: one with a and one with $|a|$.

Monotonic scalar functions. Activation functions such as ReLU, sigmoid and tanh are monotonic. Suppose $z^{\text{res}} = \sigma(z^{\text{dep}})$ where $z^{\text{res}}, z^{\text{dep}} \in \mathbb{R}$, i.e. the node applies an element-wise function to its input. The intervals can be computed trivially since z^{res} is minimized at ℓ^{dep} and maximized at u^{dep}.

$$\ell^{\text{res}} = \sigma(\ell^{\text{dep}}), \;\; u^{\text{res}} = \sigma(u^{\text{dep}}). \tag{10}$$

A.2 Numerical stability of softmax

In this section, we show how to compute interval bounds for softmax layers in a numerically stable way. We will do this by showing how to handle log-softmax layers. Note that since softmax is just exponentiated log-softmax, and exponentiation is monotonic, bounds on log-softmax directly yield bounds on softmax.

Let z^{dep} denote a vector of length m, let c be an integer $\in \{1, \ldots, m\}$, and let z^{res} represent the log-softmax score of index c, i.e.

$$z^{\text{res}} = \log \frac{\exp(z_c^{\text{dep}})}{\sum_{j=1}^m \exp(z_j^{\text{dep}})} \tag{11}$$

$$= z_c^{\text{dep}} - \log \sum_{j=1}^m \exp(z_j^{\text{dep}}). \tag{12}$$

Given interval bounds $\ell_j \le z_j^{\text{dep}} \le u_j$ for each j, we show how to compute upper and lower bounds on z^{res}. For any vector v, we assume access to a subroutine that computes

$$\text{logsumexp}(v) = \log \sum_i \exp(v_i)$$

stably. The standard way to compute this is to normalize v by subtracting $\max_i(v_i)$ before taking exponentials, then add it back at the end. logsumexp is a standard function in libraries like PyTorch. We will also rely on the fact that if v is the concatenation of vectors u and w, then $\text{logsumexp}(v) = \text{logsumexp}([\text{logsumexp}(u), \text{logsumexp}(w)])$.

Upper bound. The upper bound u^{res} is achieved by having the maximum value of z_c^{dep}, and minimum value of all others. This can be written as:

$$u^{\text{res}} = u_c^{\text{dep}} - \log \left(\exp(u_c^{\text{dep}}) + \sum_{1 \le j \le m, j \ne c} \exp(\ell_j^{\text{dep}}) \right). \tag{13}$$

While we could directly compute this expression, it is difficult to vectorize. Instead, with some rearranging, we get

$$u^{\text{res}} = u_c^{\text{dep}} - \log \left(\exp(u_c^{\text{dep}}) - \exp(\ell_c^{\text{dep}}) + \sum_{j=1}^m \exp(\ell_j^{\text{dep}}) \right). \tag{14}$$

The second term is the logsumexp of

$$\log \left(\exp(u_c^{\text{dep}}) - \exp(\ell_c^{\text{dep}}) \right) \tag{15}$$

and

$$\text{logsumexp}(\ell^{\text{dep}}). \tag{16}$$

Since we know how to compute logsumexp, this reduces to computing (15). Note that (15) can be rewritten as

$$u_c^{\text{dep}} + \log\left(1 - \exp(\ell_c^{\text{dep}} - u_c^{\text{dep}})\right) \tag{17}$$

by adding and subtracting u_c^{dep}. To compute this quantity, we consider two cases:

1. $u_c^{\text{dep}} \gg \ell_c^{\text{dep}}$. Here we use the fact that stable methods exist to compute $\text{log1p}(x) = \log(1+x)$ for x close to 0. We compute the desired value as

$$u_c^{\text{dep}} + \log 1p(-\exp(\ell_c^{\text{dep}} - u_c^{\text{dep}})),$$

since $\exp(\ell_c^{\text{dep}} - u_c^{\text{dep}})$ will be close to 0.

2. u_c^{dep} close to ℓ_c^{dep}. Here we use the fact that stable methods exist to compute $\text{expm1}(x) = \exp(x) - 1$ for x close to 0. We compute the desired value as

$$u_c^{\text{dep}} + \log(-\text{expm1}(\ell_c^{\text{dep}} - u_c^{\text{dep}})),$$

since $\ell_c^{\text{dep}} - u_c^{\text{dep}}$ may be close to 0.

We use case 1 if $u_c^{\text{dep}} - \ell_c^{\text{dep}} > \log 2$, and case 2 otherwise.[5]

Lower bound. The lower bound ℓ^{res} is achieved by having the minimum value of z_c^{dep}, and the maximum value of all others. This can be written as:

$$\ell^{\text{res}} = \ell_c^{\text{dep}} - \log\left(\exp(\ell_c^{\text{dep}}) + \sum_{1 \le j \le m, j \neq c} \exp(u_j^{\text{dep}})\right). \tag{18}$$

The second term is just a normal logsumexp, which is easy to compute. To vectorize the implementation, it helps to first compute the logsumexp of everything except ℓ_c^{dep}, and then logsumexp that with ℓ_c^{dep}.

A.3 Attack surface differences

In Alzantot et al. (2018), the adversary applies replacements one at a time, and the neighborhoods and language model scores are computed relative to the current altered version of the input. This results in a hard-to-define attack surface, as the same word can be replaced many times, leading to semantic drift. We instead pre-compute the allowed substitutions $S(x, i)$ at index i based on the original x. We define $S(x, i)$ as the set of $\tilde{x}_i \in N(x_i)$ such that

$$\log P(x_{i-W:i-1}, \tilde{x}_i, x_{i+1:i+W}) \geq \log P(x_{i-W:i+W}) - \delta \tag{19}$$

where probabilities are assigned by a pre-trained language model, and the window radius W and threshold δ are hyperparameters. We use $W = 6$ and $\delta = 5$. We also use a different language model[6] from Alzantot et al. (2018) that achieves perplexity of 50.79 on the One Billion Word dataset (Chelba et al., 2013). Alzantot et al. (2018) use a different, slower language model, which compels them to use a smaller window radius of $W = 1$.

A.4 Experimental details

We do not run training for a set number of epochs but do early stopping on the development set instead. For normal training, we early stop on normal development set accuracy. For training with data augmentation, we early stop on the accuracy on the augmented development set. For certifiably robust training, we early stop on the certifiably robust accuracy on the development set. We use the Adam optimizer (Kingma and Ba, 2014) to train all models.

On IMDB, we restrict the model to only use the 50,000 words that are in the vocabulary of the counter-fitted word vector space of Mrkšić et al. (2016). This is because perturbations are not allowed for any words not in this vocabulary, i.e. $N(w) = \{w\}$ for $w \notin V$. Therefore, the model is strongly incentivized to predict based on words outside of this set. While this is a valid way to achieve high certified accuracy, it is not a valid robustness strategy in general. We simply delete all words that are not in the vocabulary before feeding the input to the model.

For SNLI, we use 100-dimensional hidden state for the BOW model and a 3-layer feedforward network. These values were chosen by a hyperparameter search on the dev set. For DECOMPATTN, we use a 300-dimensional hidden state and a 2-layer feedforward network on top of the context-aware vectors. These values were chosen to match Parikh et al. (2016).

[5] See https://cran.r-project.org/web/packages/Rmpfr/vignettes/log1mexp-note.pdf for further explanation.

[6] https://github.com/windweller/l2w

System	κ	Learning Rate	Dropout Prob.	Weight Decay	Gradient Norm Clip Val.	T^{init}
IMDB, BOW	0.8	1×10^{-3}	0.2	1×10^{-4}	0.25	40
IMDB, CNN	0.8	1×10^{-3}	0.2	1×10^{-4}	0.25	40
IMDB, LSTM	0.8	1×10^{-3}	0.2	1×10^{-4}	0.25	20
SNLI, BOW	0.5	5×10^{-4}	0.1	1×10^{-4}	0.25	35
SNLI, DECOMPATTN	0.5	1×10^{-4}	0.1	0	0.25	50

Table 3: Training hyperparameters for training the models. The same hyperparameters were used for all training settings(plain, data augmentation, robust training)

Our implementation of the Decomposable Attention follows the original described in (Parikh et al., 2016) except for a few differences listed below;

- We do not normalize GloVe vectors to have norm 1.
- We do not hash out-of-vocabulary words to randomly generated vectors that we train, instead we omit them.
- We do randomly generate a null token vector that we then train. (Whether the null vector is trained is unspecified in the original paper).
- We use the Adam optimizer (with a learning rate of 1×10^{-4}) instead of AdaGrad.
- We use a batch size of 256 instead of 4.
- We use a dropout probability of 0.1 instead of 0.2
- We do not use the intra-sentence attention module.

A.5 Training schedule

In Table 4, we show the effect of holding ϵ or κ fixed during training, as described in Section 5.6. All numbers are on 1000 randomly chosen examples from the IMDB development set. Slowly increasing ϵ is important for good performance. Slowly increasing κ is actually slightly worse than holding $\kappa = \kappa^*$ fixed during training, despite earlier experiments we ran suggesting the opposite. Here we only report certified accuracy, as all models are trained with certifiably robust training, and certified accuracy is much faster to compute for development purposes.

A.6 Word vector bound sizes

To better understand the effect of g^{word}, we checked whether g^{word} made interval bound boxes around neighborhoods $N(w)$ smaller. For each word w with $|N(w)| > 1$, and for both the pre-trained vectors $\phi^{\text{pre}}(\cdot)$ and transformed vectors $\phi(\cdot)$, we compute

System	IBP-certified (Lower bound)
BOW	68.8
$\rightarrow$ Fixed ϵ	46.6
$\rightarrow$ Fixed κ	69.8
$\rightarrow$ Fixed ϵ and κ	66.3
CNN	72.5
$\rightarrow$ Fixed ϵ	67.6
$\rightarrow$ Fixed κ	74.5
$\rightarrow$ Fixed ϵ and κ	65.3
LSTM	62.5
$\rightarrow$ Fixed ϵ	43.7
$\rightarrow$ Fixed κ	63.0
$\rightarrow$ Fixed ϵ and κ	62.0

Table 4: Effects of holding ϵ and κ fixed during training. All numbers are on 1000 randomly chosen IMDB development set examples.

$$\frac{1}{d}\sum_{i=1}^{d}\frac{1}{\sigma_i}\left(u_w^{\text{word}} - \ell_w^{\text{word}}\right)$$

where ℓ_w^{word} and u_w^{word} are the interval bounds around either $\{\phi^{\text{pre}}(\tilde{w}) : \tilde{w} \in N(w)\}$ or $\{\phi(\tilde{w}) : \tilde{w} \in N(w)\}$, and σ_i is the standard deviation across the vocabulary of the i-th coordinate of the embeddings. This quantity measures the average width of the IBP bounds for the word vectors of w and its neighbors, normalized by the standard deviation in each coordinate. On 78.2% of words with $|N(w)| > 1$, this value was smaller for the transformed vectors learned by the CNN on IMDB with robust training, compared to the GloVe vectors. For same model with normal training, the value was smaller only 54.5% of the time, implying that robust training makes the transformation produce tighter bounds. We observed the same pattern for other model architectures as well.

