Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information

Kawin Ethayarajh [1] **Yejin Choi** [2 3] **Swabha Swayamdipta** [2]

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. However, this comparison provides little understanding of how difficult each instance in a given distribution is, or what attributes make the dataset difficult for a given model. To address these questions, we frame dataset difficulty—w.r.t. a model $\mathcal{V}$—as the lack of $\mathcal{V}$-*usable information* (Xu et al., 2019), where a lower value indicates a more difficult dataset for $\mathcal{V}$. We further introduce *pointwise* $\mathcal{V}$-*information* (PVI) for measuring the difficulty of individual instances w.r.t. a given distribution. While standard evaluation metrics typically only compare different models for the same dataset, $\mathcal{V}$-*usable information* and PVI also permit the converse: for a given model $\mathcal{V}$, we can compare different datasets, as well as different instances/slices of the same dataset. Furthermore, our framework allows for the interpretability of different input attributes via transformations of the input, which we use to discover annotation artefacts in widely-used NLP benchmarks.

1. Introduction

Datasets are designed to act as proxies for real-world tasks, yet most bear limited semblance to the tasks they purport to reflect (Torralba & Efros, 2011; Recht et al., 2019). Understanding dataset difficulty is therefore imperative to understanding progress in AI. In practice, however, estimating dataset difficulty is often limited to an informal comparison of state-of-the-art model performance to that of humans; the bigger the performance gap, the harder the dataset is said to be (Ethayarajh & Jurafsky, 2020; Ma et al., 2021). However, such performance metrics offer little understanding of the differential difficulty of individual instances, or of which attributes in the input a given model finds useful.

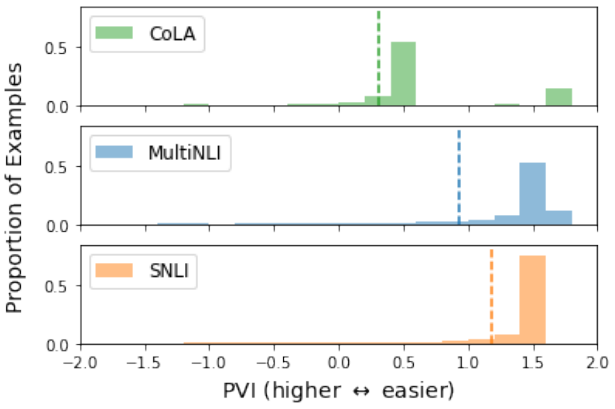

Figure 1. The Stanford NLI dataset contains more BERT-usable information than the MultiNLI and CoLA datasets, making it easier for BERT-base. Above, the distribution of instance difficulty (PVI) in the held-out sets for each; dotted lines denote the average PVI.

To understand why a dataset is difficult, we extend recent work in information theory (Xu et al., 2019). To illustrate, consider a model family $\mathcal{V}$ that can learn to map a sentence X with its sentiment Y. Even if X were to be encrypted, the information X contains about Y would not be removed; in other words, the Shannon mutual information would be unchanged (Shannon, 1948). However, encryption makes predicting the sentiment a lot more difficult for $\mathcal{V}$. But why? 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—$\mathcal{V}$-*usable information*—reflects the ease with which $\mathcal{V}$ can predict Y given X. Xu et al. (2019) show that it can be measured using the *predictive* $\mathcal{V}$-*information* framework, which generalizes Shannon information to consider computational constraints.

Our work extends the above framework by framing dataset difficulty as the lack of $\mathcal{V}$-*usable information*.[1] The higher

Work done during an internship at AI2. [1]Stanford University [2]Allen Institute for Artificial Intelligence [3]Paul G. Allen School of Computer Science, University of Washington. Correspondence to: Kawin Ethayarajh <kawin@stanford.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]We use the terms "$\mathcal{V}$-usable information" and "$\mathcal{V}$-information" from Xu et al. (2019), interchangeably.

the $\mathcal{V}$-usable information, the easier the dataset is for $\mathcal{V}$. Not only does this framework allow comparisons of models w.r.t. the same dataset, but also of different datasets w.r.t. the same model. Figure 1 illustrates that different datasets provide different amounts of usable information for the same model, even when the task is identical (i.e., natural language inference in the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) datasets).

Building on the aggregate estimate of dataset difficulty, we introduce a measure called *pointwise $\mathcal{V}$-information* (PVI) for estimating the difficulty of each instance w.r.t. a given distribution (§3). PVI estimates allow us to compare not only individual instances, but also the difficulty of slices of data w.r.t $\mathcal{V}$. On datasets containing more usable information (e.g., SNLI), PVI estimates are highly correlated (Pearson $r \geq 0.75$) across different models, seeds, and training time, and with human judgments of difficulty.

Comparisons of $\mathcal{V}$-usable information before and after isolating an input attribute shed light on *why* the dataset is easy or difficult for $\mathcal{V}$ (§4), which has significant implications for interpretability in AI[2] (Miller, 2019). Specifically, we use $\mathcal{V}$-usable information to identify some limitations in benchmarks that are widely used in NLP to test for a model's understanding of different language phenomena:

- Word ordering has a limited impact on the difficulty of a popular natural language entailment benchmark, SNLI (Bowman et al., 2015), even though entailment describes a causal relationship.
- Some of the most difficult instances in SNLI and a popular grammaticality detection benchmark, CoLA (Warstadt et al., 2018), are mislabelled.
- 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; less subtle bias may be going undetected.

2. $\mathcal{V}$-Usable Information

2.1. Background

Consider a model family $\mathcal{V}$, which can be trained to map text input X to its label Y. If we encrypted the text, or translated it into a language with a very complex grammar, it would be harder to predict Y given X using the *same* $\mathcal{V}$. How might we measure this increase in difficulty? Shannon (1948)'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 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. This quantity, called **$\mathcal{V}$-usable information**, provides an estimate of the difficulty of a dataset w.r.t. $\mathcal{V}$. It 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 $I_\mathcal{V}(X \to Y)$ is, the easier the dataset is for $\mathcal{V}$. If $\mathcal{V}$ is the set of all functions—i.e., under unbounded computation—$\mathcal{V}$-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.

2.2. Definitions

As defined in Xu et al. (2019):

Definition 2.1. 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 **predictive $\mathcal{V}$-entropy** is

$$H_\mathcal{V}(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)] \tag{1}$$

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

$$H_\mathcal{V}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)] \tag{2}$$

We use $\log_2$ to measure the entropies in bits of information, though one could also use $\log_e$ and measure them in nats instead.

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 (Eq. 2) and without the input (Eq. 1). $f[\varnothing]$ models the label entropy, so $\varnothing$ can be set to an empty string for most NLP tasks. Although *predictive family* has a technical definition[3], most neural models, provided they are finetuned without any frozen parameters, easily meet this definition. Further, as per Xu et al. (2019):

Definition 2.2. 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

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

[2] Our code and data are available here.

[3] $\mathcal{V}$ is a subset of all possible mappings from $\mathcal{X}$ to $P(\mathcal{Y})$ that satisfies *optional ignorance*: for any P in the range of some $f \in \mathcal{V}$, there exists some $f' \in \mathcal{V}$ s.t. $f'[X] = f'[\emptyset] = P$. See Xu et al. (2019) for why optional ignorance is necessary.

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:

- *Non-Negativity:* $I_{\mathcal{V}}(X \to Y) \geq 0$
- *Independence:* If X is independent of Y, $I_{\mathcal{V}}(X \to Y) = 0$.
- *Montonicity:* If $\mathcal{U} \subseteq \mathcal{V}$, then $H_{\mathcal{U}}(Y) \geq H_{\mathcal{V}}(Y)$ and $H_{\mathcal{U}}(Y|X) \geq H_{\mathcal{V}}(Y|X)$.

