Information-Theoretic Measures of Dataset Difficulty

Kawin Ethayarajh[♡*] **Yejin Choi**[♣◇] **Swabha Swayamdipta**[♣]
Stanford University[♡]
Allen Institute for Artificial Intelligence[♣]
Paul G. Allen School of Computer Science, University of Washington[◇]
kawin@stanford.edu {yejinc,swabhas}@allenai.org

Abstract

Estimating the difficulty of a dataset typically involves comparing state-of-the-art models to humans; the bigger the performance gap, the harder the dataset is said to be. Not only is this framework informal, but it also provides little understanding of how difficult each instance is, or what attributes make it difficult for a given model. To address these problems, we propose an information-theoretic perspective, framing dataset difficulty as the absence of *usable information*. Measuring usable information is as easy as measuring performance, but has certain theoretical advantages. While the latter only allows us to compare different models w.r.t the same dataset, the former also allows us to compare different datasets w.r.t the same model. We then introduce *pointwise $\mathcal{V}$-information* (PVI) for measuring the difficulty of individual instances, where instances with higher PVI are easier for model $\mathcal{V}$. By manipulating the input before measuring usable information, we can understand *why* a dataset is easy or difficult for a given model, which we use to discover annotation artefacts in widely-used benchmarks.

1 Introduction

Despite datasets being the means by which we track progress in modeling, many bear limited semblance to the real-world tasks they purport to reflect (Torralba and Efros, 2011; Recht et al., 2019). Any analysis of a dataset should involve a measurement of dataset difficulty *with respect to the models being evaluated*, yet this relationship has not been well formalized. In fact, estimating difficulty has typically been limited to comparing state-of-the-art models to humans; the bigger the performance gap, the harder the dataset is said to be (Ethayarajh and Jurafsky, 2020). Existing approaches to difficulty estimation provide little understanding of how difficult each instance is (Vodrahalli et al., 2018), or what attributes make the dataset easy or difficult for a given model (Ribeiro et al., 2016).

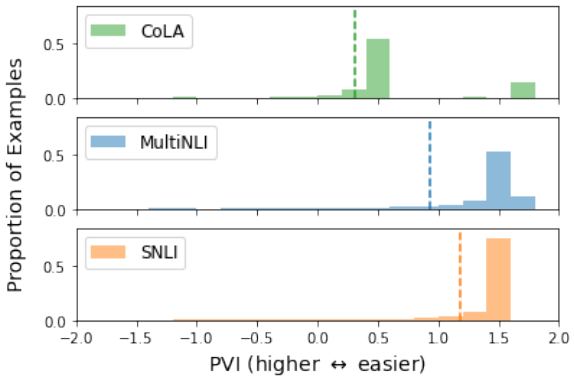

Figure 1: The Stanford NLI dataset contains more BERT-usable information than MultiNLI and CoLA, making it easier for BERT. Above, the distribution of instance difficulty (PVI) in the held-out sets w.r.t. BERT-base. The dotted lines denote the average PVI.

Heuristics such as vocabulary (Bengio et al., 2009) and input length (Spitkovsky et al., 2010; Gururangan et al., 2018) have sometimes been used as proxies for instance difficulty. However, being model-agnostic, they offer limited insight. Estimates relying on model training dynamics (Swayamdipta et al., 2020; Toneva et al., 2018), gradient magnitudes (Vodrahalli et al., 2018), or loss magnitudes (Han et al., 2018) are sensitive to factors such as variance due to initial parameterization. Other work has treated instance difficulty as learned parameters, allowing estimates to vary depending on what the generative process and parameter distributions are assumed to be (Lalor et al., 2018; Rodriguez et al., 2021). Moreover, these prior approaches do not formally relate dataset difficulty to the model being evaluated.

In this work, we provide a formal, information-theoretic treatment of dataset difficulty by framing it as the absence of *usable information* for a model (§2). For example, consider a model family $\mathcal{V}$ that can learn to map a sentence X with its sentiment

[*]Work done during an internship at AI2.

Y. If we encrypted X, predicting the sentiment would be a lot more difficult for $\mathcal{V}$. But why? The information X contains about Y would not be removed; the Shannon mutual information $I(X;Y)$ would be unchanged (Shannon, 1948). Intuitively, the task is easier when X is unencrypted because the information it contains is *usable* by $\mathcal{V}$; when X is encrypted, the information still exists but becomes unusable. This quantity—*usable information*—reflects the ease of predicting Y given X using $\mathcal{V}$, as proposed by Xu et al. (2019). It can be measured using the *predictive $\mathcal{V}$-information* framework, which generalizes Shannon information to consider computational constraints. The higher the $\mathcal{V}$-information $I_\mathcal{V}(X \to Y)$, the easier the dataset is for $\mathcal{V}$.

Measuring usable information has theoretical advantages over measuring model performance. Accuracy allows us to compare different models w.r.t. the same dataset, but not different datasets w.r.t. the same model. In contrast, $\mathcal{V}$-information permits such comparisons. In Figure 1, we can see that even datasets for the same task, i.e. natural language inference, contain different amounts of BERT-usable information, as shown in the plots for SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018).

We then introduce a new measure called *pointwise $\mathcal{V}$-information* (PVI) for estimating the difficulty of each instance in a dataset, where higher PVI instances are easier for $\mathcal{V}$ (§3). On datasets containing more information usable by large language models, such as SNLI, PVI estimates are highly correlated (Pearson $r > 0.75$) across different models, seeds, and training time. On datasets with less usable information, such as CoLA (Warstadt et al., 2018), these correlations are weaker ($0.05 \leq r \leq 0.55$)—this is intuitive, since if X and Y were independent (i.e., there were no usable information), the correlation would tend to zero.

By transforming X to isolate an attribute a before calculating $I_\mathcal{V}(X_a \to Y)$, we can also understand *why* the dataset is easy or difficult for $\mathcal{V}$ (§4). For example, by shuffling the tokens in X, we can estimate how much usable information the token identity contains about Y. Using this method, we provide several insights:

- Some attributes are more useful for certain classes. For example, in SNLI, the hypothesis-premise overlap is only useful for predicting 'entailment' instances.
- Annotation artefacts can be identified by the change in $\mathcal{V}$-information upon dropping tokens. In CoLA, auxiliary verbs and prepositions are artefacts of non-grammaticality.
- In a popular dataset for hate speech detection (Davidson et al., 2017), just 50 (potentially) offensive words contain most of the BERT-usable information about the label.

Our insights into dataset difficulty promise to offer a truer picture of the progress in natural language understanding. Moreover, a careful investigation of dataset difficulty opens up avenues for creating richer resources of data (§6).