A.7 Certifying long-term memory

We might expect that LSTMs are difficult to certify with IBP, due to their long computation paths. To test whether robust training can learn recurrent models that track state across many time steps, we created a toy binary classification task where the input is a sequence of words $x_1, \ldots, x_L$, and the label y is 1 if $x_1 = x_L$ and 0 otherwise. We trained an LSTM model that reads the input left-to-right, and tries to predict y with a two-layer feedforward network on top of the final hidden state. To do this task, the model must encode the first word in its state and remember it until the final timestep; a bag of words model cannot do this task. For perturbations, we allow replacing every middle word $x_2, \ldots, x_{L-1}$ with any word in the vocabulary. We use robust training on 4000 randomly generated examples, where the length of each example is sampled uniformly between 3 and 10. The model obtains 100% certified accuracy on a test set of 1000 examples, confirming that robust training can learn models that track state across many time steps.

For this experiment, we found it important to first train for multiple epochs with no certified objective, before increasing ϵ and κ. Otherwise, the model gets stuck in bad local optima. We trained for 50 epochs using the normal objective, 50 epochs increasing ϵ towards 1 and κ towards 0.5, then 17 final epochs (determined by early stopping) with these final values of ϵ and κ.[7] We leave further exploration of these learning schedule tactics to future work. We also found it necessary to use a larger LSTM—we used one with 300-dimensional hidden states.

[7] Note that this dataset is much smaller than IMDB and SNLI, so each epoch corresponds to many fewer parameter updates.

B Adversarial examples

In this additional supplementary material, we show randomly chosen adversarial examples found by the genetic attack. We show examples for three different models: the CNN model on IMDB trained normally, with certifiably robust training, and with data augmentation. For each model, we picked ten random development set examples for which the model was correct on the original example, but wrong after the genetic attack. Changed words are marked in bold.

Normally trained model, example 1

Original: The original is a relaxing watch , with some truly memorable animated sequences . Unfortunately , the sequel , while not the worst of the DTV sequels completely lacks the sparkle . The **biggest** letdown is a lack of a story . Like Belle 's Magical World , the characters are told through a series of vignettes . Magical World , while marginally **better** , still manages to make a mess of the story . In between the vignettes , we see the mice at work , and I personally think the antics of Jaq and Gus are the redeeming merits of this movie . The first vignette is the **best** , about Cinderella getting used to being to being a princess . This is the best , because the mice were at their funniest here . The **worst** of the vignettes , when Jaq turns into a human , is cute at times , but **has** a lack of imagination . The last vignette , when Anastasia falls in love , was also cute . The problem was , I could n't imagine Anastasia being friendly with Cinderella , as I considered her the meaner out of the stepsisters . This was also **marred** by a rather ridiculous subplot about Lucifer falling in love with PomPom . The incidental music was very pleasant to listen to ; however I hated the songs , they were really uninspired , and nothing like the beautiful Tchaikovsky inspired melodies of the original . The characters were the strongest development here . Cinderella while still caring , had lost her sincerity , and a lot of her charm from the original , though she does wear some very pretty clothes . The Duke had some truly funny moments but they were n't enough to save the **film** , likewise with Prudence and the king . As I mentioned , the mice were the redeeming merits of the movie , as they alone contributed to the film 's cuteness . I have to say also the animation is colourful and above average , and the voice acting was surprisingly good . All in all , a cute , if unoriginal sequel , that was **marred** by the songs and a lack of a story . 4/10 for the mice , the **voice** acting , the animation and some pretty dresses . Bethany Cox

Perturbed: The original is a relaxing watch , with some truly memorable animated sequences . Unfortunately , the sequel , while not the worst of the DTV sequels completely lacks the sparkle . The **greatest** letdown is a lack of a story . Like Belle 's Magical World , the characters are told through a series of vignettes . Magical World , while marginally **nicer** , still manages to make a mess of the story . In between the vignettes , we see the mice at work , and I personally think the antics of Jaq and Gus are the redeeming merits of this movie . The first vignette is the **finest** , about Cinderella getting used to being to being a princess . This is the best , because the mice were at their funniest here . The **toughest** of the vignettes , when Jaq turns into a human , is cute at times , but **possesses** a lack of imagination . The last vignette , when Anastasia falls in love , was also cute . The problem was , I could n't imagine Anastasia being friendly with Cinderella , as I considered her the meaner out of the stepsisters . This was also **tempered** by a rather ridiculous subplot about Lucifer falling in love with PomPom . The incidental music was very pleasant to listen to ; however I hated the songs , they were really uninspired , and nothing like the beautiful Tchaikovsky inspired melodies of the original . The characters were the strongest development here . Cinderella while still caring , had lost her sincerity , and a lot of her charm from the original , though she does wear some very pretty clothes . The Duke had some truly funny moments but they were n't enough to save the **cinema** , likewise with Prudence and the king . As I mentioned , the mice were the redeeming merits of the movie , as they alone contributed to the film 's cuteness . I have to say also the animation is colourful and above average , and the voice acting was surprisingly good . All in all , a cute , if unoriginal sequel , that was **tempered** by the songs and a lack of a story . 4/10 for the mice , the **voices** acting , the animation and some pretty dresses . Bethany Cox

Correct label: negative.

Model confidence on original example: 90.3.

Normally trained model, example 2

Original: When I was younger , I liked this show , but now ... BLECCH ! ! ! This show is sappy , **badly** written , and rarely **funny** . The three leads **were** all good actors and **funny** men (Saget 's stand up was a lot better than the stuff this show came up with , as was Coulier a better **stand** up , and Stamos was a better than average actor) . After a while , Stamos wanted off the show because it wanted to do more serious stuff (who could blame him ?) . The show eventually got cancelled when many of the actors demanded more money . Here are a few things that drive me crazy about the show : 1 . The **catch** phrases- How many times can one person put up with tiring **catch** phrases like with 'how rude ' , 'you got it **dude** ' , 'nerdbomber ' , 'cut it out ' and 'have mercy ' in a 24 hour time period ? 2 . Kimmy Gibler- the most **annoying** character ever written for television . 3 . The writing- **stale** and cliched as an oreo **cookie** . There is good **cliched** writing and **bad cliched** writing . Full House had **bad** cliched writing . 4 . Three men living together in San Francisco- Enough said . 5 . Unrealistic stuff- Too much to recall . 6 . Trendy kids- The **girls** had all the **latest** mall fashions and you can see posters of trendy recording **artists** they would be into . Now this show is on Nick @ Nite . I would hardly call it a classic . I have nothing bad to say about the people involved since I think many of them are **talented** in their own right . But this show was just so sugary sweet , I could n't stand it after a while .

Perturbed: When I was younger , I liked this show , but now ... BLECCH ! ! ! This show is sappy , **desperately** written , and rarely **hilarious** . The three leads **was** all good actors and **hilarious** men (Saget 's stand up was a lot better than the stuff this show came up with , as was Coulier a better **stands** up , and Stamos was a better than average actor) . After a while , Stamos wanted off the show because it wanted to do more serious stuff (who could blame him ?) . The show eventually got cancelled when many of the actors demanded more money . Here are a few things that drive me crazy about the show : 1 . The **captures** phrases- How many times can one person put up with tiring **captures** phrases like with 'how rude ' , 'you got it **bro** ' , 'nerdbomber ' , 'cut it out ' and 'have mercy ' in a 24 hour time period ? 2 . Kimmy Gibler- the most **exasperating** character ever written for television . 3 . The writing- **obsolete** and cliched as an oreo **cookies** . There is good **corny** writing and **wicked corny** writing . Full House had **wicked** cliched writing . 4 . Three men living together in San Francisco- Enough said . 5 . Unrealistic stuff- Too much to recall . 6 . Trendy kids- The **daughters** had all the **last** mall fashions and you can see posters of trendy recording **artistes** they would be into . Now this show is on Nick @ Nite . I would hardly call it a classic . I have nothing bad to say about the people involved since I think many of them are **genius** in their own right . But this show was just so sugary sweet , I could n't stand it after a while .

Correct label: negative.
Model confidence on original example: 98.9.

Normally trained model, example 3

Original: Jim Carrey **shines** in this **beautiful** movie . This is **now** one of my **favorite** movies . I read all about the making and I **thought** it was **incredible** how the **did** it . I **ca** n't wait **till** this comes out on DVD . I **saw** this in **theaters so** many **times** , I **ca** n't even count how **times** I 've **seen** it .

Perturbed: Jim Carrey **stars** in this **handsome** movie . This is **currently** one of my **favourite** movies . I read all about the making and I **figured** it was **unthinkable** how the **am** it . I **could** n't wait **unless** this comes out on DVD . I **watched** this in **theatres too** many **time** , I **could** n't even count how **period** I 've **watched** it .

Correct label: positive.
Model confidence on original example: 100.0.

Normally trained model, example 4

Original: Did anyone read the script . This has to be some of the **worst** writing and directing of the entire year . Three great **actors** , Paul Giamatti , Rachel Weisz and Miranda Richardson could n't pull this one out . About two-thirds it looked like Giamatti eyes were **saying** , I ca n't believe I signed the contract . It 's not the **worst** movie I ever saw , but it 's on the **really** really **bad** Christmas movie list . Not enough lines , but what else can be said ? Okay , the movie just does n't move with Vaughn 's con-man dialogue , his character is just a **creepy guy** that you **just** ca n't get **past** . It was just a **lackluster** walk **through** , that no one seemed to be able to get into .

Perturbed: Did anyone read the script . This has to be some of the **toughest** writing and directing of the entire year . Three great **protagonists** , Paul Giamatti , Rachel Weisz and Miranda Richardson could n't pull this one out . About two-thirds it looked like Giamatti eyes were **telling** , I ca n't believe I signed the contract . It 's not the **toughest** movie I ever saw , but it 's on the **truly** really **wicked** Christmas movie list . Not enough lines , but what else can be said ? Okay , the movie just does n't move with Vaughn 's con-man dialogue , his character is just a **terrifying buddy** that you **only** ca n't get **last** . It was just a **puny** walk **throughout** , that no one seemed to be able to get into .