Training with the cross-entropy loss finds the $f \in \mathcal{V}$ that maximizes the log-likelihood of Y given X (Xu et al., 2019). Thus, $H_{\mathcal{V}}(Y|X)$ can be easily computed by standard training or by finetuning a pre-trained model.[4] We estimate $H_{\mathcal{V}}(Y|X)$ by calculating $\mathbb{E}[-\log f[X](Y)]$ on an identically distributed held-out set, where Y is the gold label. Since training with cross-entropy ultimately aims to find the infimum over the *data distribution*, not just the training set, it is important not to overfit the model to the training instances; this is of added significance for estimating $H_{\mathcal{V}}(Y|X)$. We estimate $H_{\mathcal{V}}(Y)$ by training or finetuning another model where X is replaced by $\varnothing$, intended to fit the label distribution. As such, estimating $\mathcal{V}$-information involves training or finetuning only two models.

2.3. Assumptions

Implicit in estimating the $\mathcal{V}$-information is the assumption that the data used to find the optimal $f \in \mathcal{V}$ and the data used to estimate $H_{\mathcal{V}}(Y|X)$ are identically distributed. This dependence on the data distribution makes $\mathcal{V}$-information well-suited for estimating and interpreting dataset difficulty. However, it is still possible to estimate the difficulty of sub-populations or subsets of the data, though it would be imprecise to refer to this measure as $\mathcal{V}$-information (see §3.1 for details). We also assume that the difference between the empirical $\mathcal{V}$-information (calculated using some finite dataset) and the true $\mathcal{V}$-information (calculated over the distributions) is negligible, though this may not hold, for example, if the dataset is too small (see Appendix A).

2.4. Implications

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

i. different models $\mathcal{V}$ 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. 4; §4).

[4] Improving model calibration using more advanced methods (Kumar et al., 2019) is a possible direction of future work.

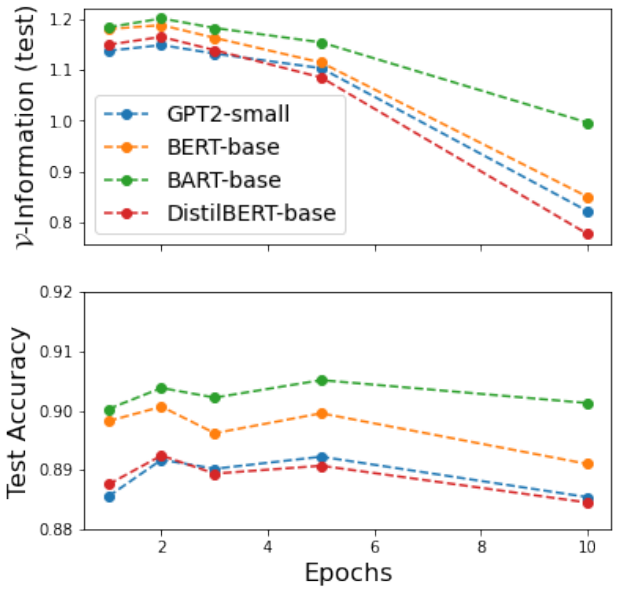

Figure 2. Comparing the $\mathcal{V}$-usable information estimate to accuracy in SNLI. In the first three epochs, estimates on the test set are similar across all models (top), but due to over-fitting, the estimates diverge and decline. The test accuracy (bottom) for each model loosely tracks the $\mathcal{V}$-information estimate for that model, since extracting information makes prediction easier.

While common classification metrics, such as accuracy or F_1 score, are often used for the above comparisons, $\mathcal{V}$-usable information offers a theoretically rigorous framework, making it better suited for interpretability. The $\mathcal{V}$-usable information is measured in bits / nats (depending on the log base), allowing for standardized 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. $\mathcal{V}$-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 *provably* zero.

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 task will at least as easy for the larger family. This obviates the need to evaluate simpler function families (e.g., linear functions) when estimating dataset difficulty. Our experiments show that this bears out in practice as well (Appendix B).

2.5. $\mathcal{V}$-Usable Information in Practice

We consider the natural language inference (NLI) task, which involves predicting whether a text hypothesis entails, contradicts or is neutral to a text premise. We first apply the $\mathcal{V}$-information framework to estimate the difficulty of

a large-scale NLI dataset, Stanford NLI (SNLI; Bowman et al., 2015), across different state-of-the-art models. 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. See Appendix B for results with larger models.

Model performance tracks $\mathcal{V}$-information. As seen in Figure 2, the model with the most $\mathcal{V}$-information on the SNLI test set is also the most accurate. 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; accuracy follows the same trend.

$\mathcal{V}$-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, as seen in Figure 2. 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 like test accuracy, $\mathcal{V}$-information can more readily inform us of over-fitting.

Different datasets for the same task can have different amounts of $\mathcal{V}$-usable information. We consider the MultiNLI dataset (Williams et al., 2018), a multi-genre counterpart of SNLI. Despite both being proxies for the NLI task, SNLI and MultiNLI have significantly different amounts of BERT-usable information, as shown in Figure 1. The $\mathcal{V}$-information framework provides a principled means of measuring this difference in levels of difficulty; MultiNLI is expected to be more difficult than SNLI due to the diversity of genres it considers. Also shown is CoLA (Warstadt et al., 2018), a dataset for linguistic acceptability where each sentence is labeled as grammatical or not; this task is seemingly more difficult than NLI for BERT.

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 (w.r.t. a given distribution). We extend the $\mathcal{V}$-information framework and introduce a new measure called **pointwise $\mathcal{V}$-information** (PVI) for individual instances. The higher the PVI, the easier the instance is for $\mathcal{V}$, under the given distribution.

Definition 3.1 (Pointwise $\mathcal{V}$-Information)**.** Given random variables X, Y and a predictive family $\mathcal{V}$, the pointwise $\mathcal{V}$-information (PVI) of an instance (x, y) is

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

where functions $g = \arg\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]$ and $g' = \arg\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)]$.

If $\mathcal{V}$ were, for instance, the BERT function family, g' and g would be the models after finetuning BERT 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 models place on the gold label. PVI is to $\mathcal{V}$-information what PMI is to Shannon information:

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

Given this relationship, our understanding of $\mathcal{V}$-information extends to PVI as well: higher PVI instances are easier for $\mathcal{V}$ and vice-versa. 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, 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). Examples with negative PVI can still be predicted correctly, as long as g' places most of the probability mass on the correct label. Algorithm 1 shows our computation of PVI and $\mathcal{V}$-information (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 $\text{PVI}(x \to y)$. 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. Implications

In addition to the comparisons that $\mathcal{V}$-information allows us to make (§2.4), PVI allows us to compare:

iv. different instances (x, y) by computing PVI $(x \to y)$ for the same $X, Y, \mathcal{V}$ (Tables 1, 4; Fig. 11)
v. different slices or subsets of the data by computing the average PVI over instances in each slice (Table 2; Fig. 5).

Algorithm 1 After finetuning on a dataset of size n, the $\mathcal{V}$-information and PVI can be calculated in $O(n)$ time.

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

Note that the average PVI of a slice of data is not its $\mathcal{V}$-information, since we optimize the model w.r.t. the entire distribution. However, since in practice one often wishes to understand the *relative difficulty* of different subpopulations w.r.t. the training distribution, calculating the average PVI—as opposed to the $\mathcal{V}$-information of the subpopulation itself—is more useful.

3.2. PVI in Practice

PVI can be used to find mislabelled instances. Correctly predicted instances have higher PVI values than incorrectly predicted ones. For the held-out sets in SNLI, MultiNLI and CoLA, the difference in mean PVI between instances correctly and incorrectly predicted by BERT-base is 3.03, 2.87, and 2.45 bits respectively. These differences are statistically significant ($p < 0.001$). Table 1 shows the most difficult (lowest PVI) instances from CoLA; we further find that some of these are in fact mislabelled (see Appendix C for an analysis of SNLI).

The PVI threshold at which predictions become incorrect is similar across datasets. In Figure 3, we plot the PVI distribution of correctly and incorrectly predicted instances in each dataset. As expected, high-PVI instances are predicted correctly and low-PVI instances are not. Notably, the point at which instances start being incorrectly predicted is similar across datasets (PVI ≈ 0.5). Such a pattern could not be observed with a performance metric because the label spaces are different, evincing why the $\mathcal{V}$-information framework is so useful for cross-dataset comparison.