2 The Absence of Usable Information

2.1 Theory

Consider a model family $\mathcal{V}$, which can be trained to map text input X to its label Y. If we encrypted the text, it would be harder to predict Y given X using the *same* $\mathcal{V}$. How might we measure this increase in difficulty? Shannon and Weaver (1949)'s mutual information $I(X;Y)$ is not an option—it would not change after X is encrypted, as it allows for unbounded computation, including any computation needed to decrypt the text.

Intuitively, the task is easier when X is *unencrypted* because the information it contains is *usable* by $\mathcal{V}$; when X is encrypted, this information still exists but becomes unusable. By measuring this quantity—**usable information**—we can thus measure the difficulty of a dataset w.r.t. $\mathcal{V}$. Usable information can be measured under a framework called **predictive $\mathcal{V}$-information**, which generalizes Shannon information to measure how much information can be extracted from X about Y when constrained to functions $\mathcal{V}$, written as $I_\mathcal{V}(X \to Y)$ (Xu et al., 2019). The greater $I_\mathcal{V}(X \to Y)$, the easier the dataset is for $\mathcal{V}$. If $\mathcal{V}$ is the set of all functions—i.e., unbounded computation—then $\mathcal{V}$-information reduces to Shannon information.

Processing the input with τ (e.g., by decrypting the text) can make prediction easier, allowing $I_\mathcal{V}(\tau(X) \to Y) \geq I_\mathcal{V}(X \to Y)$. Although this violates the data processing inequality, it explains the usefulness of certain types of processing, such as representation learning. Compared to X, the learned representations cannot have more Shannon information with Y, but they can have more usable information. As defined in Xu et al. (2019):

Definition 2.1 (Predictive $\mathcal{V}$-Entropy) *Let X, Y denote random variables with sample spaces $\mathcal{X}, \mathcal{Y}$*

respectively. Let $\varnothing$ denote a null input that provides no information about Y. Given predictive family $\mathcal{V} \subseteq \Omega = \{f : \mathcal{X} \cup \varnothing \to P(\mathcal{Y})\}$, the ***$\mathcal{V}$-entropy*** *is*

$$H_\mathcal{V}(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[\varnothing](Y)] \quad (1)$$

and the ***conditional $\mathcal{V}$-entropy*** *is*

$$H_\mathcal{V}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[X](Y)] \quad (2)$$

Put simply, $f[X]$ and $f[\varnothing]$ produce a probability distribution over the labels. The goal is to find the $f \in \mathcal{V}$ that maximizes the log-likelihood of the label data with (2) and without (1) the input. $f[\varnothing]$ models the label entropy, so $\varnothing$ can be set to an empty string for most NLP tasks. Although *predictive family* has a technical definition[1], state-of-the-art NLP models, provided they are finetuned without any frozen parameters, easily meet this definition[2]. Further, as per Xu et al. (2019):

Definition 2.2 (Predictive $\mathcal{V}$-Information) *Let X and Y denote random variables with sample spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given a predictive family $\mathcal{V}$, the* ***$\mathcal{V}$-information*** *is*

$$I_\mathcal{V}(X \to Y) = H_\mathcal{V}(Y) - H_\mathcal{V}(Y|X) \quad (3)$$

Because we are estimating this quantity on a finite dataset, the estimate can differ from the true $\mathcal{V}$-information. Xu et al. (2019) provide PAC bounds for this error, where less complex $\mathcal{V}$ and larger datasets yield tighter bounds. Xu et al. (2019) also list several useful properties of $\mathcal{V}$-information:

Proposition 2.1 *Let X and Y denote random variables with sample spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Given predictive families $\mathcal{U}$ and $\mathcal{V}$:*

1. *$I_\mathcal{V}(X \to Y) \geq 0$*
2. *If X is independent of Y, $I_\mathcal{V}(X \to Y) = 0$.*
3. *If $\mathcal{U} \subseteq \mathcal{V}$, then $H_\mathcal{U}(Y) \geq H_\mathcal{V}(Y)$ and $H_\mathcal{U}(Y|X) \geq H_\mathcal{V}(Y|X)$.*

2.2 Implications

$\mathcal{V}$-information allows us to compare

i. different models (i.e., different $\mathcal{V}$) by computing $I_\mathcal{V}(X \to Y)$ for the same X, Y (Fig. 2),
ii. different datasets $\{(x, y)\}$ by computing $I_\mathcal{V}(X \to Y)$ for the same $\mathcal{V}$ (Fig. 1), and
iii. different input variables X_i by computing $I_\mathcal{V}(X_i \to Y)$ for the same $\mathcal{V}$ and Y (Fig. 5).

Measuring usable information has many theoretical advantages over measuring performance while being almost as easy to compute.

Easy-to-Calculate: Given that training a model with cross-entropy loss finds the $f \in \mathcal{V}$ that maximizes the log-likelihood of Y given X, $H_\mathcal{V}(Y|X)$ can be easily computed by standard training or by finetuning a pre-trained model (as long as the infimum is found, it doesn't matter). We estimate $H_\mathcal{V}(Y|X)$ by calculating $\mathbb{E}[-\log f[X](Y)]$ on an identically distributed held-out set,[3] Recall that Y is not the label predicted by the finetuned model, but rather the gold label. Analogously, we estimate $H_\mathcal{V}(Y)$ by training or finetuning, but with $\varnothing$ in place of X. It is important however not to maximize the log-likelihood of the training instances to the point of over-fitting, since we ultimately care about finding the infimum over the full *data distribution*, not just the training set. This also means that it is not appropriate to use $\mathcal{V}$-information to estimate the difficulty of out-of-distribution data.

Standardized Comparisons: Common classification metrics, such as accuracy or F_1 score, allow us to compare different models w.r.t the same dataset, but not different datasets w.r.t. the same model. The usable information for every dataset is measured in bits/nats (depending on the log base), allowing for comparisons across models and datasets. Additionally, consider the case where X and Y are independent: here, model accuracy would be no greater than the majority class frequency, but this frequency varies across datasets. More generally, datasets with lower label entropy are easier to predict, all else held constant. $\mathcal{V}$-information avoids this problem by factoring in the label entropy $H_\mathcal{V}(Y)$; if X, Y are independent, then the $\mathcal{V}$-information is zero (Proposition 2.1).

Efficient Comparisons: Say we wish to compare two predictive families, $\mathcal{V}$ and $\mathcal{U}$, such that $\mathcal{U} \subseteq \mathcal{V}$. Assuming both families can model the label distribution, the larger family can extract more usable information (Proposition 2.1), making the dataset at least as easy for $\mathcal{V}$ as for $\mathcal{U}$. This provably obviates the need to evaluate simpler function

[1] We refer the reader to Xu et al. (2019) for details.