Correct label: negative.
Model confidence on original example: 100.0.

Normally trained model, example 5

Original: Yes I have rated this film as one star **awful** . Yet , it will be in my rotation of Christmas **movies** henceforth . This truly is so bad it 's good . This is another K.Gordon Murray **production** (read : buys a really cheap/bad Mexican movie , spends **zero** money getting it dubbed into English and releases it at kiddie matines in the mid 1960 's .) It 's a shame I stumbled on this so late in life as I 'm sure some " mood enhancers " would make this an even better experience . I 'm not going to rehash what so many of the other reviewers have already said , a Christmas movie with Merlin , the Devil , mechanical wind-up reindeer and some of the most **pathetic** child actors I have ever seen bar none . I plan on running this over the holidays back to back with Kelsey Grammar 's " A Christmas Carol " . Truly a **holiday** experience made in Hell . Now if I can only find " To All A Goodnight (aka Slayride) " on DVD I 'll have a triple feature that ca n't be beat . You have to **see** this movie . It moves so slowly that I defy you not to touch the fast forward button-especially on the two dance routines ! This thing reeks like an expensive bleu cheese-guess you have to get past the **stink** to enjoy the experience . Feliz Navidad amigos !

Perturbed: Yes I have rated this film as one star **horrifying** . Yet , it will be in my rotation of Christmas **cinema** henceforth . This truly is so bad it 's good . This is another K.Gordon Murray **producing** (read : buys a really cheap/bad Mexican movie , spends **nought** money getting it dubbed into English and releases it at kiddie matines in the mid 1960 's .) It 's a shame I stumbled on this so late in life as I 'm sure some " mood enhancers " would make this an even better experience . I 'm not going to rehash what so many of the other reviewers have already said , a Christmas movie with Merlin , the Devil , mechanical wind-up reindeer and some of the most **lamentable** child actors I have ever seen bar none . I plan on running this over the holidays back to back with Kelsey Grammar 's " A Christmas Carol " . Truly a **festive** experience made in Hell . Now if I can only find " To All A Goodnight (aka Slayride) " on DVD I 'll have a triple feature that ca n't be beat . You have to **admire** this movie . It moves so slowly that I defy you not to touch the fast forward button-especially on the two dance routines ! This thing reeks like an expensive bleu cheese-guess you have to get past the **scent** to enjoy the experience . Feliz Navidad amigos !

Correct label: negative.
Model confidence on original example: 97.0.

Normally trained model, example 6

Original: This show is quick-witted , **colorful** , **dark** yet fun , hip and **still** somehow clean . The cast , including an awesome rotation of special **guests** (i.e . Molly Shannon , Paul Rubens , The-Stapler-Guy-From-Office-Space) is **electric** . It 's got murder , romance , **family** , AND zombies without ever coming off as cartoony ... Somehow . You really connect with these characters . The whole production is an unlikely **magic** act that left me , something of a skeptic if I do say so myself , totally engrossed and **coming** back for more every Wednesday night . I **just** re-read this and it sounds a little like somebody paid me to write it . It really is that good . I just heard a **rumor** that it was being canceled so I **thought** I 'd send off a flare of good **will** . This is one of those shows that goes under the radar because the network suits ca n't figure out how to make it sexy and sell cars with it . Do yourself a huge favor , if you have n't already , and enjoy this gem while it lasts . OK so one more thing . This show is clever . What that means is that every armchair critic/ " writer " in Hollywood is gon na insert a stick up their youknowwhat before they sit down to watch it , defending themselves with an " I could 've written that " **type** speech to absolutely nobody in their lonely renovated Hollywood hotel room . In other words : the internet . This is a general interest/anonymous website . Before you give your Wednesday TV hour to Dirty Sexy Money or Next Hot Model reruns or whatever other out and out tripe these internet " critics " are n't commenting on , **give** my fave ' show a spin . It 's fun . Good , unpretentious fun .

Perturbed: This show is quick-witted , **colored** , **darkened** yet fun , hip and **even** somehow clean . The cast , including an awesome rotation of special **guest** (i.e . Molly Shannon , Paul Rubens , The-Stapler-Guy-From-Office-Space) is **electricity** . It 's got murder , romance , **relatives** , AND zombies without ever coming off as cartoony ... Somehow . You really connect with these characters . The whole production is an unlikely **hallucinogenic** act that left me , something of a skeptic if I do say so myself , totally engrossed and **come** back for more every Wednesday night . I **merely** re-read this and it sounds a little like somebody paid me to write it . It really is that good . I just heard a **rumour** that it was being canceled so I **figured** I 'd send off a flare of good **willpower** . This is one of those shows that goes under the radar because the network suits ca n't figure out how to make it sexy and sell cars with it . Do yourself a huge favor , if you have n't already , and enjoy this gem while it lasts . OK so one more thing . This show is clever . What that means is that every armchair critic/ " writer " in Hollywood is gon na insert a stick up their youknowwhat before they sit down to watch it , defending themselves with an " I could 've written that " **typing** speech to absolutely nobody in their lonely renovated Hollywood hotel room . In other words : the internet . This is a general interest/anonymous website . Before you give your Wednesday TV hour to Dirty Sexy Money or Next Hot Model reruns or whatever other out and out tripe these internet " critics " are n't commenting on , **lend** my fave ' show a spin . It 's fun . Good , unpretentious fun .

Correct label: positive.

Model confidence on original example: 93.7.

Normally trained model, example 7

Original: This was a hit in the South By Southwest (SXSW) Film festival in Austin last year , and features a fine **cast** headed up by E.R . 's Gloria Reuben , and a scenery-chewing John Glover . Though **shot** on a **small** budget in NYC , the film looks and **sounds** fabulous , and takes us on a behind the scenes whirl through the **rehearsal** and mounting of what **actors** call " The Scottish Play , " as a reference to the word " Macbeth " is **thought** to bring on the **play** 's **ancient** curse . The acting company exhibits all the **emotions** of the play **itself** , lust , jealousy , rage , suspicion , and a bit of fun as well . The **games begin** when an accomplished actor is replaced (in the lead role) by a well-known " pretty face " from the TV soap opera scene in order to draw bigger crowds . The green-eyed monster takes over from there , and the **drama** unfolds **nicely** . Fine soundtrack , and good performances all around . The DVD **includes** director 's **commentary** and some deleted scenes as **well** .

Perturbed: This was a hit in the South By Southwest (SXSW) Film festival in Austin last year , and features a fine **casting** headed up by E.R . 's Gloria Reuben , and a scenery-chewing John Glover . Though **murdered** on a **tiny** budget in NYC , the film looks and **sound** fabulous , and takes us on a behind the scenes whirl through the **repeating** and mounting of what **actresses** call " The Scottish Play , " as a reference to the word " Macbeth " is **thinking** to bring on the **toy** 's **old** curse . The acting company exhibits all the **thrills** of the play **yourselves** , lust , jealousy , rage , suspicion , and a bit of fun as well . The **play starts** when an accomplished actor is replaced (in the lead role) by a well-known " pretty face " from the TV soap opera scene in order to draw bigger crowds . The green-eyed monster takes over from there , and the **theatre** unfolds **politely** . Fine soundtrack , and good performances all around . The DVD **contains** director 's **remark** and some deleted scenes as **good** .

Correct label: positive.
Model confidence on original example: 97.8.

Normally trained model, example 8

Original: As a young black/latina woman I am always **searching** for movies that represent the experiences and lives of people like me . Of course when I saw this movie at the video store I thought I would enjoy it ; unfortunately , I did n't . Although the **topics** presented in the film are interesting and relevant , the story was simply not properly developed . The movie just kept dragging on and on and many of the characters that appear on screen just come and go without much to contribute to the overall film . Had the director done a better job interconnecting the scenes , perhaps I would have enjoyed it a bit more . Honestly , I would recommend a film like " Raising Victor " **over** this one any day . I just was not too impressed .

Perturbed: As a young black/latina woman I am always **browsing** for movies that represent the experiences and lives of people like me . Of course when I saw this movie at the video store I thought I would enjoy it ; unfortunately , I did n't . Although the **themes** presented in the film are interesting and relevant , the story was simply not properly developed . The movie just kept dragging on and on and many of the characters that appear on screen just come and go without much to contribute to the overall film . Had the director done a better job interconnecting the scenes , perhaps I would have enjoyed it a bit more . Honestly , I would recommend a film like " Raising Victor " **finished** this one any day . I just was not too impressed .

Correct label: negative.
Model confidence on original example: 65.4.

Normally trained model, example 9

Original: If Fassbinder has made a worse film , I sure do n't want to see it ! Anyone who complains that his films are too talky and claustrophobic should be forced to view this , to learn to appreciate the more spare style he opted for in excellent films like " The Bitter Tears Of Petra von Kant " . This film bogs down with so much arty , quasi-symbolic images it looks like a parody of an " art-film " . The scene in the slaughterhouse and the scene where Elvira 's prostitute friend channel-surfs for what seems like ten minutes are just two of the most glaring examples of what makes this film a real test of the viewer 's endurance . But what really angers me about it are the few scenes which feature just Elvira and her ex-wife and/or her daughter . These are the only moments that display any real human emotion , and prove that at the core of this **horrible** film , there was an excellent film struggling to free itself . What a waste .
Perturbed: If Fassbinder has made a worse film , I sure do n't want to see it ! Anyone who complains that his films are too talky and claustrophobic should be forced to view this , to learn to appreciate the more spare style he opted for in excellent films like " The Bitter Tears Of Petra von Kant " . This film bogs down with so much arty , quasi-symbolic images it looks like a parody of an " art-film " . The scene in the slaughterhouse and the scene where Elvira 's prostitute friend channel-surfs for what seems like ten minutes are just two of the most glaring examples of what makes this film a real test of the viewer 's endurance . But what really angers me about it are the few scenes which feature just Elvira and her ex-wife and/or her daughter . These are the only moments that display any real human emotion , and prove that at the core of this **gruesome** film , there was an excellent film struggling to free itself . What a waste .

Correct label: negative.
Model confidence on original example: 65.8.