PVI estimates are highly consistent across models, training epochs, and random initializations. The cross-model Pearson correlation between PVI estimates of SNLI instances is very high ($r > 0.80$). However, the cross-model Pearson correlation is lower for CoLA ($0.40 < r < 0.65$); see Fig. 9 in Appendix D. This is because, as visualized in Figure 1, CoLA has 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 are 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. Moreover, 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); see Table 6 in Appendix D. It also concurs with human judgments of difficulty; see Fig. 10 in Appendix D.

Sentence	Label	PVI
`Wash you!`	No	-4.616
`Who achieved the best result was Angela.`	No	-4.584
`Sue gave to Bill a book.`	No	-3.649
`Only Churchill remembered Churchill giving the Blood, Sweat and Tears speech.`	No	-3.571
`Cynthia chewed.`	No	-3.510
`It is a golden hair.`	Yes	-3.251
`I won't have some money.`	No	-3.097
`You may pick every flower, but leave a few for Mary.`	No	-2.875
`I know which book Mag read, and which book Bob said that you hadn't.`	Yes	-2.782
`John promise Mary to shave himself.`	Yes	-2.609

Table 1. The 10 hardest (lowest PVI) instances in the CoLA in-domain test set for grammaticality detection (label indicates grammaticality), according to BERT-base. Examples in red are assessed to be mislabelled by authors of this work. For e.g., '`Cynthia chewed.`' might be grammatical because the verb 'chew' could be intransitive in this usage. This suggests that PVI could be used to identify mislabelled examples. All of these examples were predicted incorrectly by BERT-base.

4. Uncovering Dataset Artefacts

A key limitation of standard evaluation metrics (e.g. accuracy) is the lack of interpretability—there is no straightforward way to understand *why* a dataset is as difficult as it is. $\mathcal{V}$-usable information offers an answer by allowing comparison of different input variables X_i under the same $\mathcal{V}$ and Y, as implicated in §2.4. We consider two approaches for this: applying input transformations (§4.1), and slicing

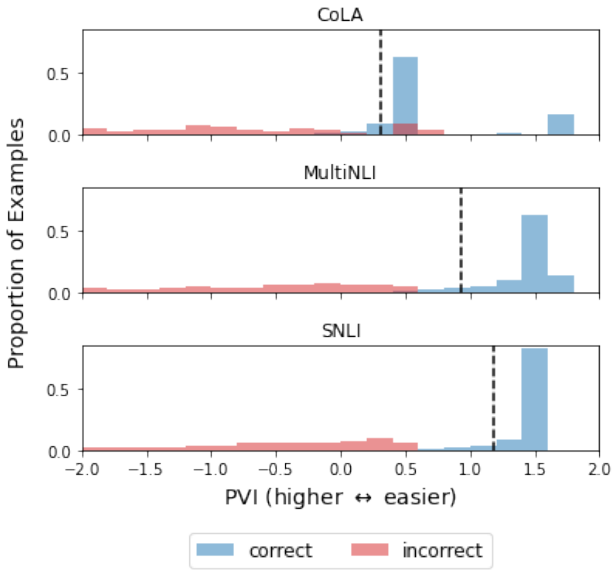

Figure 3. The distribution of PVI for correctly and incorrectly predicted instances in each dataset. Note that the point at which instances start being incorrectly predicted is similar across datasets ($\sim$ 0.5 bits). In contrast, because the label space is different across CoLA and the other two datasets, such a comparison could not be made with a performance-based metric such as accuracy.

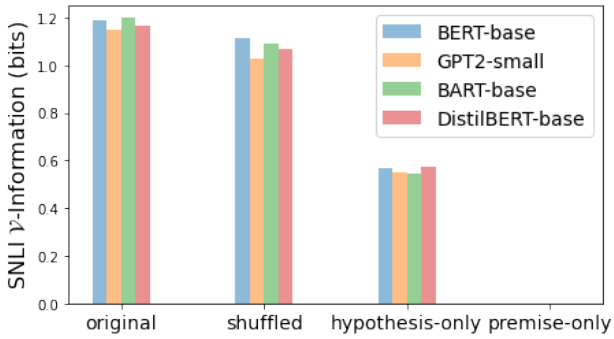

Figure 4. The amount of $\mathcal{V}$-usable information contained in different input attributes about the gold labels in SNLI. The token identity alone (regardless of order) provides most of the information for all models (see SHUFFLED). The PREMISE, which can be shared by multiple instances, is useless alone; the HYPOTHESIS, which is unique to an instance, is quite useful even without a premise, suggesting it may contain annotation artefacts.

the dataset (§4.2).

4.1. Input Transformations

Our first approach involves applying different transformations $\tau_i(X)$ to isolate an attribute a, followed by calculating $I_\mathcal{V}(\tau_i(X) \to Y)$ 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 the word order attribute.

Given that a transformation may make information more accessible (e.g., decrypting some encrypted text; c.f. §2), it is possible for $I_\mathcal{V}(\tau_i(X) \to Y) \geq I_\mathcal{V}(X \to Y)$, so the latter shouldn't be treated as an upper bound. Such transformations were applied by O'Connor & 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 our findings on SNLI, CoLA, 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.

We apply transformations to the SNLI input to isolate different attributes (see Appendix E for an example): *shuffled* (shuffle tokens randomly), *hypothesis-only* (only include the hypothesis), *premise-only* (only include the premise), *overlap* (tokens in both the premise and hypothesis).

Token identity alone provides most of the usable information in SNLI. Figure 4 shows that the token identity alone—isolated by shuffling the input—contains most of the usable information for all models. The premise, which is often shared by multiple instances, is useless alone; the hypothesis, which is unique to an instance, is useful even without a premise. This corroborates the well-known annotation artefacts in SNLI (Gururangan et al., 2018; Poliak et al., 2018), which are spurious correlations exploited by models to predict the correct answer for the wrong reasons.

Hate speech detection might have lexical biases. 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 in DWMW17, the text contains 0.724 bits of BERT-usable information about the label. Additionally, if one removed all the tokens, except for 50 (potentially) offensive ones—comprising common racial and homophobic slurs[5]—from the input post hoc, 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. Our findings corroborate prior work which shows that certain lexical items (e.g., swear words, identity mentions) are responsible for hate speech prediction (Dixon et al., 2018; Dinan et al., 2019). Allowing models to do well by simply pattern-matching may permit subtleties in hate speech to go undetected, perpetuating harm towards minority groups (Blodgett et al., 2020).

[5]These terms were manually chosen based on a cursory review of the dataset and are listed in Appendix E.

	Entailment	Neutral	Contradiction
original	1.188	1.064	1.309
shuffled	1.130	0.984	1.224
hypothesis only	0.573	0.553	0.585
premise only	0.032	-0.016	-0.016
overlap	0.415	0.177	0.298

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.

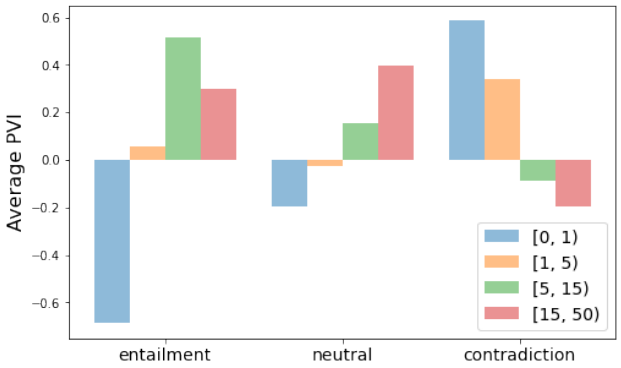

Figure 5. The mean PVI of SNLI instances according to BERT-base, broken down by the overlap length (i.e., the number of tokens shared by the hypothesis and premise). Entailment examples with no overlap are the most difficult (i.e., lowest mean PVI).

4.2. Slicing Datasets

Certain attributes are more useful for certain classes. Comparing the usefulness of an attribute across classes can be useful for identifying systemic annotation artefacts. This can be done by simply averaging the PVI over the slice of data whose difficulty we are interested in measuring. Note that the equivalence between $\mathcal{V}$-information and expected PVI only holds when the model used to estimate PVI is trained over the entire dataset, which means that the average PVI of a slice of data is not its $\mathcal{V}$-information. It would not make sense to estimate the $\mathcal{V}$-information of a slice because it would require training on examples from just one class, in which case the $\mathcal{V}$-information would be zero. Thus the only usable difficulty measure is the mean PVI.