[2] Not all models produce an explicit distribution over the output space (e.g., language generation models). Estimating $\mathcal{V}$-information would then require some assumptions about the task, which we will explore in future work.

[3] In our experiments, we use a test set whenever available. In practice, however, any identically distributed held-out set should suffice.

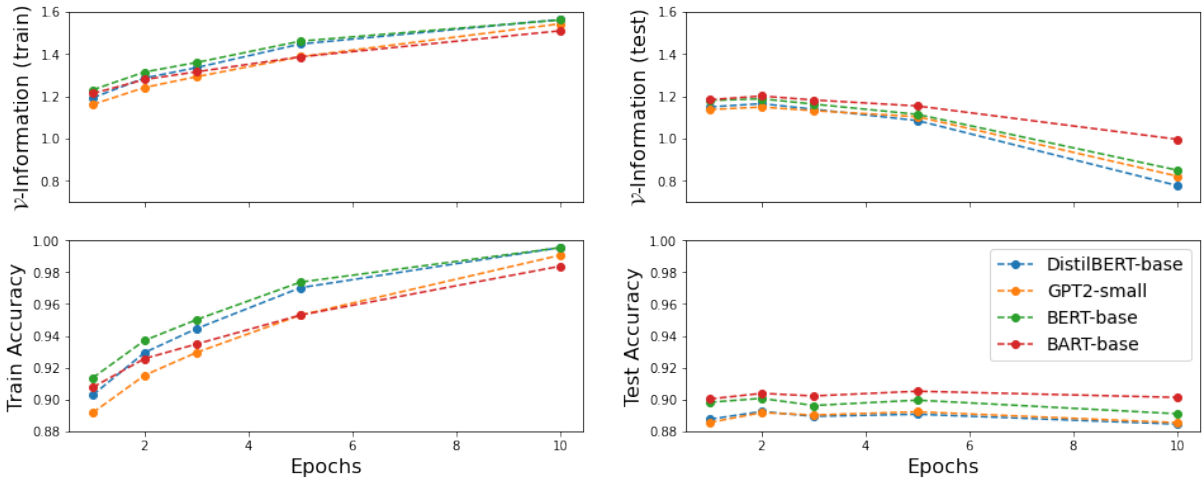

Figure 2: The usable information the input contains about SNLI (gold-standard) labels, w.r.t. various models. In the first three epochs, the $\mathcal{V}$-information estimates on the test set are similar across all models (top right), but due to over-fitting, the estimates diverge and decline by epoch 10. The test accuracy (bottom right) for each model tracks the $\mathcal{V}$-information for that model, since extracting more information makes prediction easier.

families (e.g., linear functions) when estimating dataset difficulty.

2.3 Applying $\mathcal{V}$-Information

We apply the $\mathcal{V}$-Information framework to estimate the difficulty of the Stanford natural language inference dataset (SNLI; Bowman et al., 2015), across different state-of-the-art models. SNLI is a large-scale dataset for predicting whether a text hypothesis entails, contradicts or is neutral to a text premise. The four models we use are GPT2-small (Radford et al., 2019), BERT-base-cased (Devlin et al., 2019), DistilBERT-base-uncased (Sanh et al., 2019), and BART-base (Lewis et al., 2020). Figure 2 shows the $\mathcal{V}$-information estimate for all four, as well as their accuracy on the SNLI train and held-out (test) sets, across 10 training epochs.

Model performance track $\mathcal{V}$-information. As seen in the right column of Figure 2, at each epoch, the model with the most $\mathcal{V}$-information on the SNLI test set is also the most accurate. The relative ordering of models also remains consistent across epochs. This is intuitive, since extracting more information makes prediction easier. Overall, BART-base extracts the most $\mathcal{V}$-information, followed by BERT-base, DistilBERT-base, and GPT2-small.

Figure 2 also shows why $\mathcal{V}$-information should, when possible, be estimated on a held-out set instead of the same data used for training or finetuning the model. As the model overfits to the training data, the amount of usable information in the training set grows at the expense of usable information in the test set. Ultimately, we want to maximize the log-likelihood of the random variable Y given X, not just the specific instances in the training data. Finetuning and estimating $\mathcal{V}$-information on the same set of examples thus risks the estimate being further from the true $\mathcal{V}$-information (as calculated over the actual distribution $P(X, Y)$).

$\mathcal{V}$-information is more sensitive to over-fitting than held-out performance. At epoch 10, the $\mathcal{V}$-information is at its lowest for all models, although the SNLI test accuracy has only declined slightly from its peak. This is because the models start becoming less certain about the correct label long before they start predicting the wrong label. This causes $H_\mathcal{V}(Y|X)$ to rise—and thus $I_\mathcal{V}(X \to Y)$ to decline—even while most of the probability mass is still placed on the correct label. This suggests that, compared to performance metrics, $\mathcal{V}$-information can more readily inform us of over-fitting.

3 Measuring Pointwise Difficulty

While $\mathcal{V}$-information provides an aggregate measure of dataset difficulty (§2), a closer analysis requires measuring the degree of usable information in individual instances. We extend the $\mathcal{V}$-information framework to introduce a new measure called **pointwise $\mathcal{V}$-information** (PVI) for individual instances. The higher the PVI, the easier the instance is for $\mathcal{V}$.

Definition 3.1 (Pointwise $\mathcal{V}$-Information)

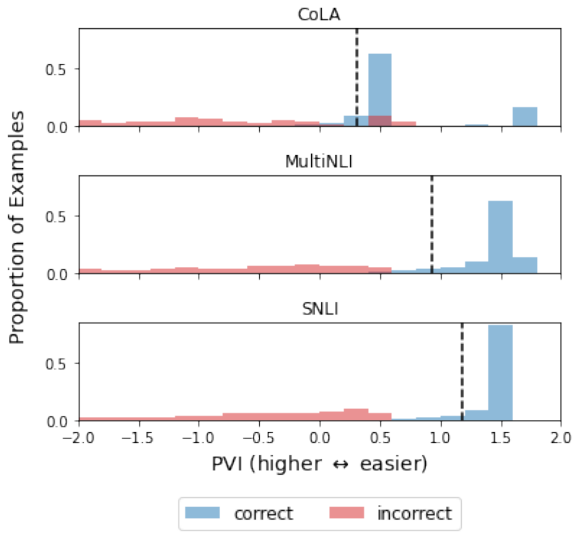

Figure 3: The normalized distribution of PVI values (w.r.t. BERT-base) for correctly and incorrectly predicted instances in CoLA, SNLI, and MultiNLI. The higher the PVI, the more likely an instance is to be predicted correctly by BERT.