Normally trained model, example 10

Original: This film is the **worst** excuse for a motion picture I have EVER seen . To begin , I 'd **like** to say the the front cover of this film is by all means misleading , if you think you are about to see a truly **scary** horror film with a monster clown , you are soooo wrong . In fact the killers face does n't even slightly resemble the front cover , it 's just an image they must have found on Google and thought it looked **cool** . Speaking of things they **found** and thought it looked cool , there is a scene in this film where some of the gang are searching for the friend in the old woods , then suddenly the screen chops to a scene where there is a mother deer nurturing it 's young in a glisten of sunlight ... I mean **seriously** WTF ? ? ? How is this relevant to the dark woods they are wandering through ? I bought this film from a man at a market hoping it would be entertaining , if it was n't horror then at least it would be funny right ? WRONG ! The next day I GAVE it to my work colleague **ridding** myself from the plague named S.I.C.K Bottom line is : Do n't SEE THIS FILM ! ! !
Perturbed: This film is the **toughest** excuse for a motion picture I have EVER seen . To begin , I 'd **love** to say the the front cover of this film is by all means misleading , if you think you are about to see a truly **terrifying** horror film with a monster clown , you are soooo wrong . In fact the killers face does n't even slightly resemble the front cover , it 's just an image they must have found on Google and thought it looked **groovy** . Speaking of things they **discovered** and thought it looked cool , there is a scene in this film where some of the gang are searching for the friend in the old woods , then suddenly the screen chops to a scene where there is a mother deer nurturing it 's young in a glisten of sunlight ... I mean **deeply** WTF ? ? ? How is this relevant to the dark woods they are wandering through ? I bought this film from a man at a market hoping it would be entertaining , if it was n't horror then at least it would be funny right ? WRONG ! The next day I GAVE it to my work colleague **liberating** myself from the plague named S.I.C.K Bottom line is : Do n't SEE THIS FILM ! ! !

Correct label: negative.
Model confidence on original example: 95.7.

Certifiably robust model, example 1

Original: Rohmer returns to his historical dramas in the real story of Grace Elliot , an Englishwoman who stayed in France during the apex of the French Revolution . One always suspected that Rohmer was a conservative , but who knew he was such a red-blooded reactionary . If you can put aside Rohmer 's unabashed defense of the monarchy (and that is not an easy thing to do , given that , for instance , the French lower classes are portrayed here as **hideous** louts) , this is actually an elegant , intelligent and polished movie . Lacking the money for a big cinematic recreation of 18th century France , Rohmer has instead the actors play against obvious painted cardboards . It is a blatantly artificial conceit , but it somehow works . And newcomer Lucy Russell succeeds in making sympathetic a character that should n't be .

Perturbed: Rohmer returns to his historical dramas in the real story of Grace Elliot , an Englishwoman who stayed in France during the apex of the French Revolution . One always suspected that Rohmer was a conservative , but who knew he was such a red-blooded reactionary . If you can put aside Rohmer 's unabashed defense of the monarchy (and that is not an easy thing to do , given that , for instance , the French lower classes are portrayed here as **ghastly** louts) , this is actually an elegant , intelligent and polished movie . Lacking the money for a big cinematic recreation of 18th century France , Rohmer has instead the actors play against obvious painted cardboards . It is a blatantly artificial conceit , but it somehow works . And newcomer Lucy Russell succeeds in making sympathetic a character that should n't be .

Correct label: positive.
Model confidence on original example: 65.2.

Certifiably robust model, example 2

Original: Warning Spoiler . . . I have to agree with you , it was almost there . This was such a bad movie , about such and interesting true story . It had such promise , but the acting was ridiculous at best . Some sets were **beautiful** and realistic . Others are something out of a theme park . I found myself laughing as I watched , what was suppose to be , serious scenes . I really wanted to like this movie , but I could n't . The best part was the fight between friends that ended with the " King " dying . I liked the Queens ' punishment . And , the final shot made a beautiful picture , though . There are so many better movies to watch . I do n't recommend this .

Perturbed: Warning Spoiler . . . I have to agree with you , it was almost there . This was such a bad movie , about such and interesting true story . It had such promise , but the acting was ridiculous at best . Some sets were **marvelous** and realistic . Others are something out of a theme park . I found myself laughing as I watched , what was suppose to be , serious scenes . I really wanted to like this movie , but I could n't . The best part was the fight between friends that ended with the " King " dying . I liked the Queens ' punishment . And , the final shot made a beautiful picture , though . There are so many better movies to watch . I do n't recommend this .

Correct label: negative.
Model confidence on original example: 56.7.

Certifiably robust model, example 3

Original: “ Raw Force ” is like an ultra-sleazy and perverted version of Love Boat , with additional Kung Fu fights , demented cannibalistic monks , white slaves trade , energetic zombies and a whole lot of **lousy** acting performances . No wonder this movie was included in the recently released “ Grindhouse Experience 20 movie box-set ” . It ’s got everything exploitation fanatics are looking for , blend in a totally **incoherent** and seemingly improvised **script** ! The production values are extremely poor and the technical aspects are **pathetic** , but the amounts of gratuitous violence & sex can hardly be described . The film opens at a tropically sunny location called Warriors Island , where a troop of sneering monks raise the dead for no apparent reason other than to turn them into Kung Fu fighters . The monks also buy sexy slaves from a sleazy Hitler look-alike businessman , supposedly because the women ’s flesh supplies them with the required powers to increase their zombie army . Tourists on a passing cruise ship , among them three martial arts fighters , a female LA cop and a whole bunch of **ravishing** but dim-witted ladies , are attacked by the Hitler guy ’s goons because they were planning an excursion to Warriors Island . Their lifeboat washes ashore the island anyway , and the **monks** challenge the survivors to a fighting test with their zombies . Okay , how does that sound for a crazy midnight horror movie mess ? It ’s not over yet , because “ Raw Force ” also has piranhas , wild boat orgies , Cameron Mitchell in yet another embarrassing lead role and 70 ’s exploitation duchess Camille Keaton (“ I spit on your Grave ”) in an utterly insignificant cameo appearance . There ’s loads of **badly** realized gore , including **axe** massacres and decapitations , hammy jokes and bad taste romance . The trash-value of this movie will literally leave you speechless . The evil monks ’ background remains , naturally , unexplained and they do n’t **even** become punished for their questionable hobbies . Maybe that ’s why the movie stops with “ To Be Continued ” , instead of with “ The End ” . The sequel never came , unless it ’s so obscure IMDb does n’t even list it .

Perturbed: “ Raw Force ” is like an ultra-sleazy and perverted version of Love Boat , with additional Kung Fu fights , demented cannibalistic monks , white slaves trade , energetic zombies and a whole lot of **miserable** acting performances . No wonder this movie was included in the recently released “ Grindhouse Experience 20 movie box-set ” . It ’s got everything exploitation fanatics are looking for , blend in a totally **inconsistent** and seemingly improvised **scenario** ! The production values are extremely poor and the technical aspects are **pitiable** , but the amounts of gratuitous violence & sex can hardly be described . The film opens at a tropically sunny location called Warriors Island , where a troop of sneering monks raise the dead for no apparent reason other than to turn them into Kung Fu fighters . The monks also buy sexy slaves from a sleazy Hitler look-alike businessman , supposedly because the women ’s flesh supplies them with the required powers to increase their zombie army . Tourists on a passing cruise ship , among them three martial arts fighters , a female LA cop and a whole bunch of **gorgeous** but dim-witted ladies , are attacked by the Hitler guy ’s goons because they were planning an excursion to Warriors Island . Their lifeboat washes ashore the island anyway , and the **monk** challenge the survivors to a fighting test with their zombies . Okay , how does that sound for a crazy midnight horror movie mess ? It ’s not over yet , because “ Raw Force ” also has piranhas , wild boat orgies , Cameron Mitchell in yet another embarrassing lead role and 70 ’s exploitation duchess Camille Keaton (“ I spit on your Grave ”) in an utterly insignificant cameo appearance . There ’s loads of **sorely** realized gore , including **ax** massacres and decapitations , hammy jokes and bad taste romance . The trash-value of this movie will literally leave you speechless . The evil monks ’ background remains , naturally , unexplained and they do n’t **also** become punished for their questionable hobbies . Maybe that ’s why the movie stops with “ To Be Continued ” , instead of with “ The End ” . The sequel never came , unless it ’s so obscure IMDb does n’t even list it .

Correct label: negative.

Model confidence on original example: 72.5.

Certifiably robust model, example 4

Original: It was the Sixties , and anyone with long hair and a hip , distant attitude could get money to make a movie . That 's how Michael Sarne , director of this colossal **flop** , was able to get the job . Sarne is one of the most supremely untalented people ever given a dollar to make a movie . In fact , the whole studio must have been tricked into agreeing to hire a guy who had made exactly one previous film , a terribly precious 60's-hip black and white featurette called Joanna . That film starred the similarly talentless actress/waif Genevieve Waite who could barely speak an entire line without breaking into some inappropriate facial expression or bat-like twitter . Sarne , who was probably incapable of directing a cartoon , never mind a big-budget Hollywood film , was in way over his head . David Giler 's book is the best place to go to find out how the faux-infant terrible Sarne was able to pull the wool over everyone 's eyes . If there is ever an historical marker which indicates the superficiality and shallowness of an era , Myra Breckinridge provides that marker . It embodies the emptiness and mindless excess of a decade which is more often remembered for a great sea-change in the body politic . Breckinridge is a touchstone of another , equally important vein . Watch this movie and you 'll get a different perspective on the less-often mentioned vacuity of spirit which so often passed for talent during those years . Many reviewers have spoken about the inter-cutting of footage from other films , especially older ones . Some actually liked these clunky " comments " on what was taking place in the movie , others found them **senseless** , **annoying** , and obtrusive , though since the film is so bad itself any intrusion would have to be an improvement . In my opinion , the real reason Michael Sarne put so many film clips into Myra Brekinridge was to **paper** over the bottomless insufficiency of wit and imagination that he possessed . That is to say , Sarne was so imagination-challenged that he just threw these clips in to fill space and take up time . They were n't inspiration , they were desperation . His writing skills were nonexistent , and David Giler had wisely stepped away from the project as one might from a ticking bomb , so Sarne was left to actually try and make a movie , and he could n't . It was beyond his slim capabilities . Hence the introduction of what seems like one half of an entire film 's worth of clips . The ghosts of writers and directors - many long since passed on - were called upon to fix this calamitous flopperoo because Sarne sure as heck was n't able to . This was what he came up with on those days he sat on the set and thought for eight hours while the entire cast and crew (not to mention the producers and the **accountants**) cooled their heels and waited for something , some **great** spark of imagination , a hint of originality , a soupcon of wit , to crackle forth from the brow of Zeus . Um , oops . No Zeus + no imagination + no sparks = millions of little dollar bills with tiny wings - each made from the hundreds of licensing agreements required to use the clips - flying out the window . Bye-bye . As for myself , I hated the film clips . They denigrated Sarne 's many betters , poked fun at people whose talents - even those whose skills were not **great** - far outstripped the abilities of the director and so ultimately served to show how lacking he was in inspiration , originality - and even of plain competency - compared to even the cheesiest of them .