We do this for SNLI in Table 2. We see that the tokens in the premise-hypothesis overlap contains much more BERT-usable information about the 'entailment' class than 'contradiction' or 'neutral'. This is unsurprising, given that the simplest means of entailing a premise is to copy it into the hypothesis and provide some additional detail. 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. Even when there is an inherent reason for an attribute to be useful for a particular slice of the data, an attribute being exceptionally useful may also be evidence of a dataset artefact. For example, if the premise-overlap hypothesis provided almost all the usable information needed for entailment, it may be because the crowdworkers who created the dataset took a shortcut by copying the premise to create the hypothesis.

In Appendix F, we show how similar comparisons can be made between instances.

Certain subsets of each class are more difficult than others. In Figure 5, we bin the examples in each SNLI class by the level of hypothesis-premise overlap and plot the average PVI. We see entailment instances with no hypothesis-premise overlap are the most difficult (i.e., lowest mean PVI) while contradiction instances with no overlap are the easiest (i.e., highest mean PVI). This is not surprising, since annotation artefacts in SNLI arise from constructing entailment and contradiction via trivial changes to the premise (Gururangan et al., 2018).

We additionally consider slices in the dataset based on dataset cartography (Swayamdipta et al., 2020), which uses training dynamics to differentiate instances via their (1) *confidence* (i.e., mean probability of the correct label across epochs), and (2) *variability* (i.e., variance of the former). The result is a dataset map revealing three regions: easy-to-learn, hard-to-learn, and ambiguous w.r.t. the trained model. Slices of the dataset based on cartographic regions have distinct ranges of average PVI (Fig. 13 in Appendix H).

4.3. Token-level Artefacts

Transforming X and then measuring the $\mathcal{V}$-information to discover all token-level signals and 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 after removing t, which yields modified input $x_{\neg t}$. We use the same model g' but evaluate only on a *slice* of the data, $\mathcal{D}_{C,t}$, which contains the token t and belongs to the class C of interest. This simplifies to measuring the increase in conditional entropy:

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

Token-level signals and artefacts can be discovered using leave-one-out. Table 3 shows that auxiliary verbs (e.g., be, did) and prepositions are most indicative of ungrammatical sentences in CoLA; in contrast, grammatical sentences have no strong indicators, with no word on average increasing

WARNING: The following content contains language from the DWMW17 dataset that is offensive in nature.

DWMW17 (Davidson et al., 2017)

Hate Speech	Offensive	Neither
f*ggots (3.844)	r*tards (2.821)	lame (4.426)
f*g (3.73)	n*gs (2.716)	clothes (0.646)
f*ggot (3.658)	n*gro (2.492)	dog (0.616)
c**ns (3.53)	n*g (2.414)	cat (0.538)
n*ggers (3.274)	c*nts (2.372)	iDntWearCondoms (0.517)

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)

Table 3. Token-level annotation artefacts in DWMW17 and CoLA. These are the tokens whose omission leads to the greatest average increase in conditional entropy for each class (given in parentheses). Note that certain racial slurs are correctly identified as 'hate speech' but in-group variants of the same terms fall under 'offensive' instead. The full lists are available in Appendix G.

the conditional entropy above 0.30 upon omission.

In DWMW17, racial and homophobic slurs are the top indicators of 'hate speech'. However, in-group variants of the same racial slur—commonly used in African-American Vernacular English (AAVE)—fall under 'offensive' instead. The fact that AAVE terms are marked as 'offensive' supports previous findings by that hate speech detection datasets may themselves be biased (Sap et al., 2019). In SNLI, we found many of the token-level artefacts matching those found using descriptive statistics in Gururangan et al. (2018). The complete word lists are available in Appendix G.

4.4. Conditioning Out Information

What if we wanted to measure how much BERT-usable information offensive words contain about the label in DWMW17 *beyond* that which is captured in the sentiment? In other words, if we already had access to the sentiment polarity of a text (positive/negative/neutral), how many *additional bits of information* would the offensive words provide? We cannot estimate this by simply subtracting $I_\mathcal{V}(\text{offensive} \to Y)$ from $I_\mathcal{V}(\text{sentiment} \to Y)$, since that difference could potentially be negative. Acquiring another random variable should not decrease the amount of information we have about the label (at worst, it should be useless).

To capture this intuition, Hewitt et al. (2021) proposed *conditional $\mathcal{V}$-information*, which allows one to condition out any number of random variables. Given a set of random variables $\mathcal{B}$ that we want to condition out, it is defined as:

$$I_\mathcal{V}(X \to Y|\mathcal{B}) = H_\mathcal{V}(Y|\mathcal{B}) - H_\mathcal{V}(Y|\mathcal{B} \cup \{X\}) \quad (6)$$

The conditional entropy with respect to multiple variables is the only new concept here. It is estimated in practice by concatenating the text inputs represented by $\mathcal{B}$ and X, which in our example is the sentiment polarity (one of 'negative'/'neutral'/'positive') and the sequence of offensive words in the input. The actual model family need not change to accommodate the longer text, as long as it remains under the input token limit.[6] We find that offensive words contain 0.482 bits of BERT-usable information about the label *beyond* that which is contained in text sentiment.[7] This is close to all of the BERT-usable information that the offensive words contain about the label (0.490 bits), suggesting that the predictive power of (potentially) offensive words is not mediated through sentiment in DWMW17.

5. Related Work

While prior literature has acknowledged that not all data instances are equal (Vodrahalli et al., 2018; Swayamdipta et al., 2020), there have been few efforts to estimate dataset difficulty formally and directly. As a notable exception, Zhang et al. (2020) proposed DIME, an information-theoretic measure to estimate a lower bound on the lowest possible (i.e., model-agnostic) 0-1 error. Model-agnostic approaches do not explain why some datasets are easier for some models, and have limited interpretability. In contrast, $\mathcal{V}$-information and PVI are specific to a model family $\mathcal{V}$.

Various techniques have been proposed to differentiate data instances within a dataset. Text-based heuristics such as word identity (Bengio et al., 2009) or input length (Spitkovsky et al., 2010; Gururangan et al., 2018) have sometimes been used as proxies for instance difficulty, but offer limited insight into difficulty w.r.t. models. Other approaches consider training loss (Han et al., 2018; Arazo et al., 2019; Shen & Sanghavi, 2019), confidence (Hovy et al., 2013), prediction variance (Chang et al., 2017), and area under the curve (Pleiss et al., 2020). Estimates relying on model training dynamics (Toneva et al., 2018; Swayamdipta et al., 2020), gradient magnitudes (Vodrahalli et al., 2018), or loss magnitudes (Han et al., 2018) are sensitive to factors such as variance during steps of training. Influence functions (Koh & Liang, 2017), forgetting events (Toneva et al., 2018), and the Data Shapley (Ghorbani & Zou, 2019; Jia et al., 2019) can all be used to assign pointwise estimates of importance to data instances based on their

[6] If the inputs were vectors, the inputs represented by $\mathcal{B}$ and $\mathcal{B} \cup \{X\}$ would need to be of the same size; to do so, we would concatenate a zero vector to the former (Hewitt et al., 2021).

[7] The sentiment was categorized as positive/neutral/negative based on the polarity estimated by spaCy's built-in sentiment classifier (Honnibal & Montani, 2017).

contribution to the decision boundary. Moreover, although these methods all capture some aspect of difficulty, they do not lend themselves to interpreting datasets as readily as the predictive $\mathcal{V}$-information framework.

Given its dependence on training behavior across time, cartography (Swayamdipta et al., 2020) offers complementary benefits to $\mathcal{V}$-information. It can be non-trivial to measure differences between, say a CoLA data map and an SNLI data map, w.r.t BERT. In contrast, $\mathcal{V}$-information provides a formal framework to make dataset difficulty estimates as an aggregate to compare datasets w.r.t a model. 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 & Hazen, 2018; Sen & Saffari, 2020).