Algorithm 1 PVI and $\mathcal{V}$-Information

Input: training data $\mathcal{D}_{\text{train}} = \{(\text{input } x_i, \text{gold label } y_i)\}_{i=1}^{m}$, held-out data $\mathcal{D}_{\text{test}} = \{(\text{input } x_i, \text{gold label } y_i)\}_{i=1}^{n}$, model $\mathcal{V}$
do
  $g' \leftarrow$ Finetune $\mathcal{V}$ on $\mathcal{D}_{\text{train}}$
  $\varnothing \leftarrow$ empty string (null input)
  $g \leftarrow$ Finetune $\mathcal{V}$ on $\{(\varnothing, y_i) \mid (x_i, y_i) \in \mathcal{D}_{\text{train}}\}$

  $H_\mathcal{V}(Y), H_\mathcal{V}(Y|X) \leftarrow 0, 0$
  for $(x_i, y_i) \in \mathcal{D}_{\text{test}}$ **do**
    $H_\mathcal{V}(Y) \leftarrow H_\mathcal{V}(Y) - \frac{1}{n} \log_2 g[\varnothing](y_i)$
    $H_\mathcal{V}(Y|X) \leftarrow H_\mathcal{V}(Y|X) - \frac{1}{n} \log_2 g'[x_i](y_i)$
    $\text{PVI}(x_i \to y_i) \leftarrow -\log_2 g[\varnothing](y_i) + \log_2 g'[x_i](y_i)$
  end for

  $\mathcal{V}$-Information $\hat{I}_\mathcal{V}(X \to Y) = \frac{1}{n} \sum_i \text{PVI}(x_i \to y_i)$
    $= H_\mathcal{V}(Y) - H_\mathcal{V}(Y|X)$
end do

Given random variables X, Y and a predictive family $\mathcal{V}$, the pointwise $\mathcal{V}$-information (PVI) of an instance (x, y) is

$$\text{PVI}(x \to y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y) \tag{4}$$

where $g, g' \in \mathcal{V}$, $\mathbb{E}[-\log g[\varnothing](Y)] = H_\mathcal{V}(Y)$ and $\mathbb{E}[-\log g'[X](Y)] = H_\mathcal{V}(Y|X)$.

If $\mathcal{V}$ were the BERT function family, g' and g would be the BERT models after finetuning on the training data with and without the input respectively. For a held-out instance (x, y), $\text{PVI}(x \to y)$ is the difference in the log-probability these finetuned models place on the gold label. PVI is to $\mathcal{V}$-information what PMI is to Shannon information:

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PMI}(x, y)] \\ I_\mathcal{V}(X \to Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)] \end{aligned} \tag{5}$$

Because of this relationship, our understanding of $\mathcal{V}$-information extends to PVI as well: higher PVI instances are easier for $\mathcal{V}$; lower PVI instances are more difficult. In Algorithm 1, we describe how the $\mathcal{V}$-information can be computed explicitly using the $\mathcal{V}$-entropies or by averaging over PVI.

The PVI of an instance (x, y) w.r.t. $\mathcal{V}$ should only depend on the distribution of the random variables. Sampling more from $P(X, Y)$ during finetuning should not change $\text{PVI}(x \to y)$ much. However, an instance can be drawn from different distributions, in which case we would expect its PVI to differ. For example, say we have restaurant reviews and movie reviews, along with their sentiment. The instance ('That was great!', *positive*) could be drawn from either distribution, but we would expect its PVI to be different in each (even though $\mathcal{V}$ is the same).

3.1 Applying PVI

We apply PVI to two datasets in addition to SNLI: MultiNLI (Williams et al., 2018) is a multi-genre counterpart of SNLI, and CoLA (Warstadt et al., 2018) is a dataset for linguistic acceptability where each sentence is labeled as grammatical or not. We summarize key findings below.

Correctly predicted instances have higher PVI values than incorrectly predicted ones. In Figure 3, we plot the normalized distribution of PVI values for CoLA, SNLI, and MultiNLI held-out examples that were correctly and incorrectly predicted by fully fine-tuned BERT-base. A higher PVI increases the odds of being predicted correctly—this is intuitive because a correct prediction of a non-majority-class instance requires that some information be extracted from the instance. Although the $\mathcal{V}$-information cannot be negative (Proposition 2.1), the PVI can be—much like how PMI can be negative even though Shannon information cannot. A negative PVI simply means that the model is better off predicting the majority class than considering X, which can happen for many reasons (e.g., mislabelling). As seen in Figure 3, examples with negative PVI can still be predicted correctly, as long as g' places most of the probability mass on the correct label.

Instances with a lower rate of annotator agreement have a lower PVI on average. In Figure 4,

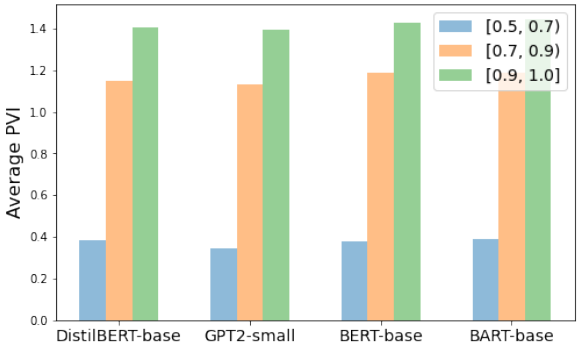

Figure 4: The average PVI for SNLI instances grouped by the level of annotator agreement (e.g., [0.9, 1.0] means that 90-100% of the annotators agreed with the gold label). The lower the annotator agreement, the lower the average PVI, suggesting a correspondence between what humans and models find difficult.

we group the SNLI test instances by the level of annotator agreement—the proportion of annotators who agree with the gold label—and plot the average PVI for each group. There is a consistent trend across all four models we evaluated: the lower the annotator agreement, the lower the average PVI. This suggests a strong correspondence between what humans and models find difficult.

When there is much usable information, PVI estimates are highly consistent across models, training epochs, and random initializations. In Table 1, we can see that the cross-model Pearson correlation between PVI estimates of SNLI instances is very high ($r > 0.80$). However, this result does not extend to all datasets: as seen in Table 1, the cross-model Pearson correlation is much lower for CoLA ($0.05 < r < 0.55$). This is because, as visualized in Figure 1, CoLA has much less usable information, making difficulty estimates noisier. In the limit, if a dataset contained no usable information, then we would expect the correlation between PVI estimates across different models and seeds to be close to zero.

It is also worth noting, however, that a high degree of cross-model correlation—as with SNLI—does not preclude comparisons between different models on the same dataset. Rather, it suggests that in SNLI, a minority of instances is responsible for distinguishing one model's performance from another. This is not surprising—given the similar complexity and architecture of these models, we would expect most instances to be equally easy.