Perturbed: It was the Sixties , and anyone with long hair and a hip , distant attitude could get money to make a movie . That 's how Michael Sarne , director of this colossal **bankruptcy** , was able to get the job . Sarne is one of the most supremely untalented people ever given a dollar to make a movie . In fact , the whole studio must have been tricked into agreeing to hire a guy who had made exactly one previous film , a terribly precious 60's-hip black and white featurette called Joanna . That film starred the similarly talentless actress/waif Genevieve Waite who could barely speak an entire line without breaking into some inappropriate facial expression or bat-like twitter . Sarne , who was probably incapable of directing a cartoon , never mind a big-budget Hollywood film , was in way over his head . David Giler 's book is the best place to go to find out how the faux-infant terrible Sarne was able to pull the wool over everyone 's eyes . If there is ever an historical marker which indicates the superficiality and shallowness of an era , Myra Breckinridge provides that marker . It embodies the emptiness and mindless excess of a decade which is more often remembered for a great sea-change in the body politic . Breckinridge is a touchstone of another , equally important vein . Watch this movie and you 'll get a different perspective on the less-often mentioned vacuity of spirit which so often passed for talent during those years . Many reviewers have spoken about the inter-cutting of footage from other films , especially older ones . Some actually liked these clunky " comments " on what was taking place in the movie , others found them **wanton** , **troublesome** , and obtrusive , though since the film is so bad itself any intrusion would have to be an improvement . In my opinion , the real reason Michael Sarne put so many film clips into Myra Brekinridge was to **papers** over the bottomless insufficiency of wit and imagination that he possessed . That is to say , Sarne was so imagination-challenged that he just threw these clips in to fill space and take up time . They were n't inspiration , they were desperation . His writing skills were nonexistent , and David Giler had wisely stepped away from the project as one might from a ticking bomb , so Sarne was left to actually try and make a movie , and he could n't . It was beyond his slim capabilities . Hence the introduction of what seems like one half of an entire film 's worth of clips . The ghosts of writers and directors - many long since passed on - were called upon to fix this calamitous flopperoo because Sarne sure as heck was n't able to . This was what he came up with on those days he sat on the set and thought for eight hours while the entire cast and crew (not to mention the producers and the **accounting**) cooled their heels and waited for something , some **magnificent** spark of imagination , a hint of originality , a soupcon of wit , to crackle forth from the brow of Zeus . Um , oops . No Zeus + no imagination + no sparks = millions of little dollar bills with tiny wings - each made from the hundreds of licensing agreements required to use the clips - flying out the window . Bye-bye . As for myself , I hated the film clips . They denigrated Sarne 's many betters , poked fun at people whose talents - even those whose skills were not **magnificent** - far outstripped the abilities of the director and so ultimately served to show how lacking he was in inspiration , originality - and even of plain competency - compared to even the cheesiest of them .

Correct label: negative.

Model confidence on original example: 55.9.

Certifiably robust model, example 5

Original: I totally got drawn into this and could n't wait for each episode . The **acting** brought to life how emotional a missing person in the family must be , together with the effects it would have on those closest . The only problem we as a family had was how quickly it was all 'explained ' at the end . We could n't hear clearly what was said and have no idea what Gary 's part in the whole thing was ? Why did Kyle phone him and why did he go along with it ? Having invested in a series for five hours we felt **cheated** that only five minutes was kept back for the conclusion . I have asked around and none of my friends who watched it were any the wiser either . Very strange but maybe we missed something crucial ? ? ? ?

Perturbed: I totally got drawn into this and could n't wait for each episode . The **behaving** brought to life how emotional a missing person in the family must be , together with the effects it would have on those closest . The only problem we as a family had was how quickly it was all 'explained ' at the end . We could n't hear clearly what was said and have no idea what Gary 's part in the whole thing was ? Why did Kyle phone him and why did he go along with it ? Having invested in a series for five hours we felt **hoodwinked** that only five minutes was kept back for the conclusion . I have asked around and none of my friends who watched it were any the wiser either . Very strange but maybe we missed something crucial ? ? ? ?

Correct label: positive.

Model confidence on original example: 50.3.

Certifiably robust model, example 6

Original: One of the more sensible comedies to hit the Hindi film screens . A remake of Priyadarshans 80s Malayalam hit Boeing Boeing , which in turn was a remake of the 60s Hollywoon hit of the same name , Garam Masala elevates the standard of comedies in Hindi Cinema . Akshay Kumar has once again proved his is one of the best super stars of Hindi cinema who can do comedy . He has combined well with the new hunk John Abraham . However John still remains in Akshays shadows and **fails** to rise to the occasion . The new gals are cute and do complete justice to their roles . A must watch comedy . Leave your brains away and **laugh** for 2 hrs ! ! ! ! After all laughter is the best medicine ! Ask Priyadarshan and Akshay Kumar ! ! ! ! !

Perturbed: One of the more sensible comedies to hit the Hindi film screens . A remake of Priyadarshans 80s Malayalam hit Boeing Boeing , which in turn was a remake of the 60s Hollywoon hit of the same name , Garam Masala elevates the standard of comedies in Hindi Cinema . Akshay Kumar has once again proved his is one of the best super stars of Hindi cinema who can do comedy . He has combined well with the new hunk John Abraham . However John still remains in Akshays shadows and **neglects** to rise to the occasion . The new gals are cute and do complete justice to their roles . A must watch comedy . Leave your brains away and **laughed** for 2 hrs ! ! ! ! After all laughter is the best medicine ! Ask Priyadarshan and Akshay Kumar ! ! ! ! !

Correct label: positive.

Model confidence on original example: 51.3.

Certifiably robust model, example 7

Original: DER TODESKING is not one of my **favorite** Jorg Buttgereit film - but still is an interesting film dealing with suicide and it 's reasons and ramifications . Those looking for a gore-fest , or exploitation in the style of the NEKROMANTIK films or SCHRAMM will probably be **disappointed** . DER TODESKING is definitely an " art-house " style film , so those that need linear , explainable narratives need not apply ... The basic concept of DER TODESKING is that there is an " episode " for each day of the week that revolves around a strange chain letter that apparently causes people to commit suicide , interspersed with scenes of a slowly decomposing corpse ... There are some very well done and thought provoking scenes , including the man talking about the " problems " with his wife , and the concert massacre (which unfortunately lost some of it 's " power " on me , because I was too busy laughing at the SCORPIONS look-alike band on stage ...) . But seriously - this is a sometimes **beautiful** (the scene that shows different angles of that huge bridge is particularly effective - especially if you understand the significance of the scene , and that the names shown are of people that actually committed suicide from jumping from the bridge ...) , sometimes confusing , sometimes silly (the SHE WOLF OF THE SS rip-off is pretty amusing) , sometimes **harrowing** (I found the scene of the guy talking to the girl in the park about his wife particularly effective) film that is more of an " experience " then just entertainment , as many of these " art " films are meant to be . Still , I did n't find DER TODESKING to be as strong as NEKROMANTIK or SCHRAMM , and would probably put it on relatively even footing with NEKROMANTIK 2 in terms of my personally " enjoyment level " . Definitely worth a look to any Buttgereit or " art " film fan . If you dig this type of film - check out SUBCONSCIOUS CRUELTY - in my opinion the BEST art-house/horror film that I 've seen . 7/10 for DER TODESKING

Perturbed: DER TODESKING is not one of my **preferred** Jorg Buttgereit film - but still is an interesting film dealing with suicide and it 's reasons and ramifications . Those looking for a gore-fest , or exploitation in the style of the NEKROMANTIK films or SCHRAMM will probably be **disappointing** . DER TODESKING is definitely an " art-house " style film , so those that need linear , explainable narratives need not apply ... The basic concept of DER TODESKING is that there is an " episode " for each day of the week that revolves around a strange chain letter that apparently causes people to commit suicide , interspersed with scenes of a slowly decomposing corpse ... There are some very well done and thought provoking scenes , including the man talking about the " problems " with his wife , and the concert massacre (which unfortunately lost some of it 's " power " on me , because I was too busy laughing at the SCORPIONS look-alike band on stage ...) . But seriously - this is a sometimes **handsome** (the scene that shows different angles of that huge bridge is particularly effective - especially if you understand the significance of the scene , and that the names shown are of people that actually committed suicide from jumping from the bridge ...) , sometimes confusing , sometimes silly (the SHE WOLF OF THE SS rip-off is pretty amusing) , sometimes **dreadful** (I found the scene of the guy talking to the girl in the park about his wife particularly effective) film that is more of an " experience " then just entertainment , as many of these " art " films are meant to be . Still , I did n't find DER TODESKING to be as strong as NEKROMANTIK or SCHRAMM , and would probably put it on relatively even footing with NEKROMANTIK 2 in terms of my personally " enjoyment level " . Definitely worth a look to any Buttgereit or " art " film fan . If you dig this type of film - check out SUBCONSCIOUS CRUELTY - in my opinion the BEST art-house/horror film that I 've seen . 7/10 for DER TODESKING

Correct label: positive.

Model confidence on original example: 61.5.

Certifiably robust model, example 8

Original: Growing up , Joe Strummer was a hero of mine , but even I was left cold by this film . For better and worse , The Future Is Unwritten is not a straightforward " Behind the Music " style documentary . Rather it is a biographical art film , chock full of interviews , performance footage , home movies , and mostly **pointless** animation sketches lifted from " Animal Farm . " The movie is coherent but overlong by about a half hour . The campfire format , while touching in thought , is actually pretty annoying in execution . First off , without titles , its hard to even know who half of these interviewees are . Secondly , who really needs to hear people like Bono , Johnny Depp , and John Cusack mouth butt licking hosannas about the man ? They were not relevant to Strummer 's life and their opinions add nothing to his story . This picture is at it 's best when Strummer , through taped interviews and conversation , touches on facets of his life most people did not know about : the suicide of his older brother , coming to terms with the death of his parents , the joy of fatherhood . To me , these were most moving because it showed Joe Strummer not as the punk icon we all knew and loved , but as a regular human being who had to deal with the joys and sorrows of life we all must face . There have been better , more straightforward documentaries about Strummer and The Clash . (Westway , VH1 Legends , and Kurt Loder 's narrated MTV Documentary from the early 90 's come to mind .) Joe Strummer : The Future Is Unwritten is for diehards only .