Item response theory (IRT; Embretson & Reise, 2013) allows the difficulty of instances to be learned via parameters in a probabilistic model meant to explain model performance (Lalor et al., 2018; Rodriguez et al., 2021). However, it does not formally relate dataset difficulty to the model being evaluated. Estimating instance difficulty is also evocative of instance selection for active learning (Lewis & Catlett, 1994; Fu et al., 2013; Liu & Motoda, 2002); however these estimates could change as the dataset picks up new instances. In contrast, PVI estimates are relatively stable, especially when the dataset has higher $\mathcal{V}$-information. Uncertainty sampling, for example, picks the instances that the partially trained model is least certain about (Lewis & Gale, 1994; Nigam et al., 2000), which could be interpreted as a measure of difficulty. However, once an instance is used for training, the model may become much more certain about it, meaning that the uncertainty values are unstable.

Interpretability of the role of certain attributes in trained models have lately led to the discovery of many dataset artefacts in NLP. Our approach to discovering dataset artefacts can also complement existing approaches to artefact discovery (Gardner et al., 2021; Pezeshkpour et al., 2021; Le Bras et al., 2020). Rissanen data analysis (Perez et al., 2021) offers a complimentary method for interpretability w.r.t attributes; it involves calculating the minimum description length (MDL): how many bits are needed to transmit the gold labels from a sender to a recipient when both have access to the same model and inputs. Since the framework depends on the order of instances (i.e., what data has been transmitted thus far), it is unsuitable for estimating dataset difficulty. In contrast, $\mathcal{V}$-information is defined w.r.t. a data distribution, so it is (in theory) agnostic to data and its ordering in fine-tuning.

$\mathcal{V}$-information (Xu et al., 2019) has had limited adoption in NLP. It has been used to study what context features Transformers actually use (O'Connor & Andreas, 2021), as well as to condition out information for probing-based interpretability techniques (Hewitt et al., 2021; Pimentel & Cotterell, 2021). However, to the best of our knowledge, ours is the first approach to use $\mathcal{V}$-usable information for estimating the difficulty of NLP datasets.

6. 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 test for in models and how pervasive annotation artefacts really are. Moreover, our framework can be used proactively: by identifying the attributes responsible for dataset difficulty, one can create useful datasets out of otherwise useless raw data. For example, the alignment of large language models is often bottlenecked by the lack of human preference data, which is slow and expensive to collect. In Appendix I, we show that starting with otherwise unremarkable web data, we can infer a dataset of collective preferences that contain as much usable information as preferences collected with paid human annotators, all the while being free and many times larger.

More immediate directions of future work include:

1. Understanding how changes to the data distribution change the difficulty of individual examples.
2. Extending $\mathcal{V}$-information to open-ended text generation, which does not induce explicit distributions over the output space. This may requiring truncating the output space (e.g., using beam search with fixed width).
3. Applying $\mathcal{V}$-information to estimate dataset difficulty in other modalities (e.g., image, audio, tabular, etc.). There is nothing limiting the use of $\mathcal{V}$-information to the NLP domain. For example, one could create a set of image filters—for different colors and objects—use them to transform the image, and then measure the drop in usable information.

7. Conclusion

We provided an information-theoretic perspective to understanding and interpreting the difficulty of various NLP datasets. 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. We showed that instances with lower PVI had lower levels of annotator agreement and were less likely to be predicted correctly. We then demonstrated how systemic and token-level annotation artefacts in a dataset could be discovered by manipulating the input before calculating these measures. Our studies indicate that $\mathcal{V}$-information offers a new, efficient means of interpreting NLP datasets.

Acknowledgements

We thank Dan Jurafsky, Nelson Liu, Daniel Khashabi, and the anonymous reviewers for their helpful comments. We thank Heidi (Chenyu) Zhang and Shabnam Behzad for helping create the SHP and SHP-2 datasets, alongside the first author, with advice from Dan Jurafsky and Yizhong Wang. This project was supported by award DMS-2134012 from the NSF and the DARPA MCS program through NIWC Pacific (N66001-19-2-4031).

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pp. 312–321. PMLR, 2019. URL https://arxiv.org/abs/1904.11238.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Bowman, S., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015. URL https://aclanthology.org/D15-1075/.

Chang, H.-S., Learned-Miller, E., and McCallum, A. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30:1002–1012, 2017. URL https://arxiv.org/abs/1704.07433.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. 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, pp. 512–515, 2017. URL https://arxiv.org/abs/1703.04009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Dinan, E., Humeau, S., Chintagunta, B., and Weston, J. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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)*, pp. 4537–4546, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1461. URL https://aclanthology.org/D19-1461.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. Show your work: Improved reporting of experimental results. In *Proc. of EMNLP-IJCNLP*, pp. 2185–2194, 2019. URL https://aclanthology.org/D19-1224/.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305. arXiv preprint arXiv:2002.06305.

Embretson, S. E. and Reise, S. P. *Item response theory*. Psychology Press, 2013. URL https://doi.org/10.4324/9781410605269.

Ethayarajh, K. and Jurafsky, D. 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)*, pp. 4846–4853, 2020. URL https://aclanthology.org/2020.emnlp-main.393/.

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

URL https://link.springer.com/article/10.1007/s10115-012-0507-8.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Gardner, M., Merrill, W., Dodge, J., Peters, M. E., Ross, A., Singh, S., and Smith, N. Competency problems: On finding and removing artifacts in language data, 2021. URL https://arxiv.org/abs/2104.08646.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251. PMLR, 2019. URL https://arxiv.org/abs/1904.02868.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation artifacts in natural language inference data. In *Proc. of NAACL*, pp. 107–112, 2018. URL https://aclanthology.org/N18-2017.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. URL https://arxiv.org/abs/1804.06872.

Hewitt, J., Ethayarajh, K., Liang, P., and Manning, C. D. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, 2021. URL https://aclanthology.org/2021.emnlp-main.122/.

Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. URL https://spacy.io/.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. 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*, pp. 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1132.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019. URL https://arxiv.org/abs/1902.10275.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017. URL https://arxiv.org/abs/1703.04730v3.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3792–3803, 2019. URL https://arxiv.org/abs/1909.10155.

Lalor, J. P., Wu, H., Munkhdalai, T., and Yu, H. 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*, pp. 4711–4716, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1500. URL https://aclanthology.org/D18-1500.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., and Choi, Y. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pp. 1078–1088. PMLR, 2020. URL https://arxiv.org/abs/2002.04108.

Lewis, D. D. and Catlett, J. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994. URL https://openreview.net/forum?id=HyEyBoWObr.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *SIGIR'94*, pp. 3–12. Springer, 1994. URL https://arxiv.org/abs/cmp-lg/9407020.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, pp. 7871–7880, 2020. URL https://aclanthology.org/2020.acl-main.703/.

Liu, H. and Motoda, H. On issues of instance selection. *Data Mining and Knowledge Discovery*, 6(2):115, 2002. URL https://link.springer.com/article/10.1023/A:1014056429969.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692. arXiv preprint arXiv:1907.11692.

Ma, Z., Ethayarajh, K., Thrush, T., Jain, S., Wu, L., Jia, R., Potts, C., Williams, A., and Kiela, D. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking, 2021. URL https://arxiv.org/abs/2106.06052. arXiv preprint arXiv:2106.06052.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, Feb 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007. URL http://dx.doi.org/10.1016/J.ARTINT.2018.07.007.

Mosbach, M., Andriushchenko, M., and Klakow, D. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2006.04884.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2):103–134, 2000. URL https://link.springer.com/article/10.1023/A:1007692713085.

O'Connor, J. and Andreas, J. 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)*, pp. 851–864, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.70. URL https://aclanthology.org/2021.acl-long.70.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Perez, E., Kiela, D., and Cho, K. Rissanen data analysis: Examining dataset characteristics via description length. In *ICML*, 2021. URL https://arxiv.org/abs/2103.03872.

Pezeshkpour, P., Jain, S., Singh, S., and Wallace, B. C. Combining feature and instance attribution to detect artifacts, 2021. URL https://arxiv.org/abs/2107.00323. arXiv preprint arXiv:2107.00323.