For all models finetuned on SNLI, the Pearson correlation (between PVI estimates made by the same model) across training epochs is above 0.80 during the first five epochs (Appendix C). Also, despite the performance of Transformer-based models varying across random initializations (Dodge et al., 2019, 2020; Mosbach et al., 2020), we find that PVI estimates are quite stable: the correlation across seeds is $r > 0.85$ (for SNLI finetuned BERT-base, across 4 seeds). In other words, SNLI examples that are more (less) difficult in one setting tend to remain more (less) difficult across models, training time, and seeds.

SNLI				
	DistilBERT	GPT2	BERT	BART
DistilBERT	1.000	0.821	**0.846**	0.815
GPT2	0.821	1.000	0.809	0.822
BERT	**0.846**	0.809	1.000	0.832
BART	0.815	0.822	0.832	1.000
CoLA				
	DistilBERT	GPT2	BERT	BART
DistilBERT	1.000	0.294	**0.515**	0.476
GPT2	0.294	1.000	0.138	0.086
BERT	**0.515**	0.138	1.000	0.248
BART	0.476	0.086	0.248	1.000

Table 1: Cross-model Pearson's r between PVI estimates made by different finetuned models, on the SNLI and CoLA test sets. For SNLI, the estimates are consistent: what one model finds difficult, others find difficult as well. Since CoLA has less usable information for all these models, the correlations are much lower. If a dataset had no usable information, we would expect the correlation to be close to zero.

4 Uncovering Dataset Artefacts

A key limitation of past work on estimating data difficulty is the lack of interpretability; there is no straightforward way to understand *why* a dataset is as difficult as it is. Readers may however recall from §2.2 that $\mathcal{V}$-information offers an approach to compare different input variables X_i under the same $\mathcal{V}$ and Y.

In this section, we apply different transformations $\tau_a(X)$ to isolate an attribute a and then calculate $I_\mathcal{V}(\tau_a(X) \to Y)$ to measure how much information (usable by $\mathcal{V}$) the attribute contains about the label. For example, by shuffling the tokens in X, we can isolate the influence of token identity. Given that a transformation may make information more accessible (e.g., decrypting some encrypted text; c.f. §2), it is possible for $I_\mathcal{V}(\tau_a(X) \to Y) \geq I_\mathcal{V}(X \to Y)$, so the latter

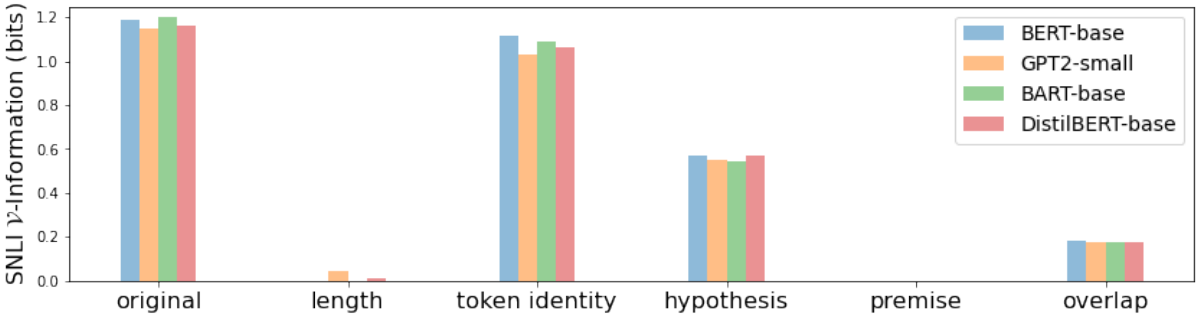

Figure 5: The amount of usable information for four models contained in different input attributes about the (gold-standard) labels in SNLI. The TOKEN IDENTITY alone (regardless of order) provides most of the information for all models. The PREMISE, which can be shared by multiple instances, is useless alone; the HYPOTHESIS, which is unique to an instance, is quite useful even without a premise, suggesting it may contain annotation artefacts. Note that different attributes can contain overlapping information.

shouldn't be treated as an upper bound. Such transformations were applied by O'Connor and Andreas (2021) to understand what syntactic features Transformers use in next-token prediction; we take this a step further, aiming to discover annotation artefacts, compare individual instances, and ultimately understand the dataset itself. We present below key findings on SNLI as well as DWMW17 (Davidson et al., 2017), a dataset for hate speech detection, where input posts are labeled as hate speech, offensive, or neither.

Token identity alone provides most of the usable information in SNLI. We apply various transformations to the SNLI input to isolate different attributes (see Appendix B for an example):

Token Identity: shuffle tokens randomly
Length: replace each token with token #
Hypothesis: only include the hypothesis
Premise: only include the premise
Overlap: only include tokens in the hypothesis that also appear in the premise

As seen in Figure 5, the token identity alone contains most of the usable information for all models. The premise, which is typically shared by multiple instances, is useless alone; the hypothesis, which is unique to an instance, is quite useful even without a premise, hinting at annotation artefacts.

Certain attributes are more useful for certain classes. Comparing the usefulness of an attribute across classes can be useful for identifying systemic annotation artefacts: we do this for SNLI in Table 2. For example, we see that the hypothesis-premise overlap contains much more BERT-usable information about the 'entailment' class than 'contradiction' or 'neutral'. If there is no inherent reason for an attribute to be more/less useful—such as overlap for entailment—there may be an artefact at work. Conversely, although input length is a known annotation artefact—with entailment instances being shorter than average in SNLI and neutral instances being longer (Gururangan et al., 2018)—all models mostly fail to exploit this artefact. This is likely due to the inability of Transformer-based architectures to count and compare numbers (Wallace et al., 2019).

	Entailment	Neutral	Contradiction
original	1.188	1.064	1.309
length	0.085	-0.074	-0.014
token identity	1.130	0.984	1.224
hypothesis	0.573	0.553	0.585
premise	0.032	-0.016	-0.016
overlap	0.271	0.049	0.224

Table 2: The average amount of usable information (i.e., mean PVI, in bits) that each attribute contains about each class in SNLI, according to BERT-base. Some attributes are more useful for a particular class: e.g., the degree of premise-hypothesis overlap is most useful for predicting 'entailment'. Note that the mean PVI for a particular class is different from the $\mathcal{V}$-information, as the latter is the mean over *all the data*.