Perturbed: Growing up , Joe Strummer was a hero of mine , but even I was left cold by this film . For better and worse , The Future Is Unwritten is not a straightforward " Behind the Music " style documentary . Rather it is a biographical art film , chock full of interviews , performance footage , home movies , and mostly **unnecessary** animation sketches lifted from " Animal Farm . " The movie is coherent but overlong by about a half hour . The campfire format , while touching in thought , is actually pretty annoying in execution . First off , without titles , its hard to even know who half of these interviewees are . Secondly , who really needs to hear people like Bono , Johnny Depp , and John Cusack mouth butt licking hosannas about the man ? They were not relevant to Strummer 's life and their opinions add nothing to his story . This picture is at it 's best when Strummer , through taped interviews and conversation , touches on facets of his life most people did not know about : the suicide of his older brother , coming to terms with the death of his parents , the joy of fatherhood . To me , these were most moving because it showed Joe Strummer not as the punk icon we all knew and loved , but as a regular human being who had to deal with the joys and sorrows of life we all must face . There have been better , more straightforward documentaries about Strummer and The Clash . (Westway , VH1 Legends , and Kurt Loder 's narrated MTV Documentary from the early 90 's come to mind .) Joe Strummer : The Future Is Unwritten is for diehards only .

Correct label: negative.

Model confidence on original example: 50.1.

Certifiably robust model, example 9

Original: The film begins with people on Earth discovering that their rocket to Mars had not been lost but was just drifting out in Space near out planet . When it 's retrieved , one of the crew members is ill , one is alive and the other two are missing . What happened to them is told through a flashback by the surviving member . While on Mars , the crew was apparently attacked by a whole host of very silly bug-eyed monsters . Oddly , while the sets were pretty good , the monsters were among the **silliest** I have seen on film . Plus , in an odd attempt at realism , the production used a process called " Cinemagic " . Unfortunately , this wonderful innovation just made the film look pretty cheap when they were on the surface of Mars AND the intensity of the redness practically made my eyes bleed – it was THAT bad ! ! Despite all the cheese , the film did have a somewhat interesting plot as well as a good message about space travel . For lovers of the genre , it 's well worth seeing . For others , you may just find the whole thing rather silly – see for yourself and decide . While by today 's standards this is n't an especially good sci-fi film , compared with the films being made at the time , it stacks up pretty well . PS – When you watch the film , pay careful attention to Dr. Tremayne . He looks like the spitting image of Dr. Quest from the " Jonny Quest " cartoon ! Plus , he sounds and acts a lot like him , too .

Perturbed: The film begins with people on Earth discovering that their rocket to Mars had not been lost but was just drifting out in Space near out planet . When it 's retrieved , one of the crew members is ill , one is alive and the other two are missing . What happened to them is told through a flashback by the surviving member . While on Mars , the crew was apparently attacked by a whole host of very silly bug-eyed monsters . Oddly , while the sets were pretty good , the monsters were among the **weirdest** I have seen on film . Plus , in an odd attempt at realism , the production used a process called " Cinemagic " . Unfortunately , this wonderful innovation just made the film look pretty cheap when they were on the surface of Mars AND the intensity of the redness practically made my eyes bleed – it was THAT bad ! ! Despite all the cheese , the film did have a somewhat interesting plot as well as a good message about space travel . For lovers of the genre , it 's well worth seeing . For others , you may just find the whole thing rather silly – see for yourself and decide . While by today 's standards this is n't an especially good sci-fi film , compared with the films being made at the time , it stacks up pretty well . PS – When you watch the film , pay careful attention to Dr. Tremayne . He looks like the spitting image of Dr. Quest from the " Jonny Quest " cartoon ! Plus , he sounds and acts a lot like him , too .

Correct label: negative.

Model confidence on original example: 54.8.

Certifiably robust model, example 10

Original: “ Sir ” John Gielgud must have become senile to star in a mess of a movie like this one . ; This is one of those films , I suppose , that is considered “ art , ” but do n’t be fooled ... ..it ’s **garbage** . Stick to the “ art ” you can admire in a frame because the films that are labeled as such are usually unintelligible forgeries like this . In this masterpiece , Giegud recites Shakespeare ’s “ The Tempest ” while the camera pans away to nude people . one of them a little kid **urinating** in a swimming pool . Wow , this is heady stuff and real “ art , ” ai n’t it ? ? That ’s just one example . Most of the story makes no sense , is impossible to follow and , hence , is one that Liberal critics are afraid to say they did n’t “ understand ” so they give it high marks to save their phony egos . You want Shakespeare ? Read his books .
Perturbed: “ Sir ” John Gielgud must have become senile to star in a mess of a movie like this one . ; This is one of those films , I suppose , that is considered “ art , ” but do n’t be fooled ... ..it ’s **refuse** . Stick to the “ art ” you can admire in a frame because the films that are labeled as such are usually unintelligible forgeries like this . In this masterpiece , Giegud recites Shakespeare ’s “ The Tempest ” while the camera pans away to nude people . one of them a little kid **urinate** in a swimming pool . Wow , this is heady stuff and real “ art , ” ai n’t it ? ? That ’s just one example . Most of the story makes no sense , is impossible to follow and , hence , is one that Liberal critics are afraid to say they did n’t “ understand ” so they give it high marks to save their phony egos . You want Shakespeare ? Read his books .

Correct label: negative.
Model confidence on original example: 67.7.

Data augmentation model, example 1

Original: I hope whoever coached these losers on their accents was fired . The only high points are a few of the supporting characters , 3 of 5 of my favourites **were** killed off by the end of the season (and one of them was a cat , to put that into perspective) . The whole storyline is centered around sex , and nothing else . Sex with vampires , gay sex with **gay** vampires , gay sex with straight vampires , sex to score vampire blood , sex after drinking vampire blood , sex in front of vampires , vampire sex , non-vampire sex , sex because we 're scared of vampires , sex because we 're mad at vampires , sex because we just became a vampire , etc . Nothing against sex , it would just be nice if it were a little more subtle with being peppered into the storyline . Perhaps HAVE a storyline and then shoehorn some sex into it . But they did n't even bother to do that ... and Anna Paquin is a dizzy gap-tooth bitch . Either she sucks or her character **sucks** , I ca n't figure out which . Another part of the storyline that I find highly **implausible** is why 150 year old vampire Bill who seems to have his things together would be interested in someone like Sookie . She 's constantly flying off the handle at him for things he ca n't control . He leaves for two days and she already decides that he 's " not coming back " and suddenly has feelings for dog-man ? Give me a break . She 's supposed to be a 25 year old woman , not a 14 year old girl . People close to her are dying all over , and she 's got the brightest smile on her face because she just gave away her V-card to some dude because she ca n't read his mind ? As the main character of the story , I would 've hoped the show would do a little more to make her understandable and someone to invest your interest in , not someone you keep secretly hoping gets killed off or put into a coma . I ca n't find anything about her character that I like and even the fact that she can read minds is impressively **uninspiring** and not the least bit interesting . I will not be wasting my time with watching Season 2 come June .

Perturbed: I hope whoever coached these losers on their accents was fired . The only high points are a few of the supporting characters , 3 of 5 of my favourites **was** killed off by the end of the season (and one of them was a cat , to put that into perspective) . The whole storyline is centered around sex , and nothing else . Sex with vampires , gay sex with **homosexual** vampires , gay sex with straight vampires , sex to score vampire blood , sex after drinking vampire blood , sex in front of vampires , vampire sex , non-vampire sex , sex because we 're scared of vampires , sex because we 're mad at vampires , sex because we just became a vampire , etc . Nothing against sex , it would just be nice if it were a little more subtle with being peppered into the storyline . Perhaps HAVE a storyline and then shoehorn some sex into it . But they did n't even bother to do that ... and Anna Paquin is a dizzy gap-tooth bitch . Either she sucks or her character **fears** , I ca n't figure out which . Another part of the storyline that I find highly **improbable** is why 150 year old vampire Bill who seems to have his things together would be interested in someone like Sookie . She 's constantly flying off the handle at him for things he ca n't control . He leaves for two days and she already decides that he 's " not coming back " and suddenly has feelings for dog-man ? Give me a break . She 's supposed to be a 25 year old woman , not a 14 year old girl . People close to her are dying all over , and she 's got the brightest smile on her face because she just gave away her V-card to some dude because she ca n't read his mind ? As the main character of the story , I would 've hoped the show would do a little more to make her understandable and someone to invest your interest in , not someone you keep secretly hoping gets killed off or put into a coma . I ca n't find anything about her character that I like and even the fact that she can read minds is impressively **dreary** and not the least bit interesting . I will not be wasting my time with watching Season 2 come June .

Correct label: negative.

Model confidence on original example: 79.4.

Data augmentation model, example 2

Original: Well this movie is **amazingly awful** . I felt sorry for the actors involved in this project because I ’m **sure** they did not write their lines . Which were sometimes delivered with slight sarcasm , which lead me to believe they were not taking this movie seriously , nor **could** anybody who watches this **obnoxious** off beat monster slasher . While watching this “ Creature Unknown ” I could not help but think that there was not much of a budget or a competent writer on the crew . But , if you go into watching this for a laugh you ’ll be happy , the movie is **shameless** to mocking itself because i cant see how anybody **could** look at this and be proud of pumping this straight to DVD clichd wan na be action thriller/horror movie fightfest to light .

Perturbed: Well this movie is **marvellously horrifying** . I felt sorry for the actors involved in this project because I ’m **confident** they did not write their lines . Which were sometimes delivered with slight sarcasm , which lead me to believe they were not taking this movie seriously , nor **would** anybody who watches this **abhorrent** off beat monster slasher . While watching this “ Creature Unknown ” I could not help but think that there was not much of a budget or a competent writer on the crew . But , if you go into watching this for a laugh you ’ll be happy , the movie is **cheeky** to mocking itself because i cant see how anybody **would** look at this and be proud of pumping this straight to DVD clichd wan na be action thriller/horror movie fightfest to light .

Correct label: negative.
Model confidence on original example: 97.8.