Pimentel, T. and Cotterell, R. A bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2869–2887, 2021. URL https://arxiv.org/abs/2109.03853.

Pleiss, G., Zhang, T., Elenberg, E. R., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2001.10528.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019. URL http://proceedings.mlr.press/v97/recht19a.html.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., and Boyd-Graber, J. 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)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346.

Rondeau, M.-A. and Hazen, T. J. Systematic error analysis of the stanford question answering dataset. In *Proceedings of the Workshop on MRQA*, pp. 12–20, 2018. URL https://aclanthology.org/W18-2602/.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108. arXiv preprint arXiv:1910.01108.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019. URL https://aclanthology.org/P19-1163/.

Sen, P. and Saffari, A. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2429–2438, 2020. URL https://aclanthology.org/2020.emnlp-main.190.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. URL https://ieeexplore.ieee.org/abstract/document/6773024/.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *Proc. of ICML*, pp. 5739–5748. PMLR, 2019. URL https://proceedings.mlr.press/v97/shen19e.html.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Proc. of NAACL*, pp. 751–759. Association for Computational Linguistics, 2010. URL https://aclanthology.org/N10-1116.

Sugawara, S., Inui, K., Sekine, S., and Aizawa, A. What makes reading comprehension questions easier? In *Proc. of EMNLP*, pp. 4208–4219, 2018. URL https://arxiv.org/abs/1808.09384.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proc. of EMNLP*, pp. 9275–9293, 2020. URL https://aclanthology.org/2020.emnlp-main.746/.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1812.05159.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011. URL https://ieeexplore.ieee.org/abstract/document/5995347.

Vodrahalli, K., Li, K., and Malik, J. Are all training examples created equal? an empirical study, 2018. URL https://arxiv.org/abs/1811.12569.

Warstadt, A., Singh, A., and Bowman, S. R. Neural network acceptability judgments, 2018. URL https://arxiv.org/abs/1805.12471. arXiv preprint arXiv:1805.12471.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/2002.10689.

Zhang, P., Wang, H., Naik, N., Xiong, C., and Socher, R. DIME: An Information-Theoretic difficulty measure for AI datasets. In *Proc. of NeurIPS 2020 Workshop for Deep Learning through Information Geometry*, October 2020. URL https://openreview.net/forum?id=kvqPFy0hbF.

A. Training Data

In Figure 6, we plot the $\mathcal{V}$-information estimate on the SNLI test set as BERT-base is trained on increasing amounts of training data. This is to test the assumption that the training set is sufficiently large to find the function $f \in \mathcal{V}$ that minimizes the conditional entropy. Although this assumption is impossible to validate with complete certainty, since we do not have access to the true distribution, if the $\mathcal{V}$-information estimate plateaus before all the training data is used, it suggests that the training set size is not a limiting factor to the estimation. We find that this is indeed the case with SNLI, where 80% of the training data on averages provides the same estimate as using the entire training set. In cases when this assumption does not hold, readers may want to consider measuring the Bayesian mutual information instead (Pimentel & Cotterell, 2021).

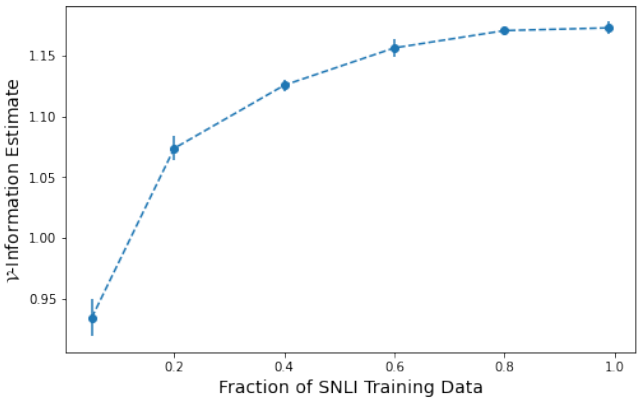

Figure 6. The $\mathcal{V}$-information estimate on the SNLI test set when BERT-base is trained on increasing fractions of the training data, drawn as a random sample (with replacement). Here we plot the average and standard deviation across four samples for each fraction.

B. Larger Models

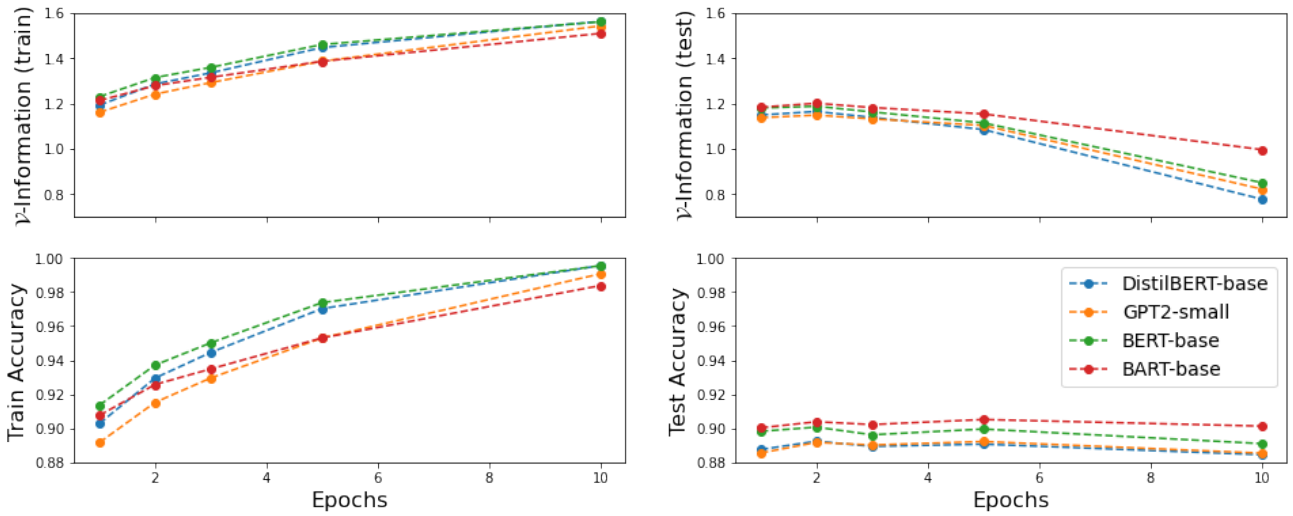

Figure 7. Comparing accuracy and the $\mathcal{V}$-information estimate on the SNLI train and test set w.r.t. various models.

In Figure 7, we plot the $\mathcal{V}$-information estimate for the SNLI test *and* train sets. In Figure 8, we plot the $\mathcal{V}$-information estimate on the CoLA in-domain held-out set for the four models that we previously studied, as well as a larger model, RoBERTa-large (Liu et al., 2019). Despite the increase in scale, the trends observed in §2 still hold.

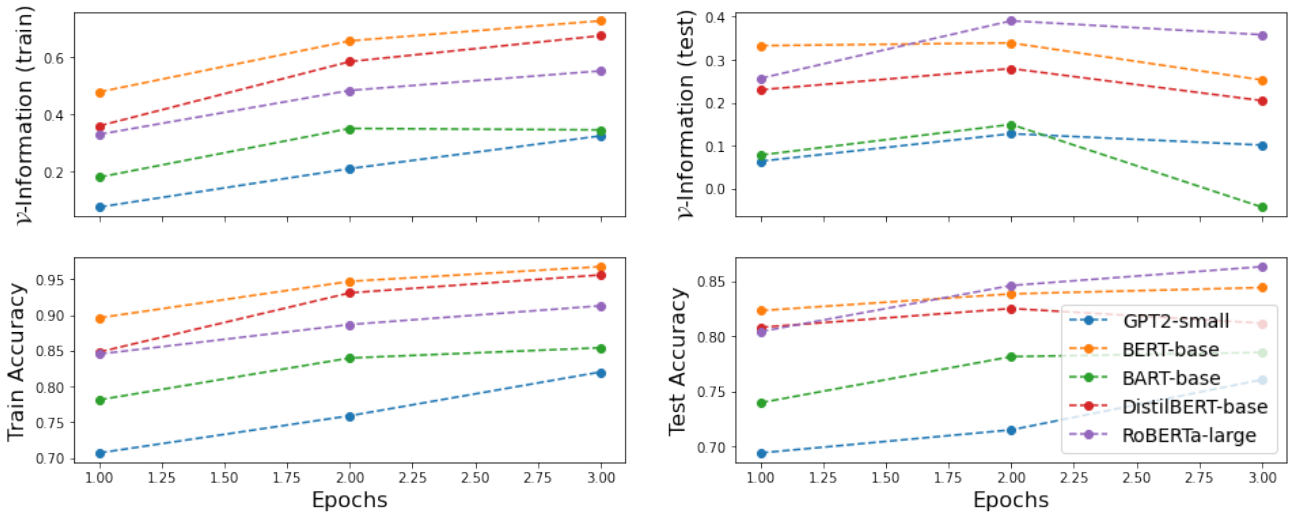

Figure 8. Comparing accuracy and the $\mathcal{V}$-information estimate on the CoLA in-domain train and held-out set w.r.t. various models.