Certain attributes are responsible for the difficulty of certain examples. Figure 6 is an example of how we might do a fine-grained comparison of instances to understand why one may be more difficult for a given model. We compare two SNLI 'neutral' instances from the test set to try to understand why #9627 is easier for BERT than #7717 (i.e., why $\text{PVI}(x_{9627} \rightarrow y_{9627}) > \text{PVI}(x_{7717} \rightarrow y_{7717})$), finding that it is likely due to the former's *hypothesis* being more informa-

#7717: *PREMISE: Little kids play a game of running around a pole. HYPOTHESIS: The kids are fighting outside.*
#9627: *PREMISE: A group of people watching a boy getting interviewed by a man. HYPOTHESIS: A group of people are sleeping on Pluto.*

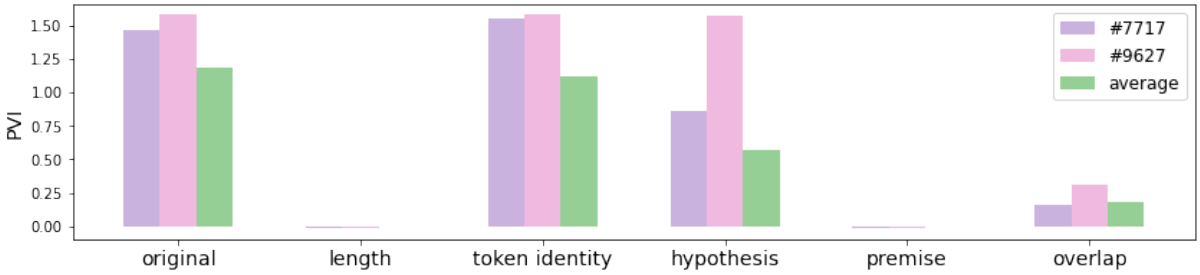

Figure 6: The PVI of two SNLI 'neutral' instances (#7717 and #9627) w.r.t. BERT-base after attribute-specific transformations, as well as the $\mathcal{V}$-information estimate (i.e., average PVI over the data) for each attribute. The latter instance is easier for BERT, likely because its hypothesis is much more informative due to being so different from its premise. Note that it makes sense to compare instances w.r.t. the same attribute, but not different attributes w.r.t. the same instance, since the models used to estimate the attribute $\mathcal{V}$-information $I_{\mathcal{V}}(\tau_a(X) \to Y)$ are chosen to maximize the likelihood of all the data.

tive. While different instances can be compared w.r.t. the same attribute, different attributes cannot be compared w.r.t. the same instance, since the models used to estimate the attribute-specific $\mathcal{V}$-information $I_{\mathcal{V}}(\tau_a(X) \to Y)$ are chosen to maximize the likelihood of *all the data*. This is why, for example, the PVI of #7717 is higher after its tokens have been shuffled even though the average PVI (i.e., dataset-level $\mathcal{V}$-information) declines after shuffling tokens.

Hate speech detection datasets may be easier than they seem. Automatic hate speech detection is an increasingly important part of online moderation, but what causes a model to label speech as offensive? We find that DWMW17 contains 0.724 bits of BERT-usable information. Additionally, if one removed all the tokens, except for 50 (potentially) offensive ones—comprising mostly of common racial and homophobic slurs–from the input post, there still remains 0.490 bits of BERT-usable information. In other words, just 50 (potentially) offensive words contain most of the BERT-usable information in DWMW17. Allowing models to do well by simply pattern-matching may permit more subtle forms of hate speech to go undetected, perpetuating harm towards minority groups.

Token-level annotation artefacts can be discovered using leave-one-out. Transforming X and then measuring the $\mathcal{V}$-information to discover token-level artefacts is untenable, since we would need to finetune one new model per token. Instead, we compute the change in the $\mathcal{V}$-information estimate using the same model g' but on a subset $\mathcal{D}_{C,t}$ of the data[4]. Let $x_{\neg t}$ denote an input after removing t. This simplifies to measuring the increase in conditional entropy:

$$\frac{1}{|\mathcal{D}_{C,t}|} \sum_{\mathcal{D}_{C,t}} [-\log_2 g'[x_{\neg t}](y) + \log_2 g'[x](y)]$$

In CoLA, auxiliary verbs (e.g., be, did) and prepositions are artefacts of ungrammatical sentences; in contrast, grammatical sentences have no artefacts, with no word on average increasing the conditional entropy above 0.30 upon omission. In DWMW17, racial and homophobic slurs are the top indicators of 'hate speech'. In SNLI, many of the token-level artefacts match those found using descriptive statistics in Gururangan et al. (2018). The word lists are available in Table 6 in Appendix D.

5 Related Work

Dataset cartography uses training dynamics to visualize datasets (Swayamdipta et al., 2020). It follows earlier work that used the training loss (Han et al., 2018; Arazo et al., 2019; Shen and Sanghavi, 2019), confidence (Hovy et al., 2013), prediction variance (Chang et al., 2017), and area under the curve (Pleiss et al., 2020) to differentiate instances. Instances are plotted on two axes: (1) *confidence* (i.e., mean probability of the correct label across epochs), and (2) *variability* (i.e., variance of the former); the former values per instance track closely with PVI values, especially when there is a high amount of usable information for a model;

[4]The instances which contain the token t and belong to the class C of interest.

see Fig. 7 in Appendix E. Given their dependence on training behavior across time, cartography offers complimentary benefits to $\mathcal{V}$-information. Indeed, $\mathcal{V}$-information provides a formal framework to make dataset difficulty estimates as an aggregate, but it is non-trivial to measure differences between, say a CoLA data map and an SNLI data map, with respect to BERT.

A more recent line of work allows the difficulty of instances to be learned via item response theory (IRT) (Embretson and Reise, 2013). Instance "difficulties" are treated as parameters in a topic model meant to explain model performance (Rodriguez et al., 2021). With IRT, however, estimates can vary depending on the generative process assumed by the topic model and the assumed parameter distributions. Because this method does not consider the input text—only whether a model's prediction was correct—it also cannot be easily adapted to understanding why an example is difficult.

Estimating instance difficulty is evocative of instance selection for active learning (Fu et al., 2013; Liu and Motoda, 2002). Uncertainty sampling, for example, picks the instances that the partially trained model is least certain about (Lewis and Catlett, 1994; Lewis and Gale, 1994; Nigam et al., 2000), which could be interpreted as a measure of difficulty. However, once an instance is sampled and used for training, the model may become much more certain about it, meaning that original uncertainty values are not stable estimates. AFLite (Le Bras et al., 2020) is an adversarial filtering algorithm for iteratively removing "predictable" instances, where predictability is determined at each iteration using a linear classifier. Although unpredictability is similar to our notion of difficulty, AFLite changes the data distribution with each iteration, causing the estimates to change as well. Its iterative nature means that it is also limited to simple models (e.g., linear classifiers).