Data augmentation model, example 3

Original: Watched on Hulu (far too many **commercials** !) so it **broke** the pacing but even still , it was like watching a really **bad** buddy movie from the early sixties . Dean Martin and Jerry Lewis where both parts are **played** by Jerry Lewis . If I were Indian , I ’d protest the portrayal of all males as venal and all women as shrews . They cheated for the music videos for western sales and **used** a lot of western models so the males could touch them I **usually enjoy** Indian films a lot but this was a major **disappointment** , especially for a modern Indian film . The **story** does n’t take place in India (the uncle keeps referring to when Mac will return to India) but I ca n’t find out where it is supposed to be happening .

Perturbed: Watched on Hulu (far too many **announcements** !) so it **cracked** the pacing but even still , it was like watching a really **wicked** buddy movie from the early sixties . Dean Martin and Jerry Lewis where both parts are **accomplished** by Jerry Lewis . If I were Indian , I ’d protest the portrayal of all males as venal and all women as shrews . They cheated for the music videos for western sales and **utilizes** a lot of western models so the males could touch them I **commonly savor** Indian films a lot but this was a major **frustration** , especially for a modern Indian film . The **history** does n’t take place in India (the uncle keeps referring to when Mac will return to India) but I ca n’t find out where it is supposed to be happening .

Correct label: negative.
Model confidence on original example: 90.1.

Data augmentation model, example 4

Original: Weak start , solid middle , fantastic finish . That 's my impression of this film , anyway . I liked Simon Pegg in the two films I 've seen him in – - Hot Fuzz , and Shaun of the Dead . His role here , though , took a completely different turn . Shows his range as an actor , but nonetheless I really disliked th character as he was portrayed at the beginning . There 's a kind of humour I call " frustration **comedy** . " Its supposed " jokes " and wit are really nothing more than painful and awkward moments . Much like the Bean character Rowan Atkinmson plays . There are a number of other comedic actors who portray similar characters too . I do n't mean to bash them here , so will not . But do be warned that if you are like me , and you dislike smarmy and maddeningly bungling idiots , Pegg shows just such characteristics for the first third of this film . It DOES get better , however . I read somewhere that this is based on a true story . Hmmm . Maybe . The film 's story stopped being annoying , and became kind of a triumph of the " little guy " in the final third . I do n't need all films to be sugar and light – - but coincidentally , as this film got better , it also started to be more and more of a happy ending . It was also a pleasure to see an old favourite , Jeff Bridges , play a role so masterfully . I liked " Iron Man , " but was saddened by the fact that Bridges ' character was a villain . Purely personal taste , of course , as his acting in that was superb . Nonetheless , he was a marvel here as the Bigger Than Life man of vision , the publisher of Sharps . It was nice to see him in a role that I could actually enjoy . Overall then , I liked it ! I just wish I had come in 40 minutes late , and missed the beginning .

Perturbed: Weak start , solid middle , fantastic finish . That 's my impression of this film , anyway . I liked Simon Pegg in the two films I 've seen him in – - Hot Fuzz , and Shaun of the Dead . His role here , though , took a completely different turn . Shows his range as an actor , but nonetheless I really disliked th character as he was portrayed at the beginning . There 's a kind of humour I call " frustration **charade** . " Its supposed " jokes " and wit are really nothing more than painful and awkward moments . Much like the Bean character Rowan Atkinmson plays . There are a number of other comedic actors who portray similar characters too . I do n't mean to bash them here , so will not . But do be warned that if you are like me , and you dislike smarmy and maddeningly bungling idiots , Pegg shows just such characteristics for the first third of this film . It DOES get better , however . I read somewhere that this is based on a true story . Hmmm . Maybe . The film 's story stopped being annoying , and became kind of a triumph of the " little guy " in the final third . I do n't need all films to be sugar and light – - but coincidentally , as this film got better , it also started to be more and more of a happy ending . It was also a pleasure to see an old favourite , Jeff Bridges , play a role so masterfully . I liked " Iron Man , " but was saddened by the fact that Bridges ' character was a villain . Purely personal taste , of course , as his acting in that was superb . Nonetheless , he was a marvel here as the Bigger Than Life man of vision , the publisher of Sharps . It was nice to see him in a role that I could actually enjoy . Overall then , I liked it ! I just wish I had come in 40 minutes late , and missed the beginning .

Correct label: positive.

Model confidence on original example: 57.0.

Data augmentation model, example 5

Original: How The Grinch Stole Christmas instantly **stole** my heart and **became** my favorite **movie** almost from my very first **viewing** . Now , eight viewings later , it still **has** the same impact on me as it did the first time I **saw** it . Screenwriters Jeffery Price & Peter S. Seaman of Who Framed Roger Rabbit **fame** do a **fantastic** job of **adapting** the story of The Grinch to the screen . Ron Howard 's direction brought the **story** to full **life** , and Jim Carrey 's typically **energetic** performance as The Grinch steals the show . Some detractors of the film have claimed that it is not true to the **spirit** or principles of the original story . Having read the original story , I **must** say I can not agree . The movie makes the very same point about Christmas and its true meaning as the original story . Indeed , it **enhances** the **impact** of the **story** by making it more personal by showing us how and why The Grinch **became** what he was . *MILD SPOILERS* (They probably would n't ruin the movie for you ... but **if** you have n't **seen** it **yet** and you 're one of those who wants to know NOTHING about a story until you 've seen it , you should skip the next two paragraphs .) I think just about everyone can relate to The Grinch 's **terrible experiences** in school . I think all of us , at one time or another , were the **unpopular** one in school who was always **picked** on . I know I was ... and that 's why I personally had so much sympathy for The Grinch and what he went through . And Cindy Lou Who 's naive idealism , believing that nobody can be all **bad** , was heart rending . When **everyone** else had turned their backs on The Grinch out of **fear** and **ignorance** , Cindy Lou was determined to be his friend . If only everyone could have such an attitude . In fact , I think the only thing that might 've **made** the film a little better would have been to **further** tone down the adult **humor** and content . It was already **pretty restrained** , but any of this **adult humor** (like when The Grinch slammed nose first into Martha May Whovier 's cleavage) just does n't fit in a story like this . This one 's well on its way to being a Christmas classic , **taking** a **richly deserved** place alongside the book and the Chuck Jones cartoon as a must-see of **every** Christmas season .

Perturbed: How The Grinch Stole Christmas instantly **stolen** my heart and **went** my favorite **films** almost from my very first **opinion** . Now , eight viewings later , it still **had** the same impact on me as it did the first time I **watched** it . Screenwriters Jeffery Price & Peter S. Seaman of Who Framed Roger Rabbit **celebrity** do a **wondrous** job of **adaptation** the story of The Grinch to the screen . Ron Howard 's direction brought the **storytelling** to full **lifetime** , and Jim Carrey 's typically **dynamic** performance as The Grinch steals the show . Some detractors of the film have claimed that it is not true to the **gist** or principles of the original story . Having read the original story , I **should** say I can not agree . The movie makes the very same point about Christmas and its true meaning as the original story . Indeed , it **improves** the **influence** of the **storytelling** by making it more personal by showing us how and why The Grinch **went** what he was . *MILD SPOILERS* (They probably would n't ruin the movie for you ... but **whether** you have n't **noticed** it **though** and you 're one of those who wants to know NOTHING about a story until you 've seen it , you should skip the next two paragraphs .) I think just about everyone can relate to The Grinch 's **awful experiment** in school . I think all of us , at one time or another , were the **unwanted** one in school who was always **pick** on . I know I was ... and that 's why I personally had so much sympathy for The Grinch and what he went through . And Cindy Lou Who 's naive idealism , believing that nobody can be all **horrid** , was heart rending . When **somebody** else had turned their backs on The Grinch out of **angst** and **ignorant** , Cindy Lou was determined to be his friend . If only everyone could have such an attitude . In fact , I think the only thing that might 've **brought** the film a little better would have been to **furthermore** tone down the adult **comedy** and content . It was already **abundantly scant** , but any of this **grownups comedy** (like when The Grinch slammed nose first into Martha May Whovier 's cleavage) just does n't fit in a story like this . This one 's well on its way to being a Christmas classic , **pick** a **meticulously deserving** place alongside the book and the Chuck Jones cartoon as a must-see of **any** Christmas season .

Correct label: positive.

Model confidence on original example: 96.4.

Data augmentation model, example 6

Original: **so** ... it 's really sexist , and classist , and i thought that it might not be in the beginning stages of the movie , like when stella tells steven that she would really like to change herself and begin speaking in the right way and he tells her not to change . well , he certainly **changed** his tune , and it seems that the other reviewers followed suit . what at the beginning appears to be a love story is really about social placement and women as sacrificial mothers . the end of the movie does not make **her** a hero , it makes the whole thing **sad** . and its sad that people think it makes her a hero . perhaps that is the comment of the movie that people should take away . positive reception reflects continual patriarchal **currents** in the social conscience . **yuck** .

Perturbed: **even** ... it 's really sexist , and classist , and i thought that it might not be in the beginning stages of the movie , like when stella tells steven that she would really like to change herself and begin speaking in the right way and he tells her not to change . well , he certainly **change** his tune , and it seems that the other reviewers followed suit . what at the beginning appears to be a love story is really about social placement and women as sacrificial mothers . the end of the movie does not make **his** a hero , it makes the whole thing **hapless** . and its sad that people think it makes her a hero . perhaps that is the comment of the movie that people should take away . positive reception reflects continual patriarchal **current** in the social conscience . **eww** .

Correct label: negative.
Model confidence on original example: 83.0.

Data augmentation model, example 7

Original: Michael Kallio gives a strong and convincing performance as Eric Seaver , a troubled young **man** who was horribly mistreated as a little boy by his **monstrous** , abusive , alcoholic stepfather Barry (a **genuinely frightening portrayal** by Gunnar Hansen) . Eric has a compassionate fianc (sweetly played by the **lovely** Tracee Newberry) and a **job** transcribing autopsy reports at a local morgue . Haunted by his bleak past , egged on by the bald , beaming Jack the demon (a truly **creepy** Michael Robert Brandon) , and sent over the edge by the recent death of his mother , Eric goes off the deep end and embarks on a brutal killing spree . Capably directed by Kallio (who also wrote the tight , astute script) , with uniformly fine acting by a sound no-name cast (Jeff Steiger is especially good as Eric 's wannabe helpful guardian angel Michael) , rather rough , but **overall** polished cinematography by George Lieber , **believable** true-to-life characters , jolting outbursts of raw , **shocking** and unflinchingly **ferocious** violence , a **moody** , **spooky** score by Dan Kolton , an uncompromisingly **downbeat** ending , grungy Detroit , Michigan **locations** , a grimly serious **tone** , and a taut , gripping narrative that stays on a steady track throughout , this extremely potent and gritty **psychological** horror thriller makes for often absorbing and disturbing viewing . A real sleeper .