C. Qualitative Analysis

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

Table 4. The 10 hardest (lowest PVI) instances in the SNLI test set, according to BERT-base. 'E' denotes entailment, 'N' neutral, and 'C' contradiction. Instances that are possibly mislabelled are colored red.

In Table 4, 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'.

D. Consistency of PVI estimates

Cross-Model Correlations Figure 9 shows a heatmap for Cross-model Pearson's r between PVI estimates made by different finetuned models, on the SNLI and CoLA test sets; these results support the findings in §2.5.

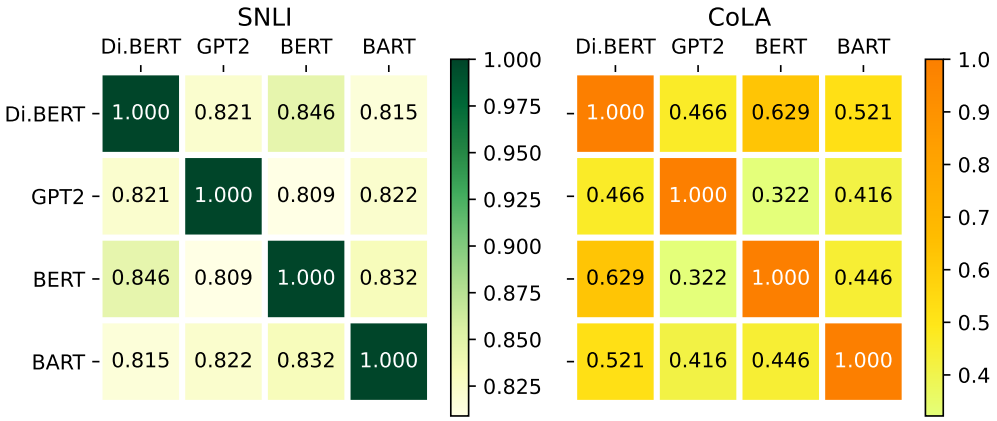

Figure 9. Cross-model Pearson's r between PVI estimates made by different finetuned models, on the SNLI and CoLA test sets. For SNLI, the estimates are consistent: what one model finds difficult, others find difficult as well. Since CoLA has less usable information for all these models, the correlations are lower. All correlations are highly statistically significant ($p < 0.001$).

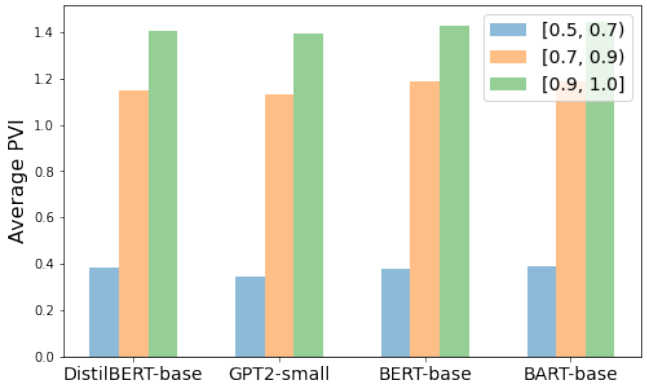

Figure 10. Examples that human annotators find easier (as measured by the fraction of annotators, in the range [0.5, 1.0], that agree with the gold label) also have higher PVI on average.

Human Agreement In Figure 10, we plot the average PVI at different levels of annotator agreement. We find that there is a concurrence between what humans find difficult and what examples are difficult according to PVI.

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 finetuning. 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).

Cross-Seed Correlations In Table 6, we list the Pearson correlation between PVI estimates made by BERT across different training runs. The correlation is high ($r > 0.87$), suggesting that what a model finds difficult is not due to chance.

E. Transformations

WARNING: The following content contains language from the DWMW17 dataset that is offensive in nature.

In Table 7, 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.

For DWMW17, we hand-picked a set of 50 potentially offensive words based on a cursory review of the dataset to see how much information these terms alone contain about the label: 'nigga', 'niggas', 'niggah', 'niggahs', 'hoe', 'hoes',

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.

'bitch', 'bitches', 'whitey', 'white trash', 'cracker', 'crackers', 'beaner', 'beaners', 'pussy', 'pussies', 'fag', 'fags', 'faggot', 'faggots', 'ho', 'hos', 'redneck', 'rednecks', 'porn', 'fuck', 'fucks', 'fucker', 'fuckers', 'motherfucker', 'motherfuckers', 'nigger', 'niggers', 'coon', 'coons', 'niggaz', 'nig', 'nigs', 'slut', 'sluts', 'wigger', 'wiggers', 'fucked', 'fucking', 'wigga', 'wiggas', 'retard', 'retards', and 'retarded'.

F. Instance-wise Comparisons

Certain attributes are responsible for the difficulty of certain examples. Figure 11 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 $\text{PVI}(x_{9627} \to y_{9627}) > \text{PVI}(x_{7717} \to y_{7717})$), finding that it is likely due to the former's *hypothesis* being more informative. 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.

G. Token-Level Artefacts

WARNING: The following content contains language from the DWMW17 dataset that is offensive in nature. In Table 8, we list the tokens in the SNLI, CoLA, and DWMW17 datasets that, when dropped out, cause the greatest decrease in

seed	Run 1	Run 2	Run 3	Run 4
Run 1	1.000	0.877	0.884	0.885
Run 2	0.877	1.000	0.887	0.882
Run 3	0.884	0.887	1.000	0.895
Run 4	0.885	0.882	0.895	1.000

Table 6. Cross-model Pearson correlation between PVI estimates made after one epoch of finetuning BERT on SNLI with different seeds. Estimates are stable: what a model finds difficult is mostly not due to chance.

Attribute	Transformation	Transformed Input
Original		PREMISE: Two girls kissing a man with a black shirt and brown hair on the cheeks. HYPOTHESIS: Two girls kiss.
Shuffled	shuffle tokens randomly	PREMISE: girls two a kissing man with a black cheeks shirt and hair brown on the . HYPOTHESIS: kiss two . girls
Hypothesis-only	only include hypothesis	HYPOTHESIS: Two girls kiss.
Premise-only	only include premise	PREMISE: Two girls kissing a man with a black shirt and brown hair on the cheeks.
Overlap	hypothesis-premise overlap	PREMISE: Two girls [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] . HYPOTHESIS: Two girls [MASK] .

Table 7. 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.

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.

H. Relation to Dataset Cartography

Swayamdipta et al. (2020) introduced dataset cartography, a method to automatically analyze and diagnose datasets with respect to a trained model. It offers a complimentary understanding of datasets and their properties, taking into account the behavior of a model towards different data instances during training. This behavior—training dynamics—helps differentiate instances via their (1) *confidence* (i.e., mean probability of the correct label across epochs), and (2) *variability* (i.e., variance of the former). The result is a dataset map revealing three regions:

- **Easy-to-learn** (high *confidence*, low *variability*) instances are the most frequent instances in the dataset, those which high capacity models like BERT predict correctly throughout training.
- **Hard-to-learn** (low *confidence*, low *variability*) instances correspond to those which are predicted incorrectly throughout training; these were shown to correspond to mislabeled examples often.
- **Ambiguous** (high *variability*) instances correspond to those which the model often changes its prediction for; these examples are the ones which are most responsible for high test performance, both in and out of distribution.