Influence functions (Koh and Liang, 2017), forgetting events (Toneva et al., 2018), and the Data Shapley (Ghorbani and Zou, 2019; Jia et al., 2019) can all be used to assign pointwise estimates of importance to data instances, but importance tends to refer to their contribution to the decision boundary. Other work has offered insight by splitting the data into "easy" and "hard" sets with respect to some attribute and studying changes in model performance, but these methods do not offer a pointwise estimate of difficulty. (Sugawara et al., 2018; Rondeau and Hazen, 2018; Sen and Saffari, 2020). In NLP, $\mathcal{V}$-information has been used to study what context features Transformers actually use (O'Connor and Andreas, 2021), as well as to condition out information during probing (Hewitt et al., 2021), but not for estimating difficulty. Our approach to discovering dataset artefacts can also complement existing approaches to artefact discovery (Gardner et al., 2021; Pezeshkpour et al., 2021).

6 Future Work

Our findings open up many lines of future work. There has been much work in the way of model interpretability, but relatively little in the way of dataset interpretability. Our framework will allow datasets to be probed, helping us understand what exactly we're testing for in models and how pervasive annotation artefacts really are. By identifying the attributes responsible for difficulty, it will be possible to build challenge sets in a more principled way and reduce artefacts in existing datasets. By studying which attributes contain information that is non-usable by existing SOTA models, model creators may also be able to make more precise changes to architectures.

Although $\mathcal{V}$-information is easy to calculate, there remains a challenge in applying it to tasks where models don't produce an explicit probability distribution over the entire output (e.g., machine translation with beam search). Addressing such cases will require breaking down text-to-text tasks into more tractable sub-tasks.

7 Conclusion

We provided an information-theoretic perspective of dataset difficulty by framing it as the absence of usable information. We extended *predictive $\mathcal{V}$-information* to estimate difficulty at the dataset level, and then introduced *pointwise $\mathcal{V}$-information* (PVI) for measuring the difficulty of individual instances. Measuring $\mathcal{V}$-information was found to have a number of theoretical advantages over measuring performance. For datasets with more usable information, PVI estimates were found to be more consistent across training time, different models, and different seeds; for datasets with less usable information, PVI estimates are less consistent across settings. We then demonstrated how systemic and token-level annotation artefacts could be discovered by manipulating the input before calculating these measures. In summary, $\mathcal{V}$-information offers

a new, efficient means of interpreting the quality of large datasets, complimenting existing methods in dataset analysis.

Acknowledgements

We thank Dan Jurafsky and Nelson Liu for their helpful comments.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30:1002–1012.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency problems: On finding and removing artifacts in language data.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Huan Liu and Hiroshi Motoda. 2002. On issues of instance selection. *Data Mining and Knowledge Discovery*, 6(2):115.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 851–864, Online. Association for Computational Linguistics.

Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron C Wallace. 2021. Combining feature and instance attribution to detect artifacts. *arXiv preprint arXiv:2107.00323*.

Geoff Pleiss, Tianyi Zhang, Ethan R Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Marc-Antoine Rondeau and Timothy J Hazen. 2018. Systematic error analysis of the stanford question answering dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 12–20.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438.

Claude E Shannon and Warren Weaver. 1949. The mathematical theory of com-munication. *Urbana: University of Illinois Press*, 96.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Yanyao Shen and Sujay Sanghavi. 2019. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR.

Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, Los Angeles, California. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Kailas Vodrahalli, Ke Li, and Jitendra Malik. 2018. Are all training examples created equal? an empirical study.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2019. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

A Examples

In Table 3, we list the 10 hardest instances in the SNLI test set according to BERT-base. All three classes—entailment, neutral, and contradiction—are represented in this list, with entailment being slightly over-represented. We see that some of the examples are in fact mislabelled—e.g., 'PREMISE: An Asian woman dressed in a colorful outfit laughing. HYPOTHESIS: The women is not laughing.' is labelled as 'entailment' even though the correct label is 'contradiction'.

B Transformations

In Table 4, we provide an instance from the SNLI test set in its original form and after various attribute-specific transformations have been applied to it. These only capture a small subset of the space of possible transformations.

C Cross-Epoch Correlations

In Table 5, we list the cross-epoch Pearson correlation between PVI estimates made by the same model on the SNLI test set over the course of fine-tuning. The correlation is high ($r > 0.80$ during the first 5 epochs), suggesting that when an instance is easy(difficult) early on, it tends to remain easy(difficult).

D Token-Level Artefacts

WARNING: The following content contains language from the DWMW17 dataset that is offensive in nature. In Table 6, we list the tokens in the SNLI, CoLA, and DWMW17 datasets that, when dropped out, cause the greatest decrease in the $\mathcal{V}$-information estimate. These are token-level artefacts of each class in the dataset. In the DWMW17 hate speech detection dataset, racial and homophobic slurs are artefacts of hate speech, while ableist and sexual slurs are artefacts of offensive speech. In-group AAVE terms are also predictive of offensive speech in DWMW17 even when they are used non-offensively, hinting at possible bias in the dataset (Sap et al., 2019). In CoLA, auxiliary verbs and prepositions are artefacts of ungrammatical sentences; grammatical sentences don't have any artefacts. For SNLI, we recover many of the token-level artefacts found by Gururangan et al. (2018) using descriptive statistics—even uncommon ones, such as 'cat' for contradiction.

E Relation to Dataset Cartography

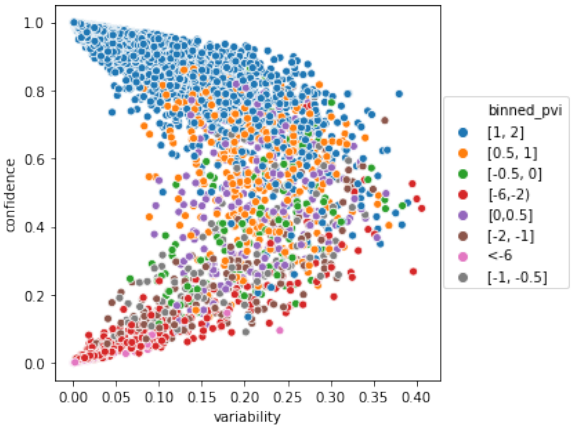

Figure 7: Relationship between PVI and the training dynamics-based data map (Swayamdipta et al., 2020) for SNLI held-out (test) set, computed for the DistilBERT-base architecture. As in , Y-axis corresponds to *confidence*, i.e. the mean probabilities of the true class across training epochs, and X-axis corresponds to *variability*, i.e. the standard deviation of the true class probabilities across the same. Colors indicate binned values of PVI. PVI estimates track closely with *confidence*.