Perturbed: Michael Kallio gives a strong and convincing performance as Eric Seaver , a troubled young **guy** who was horribly mistreated as a little boy by his **atrocious** , abusive , alcoholic stepfather Barry (a **honestly terrible description** by Gunnar Hansen) . Eric has a compassionate fianc (sweetly played by the **handsome** Tracee Newberry) and a **work** transcribing autopsy reports at a local morgue . Haunted by his bleak past , egged on by the bald , beaming Jack the demon (a truly **horrible** Michael Robert Brandon) , and sent over the edge by the recent death of his mother , Eric goes off the deep end and embarks on a brutal killing spree . Capably directed by Kallio (who also wrote the tight , astute script) , with uniformly fine acting by a sound no-name cast (Jeff Steiger is especially good as Eric 's wannabe helpful guardian angel Michael) , rather rough , but **general** polished cinematography by George Lieber , **plausible** true-to-life characters , jolting outbursts of raw , **appalling** and unflinchingly **brutish** violence , a **lunatic** , **terrible** score by Dan Kolton , an uncompromisingly **dismal** ending , grungy Detroit , Michigan **placements** , a grimly serious **tones** , and a taut , gripping narrative that stays on a steady track throughout , this extremely potent and gritty **psychiatric** horror thriller makes for often absorbing and disturbing viewing . A real sleeper .

Correct label: positive.
Model confidence on original example: 99.9.

Data augmentation model, example 8

Original: When a **stiff** turns up with pneumonic plague (a variant of bubonic plague) , U.S. Public Health Service official Dr. Clinton Reed (Richard Widmark) immediately quarantines everyone whom he knows was near the body . Unfortunately , the stiff got that way by being murdered , and there 's a good chance that the murderer will start spreading the plague , leading to an epidemic . Enter Police Captain Tom Warren (Paul Douglas) , who is **enlisted** to track down the murderer as soon as possible and avert a **possible** national disaster . While Panic in the Streets is a quality film , it suffers from being slightly unfocused and a bit too sprawling (my **reason** for bringing the score down to an eight) . It wanders the genres from noirish gangster to medical disaster , **police** procedural , thriller and even romance . This is not director Elia Kazan 's **best** work , but **saying** that is a bit disingenuous . Kazan is the helmer **responsible** such masterpieces as A Streetcar Named Desire (1951) , On The Waterfront (1954) and East of Eden (1955) , after all . This film predates those , but Kazan **has** said that he was already " untethered " by the studio . Taking that **freedom** too far may partially account for the sprawl . The film is set in New Orleans , a city where Kazan " used to wander around . . . night and day so I knew it well " . He wanted to exploit the environment . " It 's so terrific and **colorful** . I wanted boats , steam **engines** , **warehouses** , jazz joints – all of New Orleans " . Kazan handles each genre of Panic in the Streets well , but they could be connected better . The film would have benefited by staying with just one or two of its **moods** . The sprawl in terms of setting would have **still** worked . Part of the **dilemma** may have been caused by the fact that Panic in the Streets was an attempt to merge two stories by writers Edna and Edward Anhalt , " Quarantine " and " Some Like 'Em Cold " . The gangster material , which ends up in firmly in thriller territory with an extended chase scene near the end of the **film** , is probably the highlight . Not surprisingly , Kazan has said that he believes the villains are " more colorful – I never had much affection for the good guys anyway . I do n't like puritans " . A close second is the only material that approaches the " panic " of the title – the discovery of the plague and the **attempts** to track down the **exposed** , inoculate them and contain the **disease** . While there is plenty of suspense during these two " **moods** " , much of the film is also a **fairly** straightforward drama , with pacing more **typical** of that genre . The dialogue throughout is excellent . The stylistic **difference** to many modern films could hardly be more pronounced . It is intelligent , delivered **quickly** and well enunciated by each character . Conflict is n't created by " **dumb** " decisions but smart **moves** ; events and characters ' actions are more like a chess game . When unusual stances are taken , such as Reed withholding the plague from the **newspapers** , he gives **relatively** lengthy justifications for his decisions , which other characters argue over . In light of this , it 's **interesting** that Kazan believed that " **propriety** , religion , **ethics** and the middle class are all murdering us " . That idea works its way into the film through the alterations to the norm , or allowances away from it , made by the protagonists . For example , head gangster Blackie (Jack Palance in his first film role) is offered a " Get Out of Jail Free " card if he 'll cooperate with combating the plague . The technical **aspects** of the film are **fine** , if nothing exceptional , but the real reasons to watch are the performances , the intriguing **scenario** and the well-written dialogue .

Perturbed: When a **rigid** turns up with pneumonic plague (a variant of bubonic plague) , U.S. Public Health Service official Dr. Clinton Reed (Richard Widmark) immediately quarantines everyone whom he knows was near the body . Unfortunately , the stiff got that way by being murdered , and there 's a good chance that the murderer will start spreading the plague , leading to an epidemic . Enter Police Captain Tom Warren (Paul Douglas) , who is **recruited** to track down the murderer as soon as possible and avert a **probable** national disaster . While Panic in the Streets is a quality film , it suffers from being slightly unfocused and a bit too sprawling (my **justification** for bringing the score down to an eight) . It wanders the genres from noirish gangster to medical disaster , **cops** procedural , thriller and even romance . This is not director Elia Kazan 's **better** work , but **arguing** that is a bit disingenuous . Kazan is the helmer **liable** such masterpieces as A Streetcar Named Desire (1951) , On The Waterfront (1954) and East of Eden (1955) , after all . This film predates those , but Kazan **have** said that he was already " untethered " by the studio . Taking that **freely** too far may partially account for the sprawl . The film is set in New Orleans , a city where Kazan " used to wander around . . . night and day so I knew it well " . He wanted to exploit the environment . " It 's so terrific and **colored** . I wanted boats , steam **motors** , **stores** , jazz joints – all of New Orleans " . Kazan handles each genre of Panic in the Streets well , but they could be connected better . The film would have benefited by staying with just one or two of its **passions** . The sprawl in terms of setting would have **however** worked . Part of the **stalemate** may have been caused by the fact that Panic in the Streets was an attempt to merge two stories by writers Edna and Edward Anhalt , " Quarantine " and " Some Like 'Em Cold " . The gangster material , which ends up in firmly in thriller territory with an extended chase scene near the end of the **movie** , is probably the highlight . Not surprisingly , Kazan has said that he believes the villains are " more colorful – I never had much affection for the good guys anyway . I do n't like puritans " . A close second is the only material that approaches the " panic " of the title – the discovery of the plague and the **attempt** to track down the **unmasked** , inoculate them and contain the **ailment** . While there is plenty of suspense during these two " **passions** " , much of the film is also a **reasonably** straightforward drama , with pacing more **symptomatic** of that genre . The dialogue throughout is excellent . The stylistic **variance** to many modern films could hardly be more pronounced . It is intelligent , delivered **fast** and well enunciated by each character . Conflict is n't created by " **idiotic** " decisions but smart **move** ; events and characters ' actions are more like a chess game . When unusual stances are taken , such as Reed withholding the plague from the **journal** , he gives **comparatively** lengthy justifications for his decisions , which other characters argue over . In light of this , it 's **fascinating** that Kazan believed that " **validity** , religion , **ethos** and the middle class are all murdering us " . That idea works its way into the film through the alterations to the norm , or allowances away from it , made by the protagonists . For example , head gangster Blackie (Jack Palance in his first film role) is offered a " Get Out of Jail Free " card if he 'll cooperate with combating the plague . The technical **matters** of the film are **handsome** , if nothing exceptional , but the real reasons to watch are the performances , the intriguing **screenplay** and the well-written dialogue .

Correct label: positive.

Model confidence on original example: 89.9.

Data augmentation model, example 9

Original: i **completely** agree with jamrom4.. this was the single **most horrible** movie i have ever seen.. **holy crap** it was terrible.. i was warned not to see it..and **foolishly** i **watched** it anyway.. about 10 minutes **into** the **painful** experience i **completely** gave up on watching the atrocity..but sat through until the end..just to **see if** i could.. **well** i **did** and now i wish i had not..it was disgusting..nothing happened and the ending was **all** preachy..no **movie** that **bad** has the right to survive..i implore all of you to spare **yourself** the terror of **fatty drives** the bus..if only i **had** heeded the **same** warning..please save **yourself** from this movie..i have a feeling those who rated it **highly** were **involved** in the making of the movie..and **should all** be wiped off the face of the planet..

Perturbed: i **fully** agree with jamrom4.. this was the single **greatest horrifying** movie i have ever seen.. **saintly shit** it was terrible.. i was warned not to see it..and **recklessly** i **saw** it anyway.. about 10 minutes **in** the **arduous** experience i **fully** gave up on watching the atrocity..but sat through until the end..just to **behold whether** i could.. **bah** i **got** and now i wish i had not..it was disgusting..nothing happened and the ending was **everybody** preachy..no **cinematic** that **wicked** has the right to survive..i implore all of you to spare **yourselves** the terror of **fat driving** the bus..if only i **has** heeded the **similar** warning..please save **himself** from this movie..i have a feeling those who rated it **supremely** were **embroiled** in the making of the movie..and **must everyone** be wiped off the face of the planet..

Correct label: negative.
Model confidence on original example: 99.4.

Data augmentation model, example 10

Original: jim carrey can do anything . i thought this was going to be some **dumb childish** movie , and it TOTALLY was not . it was so incredibly funny for EVERYONE , adults & kids . i **saw** it once cause it was almost out of theatres , and now it 's FINALLY coming out on DVD this tuesday and i 'm way to excited , as you can see . you should definitely see it if you have n't already , it was so **great** ! Liz

Perturbed: jim carrey can do anything . i thought this was going to be some **idiotic puerile** movie , and it TOTALLY was not . it was so incredibly funny for EVERYONE , adults & kids . i **noticed** it once cause it was almost out of theatres , and now it 's FINALLY coming out on DVD this tuesday and i 'm way to excited , as you can see . you should definitely see it if you have n't already , it was so **awesome** ! Liz

Correct label: positive.
Model confidence on original example: 83.3.