Figure 12 shows that PVI values track closely to the confidence axis of a SNLI-DistilBERT-base data map[8] (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

[8]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.

#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.*

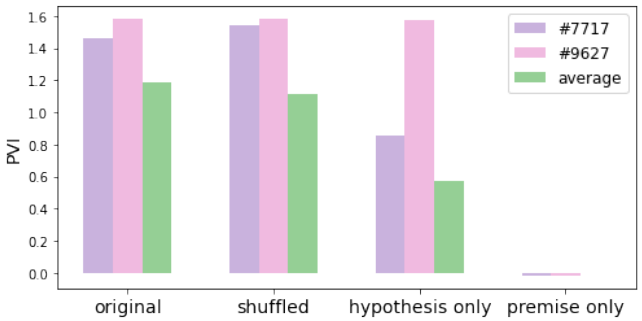

Figure 11. The PVI of two SNLI 'neutral' instances (#7717 and #9627) w.r.t. BERT-base after attribute-specific transformations, as well as the $\mathcal{V}$-information estimate (i.e., average PVI over the data) for each attribute. The latter instance is easier for BERT, likely because its hypothesis 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*.

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.

In Figure 13, we report the average PVI estimates of the three regions discovered via data maps:

- **Easy-to-learn** (high *confidence*, low *variability*) instances correspond to the highest average PVI, indicating that they have the highest amount of DistilBERT-usable information.
- **Hard-to-learn** (low *confidence*, low *variability*) instances correspond to the lowest average PVI, indicating that they have the lowest amount of DistilBERT-usable information. This is not surprising, since they also correspond to mislabeled instances, which can be difficult to extract usable information from.
- **Ambiguous** (high *variability*) instances correspond to lower average PVI, indicating that there is some usable information, but not as much as those of the easy-to-learn instances, w.r.t DistilBERT.

For each of the bars in the plot, we consider 10% of the dataset belonging to each region (with the highest corresponding measures of *confidence* and *variability*).

I. Creating Datasets

$\mathcal{V}$-information can be used to not only understand datasets but to create them anew. For example, consider the problem of aligning large language models (LLMs) with human preferences (Ouyang et al., 2022), so as to make them more helpful to users while limiting potential harm. This task is often bottlenecked by the lack of human preference data, which is slow and expensive to collect, as it relies on human annotation. In theory, online fora such as Stack Exchange and Reddit offer a free source of preferences: users collectively vote on comments in response to a question and the aggregate scores—the number of up-votes net of down-votes—ostensibly reflect a group preference. This gives us tuples (q, r_A, r_B, y), where q is a question (e.g., *How do I salt food?*), r_A and r_B are two root-level answers, and y is 1 if r_A has a higher aggregate score than r_B and 0 otherwise.

However, in their raw form, such data contain no usable information: $I_{\mathcal{V}}(\{Q, R_A, R_B\} \to Y) \approx 0$, even for capacious model families such as `GPT-3`. But why? We find that the examples with the highest PVI (i.e., the most learnable) are those where the higher-scoring comment was written *after* the lower-scoring one. In retrospect, this is intuitive: the earlier a comment is written, the more time it has to accrue votes, and since most comments end up with a positive score—up-voting being much more common than down-voting—an earlier comment is much more likely to have a higher score than a later one. If r_A has a higher score than r_B but was written earlier, we do not whether it is genuinely more preferred or simply benefits from greater visibility. However, if r_A has a higher score despite being written later, it suggests that it was so

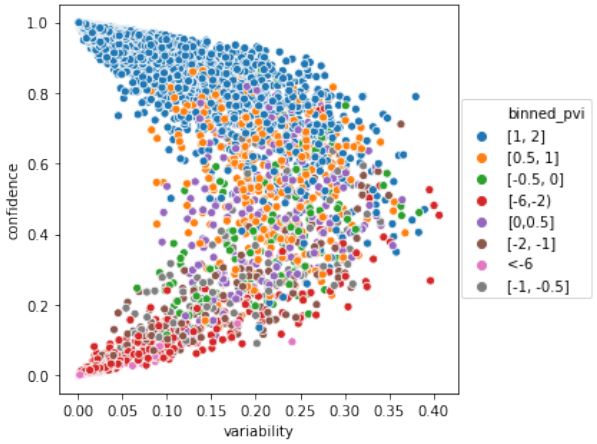

Figure 12. Relationship between PVI and the training dynamics-based data map (Swayamdipta et al., 2020) for SNLI held-out (test) set, computed for the DistilBERT-base architecture. As in Swayamdipta et al. (2020), Y-axis corresponds to *confidence*, i.e. the mean probabilities of the true class across training epochs, and X-axis corresponds to *variability*, i.e. the standard deviation of the true class probabilities across the same. Colours indicate binned values of PVI. PVI estimates track closely with *confidence*.

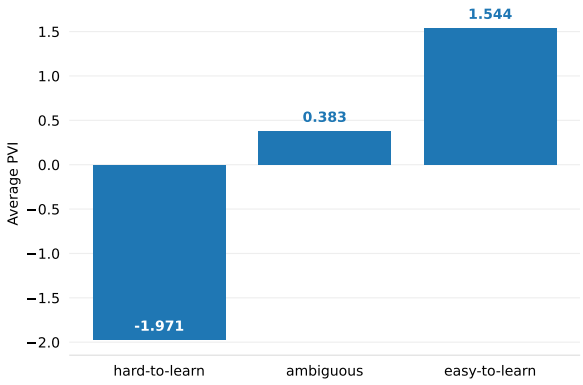

Figure 13. Average PVI of the 10% of the most ambiguous, and the 10% of the hardest-to-learn, and 10% of the easiest-to-learn regions of the SNLI / DistilBERT-base data map (Fig. 12). Hard-to-learn instances are frequently mislabeled, and therefore also reflect the lowest average PVI values. The highest average PVI values are possessed by the easy-to-learn instances, which are the most common class of instances in the SNLI dataset. Ambiguous instances are those that the model changes it decision on frequently through training; these correspond to lower average PVI values than easy-to-learn instances.

preferred by users that it was able to overcome a visibility disadvantage, meaning that it is more likely to reflect a genuine preference. By correcting for this confound and keeping only the tuples where the higher-scoring comment was written after the lower-scoring one, we get a dataset with as much `GPT-3`-usable information (≈ 0.50 bits) as HH-RLHF (Ganguli et al., 2022), one of the most widely used preference datasets in LLM alignment.

We use this filter, as well as a handful of others, to create two of the largest datasets of collective human preferences over text, which we call the Stanford Human Preferences datasets (SHP and SHP-2). SHP contains 385K preferences inferred from Reddit data, specifically advice-oriented subreddits in a set of hand-curated subject areas (e.g., `askculinary`). SHP-2 contains 4.8M preferences from both Reddit and various Stack Exchange domains (e.g., `stackoverflow`). For licensing information, schema specification, potential biases, and more details on how the specific subreddits and domains were selected, we refer the reader to our Huggingface repositories for SHP and SHP-2.

WARNING: The following content contains language from the DWMW17 dataset that is offensive in nature.

DWMW17 (Davidson et al., 2017)		
Hate Speech	Offensive	Neither
f*ggots (3.844)	r*tards (2.821)	lame (4.426)
f*g (3.73)	n*gs (2.716)	clothes (0.646)
f*ggot (3.658)	n*gro (2.492)	dog (0.616)
c*ons (3.53)	n*g (2.414)	cat (0.538)
n*ggers (3.274)	c*nts (2.372)	iDntWearCondoms (0.517)
qu*er (3.163)	p*ssies (2.29)	thank (0.47)
co*n (3.137)	qu*er (2.213)	kick (0.423)
n*gger (3.094)	r*tarded (1.997)	30 (0.345)
d*ke (3.01)	c*nt (1.919)	football (0.334)
f*gs (2.959)	b*tches (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 8. 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).