Figure 7 shows that PVI values track closely to the confidence axis of a SNLI-DistilBERT-base data map[5] (Swayamdipta et al., 2020). Data maps and PVI estimates offer orthogonal perspectives to instance difficulty, the former capturing behavior of instances as training proceeds. Moreover, $\mathcal{V}$-information can estimate dataset difficulty as an aggregate (§2), which is not the case for training dynamics metrics, which offer only point estimates. Both approaches can be helpful for discovering data artefacts. Predictive $\mathcal{V}$-information estimates, however, offer the unique capability of transforming the input to discover the value of certain attributes in an efficient manner.

[5] Data maps were originally plotted on training data; however, they can be plotted on held-out data by computing training dynamics measures on the same, after every training epoch.

hypothesis	premise	label	PVI
A man plays the trombone on the sidewalk.	Twenty five people are marching.	N	-9.966
A person is watching TV.	A woman in a striped shirt holds an infant.	N	-9.612
A person embraces the cold	A person swimming in a swimming pool.	N	-9.152
Women are playing ping pong.	Women enjoying a game of table tennis.	E	-8.713
A boy alien dressed for summer in a green shirt and kahki shorts	A boy dressed for summer in a green shirt and kahki shorts extends food to a reindeer in a petting zoo.	E	-8.486
Two snowboarders race.	Two skateboarders, one wearing a black t-shirt and the other wearing a white t-shirt, race each other.	E	-8.087
The woman is not laughing.	An Asian woman dressed in a colorful outfit laughing.	E	-7.903
An older gentleman in overalls looks at the camera while he is building a stained red deck in front of a house.	An older gentleman looks at the camera while he is building a deck.	E	-7.709
The bandana is expensive.	A man wearing black pants, an orange and brown striped shirt, and a black bandanna in a "just thrown a bowling ball" stance.	C	-7.685
Two girls kiss.	Two girls kissing a man with a black shirt and brown hair on the cheeks.	C	-7.582

Table 3: The 10 hardest (lowest PVI) instances in the SNLI test set, according to BERT-base. 'E' denotes entailment, 'N' neutral, and 'C' contradiction.

Attribute	Transformation	Transformed Input
Original		PREMISE: I am a dog. HYPOTHESIS: The dog is brown.
Token Identity	shuffle tokens randomly	PREMISE: am dog I a . HYPOTHESIS: brown the dog is .
Length	replace each token with #	PREMISE: # # # # # HYPOTHESIS: # # # # #
Hypothesis	only include hypothesis	HYPOTHESIS: I am a dog.
Premise	only include premise	PREMISE: The dog is brown.
Overlap	hypothesis-premise overlap	dog

Table 4: Given an NLI instance (see 'Original'), each transformation isolates some attribute from the input. The headers 'PREMISE' and 'HYPOTHESIS' were added by us to transform the two sentence inputs into a single text input for all models that were evaluated.

BERT-base					
Epoch/Epoch	1	2	3	5	10
1	1.000	0.908	0.871	0.838	0.762
2	0.908	1.000	0.929	0.883	0.795
3	0.871	0.929	1.000	0.879	0.796
5	0.838	0.883	0.879	1.000	0.833
10	0.762	0.795	0.796	0.833	1.000
BART-base					
Epoch/Epoch	1	2	3	5	10
1	1.000	0.925	0.885	0.853	0.754
2	0.925	1.000	0.952	0.906	0.807
3	0.885	0.952	1.000	0.914	0.814
5	0.853	0.906	0.914	1.000	0.862
10	0.754	0.807	0.814	0.862	1.000
DistilBERT-base					
Epoch/Epoch	1	2	3	5	10
1	1.000	0.928	0.884	0.828	0.766
2	0.928	1.000	0.952	0.890	0.825
3	0.884	0.952	1.000	0.900	0.819
5	0.828	0.890	0.900	1.000	0.860
10	0.766	0.825	0.819	0.860	1.000
GPT2					
Epoch/Epoch	1	2	3	5	10
1	1.000	0.931	0.887	0.855	0.747
2	0.931	1.000	0.961	0.918	0.813
3	0.887	0.961	1.000	0.933	0.827
5	0.855	0.918	0.933	1.000	0.874
10	0.747	0.813	0.827	0.874	1.000

Table 5: Cross-epoch Pearson correlation between PVI estimates made on the SNLI test set while finetuning various models on the SNLI training set. The estimates are stable: when an instance is easy(difficult) early on, it generally remains easy(difficult). For all models studied, the cross-epoch correlation does not dip below 0.80 for the first five epochs.

DWMW17 (Davidson et al., 2017)		
Hate Speech	Offensive	Neither
faggots (3.844)	retards (2.821)	lame (4.426)
fag (3.73)	nigs (2.716)	clothes (0.646)
faggot (3.658)	negro (2.492)	dog (0.616)
coons (3.53)	nig (2.414)	cat (0.538)
niggers (3.274)	cunts (2.372)	iDntWearCondoms (0.517)
queer (3.163)	pussies (2.29)	thank (0.47)
coon (3.137)	queer (2.213)	kick (0.423)
nigger (3.094)	retarded (1.997)	30 (0.345)
dyke (3.01)	cunt (1.919)	football (0.334)
fags (2.959)	bitches (1.858)	soul (0.323)
SNLI (Bowman et al., 2015)		
Entailment	Neutral	Contradiction
nap (3.256)	tall (4.246)	Nobody (7.258)
bald (3.183)	naked (2.193)	not (4.898)
crying (2.733)	indoors (1.724)	no (4.458)
Woman (2.517)	light (1.442)	naked (3.583)
asleep (2.482)	fun (1.318)	crying (2.938)
sleeping (2.416)	bed (1.006)	indoors (2.523)
soda (2.267)	motorcycle (0.993)	vegetables (2.295)
bed (2.136)	works (0.969)	sleeping (2.293)
not (2.111)	race (0.943)	jogging (2.17)
snowboarder (2.099)	daughter (0.924)	cat (2.092)
CoLA (Warstadt et al., 2018)		
Grammatical	Ungrammatical	
will (0.267)	book (2.737)	
John (0.168)	is (2.659)	
. (0.006)	was (2.312)	
and (-0.039)	of (2.308)	
in (-0.05)	to (1.972)	
’ (-0.063)	you (1.903)	
to (-0.195)	be (1.895)	
of (-0.195)	in (1.618)	
that (-0.379)	did (1.558)	
the (-0.481)	The (1.427)	

Table 6: Token-level annotation artefacts in each dataset. These are the tokens whose omission leads to the greatest average increase in conditional entropy for each class (given in parentheses).